Physical Biology

OFFICIAL JOURNAL OF
Sociedad de Biofísicos
Latino Americanos

PAPER • OPEN ACCESS

Inference of random walk models to describe leukocyte migration

Physical Biology

PAPER

Inference of random walk models to describe leukocyte migration

Phoebe J M Jones[1,2,5], Aaron Sim[1,2,5], Harriet B Taylor[1,3], Laurence Bugeon[1], Magaret J Dallman[1,2], Bernard Pereira[4], Michael P H Stumpf[1,2,6] and Juliane Liepe[1,2,6]

[1] Department of Life Sciences, Imperial College London, London, SW7 2AZ, UK
[2] Centre for Integrative Systems Biology and Bioinformatics, Imperial College London, SW7 2AZ, UK
[3] Division of Developmental Biology, MRC National Institute for Medical Research, London, UK
[4] Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK
[5] These authors contributed equally.
[6] Author to whom corresponding should be addressed.

E-mail: **m.stumpf@imperial.ac.uk** **and** **juliane.liepe08@imperial.ac.uk**

OPEN ACCESS

RECEIVED
15 February 2015

REVISED
13 July 2015

ACCEPTED FOR PUBLICATION
13 July 2015

PUBLISHED
25 September 2015

Keywords: approximate Bayesian computation, innate immune signalling, *in vivo* imaging

Supplementary material for this article is available online

Abstract

While the majority of cells in an organism are static and remain relatively immobile in their tissue, migrating cells occur commonly during developmental processes and are crucial for a functioning immune response. The mode of migration has been described in terms of various types of random walks. To understand the details of the migratory behaviour we rely on mathematical models and their calibration to experimental data. Here we propose an approximate Bayesian inference scheme to calibrate a class of random walk models characterized by a specific, parametric particle re-orientation mechanism to observed trajectory data. We elaborate the concept of transition matrices (TMs) to detect random walk patterns and determine a statistic to quantify these TM to make them applicable for inference schemes. We apply the developed pipeline to *in vivo* trajectory data of macrophages and neutrophils, extracted from zebrafish that had undergone tail transection. We find that macrophage and neutrophils exhibit very distinct biased persistent random walk patterns, where the strengths of the persistence and bias are spatio-temporally regulated. Furthermore, the movement of macrophages is far less persistent than that of neutrophils in response to wounding.

1. Introduction

Random walk models are often applied in biology to investigate movement of particles, cells or whole organisms (Codling *et al* 2008). They are often described as uncorrelated random walks with diffusion (Berg 1993) or Levy flights (Edwards *et al* 2007). Recently it was reported that living mammary epithelial cells in a tissue display a bimodal persistent random walk (Potdar *et al* 2010). The intracellular and extracellular signalling processes that lead to migration of cells have been studied over the last few decades. Examples include the PI3K signalling cascade as well as the MAPK pathway with the latter investigated by video microscopy of migrating neutrophils after p38 inhibition (?, Liepe *et al* 2012). In order to fully understand the nature of such signalling processes we need to link classical random walk descriptions via biophysical parameters to experimental data. It is then a necessity to be able to estimate parameters of random walk models from experimental data. A typical example is the estimation of the diffusion coefficient from observations of spatial displacements over time (Wilke and Lee 1955, Dohnal 1987, Neisyy 2008). In most studies, such parameter estimation is restricted to models of movement where the likelihood, i.e. the probability of observing the data for a given parameter set, can be calculated easily (typically because it is available in closed form) or in approximation (Sim *et al* 2015). However, as more realistic and complex models emerge, such as multi-scale and agent-based models (An *et al* 2009, Horstemeyer 2010, Dada and Mendes 2011, Liepe *et al* 2012), obtaining these likelihoods becomes either impossible or computationally prohibitive for the purposes of parameter inference and model selection. For this reason, one is then compelled to adopt likelihood-free methods, such as the approximate Bayesian computation (ABC) framework (Toni *et al* 2009, Turner and van Zandt 2012, Wilkinson 2013). The key challenge in such ABC approaches is to define a statistic and its accompanying distance function that most effectively captures and quantifies the differences between simulated data from models with different parameters.

To distinguish different types of random walk, such as Brownian motion, biased random walks or correlated (persistent) random walks, the most commonly used summary statistics are the distribution of step lengths, the mean square displacement and the autocorrelation function of the turning angles (Berg 1993, Codling *et al* 2008). These statistics can be applied in a straightforward manner when analytic expressions of the expected statistics are known. However, especially in biological contexts, the movement of particles, cells or animals is highly non-trivial and cannot be sufficiently well described by any single standard random walk model. At the very least, one needs to combine several such simple models to form more realistic hybrid models, albeit at the cost of a loss in analytic tractability. The commonly used statistics fail to capture the details of the walks resulting from such models. Therefore we need to define a statistic that is sufficient to describe the dynamic behaviour of the model and easy to compute. To do so we have developed transition matrices (TMs) and established their usefulness in a previous study (Taylor *et al* 2013). As we review in greater detail in the next section, these TM are summaries of the joint probabilities $P(\alpha_t, \alpha_{t+1})$ of the random walk turning angles $-\pi \leqslant \alpha_t < \pi$ at all times t. Even though the TMs are not analytically tractable, they contain more information about the random walk behaviour than previously used statistics.

One attractive feature of TMs is their simple visual representation as heat maps. For some random walk models, one can often distinguish specific patterns of movement and intuitively ascribe the approximate parameters of the underlying random walk model—all from a simple visual inspection of these heat maps. For less obvious cases and formal parameter inference purposes we need to define a metric in the space of TMs that allows us to distinguish between the details of our chosen random walk models. The choice of the metric is critical, because a poorly performing metric can lose information that was captured in the TMs.

In this study we explore different metrics for their capability to discriminate different types of random walks and evaluate their performances in an ABC inference scheme. For the purpose of validating the choice of metric, we construct a specific random walk model that has tractable analytic parameter likelihoods, but nevertheless retains some of the complexity of a hybrid model. That way we are able to compare the ABC approximated posterior distribution (using the TMs and the investigated metrics) to the exact solutions obtained through any exact Monte Carlo sampling scheme (e.g. MCMC) (Robert and Casella 2013).

We present a case study where we analyse the spatio-temporal characteristics of a biased-persistent random walk model of leukocytes in response to acute injury. Leukocytes, here macrophages and neutrophils, are white blood cells that create the first layer of defence of the innate immune system and their migration patterns are therefore subject of interest in many biomedical studies (Mathias *et al* 2009, Skinner 2011, Holmes *et al* 2012, Schiwon *et al* 2014).

The remainder of the paper is structured as follows. In section 2, we provide a description of the theoretical and experimental methods used in this paper. In particular, we give details of the random walk model, an overview of the ABC inference procedure, the list of the various candidate metrics, and details of the experimental procedures used. In section 3, we present our results for our simulations and the biological case study, and conclude with a discussion of the wider applicability of this approach.

2. Methods

2.1. The random walk model

The situation we model consists of N non-interacting point particles in $\mathbb{R}^2$. These points represent the centroid of a cell, which is often used to track cells from live imaging data. The approximation of the cell as a point mass introduces simplicity and for many problems the added complexity of particle shapes or sizes is unnecessary. In this way our model is build on the assumption that the movement results from the cells interaction with its environment. In situations where the cell density is high or the cell's compartment is small, cell–cell-interactions need to be considered. The direction of a particle's movement at any time step t is described by two random variables, which are a step length s_t and a turning angle α_t taken with respect to an arbitrary fixed reference axis. The step length follows the distribution

$$s_t \sim \sqrt{\mathrm{d}t} \times N^{+}(0, 1), \qquad (1)$$

where $N^{+}(0, 1)$ is a truncated normal distribution, truncated at 0, and $\mathrm{d}t = 0.001$. The chosen step size model is just a practical implementation; our statistical analysis is independent of step sizes.

The turning angle α_t is defined as the angle between a motion vector and a reference axis, here the negative y-axis (see figure 1(a), at time t. α_t follows the wrapped normal distributions (Breitenberger 1963) with the probability density function

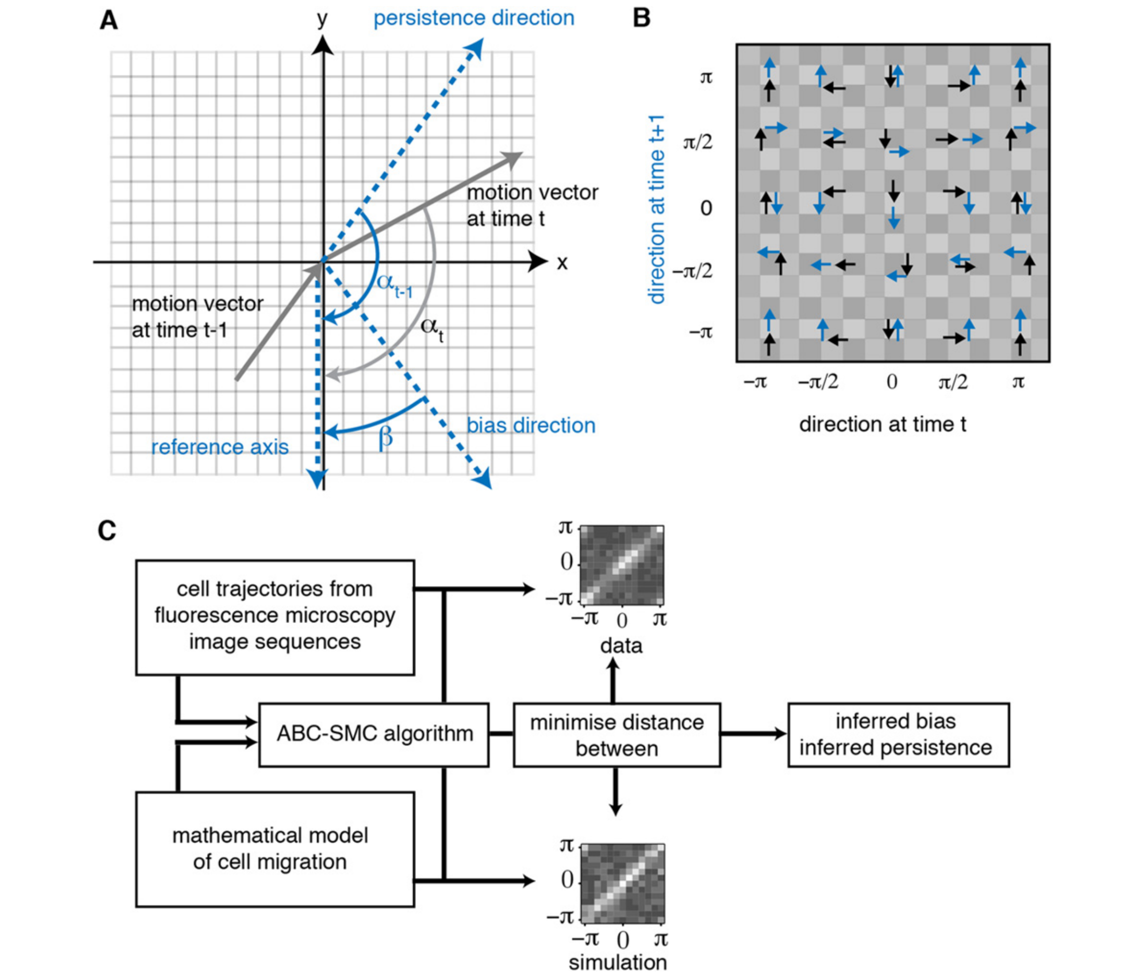


Figure 1. (A) A diagram depicting two consecutive motion vectors at times t and $t+1$ (grey), and the associated angles with respect to the reference axis (the negative y-axis). The direction of motion at time $t-1$ is described by α_{t-1}, the persistence direction at time t is given by α_{t-1}, and the bias direction is denoted by β. The combination of bias and persistence direction results in the motion vector at time t with direction α_t. (B) A key to reading heat maps of transition matrices (TMs). The TM gives the probability distribution of directional changes for a set of trajectories. The angles on rows and columns give the direction of movement with respect to the reference axis (here, the negative y-axis) in the plane containing the trajectories at time step t and $t+1$ respectively. Any entry in the TM gives the probability of a particle travelling in the direction denoted by the blue arrows (rows) at time $t+1$, given that it was travelling in the direction denoted by the black arrows (columns) at time t. (C) An overview of the inference scheme used. The ABC–SMC algorithm takes trajectories from data and numerically solves the mathematical model with varying parameters in order to minimize the distance between the transition matrices generated from the data and the simulation. This gives rise to inferred levels of bias and persistence.

$$N_w\left(\alpha_t \middle| \mu, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left(-\frac{\left(\alpha_t - \mu + 2\pi k\right)^2}{2\sigma^2}\right). \quad (2)$$

We note that we have selected the wrapped normal distribution mainly for its flexibility and relative simplicity; the inference scheme described below is, in principle, applicable to any other choice of distribution defined on a circle (e.g. von Mises distribution).

The mean μ and variance σ depend on whether the random walker follows a biased or persistent motion. For the biased motion we define $\mu = \beta$ (the direction of bias) and for the persistent distribution we have $\mu = \alpha_{t-1}$, which is the direction in the previous time step (see figure 1(a)). The variances σ for the biased and persistent motion are defined as

$$\sigma_p = -2\log(p) \quad (3)$$

and

$$\sigma_b = -2\log(b) \quad (4)$$

respectively, with the persistence and bias parameters, p and b, with $p,\ b \in [0,\ 1]$. They affect the variance of the distributions such that the closer to 1 they are, the smaller their respective variances will be, and the more likely the particle will be to sample an angle in the direction of the bias or the persistence, as appropriate. If p or b is equal to 0, then the corresponding variance will tend to infinity and thus the distribution is a wrapped uniform distribution.

The decision whether the random walker follows biased or persistent motion is based on a further random variable, which follows a Bernoulli distribution with success probability w, so that w describes the probability of a biased motion and $1 - w$ describes

Table 1. Parameters for reference transition matrices (R).

type of random walk (RW)	parameters (w, p, b)
Brownian motion (BM)	(0.5, 0.0, 0.0)
persistent RW (PRW)	(0.1, 0.5, 0.0)
biased RW (BRW)	(0.9, 0.0, 0.5)
biased persistent RW 1 (BPRW1)	(0.5, 0.7, 0.3)
biased persistent RW 2 (BPRW2)	(0.5, 0.3, 0.7)

the probability of a persistent motion; w is assumed to remain constant over time t.

At each time point, α_t and s_t are determined and, accordingly, the particle moves a distance of s_t in the direction defined by α_t. This constitutes one step in the particle's trajectory, described by three parameters (p, b and w). Five different parameter combinations are considered as reference points for specific types of random walks as can be seen in table 1. A particle that performs Brownian motion does not show either bias or persistence ($p = 0$, $b = 0$). The choice of the weight is then arbitrary and we chose without loss of generality $w = 0.5$. A persistent random walker has no bias ($b = 0$), but he has some level of persistence which is defined by p and w. Accordingly, a biased random walker has no persistence ($p = 0$) and the level of bias is defined by b and w. In the case of a biased persistent random walk (BPRW1 and BPRW2) none of the parameters should be 0. From these example parameter combinations we compute TMs, which we will refer to in the following as *query* TMs.

2.2. Transition matrices

To summarize the model output in the approximate inference framework we extract the directional transitions from the simulation and compute from these a transition matrix, first introduced in (Taylor *et al* 2013), which allows for visualization of the dynamics exhibited by the particle trajectories.

A TM T is the expected joint distribution of successively measured turning angles (α_t, $\alpha_{t+\mathrm{d}t}$). The components of the matrix T_{ij} gives the joint probability of measuring the angle $\alpha_{i,t}$ at some time t and the angle $\alpha_{j,t+\mathrm{d}t}$ at the next measurement time t+dt, where $\alpha_{i,t}$ represents any angle that lies in the half-open interval $[2\pi i/N_{\mathrm{bins}}, 2\pi(i+1)/N_{\mathrm{bins}})$ for $i = 0, 1, \ldots N_{\mathrm{bins}}$.

T_{ij} can be estimated from the data by constructing a two-dimensional histogram. First, we bin successive (and overlapping) pairs of angles of motions measured for every particle in terms of the intervals above. Then we define the sample TM as

$$\hat{T}_{ij} = \frac{1}{N'}\sum_{q=1}^{N}\sum_{r=1}^{T_q/\mathrm{d}t} I_{qr,ij}, \tag{5}$$

where

$$I_{qr,ij} = \begin{cases} 1, & \text{if } \alpha^{(q)}_{i,r*\mathrm{d}t} \in \left[\frac{2\pi i}{N_{\mathrm{bins}}}, \frac{2\pi(i+1)}{N_{\mathrm{bins}}}\right) \\ & \text{and } \alpha^{(q)}_{i,(r+1)*\mathrm{d}t} \in \left[\frac{2\pi j}{N_{\mathrm{bins}}}, \frac{2\pi(j+1)}{N_{\mathrm{bins}}}\right) \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

and $N' = \sum_{q=1}^{N} T_q/\mathrm{d}t$ the total number of successive angle-pair observations. The choice of N_{bins} is not predetermined and can, in principle, be tuned for different datasets. The more transitions are observed in a data set the larger N_{bins} can be chosen.

As can be seen in the key in figure 1(b), we have adopted the convention whereby the columns of the TM indicate the direction intervals at time t while the rows the intervals at time t+dt.

The interval between experimental observations dt_{obs} is determined by the experimental setup. In the general case, the interval between directional changes d t is an independent random variable; specifically, the number of directional changes that take place between any two observations is itself an unobserved quantity. It is precisely this hidden feature that prevents one from deriving exact closed-form expressions for the parameter likelihoods, thereby necessitating an approximate approach. In this scenario, the random walk model adopted here is an *effective* model (see (Sim *et al* 2015) for alternative velocity jump process models); concomitantly, the directional terms α_t are simply functions of the displacement data rather than the true ballistic paths of the particles.

Nevertheless in the simulation examples below we let

$$\mathrm{d}t = \mathrm{d}t_{\mathrm{obs}} = \text{constant}. \tag{7}$$

This simplifying assumption is made for the sole purpose of validating the proposed ABC approach as it allows us to determine the exact path probabilities and, hence, parameter posteriors for comparing against the approximate versions.

For the set of model parameters Θ, the likelihood resulting from the observation of N paths is simply

$$L(\Theta) = \prod_{i=1}^{N} p\left(\pi_i \mid \Theta\right), \tag{8}$$

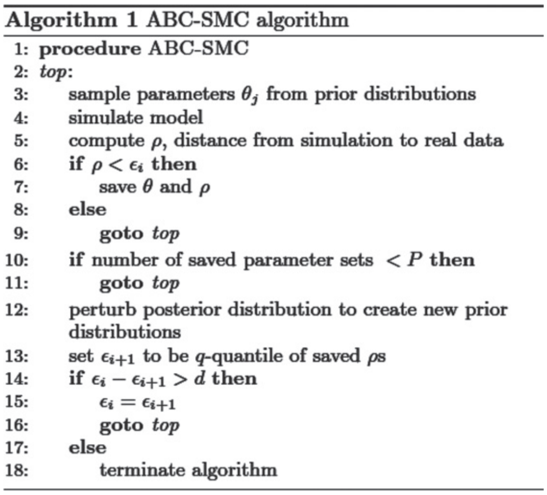
where $\pi_i = (\alpha_{i,0}, \ldots, \alpha_{i,T})$ is the ith path consisting of a sequence of T+1 observations $\alpha_{i,t} \in \mathbb{R}^2$. From equation (7) the individual terms can be factorized as

$$p\left(\pi_i \mid \Theta\right) = p\left(\alpha_{i,0} \mid \Theta\right) \prod_{t=1}^{T} p\left(\alpha_{i_t} \mid \alpha_{i,t-1}, \Theta\right), \tag{9}$$

where the individual probability terms are simply given by equation (2).

2.3. Inference using ABC

Even though for the proposed model in this study it is straightforward to define the likelihood function and perform exact inference, the motivation of this study is to provide an inference scheme that can also be applied to far more complex random walk models, for which the likelihood is not tractable. The approach taken for the parameter inference is ABC in a sequential Monte Carlo sampling scheme (ABC–SMC) (figure 1(c)) (Toni *et al* 2009). In order to perform ABC–SMC, one requires some sufficient statistic, μ, which can be calculated from both the real data and the simulated data; a distance function; a prior distribution for each of the parameters; some distance thresholds ϵ_i; and a set of perturbation kernels. The overall performance is affected by all of the factors (Filippi *et al* 2013, Silk *et al* 2013), but the choice of the summary statistic is crucial; even for good summary statistics, however, the distance function can play a major role in setting the rate and reliability of convergence. The ϵ_i are typically a series of decreasing distance thresholds where $i = 0, 1, \ldots, n$ and n. Here we choose $\epsilon_0 = 1.0$ and decrease it adaptively so that in population i the new threshold ϵ_i is based on the 10%-ile of the distances accepted in the $(i - 1)$th population. As summary statistics μ we compute the TMs; the distance function will be discussed at length later on in the paper. The prior distributions for the parameters w, p and b are all uniform distributions between 0 and 1. The ABC–SMC algorithm proceeds as follows:



Algorithm 1 ABC-SMC algorithm

```
procedure ABC-SMC
top:
    sample parameters $\theta_j$ from prior distributions
    simulate model
    compute $\rho$, distance from simulation to real data
    if $\rho < \epsilon_i$ then
        save $\theta$ and $\rho$
    else
        goto top
    if number of saved parameter sets $< P$ then
        goto top
    perturb posterior distribution to create new prior distributions
    set $\epsilon_{i+1}$ to be $q$-quantile of saved $\rho$s
    if $\epsilon_i - \epsilon_{i+1} > d$ then
        $\epsilon_i = \epsilon_{i+1}$
        goto top
    else
        terminate algorithm
```

We chose $P = 500$, $q = 10\%$. The value for d is dependent on the distance function being used since they all return values of various magnitudes, but it tends towards 0.

2.4. Exploring distance functions

In order to perform ABC–SMC as described above, it is mandatory to define a distance function that is able to efficiently discriminate between different random walk TM matrices. There is no standard method of computing the distance between two matrices and each proposed matrix metric focusses on distances between specific matrix characteristics. For this reason a number of potential distance function candidates were tested in order to find the best option for our application. The distances investigated are listed in table 2 (Deza and Deza 2006). The TM from real or simulated data is denoted by R and the query TM is denoted by Q, and $A = Q - R$. For the purposes of the following notation, all matrices are m x n. We chose $m = n = 13$ in the case study and $m = n = 15$ otherwise. These is a sufficient number of intervals to detect random walk characteristics, but also an appropriately low number of intervals to work with experimental data, as the number of data points required for reliable estimation of the TM increases dramatically with the number of intervals. The entry in the ith row and jth column of a matrix A is denoted by a_{ij}, while its singular values are written as $s_i(A)$. The singular values of a matrix A are defined as the square roots of the eigenvalues of the matrix $A^{\star}A$ where $A^{\star}$ is the conjugate transpose of A and $s_1(A) \leqslant s_2(A) \leqslant \ldots$ (Deza and Deza 2006).

Table 2. The distances that were investigated.

Distance	Formula (Deza and Deza 2006)
Frobenius norm	$\|\|A\|\|_{Fr} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\|a_{ij}\|^2}$
Hellinger distance	$H(R, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$
Infinity norm	$\|\|A\|\|^{\infty} = \max_{1 \leqslant j \leqslant n}\sum_{j=1}^{n}\|a_{ij}\|$
Kullback–Leibler div. 1	$D_{\mathrm{KL}}(R\|\|Q) = \sum_i \sum_j (r_{ij} \cdot \ln(\frac{r_{ij}}{q_{ij}}))$
Kullback–Leibler div. 2	$D_{\mathrm{KL}}(Q\|\|R) = \sum_i \sum_j (q_{ij} \cdot \ln(\frac{q_{ij}}{r_{ij}}))$
One norm	$\|\|A\|\|^{1} = \max_{1 \leqslant j \leqslant n}\sum_{i=1}^{n}\|a_{ij}\|$
Spectral norm	$\|\|A\|\|_{sp} = (\text{max eigenvalue of } A^{\star}A)^{\frac{1}{2}}$
Trace norm	$\|\|A\|\|_{tr} = \sum_{i=1}^{\min\{m,n\}} s_i(A)$

The distances include the Frobenius norm, which is also sometimes called the Euclidean norm; the Hellinger distance, which is used to calculate how similar two probability distributions are (Duan *et al* 2012); two natural norms: the infinity norm and one norm, which are defined as the maximum absolute row sum and the maximum absolute column sum, respectively; and two Ky-Fan k-norms: the spectral norm and the trace norm, which are defined to be the sum of an m x n matrix' first k singular values where $k = 1$ and $k = \min\{m, n\}$, respectively. Furthermore we test the Kullback–Leibler divergence. The latter is not strictly a distance since it is asymmetrical. However, it is often used to quantify differences between distributions. For the purpose of parameter inference in an ABC scheme, the distance function does not need to be symmetric, as long as the non-negativity, the identity of indiscernibles and the triangular inequality are fulfilled. We explore both possibilities for the Kullback–Leibler divergence ($D_{\mathrm{KL1}}(R||Q)$ and $D_{\mathrm{KL2}}(Q||R)$).

2.5. Experimental procedures

mpx:GFP/fms:RFP zebrafish embryos (Gray *et al* 2011) (5 days post fertilization) were anesthetized in 0.6 M MS-222 (Tricaine methanesulfonate, Sigma-Aldrich) and the tail fin was transacted using a sterile scalpel. The fish were then transferred to fresh system water for 2 h 28.5°C before transferral to 0.8% low-melt agarose (Flowgen, Lichfield, UK) for time-lapse imaging experiments. Images were captured using a Zeiss Axiovert 200 inverted microscope (Zeiss, Cambridge, UK) controlled by the C-Imaging Simple-PCI acquisition software (Hamamatsu, Sewickley, PA, USA) for up to 11 h post wounding. The temperature was maintained at 28.5°C throughout the experiment using a full incubation chamber with temperature control. The time gap between two consecutive images was 18 s. This resulted in time-lapse movies of GFP-labelled (green) neutrophils and mCherry-labelled (red) macrophages. Image processing was performed as described in (Taylor *et al* 2013).

3. Results

3.1. Choosing the optimal distance

In order to determine which of the seven distances is the best candidate for parameter inference we generated 500 TMs based on simulated random walk trajectories with various values for w, p and b, which we refer to as query TMs (Q). We compute each of the seven distances between these query TMs (Q) and the the five reference TMs (R) shown in table 1. The results are then represented graphically to visualize the distances' performance. The parameter values used to generate the query TMs are chosen on a lattice in parameter space with intervals 0.2 for w and 0.1 for p and b.

The first method of visualizing these results is by using an atlas of 2D heat maps. Each heat map shows the distance from (or to, in the case of $D_{\mathrm{KL}}(Q||R)$) the reference TM to (or from) query TMs with varying levels of b and p, and a fixed value of w. 'Atlases' of these heatmaps are produced for each reference TM, where each row of heatmaps uses a different distance function, and each column uses query TMs produced with a different value of w. For a complete set of these heat atlases, please see supplementary figures 1–5.

Examples from the BM and BPRW1 cases with fixed $w = 0.5$ are shown in figure 2. A good distance will have a minimum value when both the reference and query TM have been produced using the same parameters, and will increase rapidly as the parameters become less similar. We can quantify this latter property by examining the proportion of the full parameter space that reflects a distance below a threshold relative to the minimum distance value; here, a low value would indicate a good distance measure. The results for several thresholds are shown in figures 3(A) and (B) with the Hellinger distance and the trace norm performing well.

As a second approach to visualize the results we produce rank plots of the calculated distances. This means that we group all of the distances calculated to (or from) each reference TM using a particular distance function, order these numbers from lowest to highest, and plot them on a graph. This way it is possible to see the gradient of distances as the parameters change more clearly. Again, the steeper this gradient the better, because a steep gradient means this distance function will elucidate differences between TMs whose parameters are only slightly different. The results for the trace norm and Hellinger distance can be seen in figures 3(C) and (D). For other distances, please refer to supplementary figure 6.

Visual inspection of the heat atlases clearly indicates that the Hellinger distance and the trace norm outperform the other distance functions. Both the Hellinger distance and the trace norm are minimal between TMs with identical parameters, and maximal when parameters are most different. There is also a clear, strong gradient from low to high values of the distances.

3.2. TMs in an ABC inference scheme

We next explore the performance of the Hellinger distance and the trace norm in an ABC–SMC inference scheme. As data we use the five simulated reference TMs, since their parameters are known. We run ABC–SMC in order to infer for each TM the original parameters of bias, persistence and weight using both, the Hellinger distance and the trace norm, as a distance between the data and simulations. We compare the inference results to the posterior distribution obtained using the Metropolis Hastings algorithm. This comparison will indicate (i) if the applied statistics (TMs) are sufficient to describe the data, and (ii) if the chosen metric is appropriate.

We approximate the posterior parameter distribution of a TM generated from a BM with parameters $(w, p, b) = (0.5, 0.0, 0.0)$. The marginal posterior distributions based on the Hellinger distance and the trace norm, respectively, are shown in figure 4(a). While b and p are inferred well for both distances, the marginal posterior distribution of w spans the entire prior. The reason for this becomes apparent when visualizing the posterior distribution as a 3D scatterplot. All three parameters are highly correlated in a way that a high bias can reproduce the BM characteristics, as long as the weight is supporting the persistence, which itself needs to be close to 0. Vice versa, a high persistence can reproduce the BM characteristics, as long as the weight is supporting the bias, which itself needs to be close to 0. In order to reduce this dependency, we rescale the parameters by simple multiplication of the parameters. Since the weight is equivalent to the probability of the particle choosing the bias distribution over the persistence distribution, $b' = wb$ and $p' = (1 - w)p$ where b' and p' are the rescaled bias and rescaled persistence parameters, respectively.

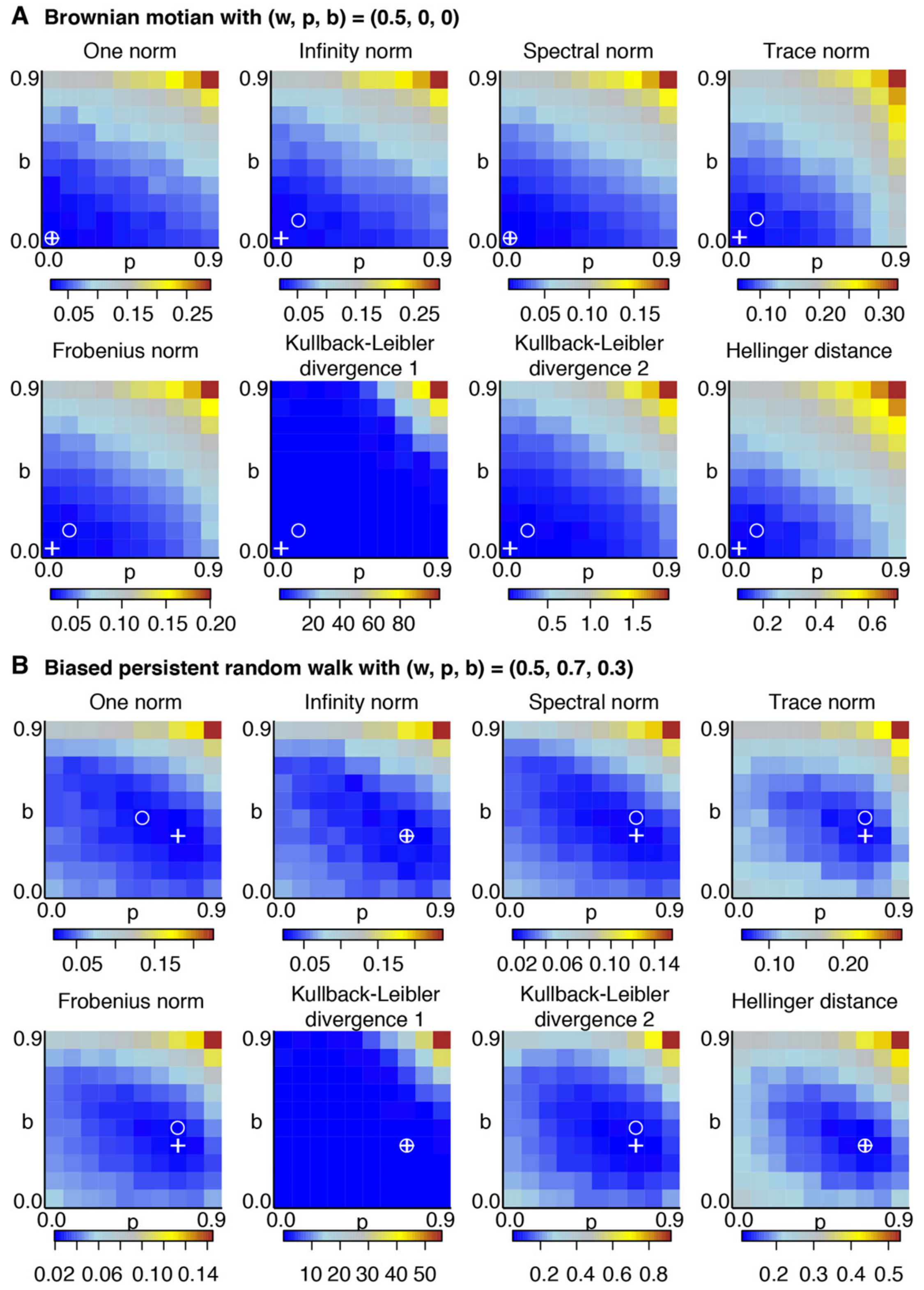


Figure 2. 100 query TMs (Q) were computed from simulations with various levels of b and p ranging from 0 to 0.9 in 0.1 increments, and with a fixed w of 0.5. Each graph is a heatmap generated by calculating the distance from/to the query TMs to/from the BM reference TM (A) and the BPRW1 reference TM (B) using various distance functions (see graph titles). Kullback–Leibler (KL) divergence 1 and 2 are the KL divergence from and to the reference TMs, respectively. Persistence (p) increases along the columns and bias (b) increases along the rows of each heat map. The white cross indicates the true parameter values; the white circle indicates the minimum of the heat map.

The rescaled parameters (p',b') used in generating the BM are (0.0, 0.0). The rescaled posterior distribution matches these values well for both Hellinger distance and the trace norm.

We repeat the inference scheme with the remaining four TMs. The results for the BPRW1 are shown in figure 4(b). The true parameters $(w, p, b) = (0.5, 0.7, 0.3)$ are inferred well by both distances.

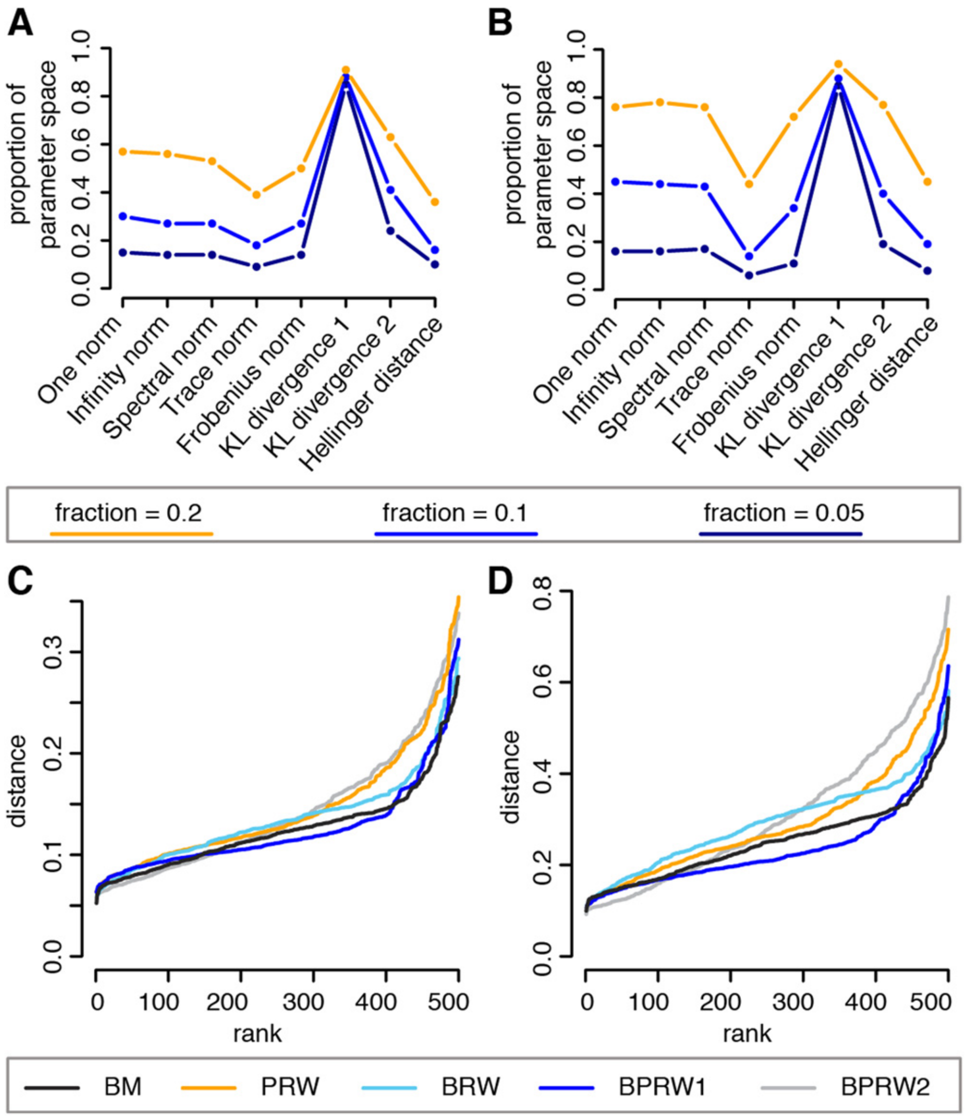


Figure 3. 500 query TMs were computed from simulations with values of w, p and b spanning the parameter space. The distances from each reference TM in table 1 to these query TMs were calculated and plotted in ascending order (rank plots). (A) and (B) The proportion of the full $p \times b$ parameter space that gives distance less than certain thresholds. The thresholds are defined in terms of fractions α of the difference between the maximum and minimum distances across the parameter space in each model. We set $\alpha = 0.05, 0.1$ and 0.2. We show the plots for the Brownian motion (A) and the BRW1 model (B). (C) Rank plot based on the trace norm. (D) Rank plot based on the Hellinger distance.

However, the marginal posterior distributions based on the Hellinger distance are narrower around the true parameter values, which shows a better performance of this distance. Note, the 3D scatterplot does not show any correlations between the three parameters, which indicates that no other parameter combinations can reproduce the characteristics of the BPRW model. Therefore it is not a surprise that also the rescaled parameters $(p', b') = (0.35, 0.15)$ are inferred well.

For all five tested scenarios the approximated posterior distributions are in good agreements with the exact posterior distributions obtained by the Metropolis Hastings algorithm. The main deviation is observed for the weight parameter w in the Brownian motion case 4(a). Furthermore, the exact marginal posterior distributions are slightly narrower than the approximated counterparts, which indicates some loss of information in the TMs and/or the applied metric.

In conclusion, the Hellinger distance is the best distance function to infer bias and persistence parameters for 2D random walks using TMs as summary statistics. The inference results are in agreement with exact methods where these are applicable.

3.3. Macrophages and neutrophils as random walkers — a case study

To test the Hellinger distance and evaluate its application on real-world problems we apply this distance to infer the random walk parameters of migrating immune cells. More specifically, we extract macrophage and neutrophil tracks from living zebrafish embryos that had undergone tail transection. Macrophages and neutrophils are the first layer of defence during inflammation, here

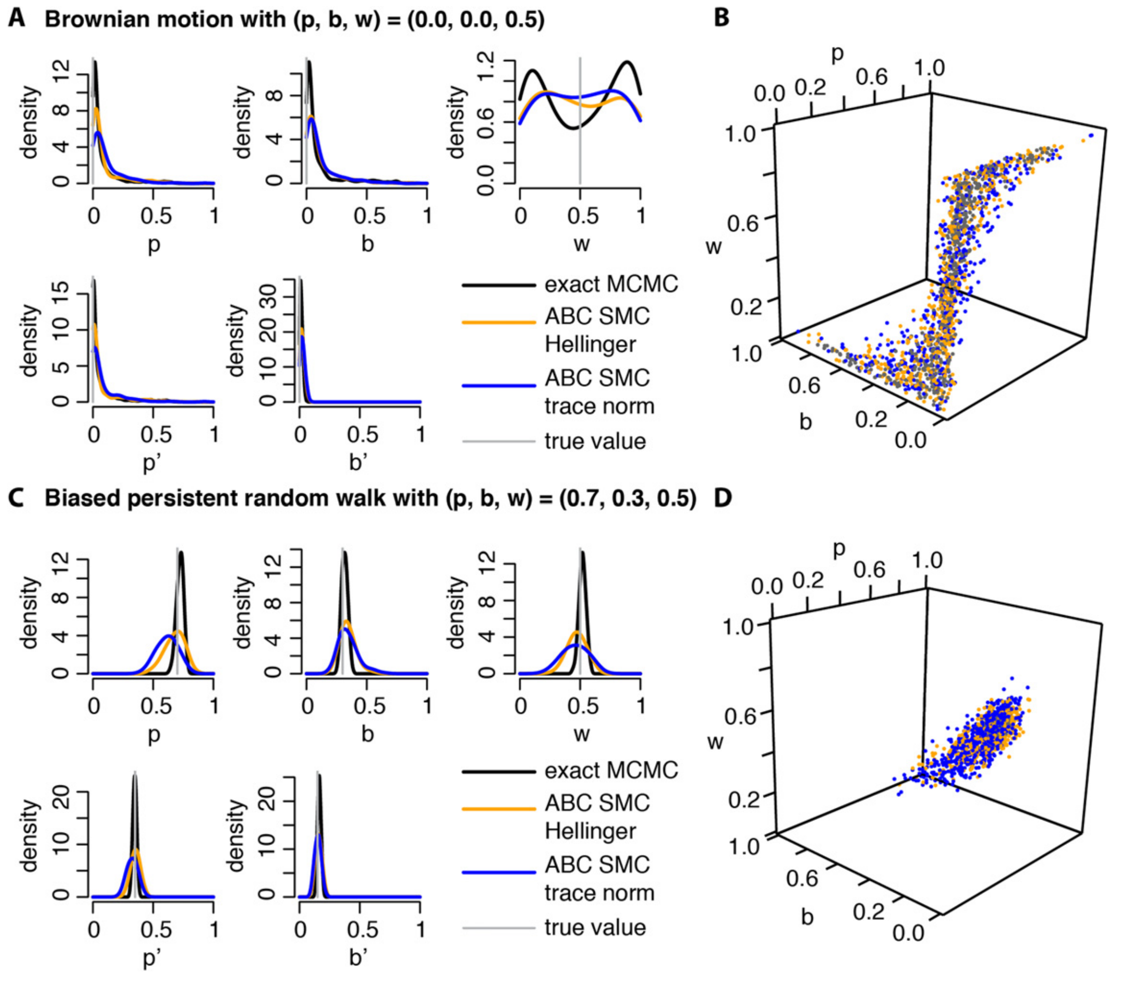


Figure 4. Results of parameter inference using ABC–SMC, with the Hellinger distance (orange) and trace norm (blue) as distance functions, performed on simulated data using the BM (A) and (B) and BPRW (C) and (D) as reference TMs. Shown are the posterior parameter distributions as marginals for p, b and w and the marginal rescaled parameters p' and b' (A) and (C). The exact posterior distributions (using Metropolis Hastings algorithm / MCMC) are shown in black; the true values used for the simulation of the reference TM are indicated by a red line. The 3D scatter plots of p, b and w are shown in (B) and (D).

mimicked by wounding. The extracted cell tracks were grouped according to how long after wounding and how far away from the wound they were detected. This results into spatial-temporal clusters, consisting of two temporal groups (T1: 1–6 h; T2: 6–11 h), each of them consisting of seven spatial groups that are generated by a sliding window approach. Figure 5 provides an overview of how the data are separated. For all 28 groups of trajectories (14 for macrophages and 14 for neutrophils) we compute the TM from the extracted cell tracks and infer the random walk parameters (figures 5 and 6). The ABC–SMC framework using the Hellinger distance successfully reproduces the TM computed from the *in vivo* data (figure 5(B)). We compare the distances from the data TM to the simulated TM to distances from the same data TMs to a given realization of a random walk with $p = b = 0$. As shown in supplementary figure 8, the former distances are significantly smaller for every TM, as is expected.

The posterior distributions of the inferred rescaled parameters $p' = (1 - w)p$ and $b' = wb$ are given in figure 6. It becomes apparent that the level of observed bias and persistence during the first 6 h after wounding is dependent on the distance from the wound for neutrohils. Macrophages, on the contrary, have a nearly constant level of persistence and only a modest increase of the bias with increasing distance from the wound. After 6 h the observed bias is decreased close to the wound, while the level of observed persistence is increased. Neutrophils show a comparable behaviour before and after 6 h. In general, we find that while the level of bias is similar in both cell types, the level of persistence is significantly higher in neutrophils than in macrophages. This case study demonstrates the potential of TMs as summary statistics combined with the Hellinger distance in an ABC framework.

4. Discussion

Traditional random walk models are often not applicable to describe *in vivo* cell migration due to the high complexity of the underlying molecular and cellular

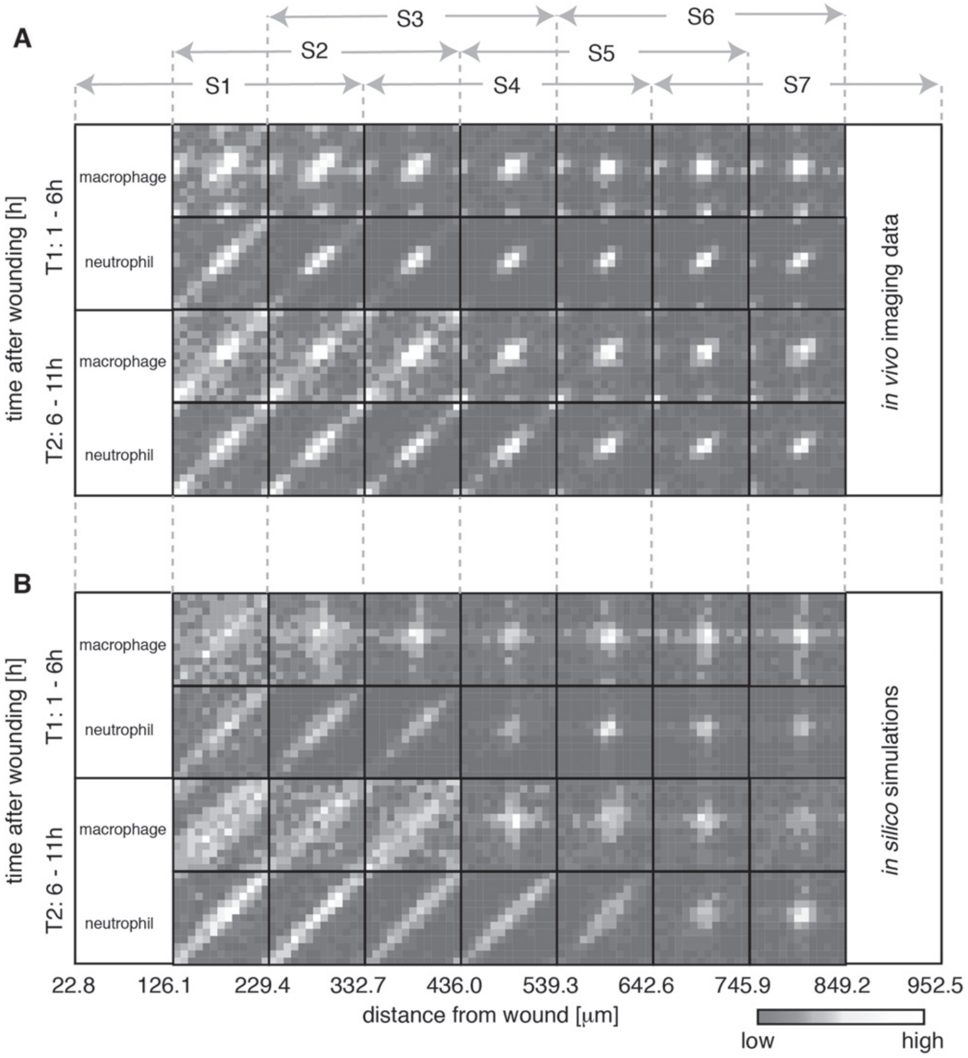


Figure 5. An overview of the TM computed from the extracted *in vivo* cell trajectories (A) and the inferred TMs (B). The inferred TMs are computed from simulated trajectories using parameters drawn from the posterior distribution (posterior predictive check). The graph depicts how the data were separated into seven spatial (S1–S7) and two temporal (T1 and T2) groups. Temporal groups are non-overlapping and are represented by individual rows whereas spatial groups span three columns and overlap the consecutive group(s) by two thirds. The TMs are computed from the extracted trajectories of the corresponding groups and are positioned in the exact row for the temporal group and the central column of the spatial group. From left to right the spatial groups read in increasing numerical order from S1 to S7, and the temporal group is T1 in the top two rows and T2 in the bottom two of each set of four. Note that the TMs are not plotted on equal scales, which was necessary for visualization purposes.

processes. In the last decade, so-called agent based models have been developed to describe pathways of the immune response (Bogle and Dunbar 2009, Liepe *et al* 2012, Chiacchio *et al* 2014). Such models can include a panoply of rules that guide the agent, here the cell. The problem is that these models can not be easily analysed. To provide an inference framework for such model, we studied a random walk model, for which the likelihood is tractable and exact inference can be applied. This provided us a reference of performance. By exploring a set of distance functions between matrices, we provide an inference scheme that is able to estimate parameters from different random walk models. The advantage of our framework is that it does not require any analytic or closed form expressions derived from the model of interest, as long as the model can be solved numerically.

We find that the Hellinger distance and the trace norm are best suited to distinguish random walk characteristics. It is interesting that the trace norm performs so well since it is just the sum of the diagonal entries of the matrix. This suggests that a lot of information about the whole TM is contained in just its diagonal entries. The level of persistence is represented by the probabilities on the diagonal on the matrix. Since the overall TM is normalized (all entries sum up to 1), the diagonal also contains information about the remaining entries in the TM. Similarly, the bias in our study is expressed as the probability in the centre of the TM, i.e. the probabilities along the diagonal of the matrix. This explains the high information content of the diagonal entries and therefore the good performance of the trace norm. However, as soon as the bias will be located at an angle other than 0, or other

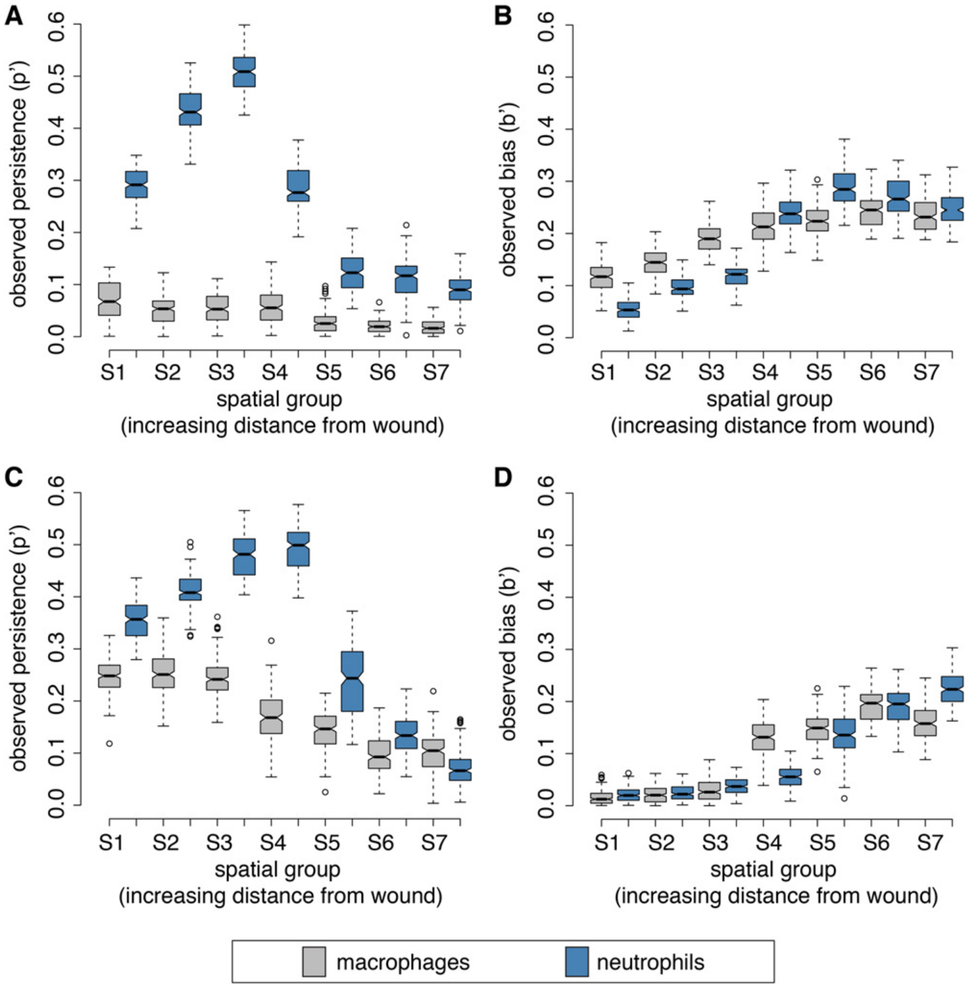


Figure 6. Shown are the posterior distributions of the rescaled persistence, p', and rescaled bias, b', following parameter inference using the data shown in figure 5. Results from 1 to 6 h of the inferred persistence (A) and bias (B) as well as the results from 6 to 11 h of the inferred persistence (C) and bias (D) are shown.

characteristics emerge, the trace norm is expected to fail. Here the Hellinger distance captures more details of the entire TM.

The framework relies on the ability of the TM to capture the information of the underlying random walk model. The TM has the advantage that it does not require the collection of data over many steps, because it does not rely on long term behaviour (in contrast to, for example, the mean square displacement). Indeed, it is enough to be able to track a cell via three points, i.e. to observe two consecutive motion vectors of the same cell. However, biological data might also bear a further problem, which is related to sampling effects. A low temporal resolution can result in a different time step in observed data compared to the reality. Rosser *et al* (2013) and Codling and Hill *et al* (2005) investigate the effects of sampling on a PRW. We briefly demonstrate such sampling effect on the TMs in supplementary figure 9. We find that strong sampling has the tendency to overestimate the level of bias, but underestimate the level of persistence. These effects may also impact the choice of the optimal metric and should be carefully considered in the interpretation of inference results.

We have demonstrated the inference framework in a case study that analyses the different migration patterns of macrophages and neutrophils in response to acute injury. Because the experimental data were extracted from only four zebrafish embryos, the biological significance of these results is debatable. However, the spatio-temporal behaviour regarding bias and persistence can be linked to cellular function in future studies. It is, for example, suggested that the behaviour of macrophages and neutrophils is regulated by a hydrogen peroxide gradient, produced at the wound site (Niethammer *et al* 2009). Such a gradient will induce the production of chemokines that interact with the cell surface receptors and in consequence regulate their motion. A similar study was conducted in zebrafish that tries to combine the migration behaviour with chemokine gradients *in vivo* (Liepe *et al* 2012). Our inference results show a clear spatio-temporal dependency of the bias and persistence of macrophages and neutrophils in response to wounding. The significantly higher persistence in neutrophils compared to macrophages results in an overall faster migration.

Although we investigate in this study the migration behaviour of innate immune cells during inflammation, the proposed inference framework can be useful in several other applications. Possible examples in biology include the migration of T-cells as part of the adaptive immune response (Masopust and Schenkel 2013), the migration of haemocytes (equivalent of macrophages) during the development of drosophila (Wood *et al* 2006, Razzell *et al* 2013) or the migration and spread of tumour cells (Boroughs *et al* 2011). The majority of studies conducted in the past decades were restricted to analyse very basic statistics, such as cell velocities and mean square displacement—choosing appropriate summary statistics is a general problem in ABC approaches, especially for model selection (Fearnhead and Prangle 2012, Prangle *et al* 2014). Here the concept of TMs is directly related to characteristics of (random) migration behaviour, and their application in an inference scheme allows us to extract more information from the available data. This is particularly important when working with animal experiments. Advanced statistical tools allow us to gain more information out of a reduced number of animals used to answer a given research question.

Acknowledgments

The project was in part granted by National Centre for the Replacement Refinement and Reduction of Animals in Research (NC3Rs) through a David Sainsbury Fellowship to JL, by BBSRC, The Leverhulme Trust and the Royal Society through a Wolfson Research Merit Award to MPHS, by BBSRC PhD studentship to PJMJ and by an EPSRC grant (EP/I017267/1) to AS. HT received funding through the Wellcome Trust Value in People Award. The zebrafish line was kindly provided by Dr Timothy Chico's lab in Sheffield (UK). Microscopy was performed in the Facility for Imaging by Light Microscopy (FILM) at Imperial College London with kind support from Martin Spitaler.

References

An G, Mi Q, Dutta-Moscato J and Vodovotz Y 2009 *Wiley Interdiscip. Rev.: Syst. Biol. Med.* **1** 159–71

Berg H C 1993 *Random Walks in Biology* (Princeton, NJ: Princeton University Press)

Bogle G and Dunbar P R 2009 *Immunology Cell Biol.* **88** 172–9

Boroughs L K, Antonyak M A, Johnson J L and Cerione R A 2011 *J. Biol. Chem.* **286** 37094–107

Breitenberger E 1963 *Biometrika* **50** 81–8

Chiacchio F, Pennisi M, Russo G, Motta S and Pappalardo F 2014 *BioMed Res. Int.* **2014** 907171

Codling E A, Plank M J and Benhamou S 2008 *J. R. Soc. Interface* **5** 813–34

Codling E and Hill N 2005 *J. Theor. Biol.* **233** 573–88

Dada J O and Mendes P 2011 *Integrative Biol.* **3** 86–96

Deza M M and Deza E 2006 *Dictionary of Distances* (Amsterdam: Elsevier)

Dohnal G 1987 *J. Appl. Probab.* **24** 105–14

Duan J, Gao T and He G 2012 arXiv:1204.0855

Edwards A M *et al* 2007 *Nature* **449** 1044–8

Fearnhead P and Prangle D 2012 *J. R. Stat. Soc.* B **74** 419–74

Filippi S, Stumpf M P H, Barnes C P and Cornebise J 2013 *Stat. Appl. Genetics Mol. Biol.* **12** 87–107

Gray C, Loynes C A, Whyte M K, Crossman D C, Renshaw S A and Chico T J 2011 *Thrombosis Haemostasis* **105** 811

Holmes G R, Dixon G, Anderson S R, Reyes-Aldasoro C C, Elks P M, Billings S A, Whyte M K, Kadirkamanathan V and Renshaw S A 2012 *Adv. Hematology* **2012** 792163

Horstemeyer M 2010 *Practical Aspects of Computational Chemistry* (Berlin: Springer) pp 87–135

Liepe J, Taylor H, Barnes C, Huvet M, Bugeon L, Thorne T, Lamb J, Dallman M and Stumpf M 2012 *Integr. Biol.* **10** 335–45

Masopust D and Schenkel J M 2013 *Nat. Rev. Immunology* **13** 309–20

Mathias J R, Walters K B and Huttenlocher A 2009 *Chemotaxis* (Berlin: Springer) pp 151–66

Neisyy A 2008 *IUST Int. J. Eng. Sci.* **19** 17–19

Niethammer P, Grabher C, Look A T and Mitchison T J 2009 *Nature* **459** 996–9

Potdar A A, Jeon J, Weaver A M, Quaranta V and Cummings P T 2010 *PLoS One* **5** e9636

Prangle D, Fearnhead P, Cox M P, Biggs P J and French N P 2014 *Stat. Appl. Gen. Mol. Biol.* **13** 67–82

Razzell W, Evans I R, Martin P and Wood W 2013 *Curr. Biol.* **23** 424–9

Ridley A J, Schwartz M A, Burridge K, Firtel R A, Ginsberg M H, Borisy G, Parsons J T and Horwitz A R 2003 *Science* **302** (5651) 1704–9

Robert C and Casella G 2013 *Monte Carlo Statistical Methods* (Berlin: Springer)

Rosser G, Fletcher A, Maini P and Baker R 2013 *J. R. Soc. Interface* **10** 20130273

Schiwon M *et al* 2014 *Cell* **156** 456–68

Silk D, Filippi S and Stumpf M P H 2013 *Stat. Appl. Gen. Mol. Biol.* **12** 603–18

Sim A, Liepe J and Stumpf M P H 2015 *Phys. Rev.* E **91** 042115

Skinner M 2011 *Nat. Rev. Mol. Cell Biol.* **13** 2–3

Taylor H, Liepe J, Barthen C, Bugeon C, Huvet M, Kirk P, Brown S, Lamb J, Dallman M and Stumpf M 2013 *Immunol. Cell Biol.* **91** 60–69

Toni T, Welch D, Strelkowa N, Ipsen A and Stumpf M P 2009 *J. R. Soc. Interface* **6** 187–202

Turner B M and van Zandt T 2012 *J. Math. Psychology* **56** 69–85

Wilke C and Lee C 1955 *Ind. Eng. Chem.* **47** 1253–7

Wilkinson R D 2013 *Stat. Appl. Gen. Mol. Biol.* **12** 129–41

Wood W, Faria C and Jacinto A 2006 *J. Cell Biol.* **173** 405–16