A Cell-Tracking Tool for Analysing Cell Behaviour During Wound Repair

Edward Antonian

Master of Science

Data Science

School of Informatics

University of Edinburgh

2019

Abstract

Inflammation is the body's first first response to external tissue damage and is vital for the healthy repair of wounds. The migration of immune cells towards the injury site during this process is mediated by unidentified chemical signals, the direct measurement of which is not currently experimentally possible. In this project we examine a set of computational techniques recently proposed by Weavers et al. (2016) for analysing the chemical environment during wound repair by observing the motion of *Drosophila* leukocytes *in vivo*. By modelling leukocyte motion as a biased-persistent random walk, one can characterise cell motion at different wound distances using Bayesian inference and, in conjunction with a chemical diffusion model, use this to deduce properties of the chemoattractant. We break down each stage of this process, exploring and testing the mathematical and computational elements in detail. In the process, we reproduce the pipeline of Weavers et al. (2016) and release the code for open source use.

Acknowledgements

I want to thank my supervisor, Dr Linus Schumacher, for his support, guidance and patience during this dissertation process. His thoughtful advice and calm demeanour throughout was invaluable. I also want to thank Professor Will Wood and his team for giving me access to the microscopy data and for providing insight and feedback. Finally I want to thank my family who always provide love and support when I need it most.

Table of Contents

Chapter 1

Introduction

1.1 Overview

The inflammatory response is a biological process which activates when tissue is damaged by an external source. It is characterised by vasodilation of local capillaries allowing plasma, containing immune cells, to cross into the inflamed tissue. Upon crossing, the cells begin migration towards the injury location where they perform various functions vital to the healthy repair of the wound such as the removal of necrotic cells and the breakdown of foreign pathogens and toxins. This behaviour is crucial for maintaining and restoring the integrity of tissue and can be observed in a wide array of animal life, from insects to mammals.

Some of the key stages involved in this phenomenon were identified as early as the mid-19th century (Cohnheim, 1867), however, despite this, many unanswered questions remain. In particular, the composition and nature of the chemical attractant that precipitates the directed motion of leukocytes through the tissue towards the wound site (chemotaxis) remains unclear. Many studies have indicated that the small, fast-diffusing compound hydrogen peroxide plays a key role in this regard (Niethammer et al., 2009; Moreira et al., 2010), however more recent work has suggested that H_2O_2 functions as an activator signal, priming leukocytes to respond to tissue damage, rather than acting as the chemoattractant itself (Evans et al., 2015). This leaves the question of what drives leukocyte chemotaxis during inflammation an open research topic.

In 2016, Weavers *et al.* released a paper titled *"Systems analysis of the dynamic inflammatory response to tissue damage reveals spatiotemporal properties of the wound attractant gradient"* (Weavers et al., 2016). In this publication the authors set out a novel methodology

for inferring properties of the chemical environment during wound repair by imaging the migration of *Drosophila* leukocytes *in vivo*. This particular organism, commonly known as the fruit fly, has a number of advantages for use in the context of studying wound repair. The optical translucency of its pupae allows for high-quality imaging while their short life cycle and relatively small genome makes identifying and modifying genes involved in the inflammatory process a more tractable task (Razzell et al., 2011). Weavers et al. (2016) construct a computational pipeline that uses Bayesian inference to examine how leukocyte motion is affected by wounding. By observing these changes at different wound distances and times, they were able to infer information about how a potential attractant might be diffusing from the wound. A key finding of this analysis was that the chemoattractant diffusion rate was likely to be considerably lower than that of H_2O_2, narrowing down the search for this unknown compound. However, the broader significance of their paper was to introduce a novel and versatile set of tools for exploring the mechanisms involved in the inflammatory response. Their computational pipeline, which can be broken down into the steps summarised in figure 1.1, is the central focus of this thesis. In it we seek to reproduce and examine the computational and mathematical tools used at each stage in detail, exploring their subtleties, capabilities and limitations.

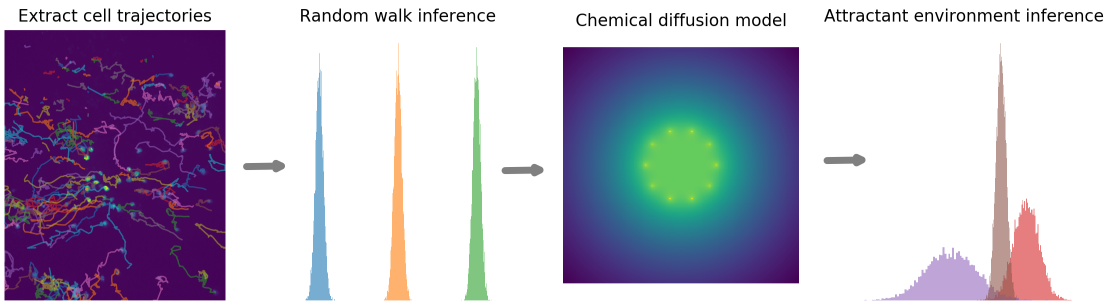

Figure 1.1: The steps needed to complete the pipeline of Weavers et al. (2016)

Continued research into the topic of inflammation has the potential to lead to results of direct clinical value. Inflammatory abnormalities have been linked to numerous medical disorders in humans including arteriosclerosis, obesity, cancer and Alzheimer disease, yet today the range of treatments available remains limited (Masayuki and Kiyoshi, 2016). A better understanding of the mechanisms that drive cell migration during healthy wound repair is a necessary precursor for the development of novel prognostic and therapeutic tools and thus we hope that, by reducing the barriers to entry for future researchers, this paper may contribute in some small way to that effort.

1.2 Objectives

As mentioned, the key objective of this project is to break down the methodology proposed by Weavers et al. (2016), exploring each stage in more detail. In addition, each piece of computational machinery is tested under a range of conditions, with important aspects and pitfalls highlighted. In that sense, this paper aims to function as a guide for future researchers new to these specific techniques. A second objective is to provide a transparent, functioning and well-documented set of computational routines for undertaking the analysis described in the project, in an open source environment. While Weavers et al. (2016) present their methodology clearly, they did not release the project code to the wider research community. Thus this may also reduce barriers for future researchers to perform further analysis into this topic. A repository containing all the relevant code, written in pure Python, can be found on GitHub, accompanied by a set of Jupyter notebooks providing an interactive environment and graphical interface to explore and develop these tools.

1.3 Thesis structure

In this report, we break down the computational pipeline of Weavers et al. (2016) into three distinct stages. We begin first, in Chapter 2, by explaining the way in which the authors model and parametrise random walks of migrating leukocytes and how one can use a Markov Chain Monte Carlo (MCMC) algorithm to infer information about their directional bias and persistence from cell trajectory data. This algorithm is tested in a variety of scenarios using simulated *in silico* trajectory data. Its robustness to detection noise, and other issues that can arise in practice, is also assessed. Chapter 3 demonstrates how information gathered about leukocyte motion at different positions in space and time can be used to infer properties of the attractant gradient. In particular, the model for attractant diffusion in terms of the heat equation is studied and an optimised formulation is proposed. The details of this second inference process are explored, and the algorithm is tested under a variety of conditions. Chapter 4 implements and tests another key stage of the computational pipeline: that of extracting raw coordinate data from *in vivo* imaging of live *Drosophila* leukocytes. Various approaches are explored and an algorithm in terms of the linear assignment problem is implemented, with some additional modifications tailored to our specific use case. Finally, chapter 5 seeks to tie together each of these three stages, and test the complete pipeline on real data. The results are critically compared with those of Weavers et al. (2016) and we suggest ways in which the analysis could be extended and improved in the future.

Chapter 2

Random walk parameter inference

A vital stage in the pipeline of Weavers et al. (2016) is inferring properties of leukocyte motion from trajectory data. In this chapter we introduce the concept of random walks, focusing on the *biased-persistent* model, and implement and test a Bayesian algorithm for inferring information about random walk parameters by observing trajectory data.

2.1 Random walks

Formally, random walks are a class of stochastic processes that describe the trajectory of an object through some mathematical space. In the typical case, one considers some state, $\mathbf{z}_t \in \mathbb{R}^d$, that is successively incremented between discrete time intervals by drawing random variables from some probability distribution.

$$\mathbf{z}_{t+1} = \mathbf{z}_t + X, \quad \text{where} \quad X \in \mathbb{R}^d \sim p(\mathbf{x}; \{\mathbf{z}_{i \leq t}\}, \mathrm{E}) \tag{2.1}$$

This distribution $p(\mathbf{x})$ may be static, or may depend on current or past states $\{\mathbf{z}_{i \leq t}\}$ and the environment E. Applications of this general formulation are numerous and far-reaching, with random walks appearing in many areas within both the natural and social sciences, as well as economics, finance and computing (Lawler, 2010). Within biology, random walks have been used extensively to represent the motion of biological objects on various scales (Codling et al., 2008). Examples include the path animals track through their local environment, the motion of biological macromolecules and, in various contexts, the path that mobile cells traverse within organisms.

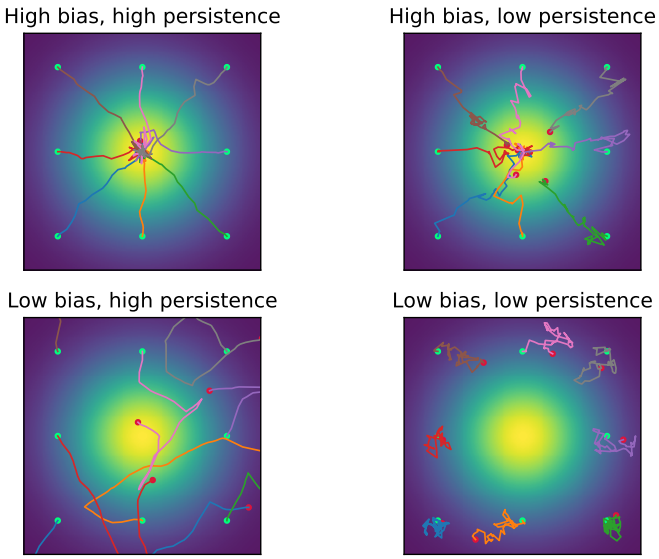

Figure 2.1: Simulated cells moving with different levels of bias and persistence are shown in the vicinity of a centrally located source. Start positions are indicated in green while end positions are indicated in red.

2.2 The biased-persistent random walk

The model that Weavers et al. (2016) opt for is known as a *biased-persistent* random walk, which first appears in the literature in Jones et al. (2015). The terms 'bias' and 'persistence' refer to two distinct phenomena observed in cell migration. Cells with a high bias have a strong tendency to drift towards a static global position, in this case the site of injury. And cells with a high persistence tend to take successive steps that are oriented in the same general direction, resulting in motion that is likely to continue along a straight line (S. Patlak, 1953). Some simulated trajectories, for cells with different levels of bias and persistence, are shown in figure 2.1.

Mathematically, the model is formulated as follows. A cell is defined by three parameters w, p and b, each of which take a value on the interval $[0, 1]$. The cell exists in an environment of two dimensional Euclidean space and its state, at any one time, is simply its x-y coordinates. At each discretised time step, the cell makes a jump of length s_t, at an angle α_t defined with reference to an arbitrary fixed axis. The step size s_t is generated by drawing from a *truncated normal* distribution, with a probability density function

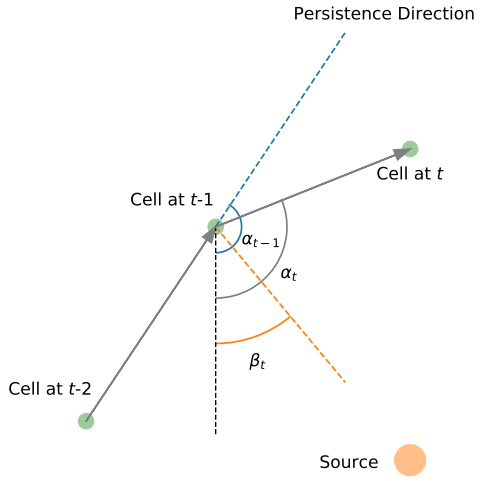

Figure 2.2: A cell tracks a path through 2D space. All angles are defined with reference to the negative y-axis.

$$\mathcal{N}^{+}(s_t;\sigma_s) = \begin{cases} 0 & \text{if } s_t < 0 \\ 2\,\mathcal{N}(s_t;0,\sigma_s) & \text{if } s_t \geq 0 \end{cases} \tag{2.2}$$

This distribution is static and does not depend on the input parameters w, p or b. The value of σ_s sets the typical step size and should be proportional to the square root of the time interval Δt. The turning angle α_t is drawn from a dynamic distribution that depends on the current angle towards the source β_t, and the previous angle α_{t-1}. At each step, the cell chooses to follow either biased motion, with probability w, or persistent motion, with probability $1-w$. The turning angle is then drawn from a *wrapped normal* distribution $\mathcal{N}_{\mathrm{w}}(\theta;\mu,\sigma)$. The probability density function for this distribution is defined as

$$\mathcal{N}_{\mathrm{w}}(\theta;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\Big(-\frac{(\theta-\mu+2\pi k)^2}{2\sigma^2}\Big), \tag{2.3}$$

(Fisher, 1993) and can be seen plotted for various values of μ and σ in figure 2.3. If biased motion is followed, the distribution parameters are set as $\mu_b = \beta_t$ and $\sigma_b = \sqrt{-2\log b}$. If persistent motion is followed, the parameters are $\mu_p = \alpha_{t-1}$ and $\sigma_p = \sqrt{-2\log p}$. Thus the entire probability density function over α_t, for a set of input parameters $\phi = \{w, p, b\}$, is

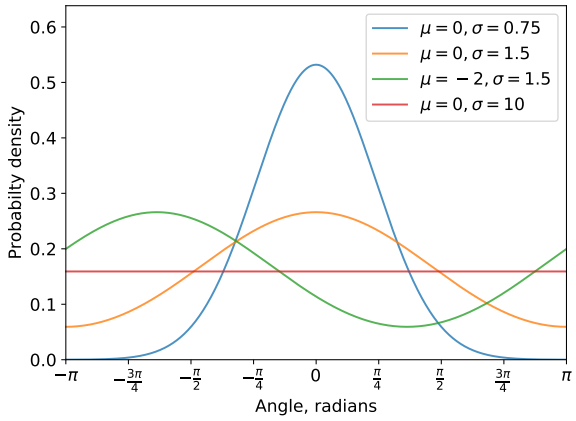

Figure 2.3: Several wrapped normal probability density functions are plotted with different value of μ and σ.

$$p_\phi(\alpha_t|\beta_t,\alpha_{t-1}) = w\mathcal{N}_w(\alpha_t;\beta_t,\sqrt{-2\log b}) + (1-w)\mathcal{N}_w(\alpha_t;\alpha_{t-1},\sqrt{-2\log p}) \tag{2.4}$$

The intuition behind the parameters w, p and b is as follows. b and p give a measure of how biased or persistent the motion is respectively. As they tend to one, the associated distribution in equation 2.4 tends to a Dirac delta function centred around the angle towards the source, β_t, and the previous angle, α_{t-1}, respectively. Conversely as they tend to zero both tend to a uniform distribution over the interval $[-\pi,\pi]$. The parameter w then mixes these two distributions. A value of one indicates fully biased motion and a value of zero indicates fully persistent motion. Notebook 1 in the supplementary materials provides an interactive GUI exploring the properties of this probability distribution.

2.3 Inferring biased-persistent parameters from observation

If one assumes a biased-persistent random walk is the underlying mechanism driving the motion of leukocytes, the next question becomes: given a set of observations of cells moving in the vicinity of a wound, what parameters values are likely to have generated this data? Given that we are dealing with a stochastic process, there is inevitably going to be uncertainty over these underlying values. Thus, the question is more accurately stated as, "given some trajectory data $\mathcal{D}$, what is the posterior distribution over the underlying values $\phi = \{w, p, b\}$ that generated it?" The formal answer can be stated using Bayes rule.

$$p(\phi|\mathcal{D}) = \frac{p(\mathcal{D}|\phi)p(\phi)}{p(\mathcal{D})} \tag{2.5}$$

In the numerator, $p(\mathcal{D}|\phi)$ is the probability of observing the data $\mathcal{D}$ given a specific set of parameters ϕ and $p(\phi)$ represents the user's prior beliefs about the distribution over parameters. In the denominator, $p(\mathcal{D})$ is the probability of observing the data independent of any set of parameters, which in principle can be found by integrating the numerator over all values of ϕ. Note also that, since we are dealing with a continuous space of trajectories and parameters, the term 'probability of' is more accurately replaced with 'value of the probability density function at', although we will continue to use the former term for convenience.

Let us assume that the data $\mathcal{D}$ is split into a set of N distinct trajectories π^i. Since all paths are assumed to be independent of one another, the probability of separately observing all of these paths given a set of parameters will then be the product of the probability of observing each one individually.

$$p(\mathcal{D}|\phi) = \prod_{i=1}^{N} p(\pi^i|\phi). \tag{2.6}$$

From each path π^i, which contains T x-y coordinate readings, one can extract T-1 angle observations, α_t^i. Correspondingly, there are associated T-1 angles towards the source, β_t^i for each step of the trajectory. Finally, there are T-2 *previous* angles observations α_{t-1}^i. If one makes the assumption that at the first step the cell must observe biased motion (since there is no previous angle for persistent motion to apply to) then the probability of observing a single path π^i is

$$p(\pi^i|\phi) = p_\phi(\alpha_1^i|\beta_1^i) \prod_{t=2}^{T-1} p_\phi(\alpha_t^i|\beta_t^i, \alpha_{t-1}^i), \tag{2.7}$$

where these distribution values are obtained from equation 2.4 (Jones et al., 2015). Putting this together with equation 2.6 gives

$$p(\mathcal{D}|\phi)p(\phi) = L(\phi) = p(\phi) \prod_{i=1}^{N} p_\phi(\alpha_1^i|\beta_1^i) \prod_{t=2}^{T-1} p_\phi(\alpha_t^i|\beta_t^i, \alpha_{t-1}^i). \tag{2.8}$$

However, in order to avoid floating point errors, it is more convenient to work with the log likelihood.

$$\log L(\phi) = \log p(\phi) + \sum_{i=1}^{N} \log p_\phi(\alpha_1^i|\beta_1^i) + \sum_{i=1}^{N} \sum_{t=2}^{T-1} \log p_\phi(\alpha_t^i|\beta_t^i, \alpha_{t-1}^i). \tag{2.9}$$

We are now in a position to evaluate the (log of the) denominator of equation 2.5, however the multidimensional integral contained within the numerator still cannot be evaluated in practice. One must therefore resort to approximate methods. Markov Chain Monte Carlo (MCMC) provides a relatively simple and thoroughly studied framework for estimating multidimensional probability distributions that can be evaluated up to a multiplicative constant. That is, MCMC can be used to sample from a probability distribution $P(x)$, given an evaluable function $P^*(x)$ where $P(x) = P^*(x)/Z$, for some unknown constant Z (MacKay, 2003). Weavers et al. (2016) opt to use the Metropolis-Hastings algorithm to sample from the posterior distribution. This method is a good choice since it is one of the more simple yet effective MCMC routines, and we are able to sample directly from the multivariate posterior. Thus alternative methods such as Gibbs sampling, for example, offer no great benefits in this case. Starting at some point in parameter space ϕ_0, the Metropolis-Hastings algorithm successively proposes a new position ϕ', drawn from a proposal distribution $Q(\phi';\phi_t)$ which is typically narrow and centred around ϕ_t. If, as is usually the case, the proposal distribution is symmetric about ϕ_t, then the new state is accepted with probability a, where

$$a = \frac{L(\phi')}{L(\phi_t)} = \exp\big(\log L(\phi') - \log L(\phi_t)\big), \tag{2.10}$$

or simply accepted outright if a is greater than one. If the new state is accepted we set $\phi_{t+1} = \phi'$. If it is rejected we set $\phi_{t+1} = \phi_t$. Repeating this many times results in a final list of parameter-space points $\Phi = [\phi_0, \phi_1, \ldots, \phi_T]$. One typically discards some of the first points as 'burn in', since they will strongly influenced by the starting position ϕ_0, although the precise number that is appropriate is problem specific.

2.4 Testing the implementation

The details of the implementation used are as follows. In the absence of any initial information about w, p or b, we begin with a uniform prior distribution $p(\phi)$ between 0 and 1 in each dimension. In practice, this means the prior term from equation 2.9 can be dropped, and instead any proposal point ϕ' that is outside of the bounds $[0,1]$ in any dimension is automatically rejected. A random point ϕ_0 is drawn from the prior to begin sampling from. We then sample for a burn in of 3000 steps (in alignment with Weavers et al. (2016)) and then record the path for a further 15,000 steps. The proposal distribution $Q(\phi';\phi_t)$ is a Gaussian centered around ϕ_t, $\mathcal{N}(\phi';\phi_t,\sigma)$. Initially, we set $\sigma = 0.02$. Then, after every 100 samples, σ dynamically

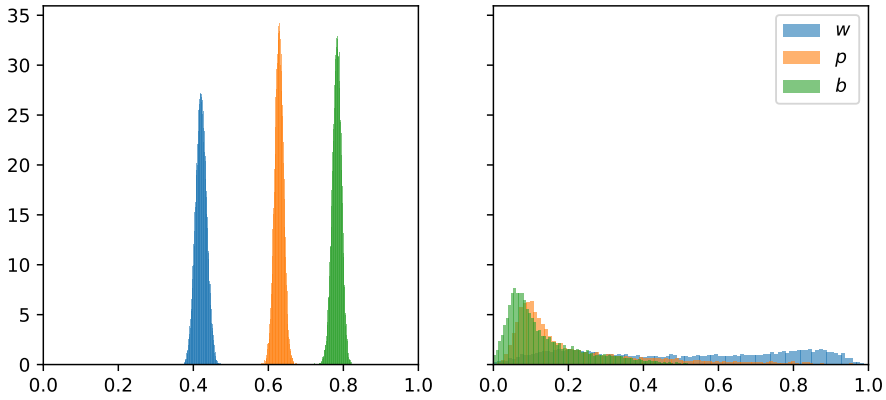

Figure 2.4: The sampled posterior distribution over w, p and b is shown for two sets of trajectory data with different characterising parameters. On the left, the input paramters were $[0.4, 0.6, 0.8]$ respectively. On the right, they were $[0.5, 0.1, 0.1]$. Note that the plotting the three distributions on a single axis is somewhat misleading as, in truth, these distributions are not independent. A full plot of the joint probability can be found in appendix section B.0.1

updated such that is is one third the standard deviation of all previously sampled points, in each dimension. This is repeated for five different starting locations and the results are combined (see Notebook 2).

In order to verify that this inference procedure is behaving as expected, we can simulate some biased-persistent trajectories, and then try to recover the input parameters that generated them. Initial experimentation reveals that, in general, the inference tends to work well. However certain input parameters are considerably 'easier' to infer back than others. Figure 2.4 demonstrates this effect by showing the final histogram over w, p and b for two different sets of input parameter values. The left plot, showing the sampled posterior for inputs $\phi = [0.4, 0.6, 0.8]$, is tightly distributed about the correct values, indicating successful inference. By contrast, the right plot, showing the posterior for inputs $\phi = [0.5, 0.1, 0.1]$ has much greater variance. p and b are distributed between roughly 0 and 0.5, and w is close to uniform between 0 and 1.

In order to investigate this phenomenon in a more systematic way, we ran the following experiment. Trajectories were simulated *in silico* for 20 leukocytes, each taking 100 steps in the vicinity of a simulated wound, following a biased-persistent random walk. The parameters dictating this motion were varied over a grid of inputs, with w, p and b each ranging from 0.1 to 0.9 inclusive in steps of 0.1, resulting in 729 sets of trajectories with unique character-

ising parameters. We then performed inference on each set following the procedure outlined above, with the mean value for each of w, p and b defining the best estimate of the underlying parameters. Table 2.1 shows some summary statistics for the error on these estimates.

	w error	p error	b error
mean	0.019	0.012	0.013
25%	0.006	0.002	0.004
50%	0.011	0.008	0.011
75%	0.042	0.022	0.031

Table 2.1: The mean and three quantiles for the inference procedure's absolute error on the estimate of the underlying parameters across all 729 runs is shown. Error is measured as the posterior mean minus the true underlying value.

In order to quantify the 'success' of a certain run, i.e. the sampler's ability to recover the input parameters, we estimate the Kullback-Leibler (KL) divergence between the posterior and the prior (Kullback and Leibler, 1951).

$$D_{\mathrm{KL}}(\, p(\phi|\mathcal{D}) \,||\, p(\phi)) = \int_{-\infty}^{\infty} p(\phi|\mathcal{D}) \log\left(\frac{p(\phi|\mathcal{D})}{p(\phi)}\right) \mathrm{d}\phi \tag{2.11}$$

Also known as the relative entropy, the KL divergence quantifies the amount of information acquired about the underlying parameters by observing the trajectory data. A high value for the KL divergence between the posterior and the prior would indicate that the inference process has significantly reduced the uncertainty over the underlying parameters. A low value conversely would indicate that the posterior remains similar to the prior, with a high uncertainty over the underlying parameters. The advantage of using relative entropy between the posterior and the prior as a measure of inference success over, say, the absolute error between the posterior mean and the underlying value is that a very wide posterior, such as that of w on the right side of figure 2.4, could easily have a mean that is close to the true value despite a high uncertainty. The KL divergence, by contrast, gives more information about the reduction in uncertainty that has occurred over the inference process.

Since the MCMC process only provides us with samples from the posterior distribution, not the analytic probability density function implied in equation 2.11, we estimate it as a multivariate Gaussian, using the empirical mean and covariance μ and Σ. This approximation suffices for our purpose since we are interested in the general trend, not the precise value (a more thorough

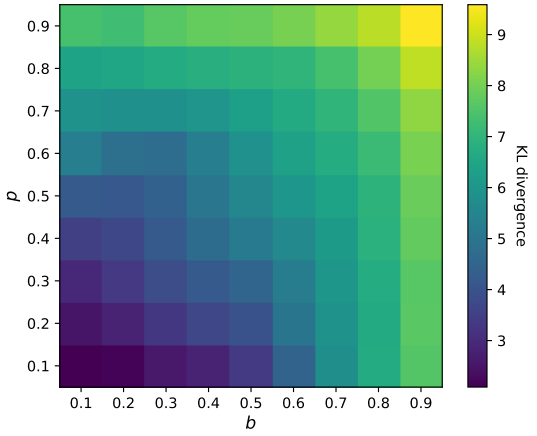

Figure 2.5: The estimated KL divergence, or relative entropy, between the posterior and the prior distribution after MCMC inference is shown as a heatmap over a grid of p and b. Each square is averaged across all nine values of w.

approach could be to use non-parametric approximations such as a kernel density estimator). Combining this with the fact that the prior is uniform over the interval $[0, 1]$ in each dimension leads to a formula for the approximate posterior/prior KL divergence.

$$D_{\mathrm{KL}}(\, p(\mathcal{D}\,|\phi) \,||\, p(\phi)) \approx \int_0^1 \mathcal{N}(\phi; \mu, \Sigma) \log\left(\mathcal{N}(\phi; \mu, \Sigma)\right) \mathrm{d}\phi, \tag{2.12}$$

where the integral over each dimension is implicit. The pattern that emerges is that when trajectories are generated from walkers with low values for p and b the amount of information gained about their true values via the inference process is small. Conversely, high values of p and b result in more information, with tightly distributed posteriors. Figure 2.5 shows the estimated KL divergence between the posterior and the prior for a range of p and b values.

2.5 Practical considerations

In the preceding analysis leukocyte paths were simulated *in silico* following a perfect biased-persistent random walk. This idealised set up enabled us to tightly control the experimental variables and evaluate the performance of the inference algorithm in isolation. However this process neglects many of the practical realities of performing inference on *in vivo* data. In the following section we identify three practical issues that are likely to arise when performing

inference on real, detected trajectories and investigate what effect these might have on the inferred posteriors. Those issues are noise on the true cell position, breaks in detected cell trajectories and choice of image frame interval.

2.5.1 Cell position uncertainty

Any real detection process will inevitably have some level of uncertainty on the true position of the cell, for instance due to image quality. In order to test the extent to which this will effect the inferred posteriors we add Gaussian noise to the x-y coordinates of the simulated trajectories with increasing variance. Inference is then performed on each set of trajectories and the difference between the true underlying parameter values and mean of the posterior is plotted as a function of the noise, measured as a fraction of the step size σ_s. This is performed for a low, medium and high value of p and b, keeping w fixed at 0.5.

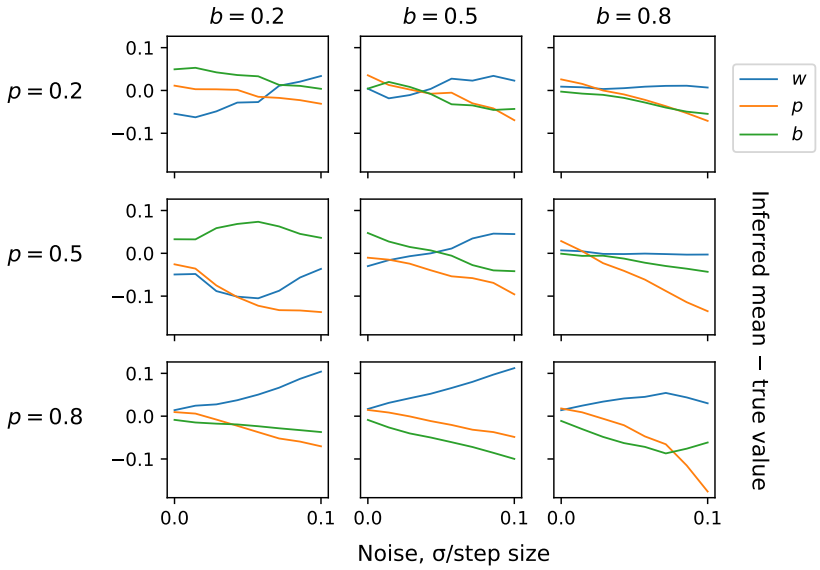

Figure 2.6: The error on the underlying parameters, as measured by the posterior mean minus the true underlying value, is plotted as a function of noise. This is performed for a low, medium and high value of p and b, while w is kept fixed at 0.5. The noise is normalised as a fraction of the step size parameter, σ_s, of equation 2.2.

The results indicate that the inferred values of w, p and b are indeed sensitive to noise. A general trend seems to be that the inferred value of persistence is consistently reduced by increasing position noise. This seems to be most pronounced when the true underlying values

of p and b are higher. Bias too tends to be reduced, although often to a lesser extent. Another interesting trend is that all parameters are most affected when persistence is higher. These results indicate that researchers should indeed take care when performing image analysis. If we assume image quality is the only factor affecting position precision, then ideally the step parameter σ_s should be greater than 20 pixels to keep this error below 0.05.

2.5.2 Trajectory breaks

A common problem that can occur during the detection process is that, on some individual frame, a certain cell or set of cells fail to be detected. This can cause a break in the output trajectories, as a single path gets interpreted as two or more separate paths. The task of linking broken trajectories is addressed in section 4.3, however, this can often be a particularly computationally intensive stage in the cell tracking process. In order to gauge the value of expending computational resources on trajectory linking one needs to ascertain the effect that broken paths have on the inferred posteriors. This is tested by simulating trajectories *in silico* and then deleting specific coordinate observations, causing a trajectory to be split, with probability P per coordinate point. This is done for increasing values of P, again, over a range of values for bias and persistence.

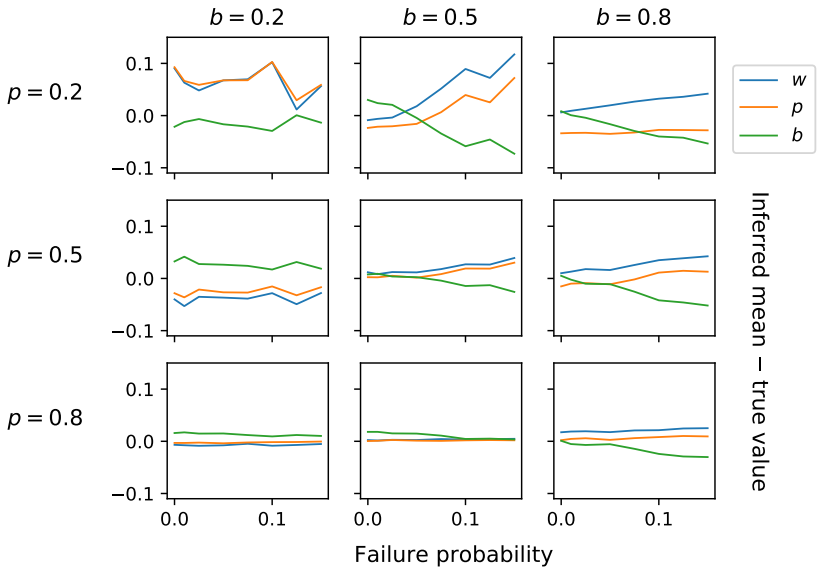

Figure 2.7: As in the previous figure, the inference error is shown for a low, medium and high value of p and b, while w is kept fixed at 0.5. This is plotted as a function of detection failure probability up to 15%, where a single failed detection causes a trajectory to be split.

Again, we see that all three parameters are impacted by trajectory breaking. However, conversely, in this case it seems that the effect is less pronounced when persistence is higher. The inferred level of bias seems to be reduced as failure probability increases, whereas w and p tend to be overestimated. This could be due to the fact that, since the first step is always considered to be constructed via pure biased motion, the effect of shortening and increasing the number of trajectories is to force steps that were in fact persistent in origin to be considered biased. This would place downwards pressure on the inferred bias values.

2.5.3 Frame interval

When imaging real cells *in vivo*, the rate at which still images are captured is an experimental design choice. Weavers et al. (2018), lay out detailed practical instructions for *Drosophila* imaging, suggest recording frames at intervals of 30 seconds or less, but leave the ultimate choice up to the practitioner. Ideally the frame rate chosen would have little effect on the posterior distributions of the underlying biased-persistent parameters. If this was the case inference could be performed on data taken at different frame rates, perhaps from different experimental groups, and remain directly comparable. On the other hand, if the posteriors are systematically altered by changing frame rate this raises issues of reproducibility and transferability. This issue has been highlighted in a number of papers (Jones et al., 2015; Sim et al., 2015) however, while Rosser et al. (2013) analyse the effect for persistent random walks, no robust analysis of the effect for biased-persistent motion has been yet provided. In order to investigate this issue further, trajectories are simulated *in silico* with different underlying biased-persistent parameters. Inference is then performed on the same paths 8 times, except on each run the frame rate is halved. That is, inference is performed on the first 200 points of the paths constructed by connecting every other point in the previous path. For each frame interval we repeat the inference procedure five times to give as accurate a posterior as possible. We also choose to vary the level of bias only while the parameters w and p are kept fixed at 0.5 and 0.5 respectively. This is because ultimately, as explained in section 3, the bias level b is the most important parameter for the purpose of inferring information about the attractant environment.

We find the effect on persistence parameter p is in general agreement with the results of Rosser et al. (2013), who demonstrate that increasing frame sample rate significantly reduces the observed persistence for correlated (unbiased) random walks. Figure 2.8 shows that, as frame interval increases, the inferred persistence is effectively reduced to zero. Conversely the opposite effect seems to be true for bias b. Here we can see that a higher frame interval leads

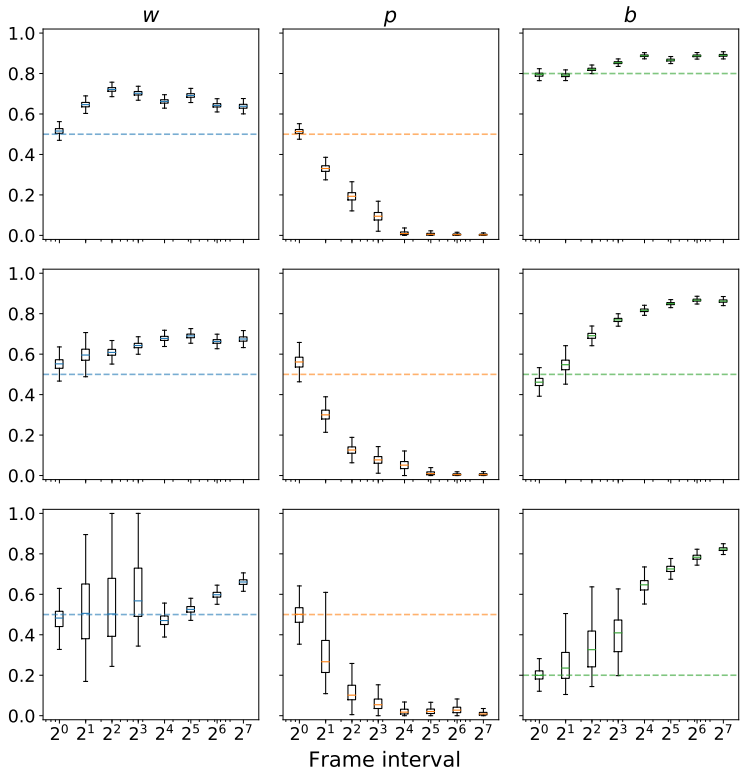

Figure 2.8: The effect of increasing the frame interval is shown for three sets of underlying parameters w, p and b at a high, medium and low level of bias, namely (0.5, 0.5, 0.8), (0.5, 0.5, 0.5) and (0.5, 0.5, 0.2) respectively. The true underlying value in each case is indicated by the dotted line. Inference is then performed at doubling frame intervals and a box plot of the posterior distribution for each parameter is plotted vertically.

to a systematic overestimation of the bias parameter. This makes some intuitive sense. A series of steps taken with low bias may not be heavily directed to the source individually, however the combined effect of many successive steps is likely to be more directed towards the wound overall. Thus cells captured at a higher frame interval may appear to be moving with a stronger bias.

At the very least, this analysis demonstrates the importance of using a consistent frame interval when performing experiments that involve inferring biased-persistent parameters. It is clear that the inferred values of w, p and b are not independent of this choice at all. One would hope that this choice would not affect inference of the important attractant dynamics parameters (the process of which is detailed in the proceeding chapter) such as attractant diffusion speed, however this is not clear. Further investigation into this would be valuable.

Chapter 3

Attractant dynamics inference

Once random walk parameters have been inferred from real *in vivo* data, the next stage is to use this information to deduce properties of the wound and its chemical environment. Chemotaxis is the process by which cells sense and swim in response to local molecules such as cytokines, nutrients and toxins. Whilst the precise details of the mechanism underlying leukocyte chemotaxis remain unknown, biologists have long observed the phenomenon of leukocytes reacting to chemical gradients of various forms (Whicher and Evans, 1992). Here we study Weavers et al. (2016)'s technique for connecting an attractant diffusion and receptor-ligand binding model to the leukocyte bias observed at different points in space and time. A similar inference process to that of the previous chapter can then be used to sample from the posterior distribution of the wound model parameters. This can give important insight into the nature of the attractant dynamics during wound repair.

3.1 Attractant diffusion and the heat equation

The first step in this process is to model how a chemoattractant might emanate from a wound. In general there may be several approaches for doing this. Liepe et al. (2012), for example, postulate three simple static distributions, where concentration decreases in a Gaussian, sigmoidal and linear fashion respectively, but stays constant with time. However, the authors remark that such models will likely fail to capture the complex spatial and temporal characteristics of real diffusion processes. More comprehensive mathematical models have since been proposed in the context of leukocyte migration such as modelling based on the heat equation; the strategy opted for by Weavers et al. (2016). This approach tries to build the model from first principles of particle diffusion, rather than asserting *ad hoc* the form of the chemical gradient.

The heat equation is a partial differential equation developed in the early 19th century to describe how the distribution of heat within some medium spreads over time (Fourier, 1822). More generally, it can be used to describe how many different substances, which tend to flow from areas of high concentration to low concentration, evolve with time. The equation can be stated as

$$\frac{\partial A(\mathbf{r},t)}{\partial t} = D\nabla^2 \left[A\left(\mathbf{r},t\right) \right], \tag{3.1}$$

where D is the diffusion coefficient, $A(\mathbf{r},t)$ is a scalar field representing the concentration of the diffusing substance at any vector position $\mathbf{r}$ and time t and ∇^2 is the Laplacian differential operator. In order to specify a solution to a partial differential equation, it must come with an associated initial condition and set of boundary conditions, which define the problem. First consider the case of a finite quantity of attractant Q_0 released at $t = 0$ at the point $\mathbf{r}'$ into empty d-dimensional euclidean space. This specifies the initial condition. The corresponding boundary condition is that the concentration must be zero at infinity in every direction. The solution for this problem, given by Pattle (1959), is

$$A(\mathbf{r},t) = \frac{Q_0}{(4\pi Dt)^{d/2}} \exp\Big(- \frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt} \Big), \tag{3.2}$$

a proof of which can be found in appendix section A.1.1. However, Weavers et al. (2016) model cells as continuous point sources, that release attractant at a rate of q from $t = 0$ to $t = \tau$. Carslaw and Jaeger (1959) demonstrate that the solution for a continuous point source can be obtained by integrating equation 3.2 with respect to time. This gives

$$A(\mathbf{r},t) = \frac{q}{(4\pi D)^{d/2}} \int_0^{\min(\tau,t)} \exp\Big(- \frac{r^2}{4D(t-t')} \Big) \frac{\mathrm{d}t'}{(t-t')^{d/2}}, \tag{3.3}$$

where $r^2 = |\mathbf{r}-\mathbf{r}'|^2$. Weavers et al. (2016) state that they use 'numerical integration' to compute the attractant concentration. However, it is possible to express this equation, for the two dimensional case, in terms of the special function $\mathrm{Ei}(x)$, the exponential integral. This allows the use of heavily optimised, pre-compiled functions at execution time instead of computationally expensive numerical integration operations, generating significant speed-ups. In terms of the exponential integral, equation 3.3 can be written as

$$A(r,t) = \begin{cases} -\frac{q}{4\pi D}\mathrm{Ei}\left(-\frac{r^2}{4Dt}\right) & \text{if } t < \tau \\ \frac{q}{4\pi D}\left(\mathrm{Ei}\left(-\frac{r^2}{4D(t-\tau)}\right) - \mathrm{Ei}\left(-\frac{r^2}{4Dt}\right)\right) & \text{if } t > \tau, \end{cases} \tag{3.4}$$

a full derivation and proof of which can be found in appendix section A.1.2. Thus, given a set of input parameters q, D and τ, one can compute the concentration of the diffusing attractant at any point in space and time, for a single continuous point source. The effect of multiple cells producing attractant from different locations can be accounted for by simply summing together the effect of individual point sources. Figure 3.1 shows a heat map of the attractant concentration over space before, at and after time τ for a single continuous point source (top) and several point sources (bottom).

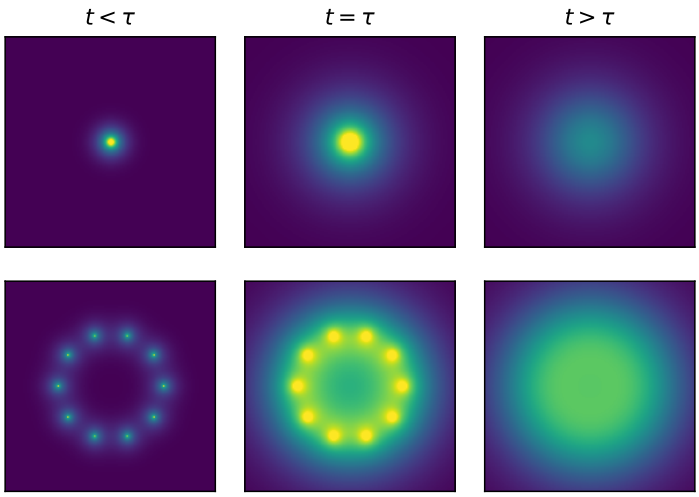

Figure 3.1: Top row: a single point source continuously producing attractant from $t = 0$ to $t = \tau$. Bottom row: the summed effect of multiple point sources of the same nature. See Notebook 3 for implementation and animation details.

3.2 Receptor-ligand binding kinetics

Once one is able to specify the concentration of attractant at any point in space and time, we must then consider how this affects the motion of leukocytes. The field of receptor-ligand binding kinetics provides a robust framework for such analysis. It lays out a set of mathematical models which attempt to quantify the rates at which chemical entities and biological receptors

bind and unbind with each other (Lauffenburger and Linderman, 1993). The concentration of bound complexes at a certain position in space will depend on several factors. These include the total concentration of receptors at that point R_0, the total concentration of ligands (attractant) at that point $A(r,t)$, and the rate at which complexes bind and unbind as a function of concentration levels. In the steady state, complex concentration is modelled as

$$C(r,t) = \frac{1}{2}(\kappa_d + R_0 + A(r,t)) - \sqrt{\frac{1}{4}\big(\kappa_d + R_0 + A(r,t)\big)^2 - R_0 A(r,t)} \tag{3.5}$$

where κ_d is the dissociation constant (Liepe et al., 2012). For a derivation of this equation see appendix section A.2.

Each leukocyte is assumed to be sensing the local attractant concentration at both its front and rear, with respect to the direction of the wound. Weavers et al. (2016) set the average radius of each leukocyte as $\Delta r = 15\mu$m, such that the concentration of bound receptor-ligand complexes at its front and rear is $C(r-\Delta r, t)$ and $C(r+\Delta r, t)$ respectively. The *observed* bias (the product of the parameters w and b from section 2.2) is then postulated to be a linear function of the difference in complex concentration.

$$wb = m[C(r-\Delta r, t) - C(r+\Delta r, t)] + b_0 \tag{3.6}$$

This is the final equation relating the attractant dynamics model to the orientation mechanism of the leukocytes. This function can be explored in notebook 4. Note that, in order to construct it, seven input variables are required, namely

$$\theta = \{q, D, \tau, R_0, \kappa_d, m, b_0\}. \tag{3.7}$$

3.3 The inference process

The previous analysis has shown how, given a set of input parameters, one can specify the observed bias that leukocytes should follow at any given point in space and time, that is, $wb = f_\theta(r,t)$. Thus, given a set of measurements of observed bias as different points in space and time, one possible approach would be to simply find the best fit for the parameters, θ, by regular gradient-based regression. However, following the Bayesian approach, it also is possible to use another MCMC process, similar to that of in section 2.3, to infer the probability distribution over these parameters.

First, the leukocyte trajectory data $\mathcal{D}$ gathered from live *in vivo* cell imaging is split into spatial and temporal clusters. For example, Weavers et al segment the trajectories into five spatial bins roughly covering the range 50-500μm away from the wound, and six temporal bins covering the range 0-125 minutes. In general, we could have S spatial clusters and T temporal clusters, indexed by i and j respectively. For each subset of the trajectory data falling within these bins $\mathcal{D}_{ij}$ we use the MCMC pipeline detailed in section 2.3 to produce an associated posterior distribution of biased-persistent parameters w, p, b and, accordingly, a posterior distribution of observed bias $p_{ij}(wb)$. The likelihood function of a certain set of attractant dynamics parameters θ, then, is the product of the value of the posterior probability density function at each of the observed biases that result from those parameters.

$$L(\theta) = \prod_{i=1}^{S} \prod_{j=1}^{T} p_{ij}(wb = f_\theta(r_i, t_j)) \tag{3.8}$$

Note that the probability density function for the posterior of the observed bias at spatial/temporal cluster ij, which must be specified to evaluate this likelihood function, could and perhaps should in principle be evaluated via non-parametric means such as a kernel density estimator. However, in practice, this will slow down the inference process dramatically. For this reason, we choose to approximate p_{ij} as a Gaussian distribution, which is sufficient in most cases. Note also that for each spatial cluster i and temporal cluster j there is associated a single distance r_i and time t_j used to represent/summarise the group. This is somewhat problematic since, as mentioned, these are clusters that span a range of distance and time. Thus deciding on a single value, which could be the midpoint of the bin, will result in some approximation error. Weavers et al. (2016) do not directly address this problem and further analysis into the uncertainty that this approximation introduces would be valuable.

3.3.1 Priors and units

Another issue that is of great practical importance in the attractant dynamics inference process is that of priors and units. The prior used in the previous inference process, for biased-persistent random walks, was an independent uniform distribution between 0 and 1 for each of w, p and b. This was bounded, unitless and fairly intuitive. However, for the second inference process, over attractant dynamics parameters, the situation is more complex. Each parameter needs specific physical units and is, in theory, unbounded. The units for each of the seven parameters can be constructed using a combination of some unit of molecule number, distance and time. For example, the flow rate q should be measured in molecule number per unit time, and the diffusion

coefficient D should be measured in units of distance squared per unit time. It is important to use units which allow the parameters to be comparable in size, since having wildly different orders of magnitude can lead to issues surrounding floating point error when performing inference. We find that units of μm for distance and minutes for time lead to fairly natural orders of magnitude. In terms of the molecular unit to use, it is not immediately clear what would be most effective. However, due to the fact that in equation 3.6, m, (which is inversely proportional to molecular units) multiplies the bound complex concentration (which is proportional to molecular units) any consistent unit will ultimately cancel and can thus be used here. Since, for the purpose of this investigation, we are not overly concerned with the absolute inferred values for q, R_0, κ_d and m, we simply choose to fix the molecular units such that the parameters which contain this unit are also of reasonable orders of magnitude. We also chose to use an independent multivariate Gaussian prior over the attractant dynamics parameters. This is chiefly due to its ease of use although further experimentation with different priors, such as log-normal or exponential would be valuable.

3.3.2 Testing the implementation

The implementation can be tested by selecting a set of attractant parameters θ, and then evaluating the function $f_\theta(r, t)$ at various points r and t to determine what the model predicts the observed bias would be, under those parameters. These output values can then be fed into the inference model to try and recover the original parameters θ. If the inference procedure is working as expected, then the true underlying attractant parameters should lie within an area of reasonable probability density in the inferred posterior. The experiment performed was as follows. A set of attractant dynamics parameters were chosen as random. These were then used to determine the implied observed bias at six distances (25, 50, 75, 100, 125 and 150 μm from the wound site) at five times (10, 30, 50, 80 and 120 minutes after wounding) giving 30 space-time pairs with an associated observed bias at each. These were then given to the inference model as normal distributions centred around this observed bias with a fixed standard deviation of 0.02. A diagonal multivariate normal prior was placed over the parameters θ, and the MCMC inference procedure as described was subsequently run, with five different starting points, to infer the posteriors. Figure 3.2 shows the results for four different sets of input parameters.

The results raise a few interesting points. Firstly, it seems that some parameters are 'easier' to infer the posterior of than other. τ, for example, was localised fairly precisely in all four of these runs, with a reasonably narrow posterior, correctly containing the true underlying value. Conversely the procedure was not able to localise R_0, m and q with the same degree

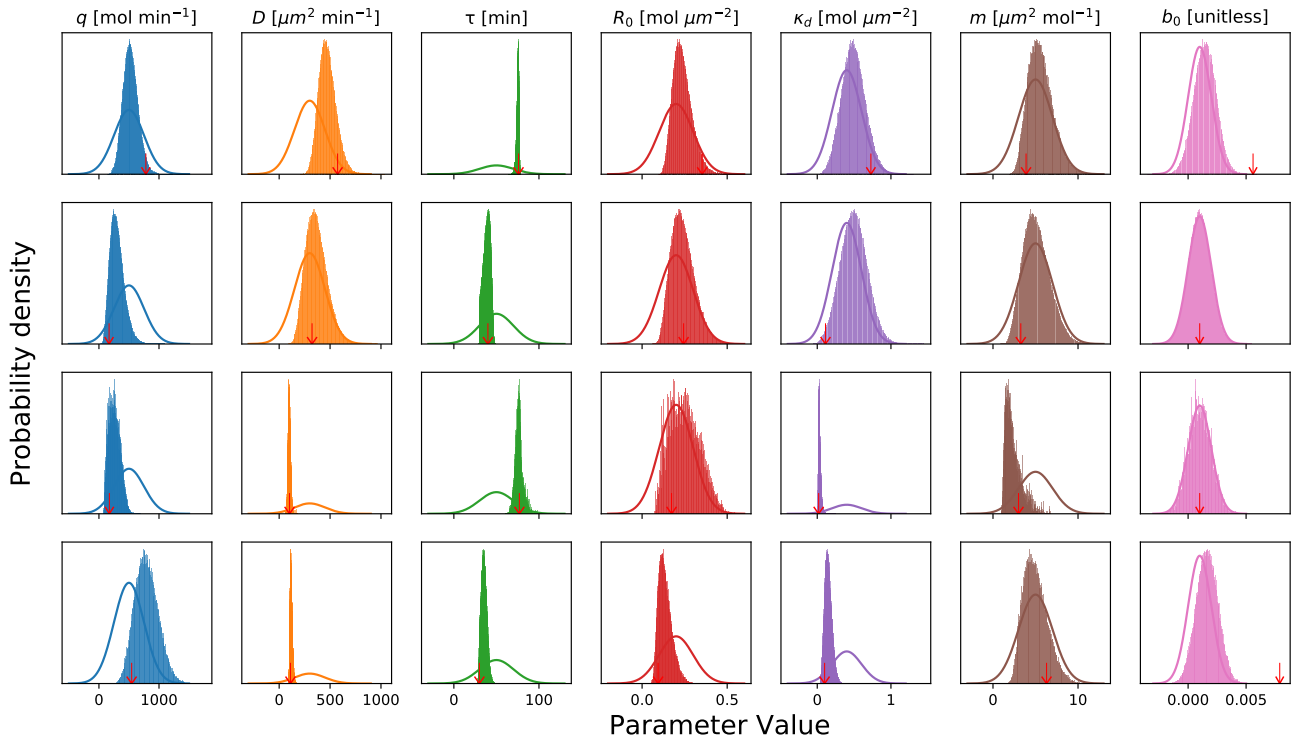

Figure 3.2: Each row shows the results of the inference procedure for a different set of input parameters. The prior distribution is shown as a solid line. The posterior histogram is shown shaded. For each parameter, the red arrow indicates the location of the true underlying value. Find a similar implementation in notebook 4.

of precision. In these cases, although the posterior does contain the true underlying values, the distribution is spread fairly widely. D and κ_d have mixed results in this regard. In particular, the diffusion coefficient, D, which is an important parameter for deducing the approximate molecular size of the unknown attractant, can be localised quite precisely when it is small (below about 250 $\mu m^2 min^{-1}$) but for higher values the posterior is considerably wider. The reason for this becomes evident when one experiments with varying D, and observing the effect this has on the observed bias at various points in time and space (see notebook 4). At low values, small changes have a dramatic effect on the observed bias landscape and that at high values the effect is not nearly as pronounced. This likely reflects the fact that D appears exclusively in denominators in equation 3.4. One final interesting point to note is that the value of b_0 seems to be very sensitive to the prior. Indeed, the posterior matches the prior almost perfectly on every run, with seemingly little regard paid to the true underlying value. This is, however, unlikely to be a major issue since its role in the model is minimal, and the true value can be confirmed to be essentially zero by inferring the bias on unwounded control footage.

Chapter 4

Cell Detection and Tracking

As previously mentioned, a vital stage in the overall computational pipeline is the extraction of cell trajectories from live imaging data; that is, the conversion of a sequence of video frames, represented as 2D arrays of pixel intensity readings, into linked chains of x-y coordinates, which track the motion of the cells. Generally, manual annotation of image data is out of the question, as human annotators are costly and are not guaranteed to be any more reliable than automated approaches. As such, many different software packages have been developed over the years to tackle this problem (Meijering et al., 2012). Two of the most prominent modern tools are OpenCV and Fiji, a distribution of ImageJ (Bradski, 2000; Schindelin et al., 2012). However, in this thesis we develop a custom tool for detecting and tracking particles. This is for two reasons. Firstly, in order to enable easy installation and usage, we are seeking to produce an application written in pure Python, with only a handful of standard accompanying packages. Secondly, it allows us to tailor the algorithm to the specific use case in hand: the detection and tracking of *Drosophila* leukocytes. Weavers et al. (2016) do not give explicit details of their detection and tracking methodology. Thus the following chapter takes a different form from the previous two. Here we approach the problem of cell tracking from fresh, describing the development process and ultimate implementation details of this algorithm. An implementation can be found in notebook 5.

The process of producing labelled trajectories from image data can be seen as a combination of two separate and distinct steps. The first is particle ('blob') detection. For each frame in the input file there will be a set of cells visible, which are generally connected regions of high pixel intensity over a dark background, with a certain level of noise overlaid. Blob detection is the process of converting each 2D array of pixel intensities to a corresponding set of static x-y

coordinates, marking the centres of the visible blobs. The second step is particle tracking. This involves linking together these extracted coordinate readings across adjacent frames to form a set of connected *x*-*y*-*t* trajectories, and then correcting for breaks in the chain due to detection failure to produce a set of fully-formed trajectories. Each of these stages is addressed in turn in the following sections.

4.1 Blob detection

Chenouard et al. (2014) provide a good overview of what may be considered state of the art detection methods, specifically in the context of cell imaging. Three techniques that appear in this paper and repeatedly in the wider literature are the *Laplacian of Gaussian*, the *difference of Gaussians* and the *determinant of Hessian* methods, often referred to by the shorthands LoG, DoG and DoH respectively. In this section we briefly introduce each method and then compare their performance on some sample cell image data.

4.1.1 Laplacian of Gaussian

The Laplacian of Gaussian method is one of the most effective and popular methods of blob detection. It can be understood as a two step process; first the input image f_{in} is smoothed by convolution with a Gaussian kernel of width σ, then the resultant image is convolved with a second derivative Laplacian kernel.

$$f_{\text{out}} = \nabla^2 * (f_{\text{in}} * g) \tag{4.1}$$

In fact, the associative nature of convolution operations means both steps can be combined into a single 'Mexican hat' kernel, which can be constructed for a given value of σ as

$$\text{LoG}(x, y, \sigma) = -\frac{1}{\pi\sigma^4}\left(1 - \frac{x^2+y^2}{2\sigma^2}\right)\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \tag{4.2}$$

(Haralick and Shapiro, 1992). The net effect of convolving an input image with this kernel is to strongly highlight areas in which pixel intensity reaches a maximum, on a length scale of roughly σ. Given that blobs occurring in the image are of an unknown size, a common procedure is to stack convolved images over a range of σ values. Blobs of different radii are then identified by finding the coordinates of the local maxima in this stacked image.

4.1.2 Difference of Gaussians

This approach involves the subtraction of the input image, blurred with a Gaussian kernel of size σ_1, from the same image blurred with a different Gaussian kernel of size σ_2.

$$f_{\text{out}} = f_{\text{in}} * g_{\sigma_1} - f_{\text{in}} * g_{\sigma_2} = f_{\text{in}} * (g_{\sigma_1} - g_{\sigma_2}). \tag{4.3}$$

As the difference between σ_1 and σ_2 becomes smaller, this convolution approaches a scaled version of the Laplacian of Gaussian kernel, however it is possible to construct such that the convolution operation is faster to compute (Lindeberg, 2013). The difference of Gaussians method, therefore, can be seen as an approximation to the more rigorously defined Laplacian of Gaussian method, which may be useful when computational resources are a primary constraint. As before, this convolution operation is generally performed over a range of Gaussian widths, with the coordinates of the local maxima extracted from the resultant three-dimensional image stack.

4.1.3 Determinant of Hessian

The final approach involves constructing a kernel from the determinant of a 2×2 Hessian matrix, $H_{ij} = \partial^2 / \partial x_i \partial x_j$, of the input image, after Gaussian blur convolution.

$$f_{\text{out}} = L_{xx} L_{yy} - L_{xy}^2, \quad \text{where} \quad L = f_{\text{in}} * g_{\sigma} \tag{4.4}$$

(Xu, 2014). In two dimensions, the determinant of the Hessian matrix selects not only for intensity maxima, but also minima and saddle points on length scales similar to σ, making it slightly more general. As with the two previous approaches, the input image is convolved with the DoH kernel over a range of values for σ, stacked into an image block, and the coordinates of the local maxima are extracted to produce the locations of the blobs.

In order to compare the performance of the three approaches on real data, each method was implemented on a single image frame with 73 cells, which had been manually annotated carefully to provide a ground truth. The output can be seen plotted in figure 4.1 and the true positive, false positive and false negative count can be seen in table 4.1.

One clear result that emerges from this analysis is that the determinant of Hessian method is not well-suited to this dataset. The large false positive count is likely due to the minima and

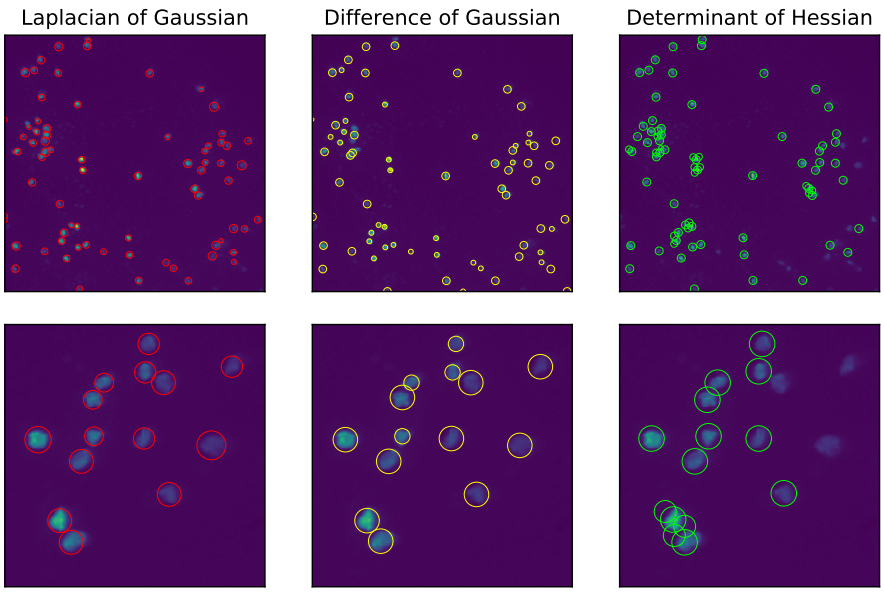

Figure 4.1: Results from the three detection methods are compared on the same 732 $\times$ 893 pixel image. The second row shows the same image, zoomed-in on a particular area. Data courtesy of Wood Lab, Edinburgh.

saddle-point finding properties of this particular approach, which are not necessary for this data, in which cells are purely regions of high intensity on a dark background. This method also has a high false negative count, indicating that it also fails to correctly identify cells in many cases. The Laplacian of Gaussian and difference of Gaussians methods, by contrast, both perform fairly well, with a low false positive count and fairly low false negative count. The Laplacian of Gaussian method slightly outperforms the difference of Gaussians method although, on such a small sample size, it is difficult to say conclusively whether it is a significantly superior method. On the other hand, the computation time for the difference of Gaussians method is significantly lower, taking roughly a quarter of the total time.

As such, the Laplacian of Gaussian method is used in most of the analysis going forward, as it seems to provide marginally better performance. However, the option is provided in the software package to use a difference of Gaussians detector if computation time is a key priority, or hardware capacity is limited.

	TP	FP	FN	Compute time (ms)
Laplacian of Gaussian	69	1	4	681 ± 8
Difference of Gaussian	67	1	6	172 ± 2
Determinant of Hessian	60	18	13	358 ± 8

Table 4.1: The raw number of true positives, false positives and false negatives are shown for each blob detection approach on a single frame of cell imaging data, which contained 73 true cells as labelled manually. The average and standard deviation of the compute time across 50 separate runs is also shown in milliseconds, performed on an intel 7th gen i7 quad core processor.

4.2 Linking detections across frames

Having detected the coordinates of each cell in each frame, the next stage is to link cells together across frames to form trajectories. Numerous methods exist in the literature for this task, ranging from simple nearest neighbour models to the highly sophisticated and accurate (and often prohibitively computationally expensive) Multiple Hypothesis Tracking (MHT) Pulford (2005). In this project we choose to implement a modified version of the Linear Assignment Problem (LAP) approach of Jaqaman et al. (2008), which has been used in cell-specific contexts with success in projects such as TrackMate, a popular extension to ImageJ focussed on computer vision for cell imaging (Tinevez et al., 2017). This approach formulates the task in terms of a series of particle linking problems between consecutive frames by constructing a square 'cost matrix', where the objective is to choose a single element from each column where no two elements are chosen from the same row, such that the sum of the elements is minimised. While a naive solution to this problem would have factorial complexity, Kuhn (1955) developed an algorithm that was able to produce solutions in polynomial time, making fairly large problems viable.

Consider two consecutive frames A and B for which n_A and n_B cell have been detected at 2D coordinates $[\mathbf{x}_1^A, \mathbf{x}_2^A, \ldots, \mathbf{x}_{n_A}^A]$ and $[\mathbf{x}_1^B, \mathbf{x}_2^B, \ldots, \mathbf{x}_{n_B}^B]$ respectively. First consider a simplified case in which $n_A = n_B = n$ and we are certain that no cells have entered or exited the camera view, i.e. there is a one-to-one correspondence between cells in the two frames (we also exclude the possibility of splitting and merging events). For a given motion model, a cell located at position $\mathbf{x}_0 = (x_0, y_0)$ in frame A has a well-defined probability distribution over where it will be in the next frame, $p(\mathbf{x}; \mathbf{x}_0)$. One could, in principle, use any probability distribution here. An

interesting avenue of research would be to experiment with using a biased-persistent random walk model to construct this distribution, however, for simplicity, and because we initially have no information about the parameters w, p and b we follow the methodology of Tinevez et al. (2017) and assume simple two-dimensional Brownian motion.

$$p(\mathbf{x};\mathbf{x}_0) = \mathcal{N}^+(\,|\mathbf{x}-\mathbf{x}_0|;\,\sigma_s) \tag{4.5}$$

From this we construct a square matrix P, where each element P_{ij} is the value of the probability density function for cell i from frame A moving to the position of cell j in frame B, $p(\mathbf{x}_j^B;\mathbf{x}_i^A)$.

$$P = \begin{bmatrix} p(\mathbf{x}_1^B;\mathbf{x}_1^A) & p(\mathbf{x}_2^B;\mathbf{x}_1^A) & \dots & p(\mathbf{x}_n^B;\mathbf{x}_1^A) \\ p(\mathbf{x}_1^B;\mathbf{x}_2^A) & p(\mathbf{x}_2^B;\mathbf{x}_2^A) & \dots & p(\mathbf{x}_n^B;\mathbf{x}_2^A) \\ \vdots & \vdots & \ddots & \vdots \\ p(\mathbf{x}_1^B;\mathbf{x}_n^A) & p(\mathbf{x}_2^B;\mathbf{x}_n^A) & \dots & p(\mathbf{x}_n^B;\mathbf{x}_n^A) \end{bmatrix} = \begin{bmatrix} p_{1\to1} & p_{1\to2} & \dots & p_{1\to n} \\ p_{2\to1} & p_{2\to2} & \dots & p_{2\to n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n\to1} & p_{n\to2} & \dots & p_{n\to n} \end{bmatrix} \tag{4.6}$$

Under the assumption that all cells move independently, the total probability of any cell link configuration will then be the product of n elements, under the condition that no column or row is selected twice. Furthermore, the most likely configuration will be the one, out of all $n!$ possibilities, which maximises this product. This is, in essence, the linear assignment problem. The only differing detail is that the LAP is typically formulated as a sum minimisation rather than product maximisation problem, but this can be adjusted for by taking a negative element-wise logarithm of this probability matrix. Thus, the cost matrix is

$$C = -\log P = \begin{bmatrix} -\log p_{1\to1} & -\log p_{1\to2} & \dots & -\log p_{1\to n} \\ -\log p_{2\to1} & -\log p_{2\to2} & \dots & -\log p_{2\to n} \\ \vdots & \vdots & \ddots & \vdots \\ -\log p_{n\to1} & -\log p_{n\to2} & \dots & -\log p_{n\to n} \end{bmatrix} \tag{4.7}$$

However, in any two frames there are unlikely to be the exact same number of cell detections. The detection method may fail for a given cell on either frame, or a cell may move in or out of the camera's field of vision. Jaqaman et al. (2008) resolve this issue by expanding the cost matrix in both directions, appending an $n_A \times n_A$ matrix as an upper right quadrant and an $n_B \times n_B$ matrix as a lower left quadrant, where, in both cases, the off-diagonal elements are

infinite. The lower right quadrant is then the transpose of the upper left quadrant, resulting in a square $(n_A + n_B) \times (n_A + n_B)$ cost matrix.

$$C' = \left[\begin{array}{cccc|cccc} & & & & d & \infty & \dots & \infty \\ & & & & \infty & d & \dots & \infty \\ & -\log P & & & \vdots & \vdots & \ddots & \vdots \\ & & & & \infty & \infty & \dots & d \\ \hline b & \infty & \dots & \infty & & & & \\ \infty & b & \dots & \infty & & -\log P^\top & & \\ \vdots & \vdots & \ddots & \vdots & & & & \\ \infty & \infty & \dots & b & & & & \end{array}\right]$$

Consider the meaning of the new LAP problem on cost matrix C'. Since the off-diagonal elements in the upper right and lower left quadrants are infinite, they cannot be chosen since this would result in infinite cost. However, selecting the diagonal element in row i from the upper right quadrant indicates that the ith cell from frame A has disappeared (due to either detection failure or moving out of the camera's field of view). Similarly, selecting the diagonal element from the lower left quadrant in column j indicates that the jth particle in frame B has just appeared, and was not present in frame A. As before, selecting element C'_{ij} in the upper left quadrant indicates that cell i from frame A has transitioned to cell j from frame B, however we no longer have the restraint that $n_A = n_B$. The effect of setting the lower right quadrant to the transpose of the upper left quadrant is that, by symmetry, the same transition will be selected (i.e. selections in this quadrant can be ignored).

Jaqaman et al. (2008) suggest setting the diagonal elements in the upper right and lower left quadrants as d (for *death*) and b (for *birth*) respectively, indicating a flat cost across all cells for disappearance and appearance. Under their implementation, the values d and b are hyperparameters which can be varied to encourage or discourage these events. However, the proposed modification we make is to adjust these costs on a per-particle basis. While the probability of detection failure may be fairly constant across cells, the probability of stepping in or out of the camera's field of view will clearly be higher for cells located near the edge. In order to correct for this, we propose setting the diagonal elements as

$$C'_{i,n_B+i} = \log\left[p\left(\mathbf{x}_{\text{out}}(\mathbf{x}_i^A); \mathbf{x}_i^A\right) + d\right], \tag{4.8}$$

and

$$C'_{n_A+j,j} = \log\left[p\left(\mathbf{x}_{\text{out}}(\mathbf{x}_j^B);\mathbf{x}_j^B\right) + b\right] \tag{4.9}$$

where $\mathbf{x}_{\text{out}}(\mathbf{x}_i)$ is the position of the closest edge to the cell located at position $\mathbf{x}_i$. This has the property that as $d \to 0$, we are suggesting that disappearances due to detection failure are impossible, and only movement out of the field of vision could account for differing numbers of cells. Conversely, as d becomes large, we are suggesting that detection failure is more likely (and, correspondingly, the same for b).

4.3 Correcting for chain breaks

As mentioned, on any given frame the detection method may fail to identify a particular cell. This could occur for several reasons. Firstly, cells will have some small motion in the z-direction and thus may fall out of the focal range of the microscope. Additionally, during genetic modification stage, cells may have different uptakes of the Red Fluorescent Protein and thus certain cells may sit on the borderline between being detectable and undetectable (Piatkevich and Verkhusha, 2011). The result is that, since the previous linking procedure only considered adjacent frames, breaks may occur in the trajectories.

We can attempt to correct for this by linking trajectory stops and starts across multiple frames. This too can be achieved in terms of the LAP, by introducing a slight modification. Consider the points $\mathbf{x}_t^d$ and $\mathbf{x}_t^b$ representing the set of coordinates where trajectories end and begin respectively, indexed by the frame number t. We can again construct a matrix of pairwise distances in space, and also a matrix of the same shape holding the pairwise time separations, some of which will be negative. Since the standard deviation of a Brownian motion distribution scales with the square root of time, we can modify the probability density function at position ij such that it accounts for the temporal difference between stop i and start j, by multiplying σ_s by $\sqrt{\Delta t}$. If this difference is negative, the corresponding element of the cost matrix is automatically infinite. We can thus construct a full cost matrix as before and perform the LAP.

4.4 Evaluating the linking algorithm

The full linking algorithm can be tested by generating synthetic data. The procedure is as follows. 70 leukocytes are placed at random coordinates within a box of fixed width and height.

They are then set to walk using a given motion model for 100 steps. If any cell leaves the box boundaries, it is removed from the observations list at that frame. Similarly, there is a random 5% chance at any one frame that a cell is removed from the observation list, corresponding to random detection failure. The linking algorithm is then given 100 sets of cell observations and asked to link the trajectories together.

The performance of the algorithm is tested in two ways. Firstly, the adjacent frame linking stage is evaluated. Each cell observation at frame t can be a trajectory birth, a trajectory death, or can link be linked to a cell in frame $t+1$. The algorithm will either identify a link and connect it with the correct cell, correctly identify a trajectory death, correctly identify a trajectory birth or make a mistake. Thus, the performance can be summarised by a single accuracy metric. Similarly, the broken trajectory linking stage can be evaluated in the same way.

The experiment was run for five different models driving cell motion. First, Brownian motion is tested. Then biased-persistent motion with four sets of underlying parameters w, p and b. For each motion model, data is generated and linked five times, and the mean accuracy is quoted. The results are shown below in table 4.2

	Single-frame linking	Multi-frame linking
Brownian motion	$97.2 \pm 0.5\%$	$87 \pm 1\%$
Biased-persistent (0.5, 0.8, 0.8)	$87 \pm 2\%$	$76 \pm 1\%$
Biased-persistent (0.5, 0.5, 0.5)	$91.5 \pm 0.2\%$	$84 \pm 3\%$
Biased-persistent (0.5, 0.2, 0.8)	$90 \pm 1\%$	$78 \pm 2\%$
Biased-persistent (0.5, 0.8, 0.2)	$91 \pm 3\%$	$77 \pm 3\%$

Table 4.2: The results of the two stages in the linking algorithm are shown. Each experiment is repeated five times and the mean and standard deviation of the accuracy are quoted.

The highest accuracy in both cases is for Brownian motion. This is to be expected, since the underlying assumption in the linking algorithm is that the cells move with Brownian motion.

Chapter 5

Completing the pipeline

In this section we combine the various stages outlined in the project to create an end-to-end pipeline for performing wound repair analysis, and execute it on a real data set. The results are analysed in the context of the results from Weavers et al. (2016).

5.1 The dataset

The live *in vivo* imaging data for this stage of the project was provided by the research laboratory of Professor Will Wood, Edinburgh. A detailed overview of the preparation and collection procedure that was used for this data can be found in Weavers et al. (2018). Each video file was composed of consecutive frames, taken at regular 30 second intervals, containing the microscope's raw pixel intensity readings over an area on the *Drosophila* pupal wing. The leukocytes were genetically modified such that their nuclei contained fluorescent proteins emitting light at a fixed wavelength, and the microscope was adjusted such that it had a narrow window of spectral sensitivity around this frequency, thus the nucleus of each leukocyte appeared as a bright spot. Each image had a width of 732 pixels and a height of 893 pixels, with each pixel spanning 0.2687μm, giving a total real space size of 196.7 $\times$ 240.0μm. In total, footage from eight separate *Drosophila* pupae was provided. Two files were part of a control group for which no wound was present. A further four were wounded by laser incision at a point roughly centrally located. Three of these were filed for a period of two hours (241 frames) post-wounding and the other for one hour (121 frames). In addition, there were two wounded specimens for which the leukocyte RNA had been modified. Unfortunately the details of this genetic procedure were not provided, but we do know that the modification was aimed at impacting the leukocyte chemoattractant signal receptors. Both these files contained 121 frames of footage.

5.2 Methodology and Results

For each separate file we first extract the cell trajectories using the methodology described in chapter 4. As a preliminary experiment, we ran the biased-persistent inference procedure of chapter 2 on the complete trajectories of the two unwounded pupae (with a 'wound' assumed to be located at the central pixel). The posterior distribution, which can be found in appendix section B.0.2, shows no sign of pre-existing bias, as expected, with the b parameter very tightly distributed around zero.

For the four wounded pupae with no genetic modification, we then segment the paths into sub-trajectories falling into specific spatial and temporal bins. Sub-trajectories falling into the same bins across files are then combined, and random walk inference is run on these combined paths to produce a posterior distribution for each of w, p and b for each cluster. Following the methodology of Jones et al. (2015), we chose to use bins which overlapped in both space and time providing a moving window for trajectories. The full posterior distributions, and the corresponding bins, for the random walk parameters can be found in appendix section B.0.2. One notable feature here is that the inferred persistence is highly consistent across all clusters for the wounded specimens, and also with the unwounded data. This is in line with the findings of Weavers et al. (2016), who detect no change in leukocyte persistence upon wounding.

Having measured the random walk parameter posterior distributions at each of these spatial/temporal locations, we can associate a distribution over observed bias, the product of w and b. Each distribution is approximated as an independent Gaussian, and these distributions are passed to the attractant dynamics inference pipeline, outlined in section 3, to infer a posterior distribution over the attractant dynamics parameters. Inference is performed using five separate starting positions, selected randomly from the prior, and run for 500,000 steps with an additional burn-in of 300,000 steps. Three separate wounds types are tested at this point. As mentioned in section 3.1, the effect of multiple cells emitting attractant can be accounted for simply by summing several point sources. First, a single continuous point source located at the wound centre is tested, then a ring of cells located on the cell margin, and finally a lattice of cells within the wound area. The resultant parameter distributions are shown in figure 5.1.

One notable feature of this plot is that the posterior distributions across all three wound types are very consistent. One of the key results of Weavers et al. (2016) is that, by comparing the posterior distributions of flow rate q between small and large wounds and observing that its increase is proportional to wound circumference rather than area, attractant is likely to be emitted from the cell margin. This analysis shows that it is likely impossible to disentangle the

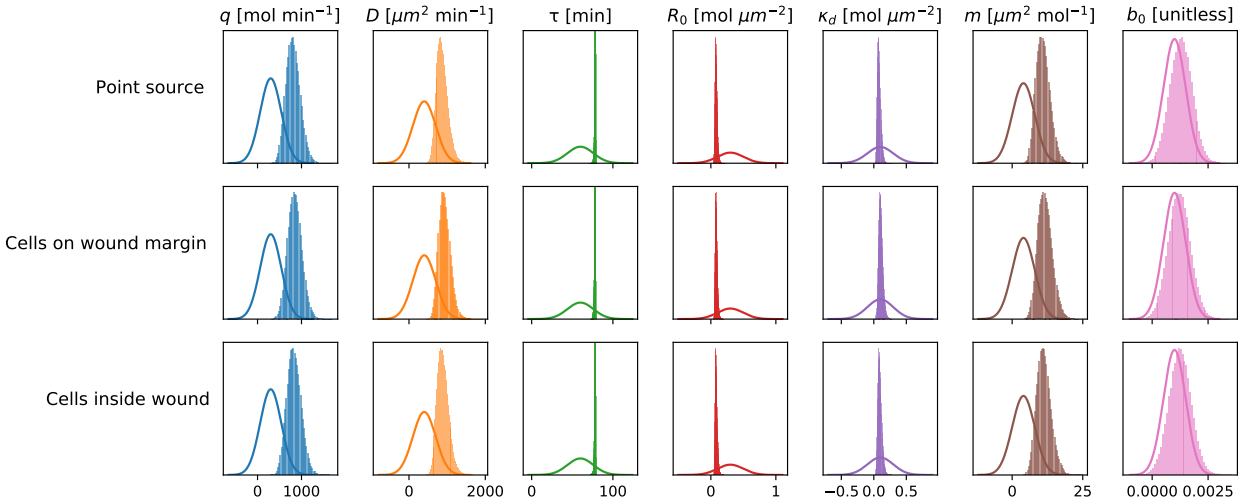

Figure 5.1: The posterior distributions over attractant dynamics parameters are shown for the three wound types. The prior distribution is shown as a solid line. The posterior histogram is shown shaded.

effects of different wound types using data from wounds of a single size.

A second key feature is that for both variables of particular importance, the diffusion coefficient D and production time τ, this analysis is in disagreement with the results of Weavers et al. (2016). They find D to be distributed about a mean value of approximately 200 $\mu m^2 min^{-1}$, whereas the distribution we find is significantly higher, at around 900 $\mu m^2 min^{-1}$. This is owing to the fact that, as is visible in figure 5.2 which shows a boxplot of the observed bias distributions, the observed bias we detect at each time interval is relatively flat across distances. This would imply that the attractant gradient is fairly constant, as would be expected for a fast diffusing chemical. However, a key point to note here is that the magnification of the data we had access to was significantly higher than that of Weavers et al. (2016), meaning they were able to measure leukocyte bias out to distances of 500 μm away from the wound - over three times higher than we were able to. Indeed, one can see from figure 2F-G from the original paper, that their measurements of observed bias too were relatively constant between 0 and 150 μm from the wound. It was only by detecting the significant decline in observed bias at further distances that they were able to fit this function accurately. The posterior we find for the diffusion coefficient D should, therefore, be treated with some degree of scepticism.

The second important parameter, the production time τ, is also at odds with the value found by Weavers et al. (2016). Our posterior is tightly distributed around a value of approximately 78 minutes whereas their posterior is centred around 30 minutes. The reason for this discrepancy

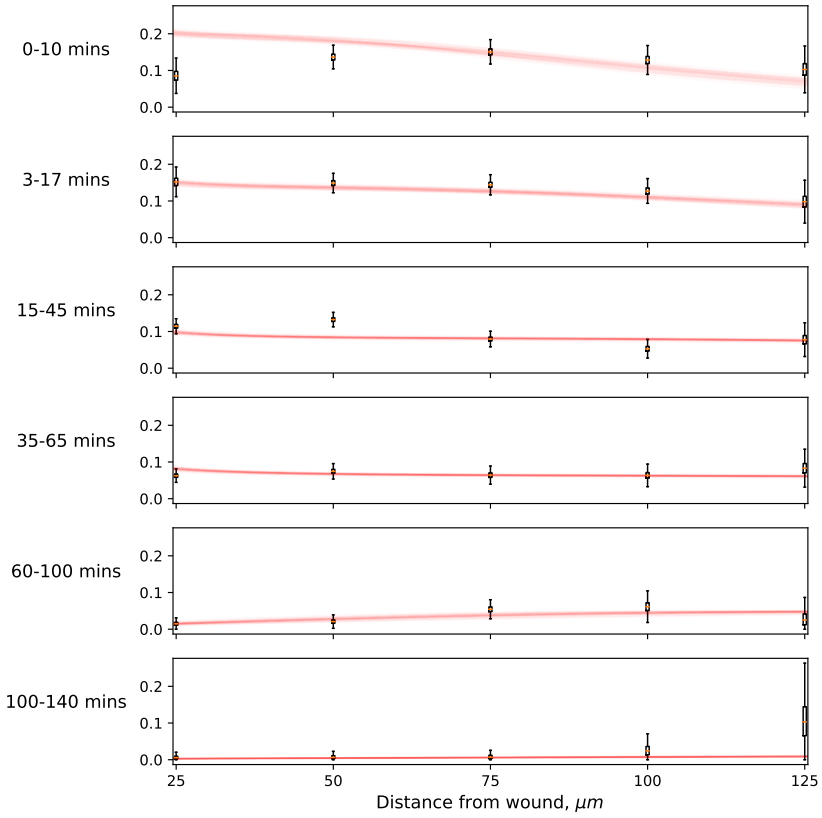

Figure 5.2: The posterior distributions for the observed bias are shown as box plots at each of the temporal bins. For each, 100 samples from the posterior of the attractant dynamics parameters are used to construct the implied observed bias, shown in red.

is less clear. Since we do measure relatively strong bias for leukocytes 60-100 minutes post wounding, our results do seem to indicate a longer attractant production time. However, this may also be an artefact of our probable over-estimation of the diffusion coefficient parameter. If, indeed, D is lower than we predict then bias could still be observed for a longer period post-wounding as a slow-moving chemical diffuses outwards.

We also run the same procedure on the two files showing the RNA-modified leukocytes. In the absence of detail about the modification procedure it is difficult to draw strong conclusions from these results, nevertheless, the observed bias does seem to be somewhat lower in this case. However, due to the fact that data was more limited, the posterior distributions over random walk parameters were considerably wider making differences difficult to discern accurately. The full posterior distributions at each spatial and temporal bin can again be found in appendix section B.0.2. Figure 5.3 shows the attractant parameter posteriors and observed bias box plots for the RNA-modified leukocytes.

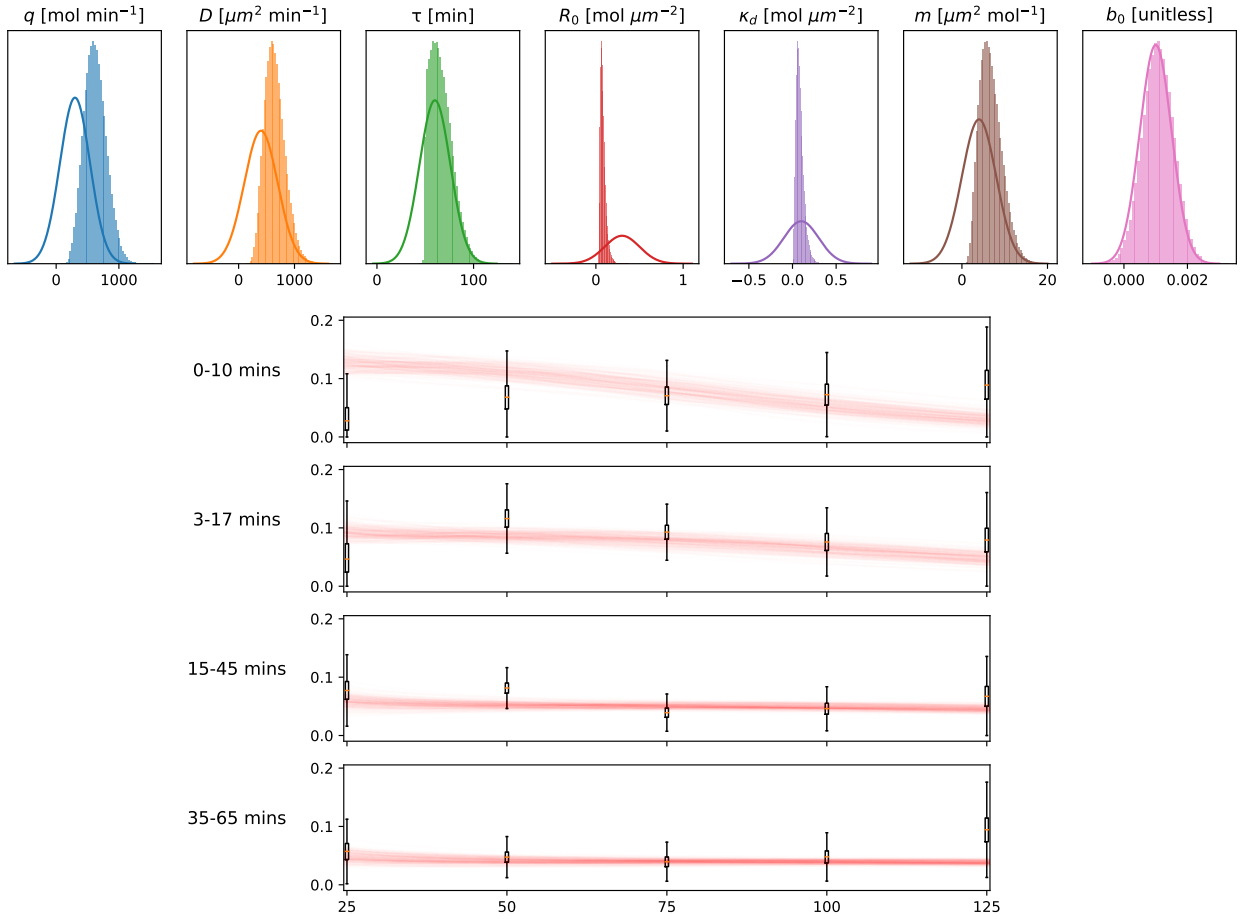

Figure 5.3: (Top) The posterior distributions over attractant dynamics parameters are shown for the leukocytes with modified RNA, this time for a point wound only. (Bottom) The posterior distributions for the observed bias for the RNA-modifed leukocytes are shown as box plots at each of the temporal bins. For each, 100 samples from the posterior of the attractant dynamics parameters are used to construct the implied observed bias, shown in red.

5.3 Discussion

One issue worth mentioning here is that there may be a systematic underestimation of the true effective bias level at short distances. Note the upper plot in figure 5.2. The first spatial bin, which spans 5-45μm from the wound, is particularly poorly fit by the model, appearing much lower than the parameters would imply. This could be due to the fact that, in reality, when leukocytes reach the wound margin they are unable to continue taking steps towards the injury site. Thus, the inferred bias for cells in this region may be low despite a large difference in bound complexes between their front and rear. This is likely true at all times, introducing unknown error on the posterior distributions. One possible solution could be reject trajectories

that fall within a certain radius of the wound position.

Another point of concern relates to the cell position uncertainty analysis of section 2.5.1. The detection method we used is able to localise the position of a cell to within one pixel. However, the step size parameter σ_s, measured in pixels is ~ 8.5. This corresponds to a roughly 12% position uncertainty. At this level, we could expect some distortions on the posteriors for the random walk parameters. We have mentioned already the benefit of using data at a lower magnification as this would allow levels of bias to be analysed at greater distances from the wound. However, there would likely be a trade off between gathering data over a wide area and localising cells to within a reasonable degree of certainty. Ultimately, this could be addressed by using microscopes with a greater resolution, although this will be limited by cost constraints.

Chapter 6

Conclusions

In this project we have reconstructed the computational pipeline, set out in Weavers et al. (2016), for performing inflammatory response analysis using videos of *Drosophila* leukocytes migrating during wound repair. This involved three distinct stages: extracting cell trajectories from the raw microscope footage, inferring properties of the leukocyte motion at different positions and times, and finally using this information to determine properties of the attractant environment.

For the trajectory extraction stage we propose a Laplacian of Gaussian or Difference of Gaussian cell detection method. Our analysis showed that these two methods had similar performance and both outperformed Difference of Hessian. Following the methodology of Jaqaman et al. (2008), we propose formulating the tracking process in terms of the linear assignment problem. One benefit of this method is that it enables different random walk models to describe location transition probabilities. Although we implement this algorithm under the assumption of Brownian cell motion, an interesting future avenue of research here could be to assume biased-persistent motion. We also propose an enhancement of the algorithm of Jaqaman et al. (2008) that accounts for cells' relative probability of leaving the microscope's field of view.

The topic of inferring biased persistent random walk parameters from trajectory data is covered in chapter 2. Here we find that the inference process has greater difficulty localising low values for b and b. We also analyse the effect of position uncertainty, trajectory breaks and frame interval and find that all three can have an impact on the parameter posteriors. We show that cell position uncertainty seems to reduce the inferred bias and persistence, although further experimentation should be done here to quantify this phenomenon more precisely. This highlights the importance of using sufficiently high-resolution microscopy. The effect of using

data captured at different frame intervals would also benefit from further experimentation. In particular, future researchers could explore the possibility of a deterministic map, that would allow posteriors of random walk parameters sampled at different frame rates to be directly compared.

Chapter 3 explored the topic of inferring the posteriors of the attractant dynamics parameters. Here we provide a formula for attractant diffusion in terms of the exponential integral, which can be precompiled for optimised inference speed. We also highlight the importance of prior distributions and physical units. We only perform experiments using a diagonal covariance multivariate prior here, but future research into alternative priors would be highly valuable.

Finally, we execute the computational pipeline on a real dataset. We are not able to reproduce the results of Weavers et al. (2016), finding values for the attractant diffusion rate and signal production time to be considerably higher. However we believe this could be due to the fact that the video footage we had access to was taken at a magnification such that it only showed leukocytes moving up to a distance of approximately 150 μm away from the wound, significantly less than that of Weavers et al. (2016). We suggest that future researchers gather data at greater wound distances in order to fit the attractant dynamics parameters more accurately, but also draw attention to the impact this may have on resolution and therefore cell localisation.

Bibliography

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Carslaw, H. S. and Jaeger, J. C. (1959). *Conduction of heat in solids*. Clarendon Press, Oxford, second edition.. edition.

Chenouard, N., Smal, I., Chaumont, F. D., Maka, M., Sbalzarini, I. F., Gong, Y., Cardinale, J., Carthel, C., Coraluppi, S., Winter, M., Cohen, A. R., Godinez, W. J., Rohr, K., Kalaidzidis, Y., Liang, L., Duncan, J., Shen, H., Xu, Y., Magnusson, K. E. G., Jaldn, J., Blau, H. M., Paul-Gilloteaux, P., Roudot, P., Kervrann, C., Waharte, F., Tinevez, J.-Y., Shorte, S. L., Willemse, J., Celler, K., Wezel, G. P. V., Dan, H.-W., Tsai, Y.-S., Solrzano, C. O. D., Olivo-Marin, J.-C., and Meijering, E. (2014). Objective comparison of particle tracking methods. *Nature Methods*, 11(3).

Codling, E. A., Plank, M. J., and Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal Society, Interface*, 5(25):813–34.

Cohnheim, J. (1867). Ueber entzündung und eiterung. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, 40(1):1–79.

Evans, I. R., Rodrigues, F. S. L. M., Armitage, E. L., and Wood, W. (2015). Draper/ced-1 mediates an ancient damage response to control inflammatory blood cell migration in vivo. pages 1606–1612. ISSN 0960–9822.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge.

Fourier, J. B. J. (1822). *Thorie Analytique de la Chaleur*. Cambridge library collection. Mathematics. publisher not identified, Place of publication not identified.

Haralick, R. M. and Shapiro, L. G. (1992). *Computer and robot vision*, volume 1. Addison-wesley Reading.

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S. L., and Danuser,

G. (2008). Robust single-particle tracking in live-cell time-lapse sequences. *Nature Methods*, 5(8).

Jones, P. J. M., Sim, A., Taylor, H. B., Bugeon, L., Dallman, M. J., Pereira, B., Stumpf, M. P. H., and Liepe, J. (2015). Inference of random walk models to describe leukocyte migration. *Physical Biology*, 12(6):066001.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lauffenburger, D. A. and Linderman, J. J. (1993). *Receptors: models for binding, trafficking, and signaling*. Oxford University Press.

Lawler, G. F. (2010). *Random Walk: A Modern Introduction*. Cambridge Studies in Advanced Mathematics ; no. 123. Cambridge University Press, Cambridge.

Liepe, J., Taylor, H., Barnes, C. P., Huvet, M., Bugeon, L., Thorne, T., Lamb, J. R., Dallman, M. J., and Stumpf, M. P. (2012). Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate Bayesian computation. *Integrative biology : quantitative biosciences from nano to macro*, 4(3):335–345.

Lindeberg, T. (2013). Image matching using generalized scale-space interest points. *Scale Space And Variational Methods In Computer Vision: 4th International Conference*, 7893:355–367.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Masayuki, M. and Kiyoshi, T. (2016). *Chronic Inflammation: Mechanisms and Regulation*.

Meijering, E., Dzyubachyk, O., and Smal, I. (2012). Methods for cell and particle tracking. *Methods in enzymology*, 504:183–200.

Moreira, S., Stramer, B., Evans, I., Wood, W., and Martin, P. (2010). Prioritization of competing damage and developmental signals by migrating macrophages in the drosophila embryo. *Current Biology*, 20(5):464–470.

Niethammer, P., Grabher, C., Look, A. T., and Mitchison, T. J. (2009). A tissue-scale gradient of hydrogen peroxide mediates rapid wound detection in zebrafish. *Nature*, 459(7249):996.

Pattle, R. (1959). Diffusion from an instantaneous point source with a concentration-dependent coefficient. *Quarterly Journal of Mechanics and Applied Mathematics*, 12(4):407–409.

Piatkevich, K. D. and Verkhusha, V. V. (2011). Guide to red fluorescent proteins and biosensors for flow cytometry. *Methods in Cell Biology*, 102:431–461.

Pulford, G. (2005). Taxonomy of multiple target tracking methods. *Iee Proceedings-Radar Sonar And Navigation*, 152(5):291–304.

Razzell, W., Wood, W., and Martin, P. (2011). Swatting flies: modelling wound healing and inflammation in drosophila. *Disease Models & Mechanisms*, 4(5):569–574.

Rosser, G., Fletcher, A. G., Maini, P. K., and Baker, R. E. (2013). The effect of sampling rate on observed statistics in a correlated random walk. *Journal of the Royal Society, Interface*, 10(85).

S. Patlak, C. (1953). Patlak c.s.: Random walk with persistence and external bias. bull. math. biophys. 15, 311-338. *Bulletin of Mathematical Biology - BULL MATH BIOL*, 15:311–338.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676.

Sim, A., Liepe, J., and Stumpf, M. P. H. (2015). Goldstein-kac telegraph processes with random speeds: Path probabilities, likelihoods, and reported lévy flights. *Phys. Rev. E*, 91:042115.

Tinevez, J.-Y., Perry, N., Schindelin, J., Hoopes, G. M., Reynolds, G. D., Laplantine, E., Bednarek, S. Y., Shorte, S. L., and Eliceiri, K. W. (2017). Trackmate: An open and extensible platform for single-particle tracking. *Methods*, 115:80 – 90. Image Processing for Biologists.

Weavers, H., Franz, A., Wood, W., and Martin, P. (2018). Long-term In Vivo Tracking of Inflammatory Cell Dynamics Within Drosophila Pupae. *Journal of Visualized Experiments*, (136).

Weavers, H., Liepe, J., Sim, A., Wood, W., Martin, P., and Stumpf, M. P. (2016). Systems Analysis of the Dynamic Inflammatory Response to Tissue Damage Reveals Spatiotemporal Properties of the Wound Attractant Gradient. *Current Biology*, 26(15):1975–1989.

Whicher, J. T. and Evans, S. W. (1992). *Biochemistry of inflammation*. Immunology and medicine series; v.18. Kluwer Academic, Dordrecht ; London.

Xu, X. (2014). Blob detection with the determinant of the hessian. *Communications in Computer and Information Science*, 483:72–80.

Appendix A

Derivations and Proofs

A.1 The heat equation

A.1.1 Diffusion of a fixed, finite quantity of attractant

For a finite quantity of attractant Q_0 released at $t = 0$ at the point $\mathbf{r}'$ into an empty environment, the stated evolution of the system as per the heat equation is

$$A(\mathbf{r},t) = \frac{Q_0}{(4\pi Dt)^{d/2}} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big). \tag{A.1}$$

with boundary conditions of zero concentration at infinity. First, as a simple sanity check, observe that the total quantity of the substance present in the system at any time is conserved.

$$
\begin{aligned}
Q &= \frac{Q_0}{(4\pi Dt)^{d/2}} \int_{-\infty}^{\infty} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big)\,\mathrm{d}\mathbf{r} \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \int_{-\infty}^{\infty} \exp\big(-\frac{(x_1-x_1')^2+(x_2-x_2')^2+\ldots+(x_d-x_d')^2}{4Dt}\big)\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_d \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \int_{-\infty}^{\infty} \Big[\prod_{j=1}^{d} \exp\big(-\frac{(x_j-x_j')^2}{4Dt}\big)\Big]\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_d \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \prod_{j=1}^{d} \int_{-\infty}^{\infty} \exp\big(-\frac{(x_j-x_j')^2}{4Dt}\big)\,\mathrm{d}x_j \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} (\sqrt{4\pi Dt})^d \\
&= Q_0 \quad (\text{for } t \geq 0)
\end{aligned}
\tag{A.2}
$$

This also satisfies the initial condition. Note also that the boundary conditions are always satisfied since, as $|\mathbf{r}-\mathbf{r}'|^2$ goes to infinity, the concentration goes to zero. Now consider the partial derivative the expression with respect to time.

$$
\begin{aligned}
\frac{\partial A(\mathbf{r},t)}{\partial t} &= \frac{Q_0}{(4\pi D)^{d/2}} \frac{\partial}{\partial t}\Big[t^{-d/2}\exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big)\Big] \\
&= \frac{Q_0}{(4\pi D)^{d/2}} \Big(\frac{\partial}{\partial t}\big[t^{-d/2}\big]\exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) + t^{-d/2}\frac{\partial}{\partial t}\big[\exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big)\big]\Big) \\
&= \frac{Q_0}{(4\pi D)^{d/2}} \Big(-\frac{d}{2}t^{-d/2-1}\exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) + t^{-d/2}\frac{|\mathbf{r}-\mathbf{r}'|^2}{4D}t^{-2}\exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big)\Big) \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \big(\frac{|\mathbf{r}-\mathbf{r}'|^2}{4D}t^{-2} - \frac{d}{2}t^{-1}\big)\exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) \\
&= \Big(\frac{|\mathbf{r}-\mathbf{r}'|^2}{4D}t^{-2} - \frac{d}{2}t^{-1}\Big)A(\mathbf{r},t)
\end{aligned}
\tag{A.3}
$$

And now consider the Laplacian of the concentration.

$$
\begin{aligned}
\nabla^2 A(\mathbf{r},t) &= \frac{Q_0}{(4\pi Dt)^{d/2}} \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \prod_{j=1}^{d} \exp\big(-\frac{(x_j-x_j')^2}{4Dt}\big) \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \big[(-2\frac{x_i-x_i'}{4Dt}) \prod_{j=1}^{d} \exp\big(-\frac{(x_j-x_j')^2}{4Dt}\big)\big] \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \sum_{i=1}^{d} \Big[\frac{\partial}{\partial x_i}\big[-\frac{x_i-x_i'}{2Dt}\big] \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) - \frac{x_i-x_i'}{2Dt}\frac{\partial}{\partial x_i}\big[\prod_{j=1}^{d} \exp\big(-\frac{(x_j-x_j')^2}{4Dt}\big)\big]\Big] \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \sum_{i=1}^{d} \Big[-\frac{1}{2Dt} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) + \frac{(x_i-x_i')^2}{4D^2t^2} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big)\Big] \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) \frac{1}{D} \sum_{i=1}^{d} \Big[\frac{(x_i-x_i')^2}{4D} t^{-2} - \frac{1}{2} t^{-1}\Big] \\
&= \frac{Q_0}{(4\pi Dt)^{d/2}} \exp\big(-\frac{|\mathbf{r}-\mathbf{r}'|^2}{4Dt}\big) \frac{1}{D} \Big(\frac{|\mathbf{r}-\mathbf{r}'|^2}{4D} t^{-2} - \frac{d}{2} t^{-1}\Big) \\
&= \frac{1}{D}\Big(\frac{|\mathbf{r}-\mathbf{r}'|^2}{4D} t^{-2} - \frac{d}{2} t^{-1}\Big) A(\mathbf{r},t)
\end{aligned}
\tag{A.4}
$$

Therefore, one can see that

$$
\frac{\partial A(\mathbf{r},t)}{\partial t} = D\nabla^2 \left[A\left(\mathbf{r},t\right)\right]. \tag{A.5}
$$

Thus, the heat equation and all boundary and initial conditions are satisfied.

A.1.2 Diffusion from a continuous point source.

As stated in section 3.1, the continuous point source case can be found by integrating the finite quantity case over time.

$$
A(\mathbf{r},t) = \frac{q}{(4\pi D)^{d/2}} \int_0^{\min(\tau,t)} \exp\Big(-\frac{r^2}{4D(t-t')}\Big) \frac{\mathrm{d}t'}{(t-t')^{d/2}} \tag{A.6}
$$

Carslaw and Jaeger (1959) note that, in any number of dimensions other than $d = 2$, a solution in terms of the error function $\mathrm{Erf}(x)$ may be found by making the substitution

$$\mathrm{d}u = \frac{\mathrm{d}t'}{(t-t')^{d/2}}, \rightarrow u = \frac{2}{d-2}(t-t')^{1-\frac{d}{2}}. \tag{A.7}$$

However, one can see that this fails for the case of interest to us: two dimensions. Here

$$A(r,t) = \frac{q}{4\pi D}\int_0^{\min(\tau,t)} \exp\Big(-\frac{r^2}{4D(t-t')}\Big)\frac{\mathrm{d}t'}{t-t'}. \tag{A.8}$$

As stated, this can in fact be reformulated in terms of the exponential integral $\mathrm{Ei}(x)$, defined as

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty}\frac{\mathrm{e}^{-z}}{z}\mathrm{d}z \;=\; \int_{-\infty}^{x}\frac{\mathrm{e}^{z}}{z}\mathrm{d}z. \tag{A.9}$$

In order to write this in terms of the exponential integral $\mathrm{Ei}(x)$, first make the substitution

$$u = -\frac{r^2}{4D(t-t')} \;\rightarrow\; \mathrm{d}t' = -\frac{r^2}{4D}u^{-2}\mathrm{d}u. \tag{A.10}$$

Then

$$
\begin{aligned}
A(\mathbf{r},t) &= \frac{q}{4\pi D}\int_{t'=0}^{t'=\min(\tau,t)} e^{u}(-u\frac{4D}{r^2})(-u^{-2}\frac{r^2}{4D})\,\mathrm{d}u \\
&= \frac{q}{4\pi D}\int_{t'=0}^{t'=\min(\tau,t)} \frac{e^{u}}{u}\,\mathrm{d}u \\
&= \frac{q}{4\pi D}\Big[\int_{t'=-\infty}^{t'=\min(\tau,t)} \frac{e^{u}}{u}\,\mathrm{d}u - \int_{t'=-\infty}^{t'=0} \frac{e^{u}}{u}\,\mathrm{d}u\Big] \\
&= \frac{q}{4\pi D}\Big[\mathrm{Ei}\big(-\frac{r^2}{4D(t-\min(\tau,t))}\big) - \mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big] \\
&= \begin{cases} \frac{q}{4\pi D}\Big(\mathrm{Ei}(-\infty) - \mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big) & \text{if } t<\tau \\ \frac{q}{4\pi D}\Big(\mathrm{Ei}\big(-\frac{r^2}{4D(t-\tau)}\big) - \mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big) & \text{if } t>\tau \end{cases} \\
&= \begin{cases} -\frac{q}{4\pi D}\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big) & \text{if } t<\tau \\ \frac{q}{4\pi D}\Big(\mathrm{Ei}\big(-\frac{r^2}{4D(t-\tau)}\big) - \mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big) & \text{if } t>\tau \end{cases}
\end{aligned}
$$

This satisfies the boundary conditions, since

$$\lim_{r\to\infty}\Big[\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big] = 0 \tag{A.11}$$

In order to prove that it satisfies the initial condition, we must show that the total quantity of attractant in the system for $t<\tau$ is qt, and, for $t>\tau$, is $q\tau$. Again, the total quantity of attractant is found by integrating the concentration over all space. For $t<\tau$,

$$
\begin{aligned}
Q_{t<\tau} &= -\frac{q}{4\pi D}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\,\mathrm{d}x\,\mathrm{d}y \\
&= -\frac{q}{4\pi D}\int_{0}^{2\pi}\int_{0}^{\infty}\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\,r\,\mathrm{d}r\,\mathrm{d}\theta \\
&= -\frac{q}{4\pi D}2\pi\Big[\frac{1}{2}\Big(r^2\,\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big) + 4Dt\,\exp\big(-\frac{r^2}{4Dt}\big)\Big)\Big]_{r=0}^{r=\infty}
\end{aligned}
$$

Note that both

$$\lim_{r\to\infty}\Big[r^2\,\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big] = 0, \quad \text{and} \quad \lim_{r\to 0}\Big[r^2\,\mathrm{Ei}\big(-\frac{r^2}{4Dt}\big)\Big] = 0$$

Therefore

$$\begin{aligned} Q_{t<\tau} &= -\frac{q}{4\pi D} 2\pi \frac{1}{2}(-4Dt) \\ &= qt. \end{aligned}$$

Using the same method, for $t > \tau$

$$\begin{aligned} Q_{t>\tau} &= -q(t-\tau) - qt \\ &= q\tau \end{aligned}$$

To complete the proof, we must now show that the solution satisfies the heat equation. As before, first consider the derivative of the concentration with respect to time. Since it is defined in terms of an integral with respect to time, the answer is trivial.

$$\frac{\partial A(r,t)}{\partial t} = \frac{q}{4\pi Dt} \exp\big(-\frac{r^2}{4Dt}\big) \tag{A.12}$$

Next we must compute the Laplacian of the concentration. First consider the second partial derivative with respect to x. For $t < \tau$,

$$\begin{aligned} \frac{\partial^2}{\partial x^2} A(r,t) &= -\frac{q}{4\pi D} \frac{\partial^2}{\partial x^2} \mathrm{Ei}\big(-\frac{x^2+y^2}{4Dt}\big) \\ &= -\frac{q}{4\pi D} \frac{\partial}{\partial x} \frac{2x}{x^2+y^2} \exp\big(-\frac{r^2}{4Dt}\big) \\ &= -\frac{q}{4\pi D} \Big(\frac{2(x^2+y^2)-2x(2x)}{(x^2+y^2)^2} - \frac{2x(2x)}{4Dt(x^2+y^2)}\Big) \exp\big(-\frac{r^2}{4Dt}\big) \end{aligned}$$

Therefore the Laplacian, as the sum of the second partial derivatives with respect to x and y, is

$$\begin{aligned}
\nabla^2 A(r,t) &= -\frac{q}{4\pi D}\Big(\frac{2r^2-4x^2}{r^4} - \frac{4x^2}{4Dtr^2} + \frac{2r^2-4y^2}{r^4} - \frac{4y^2}{4Dtr^2}\Big)\exp\big(-\frac{r^2}{4Dt}\big) \\
&= -\frac{q}{4\pi D}\Big(\frac{4r^2-4r^2}{r^4} - \frac{4r^2}{4Dtr^2}\Big)\exp\big(-\frac{r^2}{4Dt}\big) \\
&= -\frac{q}{4\pi D}\big(-\frac{1}{Dt}\big)\exp\big(-\frac{r^2}{4Dt}\big) \\
&= \frac{1}{D}\frac{q}{4\pi Dt}\exp\big(-\frac{r^2}{4Dt}\big).
\end{aligned}$$

The solution thus satisfies the heat equation and all boundary and initial conditions.

A.2 Receptor-ligand binding kinetics

The following section is extracted directly from the appendix of my IPP report.

Equation 3.5 can be derived from the assumption that single attractant particle binds with a single receptor to form a complex. At any point in space one can denote the concentration of free receptors as R, the concentration of free attractant as A, and the concentration of bound complexes as C. In addition, one can denote the total (both bound and unbound) concentration of receptors and attractant as R_0 and A_0 respectively. Thus, by definition,

$$R_0 = R + C \quad \text{and} \quad A_0 = A + C$$

which are, in equilibrium, both conserved quantities, as attractant and receptors bind and unbind. One can model the rate at which bound complex concentration is changing as

$$\begin{aligned}
\frac{\mathrm{d}C}{\mathrm{d}t} &= k_{\text{on}} RA - k_{\text{off}} C \\
&= \underbrace{k_{\text{on}}(R_0 - C)(A_0 - C)}_{\text{new pairs bind}} \underbrace{- k_{\text{off}} C}_{\text{pairs unbind}},
\end{aligned}$$

where k_{on} and k_{off} are constants quantify the rate at which attractant-receptor pairs bind and unbind respectively. When the system reaches equilibrium this rate of change will be zero. Denoting a new constant $\kappa_d = k_{\text{off}}/k_{\text{on}}$ we have

$$\begin{aligned} 0 &= (R_0 - C)(A_0 - C) - \kappa_d C \\ &= C^2 - C(R_0 + A_0 + \kappa_d) + R_0 A_0. \end{aligned}$$

Applying the quadratic formula gives

$$\begin{aligned} C &= \frac{1}{2}\big((R_0 + A_0 + \kappa_d) \pm \sqrt{(R_0 + A_0 + \kappa_d)^2 - 4R_0A_0}\big) \\ &= \frac{1}{2}(R_0 + A_0 + \kappa_d) \pm \sqrt{\frac{1}{4}(R_0 + A_0 + \kappa_d)^2 - R_0A_0} \end{aligned}$$

Both the positive and negative solutions represent valid steady states, however, only the negative solution is actualised since, as complex concentration increases, the lower value is reached first. As previously stated, this equation is true for any position in space. Thus, now denoting the total attractant concentration at position x as $A(x)$, and complex concentration as $C(x)$, we arrive at equation 4.

$$C(r,t) = \frac{1}{2}(\kappa_d + R_0 + A(r,t)) - \sqrt{\frac{1}{4}\big(\kappa_d + R_0 + A(r,t)\big)^2 - R_0 A(r,t)}$$

Appendix B

Supplementary figures

B.0.1 Random walk inference: joint distributions

The following figures accompany figure 2.4 from section 2.4 on testing parameter inference on random walks.

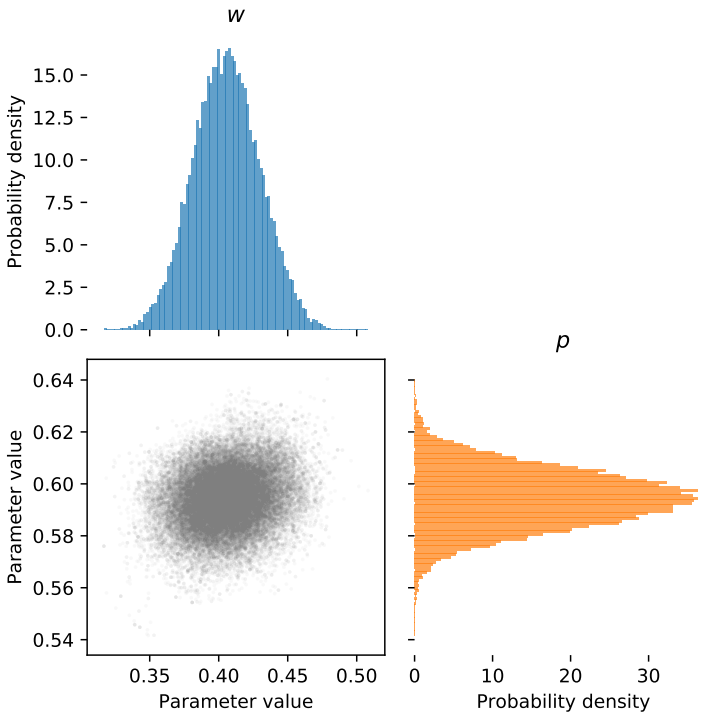

Figure B.1: The joint posterior distribution for w and p, when the true underlying parameters are 0.4 and 0.6 respectively.

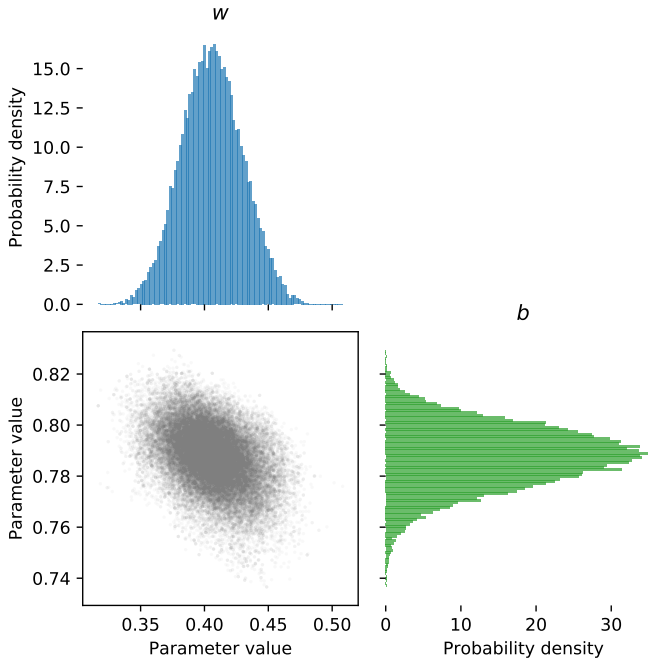

Figure B.2: The joint posterior distribution for w and b, when the true underlying parameters are 0.4 and 0.8 respectively.

B.0.2 Random walk posteriors

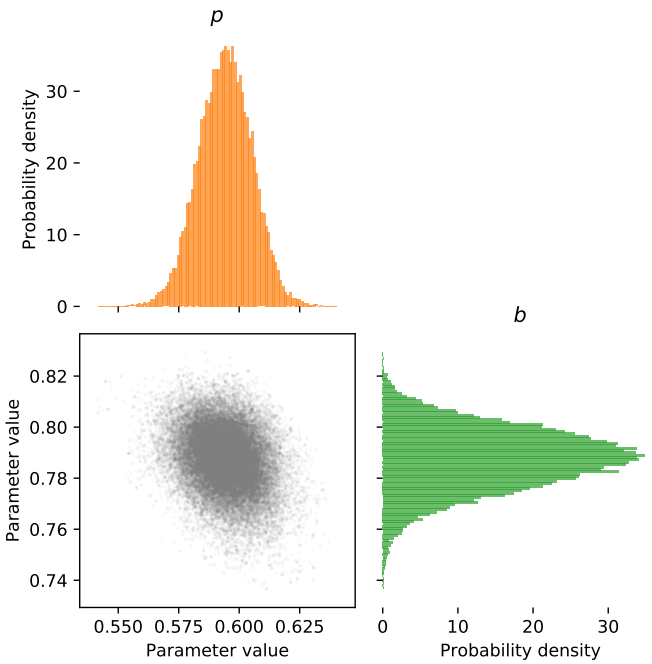

Figure B.3: The joint posterior distribution for p and b, when the true underlying parameters are 0.6 and 0.8 respectively.

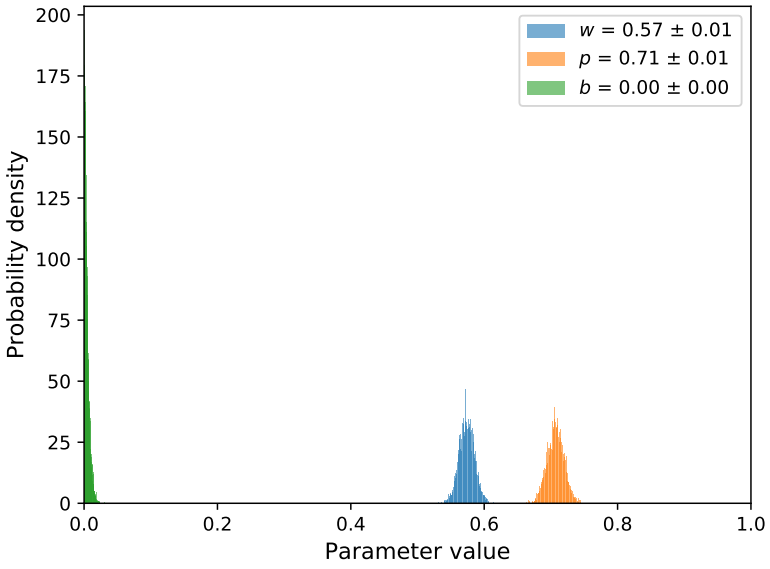

Figure B.4: The posterior distributions over the random walk parameters w, p and b are shown for the trajectories taken from the two unwounded movies.

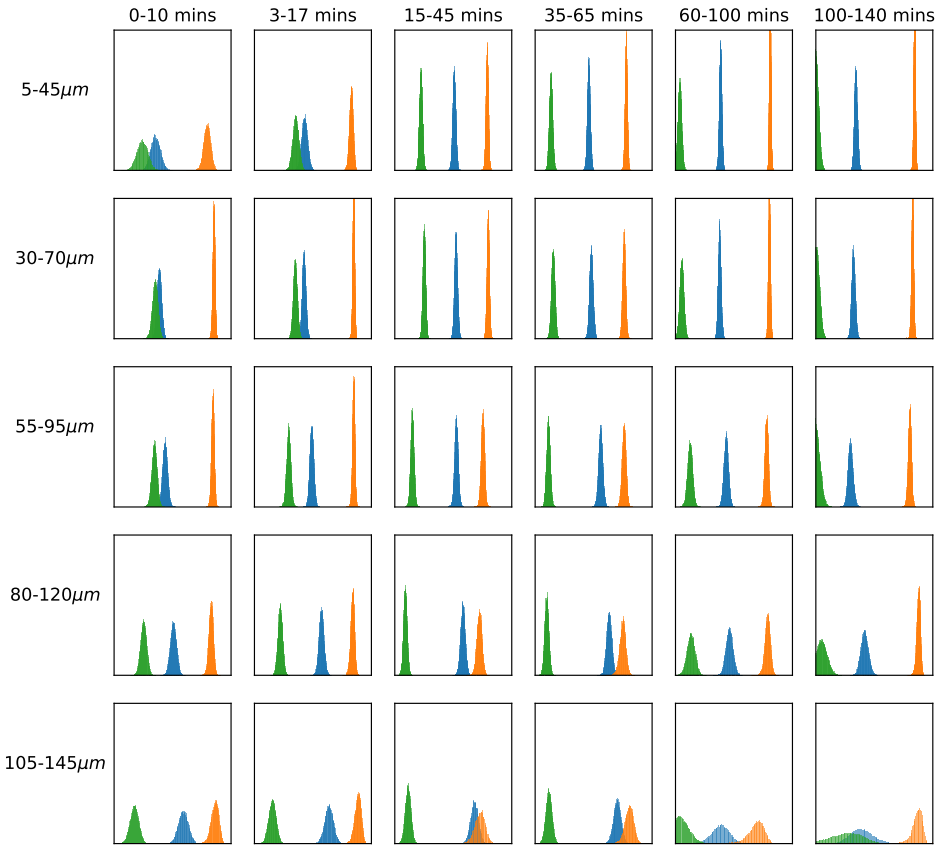

Figure B.5: The posterior distributions over the random walk parameters w (blue), p (orange) and b (green) for the regular, wounded specimens are shows for each of the temporal and spatial bins. Each x-axis runs from 0-1, and each y-axis, indicating probability density, has been set to the same scale.

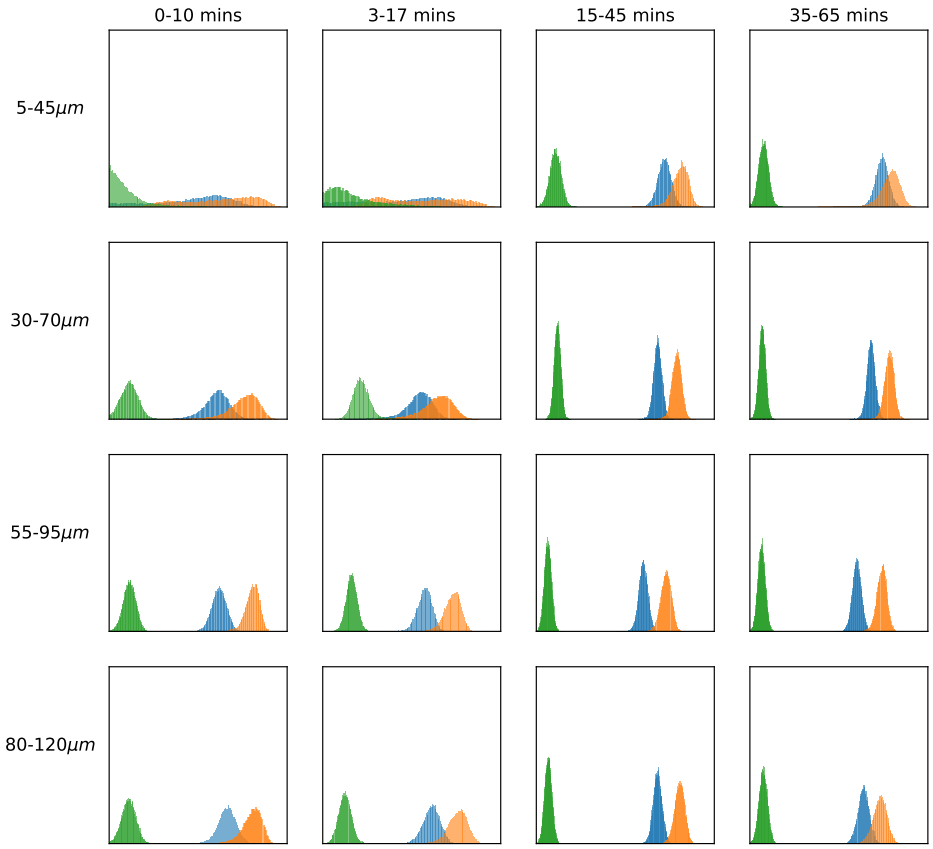

Figure B.6: The posterior distributions over the random walk parameters w (blue), p (orange) and b (green) for the specimens with modified RNA are shows for each of the temporal and spatial bins. Each x-axis runs from 0-1, and each y-axis, indicating probability density, has been set to the same scale.