35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Bayesian framework for identifying cell migration dynamics

Geoffrey R. Holmes[1], Sean R. Anderson[1], Giles Dixon[2], Stephen A. Renshaw[2], Visakan Kadirkamanathan[1]

Abstract— **Cell migration is a vital process in living organisms. In particular we are interested in the way that white blood cells such as neutrophils migrate during episodes of inflammation which are important events in the working of the innate immune system. Migration of populations of many kinds can be modelled using drift-diffusion models by drawing analogies between the individual agents and the molecules in a fluid. It is challenging to arrive at a data-driven estimate of the parameters of this kind of process, particularly so if the individual agents have time varying properties that are not uniform over the population. In this paper, we offer a novel framework to estimate migration dynamics in this context. It makes use of the Approximate Bayesian Computation approach for parameter estimation and model selection. The Framework is applied to zebrafish neutrophil dynamics but is applicable for general migration scenarios.**

I. INTRODUCTION

Cell migration is a key process in complex living organisms. It occurs naturally during embryogenesis and wound healing, pathologically in tumour metastasis, and is essential in the processes used for tissue engineering [1]. Our particular interest is in the migration of neutrophils during episodes of inflammation and their resolution [2]. The neutrophil is a type of white blood cell and a key agent in the innate immune system [3]. Neutrophils migrate rapidly to any site of injury and infection where they destroy harmful bacteria and contribute to the healing processes. However, neutrophils are harmful when wrongly activated or when their inflammatory response fails to resolve normally [2]. Novel modelling and identification methods are needed to better understand the migration dynamics of these cells, so that preventative and therapeutic strategies can be properly developed for inflammatory diseases.

Drift-diffusion models are often used to model migration in diverse settings [4], [5]. The distribution of a population with drift-diffusion dynamics can be described by a partial differential equation. In simple cases the solution to this equation could be fitted to observed data in order to arrive at the dynamic coefficients. However, if the individuals of the population have non-uniform characteristics this approach becomes problematic. Furthermore it provides no natural mechanism for model selection. Hence, there is a need to develop more sophisticated data-driven modelling methods to handle the complex scenarios of cell migration.

The novel contribution we make here is to develop a framework for modelling cell migration using approximate Bayesian computation (ABC) [6]. ABC is a family of methods that arose within the field of population genetics and is rapidly gaining acceptance and application within a wide range of research areas. It is a rigorous simulation based approach for processes where a likelihood function is not available or not feasible for computation. ABC comes with the benefits of the Bayesian method [7]: the incorporation of prior knowledge, automatic regularisation, model selection, and with uncertainty of identified models and parameters as a standard output. We are particularly interested in investigating whether neutrophil migration during inflammation resolution is directed [8] or random (corresponding to a non-zero or zero drift term respectively) and we make a novel use of Bayes factor analysis [9] to address this question.

Toni et al. [10] proposed an ABC method for parameter estimation and model selection which uses standard error metrics for vector time series data from dynamical systems. We develop this into a framework for identifying cell migration dynamics by using a distributional summary of cell positions and the Cha-Srihari distance [11] to compare the resulting distributions. These novel modifications allow us to handle the challenging problem of quantifying the similarity of two sets of cell observation data and thus estimate the underlying dynamics. Whilst our interest particularly focuses on immune cell migration, this framework could be applied to any migration process where spatio-temporal observations are at the population scale.

II. IDENTIFICATION FRAMEWORK

Our ABC-Sequential Monte-Carlo (ABC-SMC) parameter estimation and model selection algorithm, based largely on that in [10] is set out in the Appendix. In order to utilise this in a framework for estimating cell migration dynamics the following components need to be specified,

- a simulation algorithm, tailored to the application,
- a way of summarising observations from the system,
- a distance measure to compare simulated observations to those of the real system.

The simulation algorithm is particular to the migration models and will be developed in the applied section, Section III.

A. Observation Summary

The system observations consist of time indexed sets of cell positions. In our particular application we have a constant cell population size, which is achieved by fluorescent cell labelling [5]. However, the number of observed cells will tend to vary stochastically from time to time due to a

[1] G.R. Holmes, S.R. Anderson and V. Kadirkamanathan are with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, UK {geoff.r.holmes, s.anderson, visakan} at sheffield.ac.uk

[2] G. Dixon and S.A. Renshaw are with the Academic Unit of Respiratory Medicine, Department of Infection and Immunity, University of Sheffield, Sheffield, S10 2RX, UK {mda07gd, s.a.renshaw} at sheffield.ac.uk

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

cell being temporarily indistinguishable from or occluded by another. Also, to facilitate long experimental run times, the time between observations is relatively long (compared to characteristic cells migration times) so that tracking of cells is not possible. Though, in fact, a strength of the method is that all available cell observations are included in the analysis, whereas tracking analysis almost certainly implies selectivity. The observations of cells in the experimental data at time t can be described by,

$$\mathcal{Y}_t = \{x_{t,i}\}_{i=1}^{M_t}, \tag{1}$$

where $M_t \leqslant N_\mathrm{c}$ is the number of observed cells, N_c the total number of cells in the system, and the $x_{t,i}$ are the observed cell positions. If T_obs is the total number of observations a complete observation set may be defined as,

$$\mathcal{Y}^\mathrm{obs} = \{\mathcal{Y}_t\}_{t=1}^{T_\mathrm{obs}}. \tag{2}$$

It is clear that for a simulation run to reproduce an observed set of cell positions subsequent to the initial observation has probability zero. It is therefore necessary to construct a summary of the observations which is of low dimension and yet preserves as much of the information content in the data as possible. We did this by summarising the cell positions as a discrete distribution over space, equivalent to a normalised histogram as follows,

$$\mathbf{V}_t = \begin{pmatrix} \sum_{i=1}^{M_t} \chi_{B_1}(x_{t,i}) \\ \vdots \\ \sum_{i=1}^{M_t} \chi_{B_b}(x_{t,i}) \end{pmatrix} \tag{3}$$

$$\mathbf{Y}_t = \frac{1}{\sum_i V_{t,i}} \mathbf{V}_t, \tag{4}$$

where $B_j, j = 1, \ldots, b$ is a set of spatial intervals forming a partition of the range of the $x_{t,i}$; χ_{B_j} is the indicator function of interval B_j; and $\mathbf{Y}_t$ is thus the normalised form of $\mathbf{V}_t$.

B. Distance Measure

Commonly used methods for measuring the distance between two distributions include the Kullback-Liebler distance (KLD) and the Bhattacharyya divergence (BD) [12]. However, KLD is problematic if the distributions compared do not have identical support. Furthermore, both KLD and BD (which we used previously [5]), whilst often used for comparing discrete distribution data, have limitations when the histograms bars have an inherent order [11]. This is because both KLD and BD consider only the differences between corresponding histogram bars and not the amount of 'work' needed to transform one histogram into the other.

Hence, we improve the methodology here, making it more robust, by using an alternative metric, the Cha-Srihari distance, which takes account of the work needed to transform one histogram into another. A naïve way of taking this into account is to consider the minimal pairwise difference between all samples making up the histogram data. Computing this is exponential in time as there are $n!$ possible pair assignments if n is the number of samples. Cha and Srihari derive an algorithm which is linear in time by noting that the minimum difference of pairwise assignments is equivalent to the minimum cost of moving cells (the basic histogram bar size units) to transform from one histogram to the other. Their algorithm for computing it is as follows [11].

```
Require: Histograms $A, B$ with bar sizes $A_i, B_i$,
            $i = 1, \ldots, n$
Ensure: $D_\mathrm{CS}(A, B)$ the Cha-Srihari distance between the
    two histograms.
  for $i = 1$ to $n$ do
    Compute the bar size differences, $d_i = A_i - B_i$.
    Compute the cumulative sums of the differences,
        $c_i = \sum_{j=1}^{i} d_j$.
  end for
  Compute $D_\mathrm{CS}(A, B) = \sum_{i=1}^{n} |c_i|$.
```

The data was summarised in (4) as a separate discrete distribution for each of the T timepoints. Therefore, the Cha-Srihari derived distance between two complete observation sets was defined as follows.

$$\rho_\mathrm{C}(\mathcal{Y}^{(p)}, \mathcal{Y}^{(q)}) = \sum_{t=1}^{T_\mathrm{obs}} D_\mathrm{CS}(\mathbf{Y}_t^{(p)}, \mathbf{Y}_t^{(q)}) \tag{5}$$

III. Identifying Neutrophil Migration

The framework was applied to the migration of inflammatory neutrophils in zebrafish. The data is described in [4]. We observed in that paper that neutrophils continued to be recruited to the inflammation site even while earlier recruited cells were migrating away. In this study, therefore we propose a more general model to that used in [5] which includes a simple form of attractant ligand receptor dynamics to model how individual cells switch between recruitment and resolution mode. The concept is illustrated in Fig. 1.

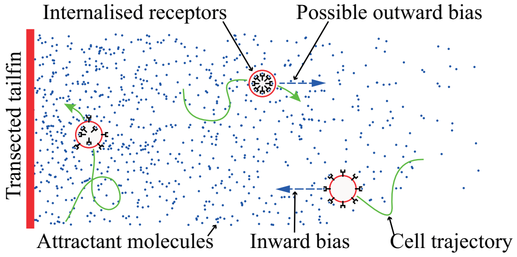

Fig. 1. Tailfin transection of zebrafish larvae 3 days post fertilization induces inflammation [13]. Attractant proteins recruit neutrophils. A neutrophil entering the field of attractant has a full complement of receptors available on the cell surface for binding attractant molecules. Binding events result in bias of the cells migration towards the wound. Bound receptors also become internalised and this weakens the response. We call this receptor depletion. Eventually, the cell loses its response to the attractant field. It may now recognises other guidance cues which bias its migration away from the wound region. The green curves show the cell paths from which observations are sampled. The blue arrows show recruitment and resolution biases which are to be estimated together with the coefficient of random diffusive motion and the receptor depletion rate.

A. Model Description

We formalised the model shown in Fig. 1, by including receptor depletion terms alongside a basic drift-diffusion model. In discrete time this is described as follows,

$$x_{t+1}^{(i)} = \max\left(0, x_t^{(i)} + \left(b_{\text{out}} - R_t^{(i)} b_{\text{in}}\right)\Delta t + \omega_t^{(i)}\sqrt{2D\Delta t}\right) \tag{6}$$

$$R_{t+1}^{(i)} = R_t^{(i)} - \lambda \max\left(0, \frac{L - x_t^{(i)}}{L}\right) R_t^{(i)} \Delta t, \tag{7}$$

where $\hat{x}_{t+1}^{(i)}$ is the position $x_t^{(i)}$ of the i^{th} neutrophil at time t; b_{in} and b_{out} are respectively bias velocities, or drifts, towards and away from the wound; $R_t^{(i)}$ is the proportion of receptors available for the i^{th} cell at time t; $\omega_t^{(i)} \sim \mathcal{N}(0,1)$ are a family of independent white noise processes; D is the underlying diffusivity constant or magnitude of random movement of the neutrophils; Δt is the time increment; λ is the composite depletion constant described above; L is the range of the chemoattractant field. This model is straightforward to simulate using the output of a random number generator to sample the $\omega_t^{(i)}$. We applied the ABC-SMC algorithm (see Appendix) with three candidate models which are variants of (6), (7),

- Model 1 : b_{in} and $b_{\text{out}} = 0$, a simple diffusion model.
- Model 2 : $b_{\text{out}} = 0$, a model with no outward bias.
- Model 3 : The full model which includes both inward (recruiting) and outward (resolution) biases.

The models were compared by the algorithm on a pairwise basis. Uniform priors were applied for each parameter over ranges corresponding where possible to physical plausibility, i.e. $b_{\text{in}} : [0 - 5\,\mu\text{m}\,\text{min}^{-1}], b_{\text{out}} : [0 - 2\,\mu\text{m}\,\text{min}^{-1}], D : [0 - 200\,\mu\text{m}^2\,\text{min}^{-1}], \lambda : [0 - 0.1]$. Furthermore, in comparing models 2 and 3 the algorithm was applied repeatedly with increasing values on the lower prior range for the outward bias, b_{out}. This allowed Bayes factor calculations to determine the evidence between theses models given what amount of outward bias is considered significant.

B. Results and Discussion

The model comparison results in the posterior model marginals shown in Fig. 2. The corresponding Bayes factors and model evidence [9] are $B_{21} = 24$ (strong evidence) $B_{31} = 2.4$ (no significant evidence) with $B_{23} = 3.3$ (substantial evidence). Thus Model 2 is the preferred model and this corresponds to a purely stochastic, non-directed migration of neutrophils away from the inflammation site during resolution.

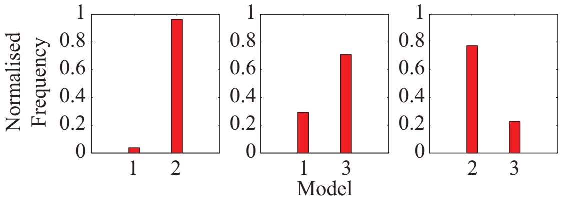

Fig. 2. Model selection Algorithm IV was applied to pairs of candidates models to identify the preferred model. The posterior model marginal is given simply by the number of parameter sets representing each model in the final generation of samples.

To further explore this result the analysis was repeated several times with increasing minimum allowed values of outward drift. The logarithm of the Bayes factor for model 2 with respect to model 3 is plotted against this minimum allowed value in Fig. 3. The evidence for the zero drift model increases as the minimum allowed values increases, showing that if very small values of drift are considered insignificant the evidence is stronger.

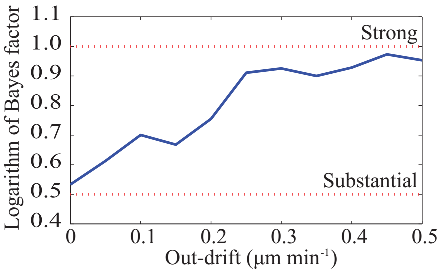

Fig. 3. The Bayes factor for model 2 with respect to model 3 is plotted on a logarithmic scale against the lower limit on the prior for the drift coefficient in Model 3. Interpretation of Bayes factor evidence is taken from [9].

The algorithm calculates parameter estimates as well as model identification and those for the preferred model, model 2, are shown in Fig. 4. This combination of drift and diffusivity parameters corresponds to average cell speeds of approximately $10\,\mu\text{m}\,\text{min}^{-1}$ in keeping with observed neutrophil speeds [8].

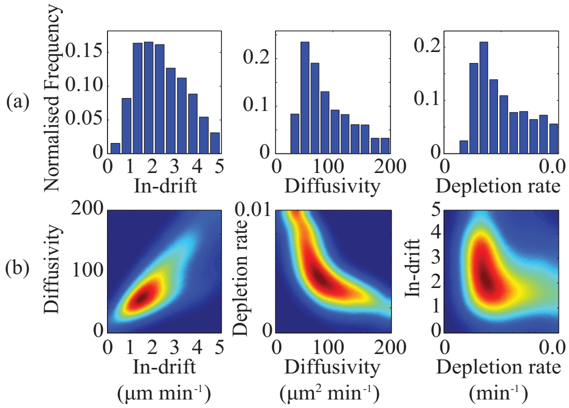

Fig. 4. Parameter estimation results for the preferred model, model 2. (a) The posterior distribution over the individual parameters. (b) The joint distributions over each pair of parameters. The maximum a posteriori parameter set, calculated via optimization of a kernel density estimate [14] is $b_{\text{in}} = 2.1\,\mu\text{m}\,\text{min}^{-1}, D = 83\,\mu\text{m}^2\,\text{min}^{-1}, \lambda = 0.0046\,\text{min}^{-1}$.

Our finding that zebrafish neutrophil dynamics during inflammation resolution are purely stochastic conflicts with an emerging consensus: that the neutrophils exhibit directed migration during resolution similar to that during recruitment [8]. The new result in this study suggests, not only that neutrophil migration during inflammation resolution is purely stochastic but that receptor depletion dynamics may be the key to understanding the change of migration mode. This in turn suggests that the search for ways of influencing neutrophil behaviour when things go wrong should be directed at the cell itself as well as at its external environment.

IV. CONCLUSION

We have developed a robust framework for estimating migration models and demonstrated its effectiveness when applied to zebrafish neutrophil inflammation dynamics. Whilst the candidate models in this paper are nested, a strength of the framework is that it is equally applicable to arbitrarily related models. This gives it an advantage compared to classical model selection tools. Key to our framework is the summarising of cell observations using a distributional description of the cell population adopting the Cha-Srihari distance to compare the similarity or difference between two sets of summarised observations.

APPENDIX

The ABC-SMC algorithm was introduced in [15] and developed for model selection in [10] on which our implementation is based. We have included a non-uniform acceptance kernel as suggested in [16]. Briefly, a model and its parameters are sampled from their current joint distribution. This model is simulated. If the distance between simulated and experimental observations is within the current error tolerance this model / parameter set, appropriately weighted, is accepted as a sample from the target distribution.

```
Require: data, $\mathcal{Y}^{\text{obs}}$; Monte Carlo population size, N; iterations, T; priors on models $\pi(m)$ and on model parameters $\pi(\theta|m)$; simulation algorithm, $\mathcal{Y} \sim p(\mathcal{Y}|m,\theta)$; distance metric $\rho$, model and parameter perturbation kernels $M$, $K$; decreasing error schedule $\epsilon_1, \dots, \epsilon_T$.
Ensure: a set of parameter vectors $\theta_i$ augmented with model indicator $m_i$, with importance weights $\omega_i$, that together form a weighted sample from the joint posterior distribution, $p(\theta, m|\mathcal{Y}^{\text{obs}})$.
  for $i = 1$ to $N$ do
    Simulate $m_i \sim \pi(m)$, $\theta_i \sim \pi(\theta|m_i)$ and
    $\mathcal{Y} \sim p(\mathcal{Y}|m_i,\theta_i)$ until $e_i = \rho(\mathcal{Y},\mathcal{Y}^{\text{obs}}) \leqslant \epsilon_1$.
  end for
  Set each $\omega_i^{(1)} \propto \frac{1}{\epsilon_1}\left(1-\left(\frac{e_i}{\epsilon_1}\right)^2\right)$, such that $\sum \omega_i^{(1)} = 1$.
  for $t = 2$ to $T$ do
    For each model, m, set $\tau(m)^2 = 2\text{Var}(\{\theta_i : m_i = m\})$.
    for $i = 1$ to $N$ do
      Choose k from $\{1 \dots \text{N}\}$ with probabilities $\{\omega_1 \dots \omega_N\}$ and set $m^* = m_k$ and $\theta^* = \theta_k$.
      Simulate $\hat{m}_i \sim M(m|m^*)$.
      Re-choose $\theta^*$ from $\{\theta_j : m_j = \hat{m}_i\}$ with probabilities $\{\omega_j : m_j = \hat{m}_i\}$.
      Simulate $\hat{\theta}_i \sim K(\theta|\theta^*;\tau(\hat{m}_i)^2)$ and $\mathcal{Y} \sim p(\mathcal{Y}|\hat{m}_i,\hat{\theta}_i)$
      until $e_i = \rho(\mathcal{Y},\mathcal{Y}^{\text{obs}}) \leqslant \epsilon_t$.
      Set $\tilde{\omega}_i = \frac{1}{\epsilon_t}\left(1-\left(\frac{e_i}{\epsilon_t}\right)^2\right)$.
    end for
    Set each $\hat{\omega}_i \propto \frac{\tilde{\omega}_i \pi(\hat{\theta}_i)}{\sum_{j:m_j=\hat{m}_i} \omega_j K(\hat{\theta}_i|\theta_j;\tau(\hat{m}_i)^2)}$,
    such that $\sum \hat{\omega}_i^{(t)} = 1$.
    Set each $m_i = \hat{m}_i$, $\theta_i = \hat{\theta}_i$, $\omega_i = \hat{\omega}_i$.
  end for
```

The distance $\rho(.,.)$ was defined as in (5) and thus includes the summarising of the data as a distribution via (4). Benchmarking tests indicated that $T = 4$ was the best choice for efficiency and we chose $N = 4000$ to give a good balance between posterior coverage and computation time. Error tolerances were chosen automatically: an initialisation run chose model / parameter sets from the joint prior to form a set of N parameter vectors with associated errors, e_i. ϵ_1 was chosen as $0.5\text{max}(e_i)$ and ϵ_T was chosen as the first percentile of the e_i. Parameter sets from the initialisation run were recycled into the first iteration if their associated error was less than ϵ_1. Then we set $\epsilon_i = \epsilon_1 e^{-\alpha i}, \quad i = 2, \dots, N-1$ with $\alpha = (\log \epsilon_1 - \log \epsilon_N)/N$.

The parameter perturbation kernel was chosen to be zero mean Gaussian with variance computed as in the algorithm to be twice the weighted empirical variance of the previous population. Also, a model perturbation kernel was used in which the original model was kept with probability 0.6 and one of the r remaining alternative candidate models with probability $\frac{0.4}{r}$.

REFERENCES

[1] D. A. Lauffenburger and A. F. Horwitz, "Cell migration: A physically integrated molecular process," *Cell*, vol. 84, no. 3, pp. 359–369, 1996.

[2] B. Amulic, C. Cazalet, G. L. Hayes, K. D. Metzler, and A. Zychlinsky, "Neutrophil function: From mechanisms to disease," *Annual Review of Immunology*, vol. 30, no. 1, pp. 459–489, 2012.

[3] C. Nathan, "Neutrophils and immunity: challenges and opportunities," *Nature Reviews. Immunology*, vol. 6, no. 3, pp. 173–182, 2006.

[4] G. Holmes, G. Dixon, S. Anderson, C. Reyes-Aldasoro, P. Elks, S. Billings, M. Whyte, V. Kadirkamanathan, and S. Renshaw, "Drift-diffusion analysis of neutrophil migration during inflammation resolution in a zebrafish model," *Advances in Hematology*, 2012.

[5] G. R. Holmes, S. R. Anderson, G. Dixon, A. L. Robertson, C. C. Reyes-Aldasoro, S. A. Billings, S. A. Renshaw, and V. Kadirkamanathan, "Repelled from the wound, or randomly dispersed? reverse migration behaviour of neutrophils characterized by dynamic modelling," *Journal of The Royal Society Interface*, 2012.

[6] M. A. Beaumont, "Approximate Bayesian computation in evolution and ecology," *Annual Review of Ecology, Evolution, and Systematics*, vol. 41, pp. 379 – 406, 2010.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman & Hall, 2004.

[8] J. R. Mathias, B. J. Perrin, T. X. Liu, J. Kanki, A. T. Look, and A. Huttenlocher, "Resolution of inflammation by retrograde chemotaxis of neutrophils in transgenic zebrafish," *Journal of Leukocyte Biology*, vol. 80, no. 6, pp. 1281–1288, 2006.

[9] H. Jeffreys, *Theory of Probability*, 3rd ed. Oxford University Press, 1998.

[10] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society Interface*, vol. 6, pp. 187–202, 2009.

[11] S.-H. Cha and S. N. Srihari, "On measuring the distance between histograms," *Pattern Recognition*, vol. 35, no. 6, pp. 1355–1370, 2002.

[12] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, 1989.

[13] J. S. Martin and S. A. Renshaw, "Using in vivo zebrafish models to understand the biochemical basis of neutrophilic respiratory disease," *Biochemical Society Transactions*, vol. 37, no. Pt 4, pp. 830–837, 2009.

[14] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, 1st ed. Chapman and Hall/CRC, 1986.

[15] S. A. Sisson, Y. Fan, and M. M. Tanaka, "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 6, pp. 1760–1765, 2007.

[16] M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate Bayesian computation in popoulation genetics," *Genetics*, vol. 162, pp. 2025–2035, 2002.