Using mathematical modelling to explore mRNA localization

Jonathan Harrison [†], Richard Parton [‡], Ilan Davis [‡], Ruth Baker [†]

[†]Mathematical Institute, University of Oxford; [‡] Department of Biochemistry, University of Oxford

Abstract

In early *Drosophila* development, mRNA localization plays a key role in determination of the body axes. Maternal mRNA is transported from the nurse cells into the oocyte itself, where it is localized to specific locations in order to target proteins to regions where they can function. By employing a simple velocity jump process model of the transport of mRNA, we describe this transport process and demonstrate how it depends on certain parameters of the model. The model used represents the active transport and diffusion phases observed experimentally. Based on simulations from the model with inferred parameters, we estimate a mean first passge time from the nucleus of the nurse cell to the ring canal joining to the oocyte of 18 min. Since the timescales of oogenesis are at least an order of magnitude longer than this, we suggest that production is the rate-limiting step in mRNA localization. Finally, we incorporate production of mRNA complexes directly into our model and are thus able to reproduce a spatial distribution of mRNA complexes similar to experimental observations. In future, this modelling approach could be applied to improve tracking of individual mRNA complexes.

1   Introduction

The localization of mRNA is crucial in a variety of biological contexts for the targeting of proteins to their site of function. This method of gene expression is particularly relevant for polarized cells such as oocytes and early embryos. For example, the axes of *Drosophila Melongaster* are established through regulation of gradients in *bicoid* (*bcd*), *gurken* (*grk*), *oskar* (*osk*) and *nanos* (*nos*) mRNA (**?**). mRNA localization has been observed in a variety of species and cell types, including *Drosophila* and *Xenopus* oocytes, chicken fibroblasts, yeast and bacteria (**? ? ? ? ?**), thus demonstrating that this process is ubiquitous and not limited to large cells. *Drosophila* relies heavily on the asymmetric locatization of mRNAs to coordinate early development processes both spatially and temporally.

Recent advances in imaging techniques and image analysis technologies (**? ? ?**) have allowed an advancement of our understanding of the mechanisms of mRNA localization. The use of flourescence *in situ* hybridisation (FISH) enables single molecules of mRNA to be labelled with high levels of sensitivity and specificity. The development of the MS2-MCP system, which consists of a MS2 bacteriophage RNA stem loop bound by MS2 coat protein fusion to a flourescent protein (**?**), has provided great benefits for the imaging of live cells *in vivo*. The MS2-MCP system has been used successfully in the visualisation of *nos*, *grk*, *bcd* and *osk* mRNAs in Drosophila (**? ? ? ?**). Although there are limitations to these technologies dependent on cell type, they have

permitted an improvement of our understanding of intracellular motility and structure. What has become clear is that a variety of different mechanisms are used to ensure localization of mRNAs.

What are not yet well understood are the molecular mechanisms used in the mRNA localization process and why exactly such a variety of transport methods are present.

1.1   Outline of mRNA localization process

Once mRNA has been transcribed from DNA in the nucleus, the process by which it reaches its final site of localization, where it is then translated into protein, can be divided into four main stages: particle formation; nuclear export; transport; and anchoring. We will address each of these processes in turn.

mRNA does not exist in isolation inside the cell. Instead the mRNA molecules bind to proteins to form particle complexes, which are also known as ribonuclearproteins (RNPs). These proteins perform a range of different functions, including translational regulation preventing translation while the mRNA is in transit and may determine the final destination of the particle (**?**). There is also evidence that RNPs are dynamically remodelled during the transport process, to support translational regulation (**?**).

After the mRNA has assembled into RNPs, it must diffuse through the crowded interchromatin spaces in the nucleus until it reaches a nuclear pore complex (NPC). The RNP is then exported out of the NPC into the cytoplasm via

interactions with co-factors. Certain NPCs are more active than others at different times (**?**), but the overall direction of export out of the nucleus is unbiased (**?**).

The transport stage of mRNA localization occurs by different mechanisms for different mRNAs. The most common method observed is transport via molecular motors moving along the cytoskeleton (either microtubules or actin filaments). However, *nos* mRNA localizes using a diffusion and trapping technique (**?**) rather than by active transport. Active transport results in fast, directed motion with velocities of the order of $1 \mu \mathrm{ms}^{-1}$ (**? ? ?**), which is an order of magnitude faster than movement by free diffusion. Bidirectionality was observed by **?**) in the motion of RNP complexes on microtubules in *Drosophila* blastoderm embryos. It should be noted that, conventionally, individual types of molecular motors are thought to move unidirectionally, with Dynein directed towards the minus end of microtubules and Kinesin directed towards the plus end (**?**). The number of motors required for a given RNP complex can vary and may be governed by cis-acting localization elements in the mRNA (**?**). Possible explanations for the bidirectionality include: the bidirectional disorganized distribution of the microtubule network; a tug of war between different molecular motors moving in opposing directions on different microtubules; a tug of war between different motor species all bound to the same RNP complex; reversal of a molecular motor moving on a single microtubule, possibly due to regulation by microtubule associated proteins (MAPs) (**?**). The molecular motors must be correctly joined to their cargo and the motor cargo complex secured to the cytoskeleton. This is ensured by the linkers *Bicaudal D* (*BicD*) and *Egalitarian* (*Egl*) (**?**). Although the cytoskeleton was initially thought to be a stable static network, due to how it was analysed in fixed material, it has recently been revealed to be a dynamic network with a biased random orientation of microtubules (**?**). This underlying structure had been suggested by the work of **?**) who observed RNPs containing *osk* mRNA moving in a biased random walk.

Once the RNP complex has reached its destination, it must be maintained in position by some anchoring mechanism to keep the mRNA localized within the cell. *grk* mRNA is anchored by the molecular motor Dynein (**?**) in the *Drosophila* oocyte and *nos* is trapped by actin (**?**). Dynein was observed by **?**) to act as a static motor without requiring further energy in the form of ATP to function. An alternative hypothesis is that continuous active transport is required to ensure the localization of mRNA, since **?**) found that Dynein motor activity is required to ensure *bcd* localization in the anterior cortex of the Drosophila embryo. The transport process may ensure that RNP complexes are kept near their targeted region of localization by directing them back if they move away.

1.2 Motivation

After transcription and transport out of the nucleus, many mRNAs are localized and translated in the distinct cytoplasmic domains where they function (**? ?**). This process is highly regulated and involves active transport by intricate molecular motors. mRNA localization is crucial for the establishment of polarity of cells and the formation of the basic animal body plan during development (**?**). It may have importance for a wide range of functions involving memory and learning in the nervous system. However, we are a long way from a complete understanding of the mechanisms governing the mRNA cargo transport process.

In particular, the mRNA transcripts that set up the *Drosophila* body axes originate in the nurse cells (see Figure 1) and are actively transported on specialized subpopulations of microtubules into the oocyte through the ring canals (**?**). We will focus on this maternal transfer of transcripts into the oocyte, since this area has been less well studied and it is tractable both to modelling and data collection.


```
/home/harrison/Documents/mRNA/Figures_for_writeu
```

Figure 1: Stages of *Drosophila* development, showing nurse cells labelled as nc, follicle cells denoted by fc and the oocyte marked as oo. The localization of *grk* is shown in blue. Adapted from **?**).

One of the challenges involved in live imaging of cells to examine this transport process is the need to scan a sample in three spatial dimensions as well as in time. This places restrictions on the resolution of data that can be obtained temporally and in the z direction due to the limitations of current imaging technologies (**?**). Although there are physical restrictions on what can be achieved by improvements

in imaging technologies, these could be overcome by modelling of the transport of mRNA cargoes, which may permit sparser sampling in time and finer sampling in space. By simulating from our model, assuming we have knowledge of the current state of a particle and appropriate values for model parameters, we can predict a likelihood for the state of a particle after a short time Δt. This could be used to inform the linkage step to identify particles between frames and may allow larger timesteps Δt to be taken, although a hidden Markov model (HMM) framework may be necessary to update the current state of the particle. An HMM approach has previously been applied to tracking in the context of bacteria (**?**) and for single-molecule diffusive protein data (**?**).

Through mathematical modelling of mRNA transport in the *Drosophila* nurse cell, we aim to better understand the limiting factors in mRNA localisation. This should improve our understanding of which parts of the process are important biologically for its function and could give insight into the molecular mechanisms used. We see also the potential for this modelling framework to be applied using an HMM framework to improve tracking of RNP complexes over longer time periods in three dimensional data.

In this work, we present a simple model for transport of mRNA particles through nurse cells during localization and demonstrate how this model can be effectively parametrised. In Section 2, we construct the velocity jump process model and explain the approximate Bayesian computation (ABC) approach taken in inference of the model parameters. We exhibit the posteriors over the model parameters generated via our inference approach for both *in silico* and *in vivo* data in Section 3. Finally, in Section 4, we discuss some of the implications of our modelling and parameter inference, before exploring possible future directions.

2 Materials and Methods

2.1 Modelling approach

In this context, modelling involves breaking down the mRNA localization process into its key components and identifying the biologically important aspects. We will put in place appropriate, simple modelling assumptions to describe the most important parts of the system biologically. These assumptions should as far as possible be supported by existing experimental observations.

Previous modelling approaches in this area have focused on descriptions of the mechanisms driving molecular motors at smaller spatial and temporal scales (**?**) or with continuous densities of mRNA (**?**) rather than representing single RNP particles which would be more appropriate at low densities. Other studies have generally considered mRNA dynamics in the oocyte in terms of cytoplasmic streaming or diffusion without direct reference to active transport (**?** **?**). Instead we shall describe the active transport of mRNA in the nurse cells

surrounding the ooctye, which are connected to each other and to the oocyte via ring canals, using an individual-based stochastic description.

We propose two different approaches to parameterise our model. The first will involve direct measurement of parameters, such as the average speed of cargo complexes during periods of active transport. The second method requires application of an ABC framework (**?** **?** **?**) to parameterise the model via comparison with certain summary statistics of the data. This will enable comparison between the experimentally measured parameter values and the posterior for these parameters obtained via ABC.

We aim to produce a testable model in the sense that we are able to make predictions from the model based on varying certain parameters, such as the speed of the motor carrying the cargo, which could be tested experimentally using genetic mutants that have behaviour that corresponds to a different value of a given parameter. This should inform our understanding of how transcripts transport and localize.

2.2 Velocity jump process

The movement of RNP cargoes is often described as a biased random walk (**?**). Rather than modelling this transport process as a simple random walk in position, we let the direction of motion vary. This type of model is known as a velocity jump process and has been successfully applied to directed migration of animals and cells (**?** **?**).

/home/harrison/Documents/mRNA/Figures_for_writeu

Figure 2: States and transitions between the states in the velocity jump process model. The speeds in active transport and diffusion are given by ν_1 and ν_2 respectively. Transition rates between states are given by ω_1, ω_2 and λ as shown.

We model a two-dimensional slice through a nurse cell as a rectangular region $[0, L_x] \times [-L_y/2, L_y/2]$, with the x axis aligned along the anterior-posterior axis of the oocyte. The cell nucleus is modelled as a circular region of given radius, R, placed centrally in the cell. This is excluding to the RNP complexes in the model which are reflected by it. The remainder of the nurse cell is accessible to the particles, which seems reasonable since experimentally particles are observed in all parts of the nurse cell (data not shown). Initially, RNP cargoes are placed randomly at time $t = 0$ on the circumference of the nucleus to represent their export from nuclear pore complexes. In the model therefore, all mRNAs are produced simultaneously rather than directly modelling continual production and export over time. Cargoes move randomly, as described below, until eventually they hit a small target at the posterior end of nurse cell representing the ring canal joining to the oocyte, where they are absorbed and removed from the system. Throughout all simulations in this work, we fix $L_x = 52\mu$m, $L_y = 37\mu$m, $R = 10\mu$m and the size of the ring canal as 2μm. The geometry of the system can be seen in Figure 5.

There are two phases of motion included in the model: an active transport phase when the motor is moving on the microtubules with constant speed ν_1; and a slower diffusive phase with constant speed ν_2. Switching occurs between these two phases of motion with exponential waiting times between switching events. Biologically this switching corresponds to the molecular motor falling off and reattaching to the microtubule. As illustrated in Figure 2, switching occurs from the active transport state to diffusion with rate ω_1. From the diffusive phase, reorientations within the same phase occur with rate λ and switching to the active transport phase takes place at rate ω_2. At a reorientation event, the particle remains in the diffusive phase at speed ν_2 and a new direction is picked uniformly on $[0, 2\pi]$. After each switching event from diffusion to the active transport state, the RNP complex also changes direction by picking a new direction of travel at random, where the angle θ is drawn from some distribution $T(\theta)$. Here we choose to define $T(\theta)$ empirically as uniform on $[0, \pi]$ and $[\pi, 2\pi)$ with a bias in favour of $[0, \pi]$ based on biological data. We define a proportion of microtubules, ϕ, aligned in the posterior direction (parallel to the x direction in our model) and thus take the following for $T(\theta)$:

$$T(\theta) = \begin{cases} \frac{\phi}{\pi} & \theta \in [0, \pi] \\ \frac{1-\phi}{\pi} & \theta \in [\pi, 2\pi). \end{cases}$$

2.3   Approximate Bayesian computation (ABC)

In a statistical inference context, the likelihood of data given a certain set of parameters is a central quantity. In particular, it is essential in calculation of the posterior over the parameters given certain data. The posterior is desirable as it offers more information than just point estimates of the model parameters. Although for simple models it may be possible to evaluate the likelihood analytically, often for more complex models the likelihood is not tractable or is very expensive to compute. ABC techniques have been developed to address this issue (**?**). Instead of directly evaluating the likelihood, ABC techniques assume it is possible to cheaply simulate from the model and use this to approximate the likelihood.

In an ABC rejection sampling approach we follow the procedure shown in Algorithm 1, where $\pi(\theta)$ is a prior on the parameters, ρ is a distance metric, $S(x)$ is a summary statistic used to summarize the data, x, and ϵ is a maximum tolerance for acceptance.

Algorithm 1 ABC Rejection Sampling

1: **for** $i = 1$ to n {
2: Sample parameters θ from a prior on those parameters $\pi(\theta)$.
3: Simulate data x from the model $M(\theta)$ using those parameters.
4: Calculate distance from observed data y:
5: **if** $\rho(S(x), S(y)) < \epsilon$ **then** accept the parameters θ
6: **else** reject the parameters θ.
7: }

One drawback of this algorithm is that it depends on appropriate choice of the distance metric, ρ, the summary statistic, $S(x)$, and the tolerance, ϵ. Clearly the quality of the posterior will depend on the choice of these hyperparameters, as, for example, increasing ϵ will decrease the quality of the corresponding posterior. In some settings it may be possible to take the full data, rather than a summary statistic, and to set $\epsilon = 0$ which would give an exact sample from the posterior, but in general this is not possible computationally. We can eliminate the importance of the tolerance to some extent by simulating N samples from the prior, storing all the distances and keeping the closest α quantile of the sampled parameters to the observed data. The quality of the posterior still depends on N and α but the choice of ϵ, which may be dependent on other model parameters, is removed.

The efficiency of the rejection sampling method is low, particularly for small values of ϵ as they entail tiny acceptance rates meaning many wasted samples. More efficient ways of sampling possible parameters θ have been suggested including Sequential Monte Carlo and Population Monte Carlo techniques (**????**). As such, we have opted to use an adaptive Population Monte Carlo (APMC) method (**?**). This is implemented as shown in Algorithm 2, where N is the number of parameter samples to simulate, $N_\alpha = \lfloor \alpha N \rfloor$ is the number to keep at each iteration, $p_{acc_{min}}$ is the minimum acceptance rate at which the algorithm stops, and a Gaussian

kernel $K_\sigma(x, y)$, where:

$$K_\sigma(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-y)^2}{2\sigma^2}}.$$

Algorithm 2 ABC Adaptive Population Monte Carlo

1: Initialise by setting N,α, N_α and $p_{acc_{min}}$.

2: Set $t = 1$.

3: **for** $i = 1$ to N {

4: Sample form prior: $\theta \sim \pi(\theta)$.

5: Simulate data $x \sim f\left(x|\theta_i^{(0)}\right)$.

6: Calculate distance from observed data y via $\rho_i^{(0)} = \rho(S(x), S(y))$

7: Set $w_i^{(0)} = 1$.

8: }

9: Take ϵ_1 as the first α-quantile of $\rho^{(0)}$.

10: Let $\{ (\theta_i^{(1)}, w_i^{(1)}, \rho_i^{(1)}) \} = \{ (\theta_i^{(0)}, w_i^{(0)}, \rho_i^{(0)})|\rho_i^{(0)} \leq \epsilon_1, 1 \leq i \leq N \}$

11: Take σ_1^2 as twice the weighted empirical variance of $\{(\theta_i^{(1)}, w_i^{(1)})\}_{1 \leq i \leq N_\alpha}$

12: Set $p_{acc} = 1$.

13: Update $t := t + 1$

14: **while** $p_{acc} > p_{acc_{min}}$ {

15: **for** $i = N_{\alpha+1}$ to N {

16: Pick θ_i^* from $\theta_j^{(t-1)}$ with probability $\frac{w_j^{(t-1)}}{\sum_{k=1}^{N_\alpha} w_k^{(t-1)}}$, $1 \leq j \leq N_\alpha$.

17: Generate new parameters $\theta_i^{(t-1)}|\theta_i^* \sim \mathcal{N}\left(\theta_i^*, \sigma_{(t-1)}^2\right)$ and data $x \sim f\left(x|\theta_i^{(t-1)}\right)$

18: Set$\rho_i^{(t-1)} = \rho(S(x), S(y))$

19: Set $w_i^{(t-1)} = \frac{\pi(\theta_i^{(t-1)})}{\sum_{j=1}^{N_\alpha}(w_j^{(t-1)}/\sum_{k=1}^{N_\alpha} w_k^{(t-1)})K_{\sigma_{(t-1)}}\left(\theta_i^{(t-1)}, \theta_j^{(t-1)}\right)}$

20: }

21: Set $p_{acc} = \frac{1}{N-N_\alpha}\sum_{k=N_\alpha+1}^{N} \mathbb{1}_{\rho_i^{(t-1)} < \epsilon_{t-1}}$

22: Take ϵ_1 as the first α-quantile of $\rho^{(t-1)}$.

23: Let $\{ (\theta_i^{(t)}, w_i^{(t)}, \rho_i^{(t)}) \} = \{ (\theta_i^{(t-1)}, w_i^{(t-1)}, \rho_i^{(t-1)})|\rho_i^{(t-1)} \leq \epsilon_t, 1 \leq i \leq N \}$

24: Take σ_t^2 as twice the weighted empirical variance of $\{(\theta_i^{(t)}, w_i^{(t)})\}_{1 \leq i \leq N_\alpha}$

25: Update t:=t+1

26: }

Inference on model parameters using ABC has been successfully applied to a variety of types of model in a biological context, often helping to offer biological insight into the model system. Cell migration models have used ABC to process data from scratch assays (**?**) and *in vivo* data (**?**), while an individual-based ecological models with 14 parameters has also applied ABC for parameterisation (**?**).

When using ABC methods, it is important to be aware also of their limitations. It has been highlighted via **?**) that, although the error from the tolerance ϵ can be eliminated in the limit $\epsilon \to 0$, the use of summary statistics of the data may have larger effects, particularly when comparing models using ABC. Some more systematic methods for choice of summary statistics have been suggested, which use minimization of certain measures of entropy to determine which summary statistics are most informative (**?**).

2.3.1 Dependence of weights on distance

At each generation of the APMC algorithm, we calculate a distance for each set of parameters. These ought to be informative for which particles should contribute most heavily to the next generation, but are not used directly in the ABC-APMC algorithm. We have considered how we could use weights in the ABC-APMC algorithm that depend directly on the distances, using the update step:

$$w_i^{(t-1)} = \frac{\pi(\theta_i^{(t-1)})}{\rho_i^{(t-1)}} \frac{w_i^{(t-1)}}{\sum_{j=1}^{N_\alpha} w_j^{(t-1)}}. \qquad (1)$$

However, the weights used in the ABC-APMC algorithm are required to ensure that the importance sampling is performed correctly when sampling from the weighted distribution at the next time step (**?**). Without this exact choice, there is no guaruntee that the resulting posterior would be unbiased; a similar bias due to incorrect choice of weights was highlighted by **?**) in the work of **?**). The appropriate weight for importance sampling to be effective uses the density $d_i^{(t)}$ from which parameter values are generated. Therefore the update step must be $w_i^{(t-1)} = \frac{\pi(\theta_i^{(t-1)})}{d_i^{(t-1)}}$, where $d_i^{(t)}$ is given by the probability to arrive at $\theta_i^{(t)}$ from one of the previous weighted parameter values. Although heuristically, weights directy dependent on distance seem an attractive option, statistically they may result in poor quality posteriors.

2.4 Imaging

We observed *grk* mRNA in *Drosophila* nurse cells tagged using Dendra. Imaging was performed with a wide-field deconvolution system (Delta Vision CORE) from Applied Precision. Image stacks were deconvolved with the Soft-WoRx Resolve 3D constrained iterative deconvolution algorithm (Applied Precision). Segmentation of images was performed using Felzenszwalb's algorithm (**?**) after denoising and deconvolution of the images, with coordinates of particles obtained from centroids of appropriately sized segments. The age of embryos was determined by measuring oocyte size and comparing to a logistic growth model.

3   Results and Discussion

3.1   Sensitivity analysis

We note that our model is dependent on six parameters: $\nu_1, \nu_2, \omega_1, \omega_2, \lambda, \phi$. We are able to explore the dependence of our model on these parameters by performing a sensitivity analysis, varying each paramters over several orders of magnitude. We show in Figure 3 the effect of varying key parameters on the mean first passage time (MFPT) from the nucleus to the posterior ring canal, the mean number of jumps on each path up to absorption, and the mean over different particles of the median jump distance on a single path. The results of this sensitivity analysis demonstrate that the turning rate within the diffusive state, λ, has no noticeable effect on any of these summary statistics (Panel 3e). Decreasing the speed parameters, v_1 and v_2, or the microtubule bias, ϕ, leads to a sharp rise in the MFPT (Panels 3b, 3a). The transition rates ω_1 and ω_2 have opposite effects with an increase in ω_1 increasing the MFPT, and a decrease in ω_2 increasing the MFPT (Panels 3c, 3d). Similar effects are seen for the mean number of jumps per simulation and the mean of the median jump distance.

In addition, we exhibit the dependence of the spatial distribution parallel to the anterior-posterior axis of particles simulated from our model, as shown in Figure 4. For increased motor speeds, ν_1 and ν_2, the particles spread out faster from the nucelus and are directed towards the posterior of the nurse cell. By increasing the microtubule bias, ϕ, the motion of the cargoes becomes much more directed towards to posterior with accumulation near the posterior and absorption upon reaching the ring canal there. For high values of the switching rate ω_1, the distribution of particles is more spread and has not reached the posterior, as little time is spent in the active transport state. For low values of ω_1, the particles do not change direction often in the active transport state so greater accumulation is seen for intermediate values of ω_1. We note that, for large values of ω_2, there is an accumulation of particles near the posterior, but that these are not quickly absorbed, since they spend little time in the diffusive phase of motion. The effect that the turning rate λ has on the spatial distribution appears to be unclear.

Based on direct measurements from manually tracked grk-GFP mRNA complexes in *Drosophila* nurse cells, we are able to make direct preliminary estimates of the parameters of our model. All data were obtained from **?**). Particles are classed as static, paused or active. Paused and active particles will correspond to those in the diffusive and active transport phases of our model respectively. Static particles will be assumed to be either anchored or undergoing anomolous behaviour. The number of total particles counted was $N = 340$ in a 40×40 μm area. The proportions of particles in each of the classes is 25% active, 50% paused and 21% static. Over the timescale observed, particles did not transition between states and these are proportions of particles. Particle speeds were assessed by observing a $10 \times$

10 μm area for 50 timepoints, with images taken in a single z slice at three frames per second. The average speed in the active transport state ν_1 was $1.163 \pm 0.08\mu\mathrm{ms}^{-1}$ from $n = 33$. The average run length was $2.785 \pm 0.66\mu$m. In the paused phase, an average speed, ν_2, of $0.798 \pm 0.6\mu\mathrm{ms}^{-1}$ was observed from 58 particles with an average run length of $0.84 \pm 0.06\mu$m.

Considering the process as a continuous time Markov chain, with active transport and diffusive states, we obtain that the steady state of the continuous time Markov chain is $[\omega_2/(\omega_1 + \omega_2), \omega_1/(\omega_1 + \omega_2)]$. Then assuming ergodicity and neglecting static particles, this approach gives us $\omega_2/(\omega_1+\omega_2) = 2/3$ and $\omega_1/(\omega_1+\omega_2) = 1/3$ by comparing to the proportions of particles observed in each state.

Assuming there are no internal transitions back to the active transport state, then we have $\omega_1 = 0.42\mathrm{s}^{-1}$ and hence $\omega_2 = 0.84\mathrm{s}^{-1}$. We can also deduce that the rate $\lambda = 0.95 - 0.84 = 0.11\mathrm{s}^{-1}$. In addition, we assume the distribution of microtubule orientations in the nurse cells is similar to that in the oocyte, since there is data available for the oocyte (**?**) but not for the nurse cell. Therefore we take $\phi = 0.58$.
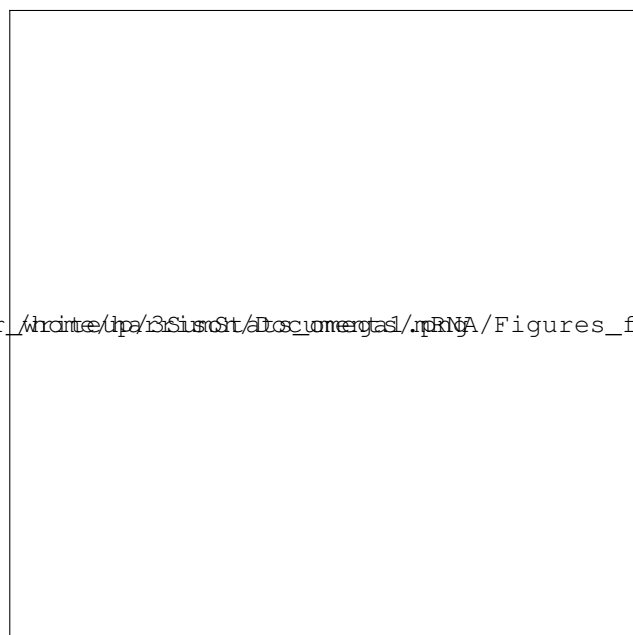
This gives parameter estimates of $\nu_1 = 1.16\mu\mathrm{ms}^{-1}$, $\nu_2 = 0.80\mu\mathrm{ms}^{-1}$, $\omega_1 = 0.42\mathrm{s}^{-1}$, $\omega_2 = 0.84\mathrm{s}^{-1}$, $\lambda = 0.11\mathrm{s}^{-1}$, $\phi = 0.58$.

Based on simulations from the model, we obtain a MFPT, averaged across 1000 particles, of $1070 \pm 840\mathrm{s} \approx 18$min. We plot the evolution of a typical path in Figure 5. Given that stage 5 in *Drosophila* oocytes lasts five hours, this timescale for the MFPT appears to be an order of magnitude shorter than the timescales over which *grk* mRNA transport is observed in oogenesis, suggesting that the limiting step in the transport and localization of the mRNAs is their production from the nucleus rather than the transport step. We are also able to make predictions on what would happen if we halved the speed of the molecular motors via a genetic mutation. We assume that altering the speed of the molecular motor, ν_1, would leave the speed in diffusive motion, ν_2, unchanged. The resulting distribution is shown in Figure 6 and the MFPT would be $2150 \pm 1600\mathrm{s} \approx 36$min.

3.2   Evaluation of ABC methods for in silico data

We demonstrate the effectiveness of the ABC inference techniques on in silico data, before applying them directly to real data sets. We present first the results of applying ABC methods to our model with a broad prior and known parameter values, as shown in Figure 7.

For each experiment, we simulate a data set from chosen parameters and evaluate its summary statistic. That summary statistic is compared to summaries of data generated from proposed parameters, as described above. We average the results obtained across a number of experiments denoted

/home/harrison/Documents/mRNA/Figures_for_writeup/3SumStats_phi.png

(a) Summary statistics as a function of ϕ.

/home/harrison/Documents/mRNA/Figures_for_writeup/3SumStats_nu.png

(b) Summary statistics as a function of ν_1 and ν_2.

/home/harrison/Documents/mRNA/Figures_for_writeup/3SumStats_omega1.png

(c) Summary statistics as a function of ω_1.

/home/harrison/Documents/mRNA/Figures_for_writeup/3SumStats_omega2.png

(d) Summary statistics as a function of ω_2.

/home/harrison/Documents/mRNA/Figures_for_writeup/3SumStats_lambda.png

./Q_distn_phi-eps-converted-to.pdf

./Q_distn_nu-eps-converted-to.pdf

(a) Spatial distribution as a function of ϕ.

(b) Spatial distribution as a function of ν_1 and ν_2.

./Q_distn_omega1-eps-converted-to.pdf ./Q_distn_omega2-eps-converted-to.pdf

(c) Spatial distribution as a function of ω_1.

(d) Spatial distribution as a function of ω_2.

./Q_distn_lambda-eps-converted-to.pdf

(e) Spatial distribution as a function of λ.

Figure 4: The spatial distribution in the x direction of particles simulated for parameters as in Figure 3. For each of (a) to (e), we vary a parameter in turn. For (b), $\nu_2 = \nu_1/2$ while varying ν_1. Results are shown for 100 particles at time $t = 270s$. Note the final spatial box (indexed as box 52) represents particles that have been absorbed and removed from the system.

/home/harrison/Documents/mRNA/Figures_for_writeup/Posterior_Scatter_N_2000_RS

(a) Posterior obtained for $N = 2000, \alpha = 0.1$.

/home/harrison/Documents/mRNA/Figures_for_writeup/Hist_of_posterior_N_2000_RS_

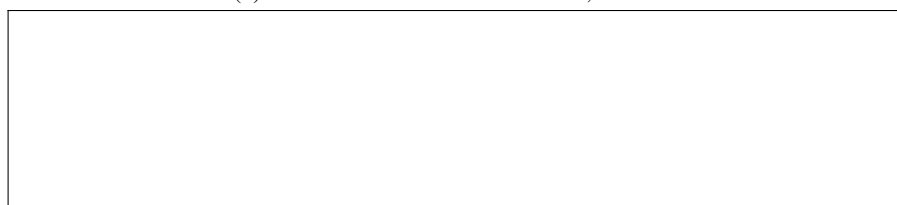(b) Posterior obtained for $N = 2000, \alpha = 0.1$.

Figure 5: The evolution of a typical path until absorption at the posterior ring canal, showing also the geometry in the model. Small black circles on the path show switching events between the active transport and diffusion states. The dashed box represents the boundary of the nurse cell, the large filled black circle shows the nurse cell nucleus, and the thick black line at the posterior represents the ring canal into the oocyte.

$n_{repeats}$. As summary statistics we take the spatial distribution of the cargoes averaged in the x direction at given times $t_1, ..., t_\kappa$ and use a symmetric version of the Kullback-Leibler divergence as our distance to compare distributions. Uniform priors are used for each parameter between biologically realistic maximum and minimum values. In Figure 7, scatterplots and heatmaps of the pairwise parameter distributions are shown, along with histograms for each parameter.

We obtain most information from the posterior for the microtubule bias, ϕ, while the speed parameters, ν_1 and ν_2, are also identifiable. The remaining rate parameters have posteriors similar to the uniform prior and it is harder to recover the original parameter values used. This suggests that the microtubule bias and the speeds have most effect on the efficiency of the summary statistic chosen, which agrees with our earlier sensitivity analysis.

3.3   Incorporation of production into the model

Since we have suggested that production is an important part of the mRNA localization process, we choose to examine the effect of incorporating this into our model set out previously in Section 2.2. We assume a constant rate of production,

Figure 6: The spatial distribution resulting at time $t = 270s$ from 1000 model simulations with parameters $\phi = 0.58$, $\nu_2 = 0.8\mu$ms^{-1}, $\omega_1 = 0.42$s^{-1}, $\omega_2 = 0.84$s^{-1}, $\lambda = 0.11$s^{-1}, varying ν_1 as 0.58μms^{-1} and 1.16μms^{-1}. This represents the effect of introducing a genetic mutation to alter the speed of molecular motors.

γ, modelling the export of RNP complexes from the NPC as discrete events that occur as a Poisson process with rate γ. Although there is some evidence that the production rate varies during oogenesis, we will assume here that is remains constant for simplification. Upon export of a particle in our model, it is placed randomly on the circumference of the nucleus and the model evolves as outlined previously. We begin at time $t = 0$ with a single particle that has just been exported, and note that since the production rate is constant

By introducing the extra parameter, γ, into our model, there is now an extra parameter to estimate. To estimate this directly, we note that rate of production of mRNA is limited by its rate of transcription in the nucleus. Since the rate of transcription for Drosophila at 22°C is 25 nt/s (BNID 111484) (**?**), and the length of the *grk* gene is 295 nt (**?**), we can obtain an estimate for the production rate of *grk* mRNA of $\gamma = 0.08$s^{-1}. Simulation from the extended model with production at rate $\gamma = 0.08$s^{-1} suggests that a spatial steady state is reached after a short initial transient, with a constant rate of absorption at the posterior ring canal. Results are shown in Figure 8 and exhibit spatial profiles similar to those seen in experimental data exhibited in Figure 9.

3.4   Application of ABC to Drosophila nurse cell data

Although comparative simulation results with ABC suggest that the detailed summary statistics using MFPT, number of

Figure 8: The spatial distribution evolving over time up to 25 hours from a single simulation of the model with production included and parameters of $\phi = 0.58$, $\nu_1 = 1.16\mu\text{ms}^{-1}$, $\nu_2 = 0.8\mu\text{ms}^{-1}$, $\omega_1 = 0.42\text{s}^{-1}$, $\omega_2 = 0.84\text{s}^{-1}$, $\lambda = 0.11\text{s}^{-1}$, $\gamma = 0.08\text{s}^{-1}$. A logarithmic scale for the colourmap is used to emphasise the presence of small numbers of particles in the nurse cell.

(a) Raw data after denoising and deconvolution.

(b) Spatial distribution of *grk* particles seen in raw data.

Figure 9: Distribution of *grk* MCP Dendra in a stage 5 Drosophila oocyte and surrounding nurse cells. Stage 5 is equivalent to a time point of around $t = 25$ hours, assuming *grk* mRNA production begins at stage 2. A logarithmic scale for the colourmap is used in (b).

jumps and jump distances result in a higher quality posterior, there are limitations to what data can be collected experimentally. In particular, these summary statistics would require detailed tracking data of individual particles over an extended period of time, which would be difficult to obtain. Instead, we use the spatial distribution of particles which is more tractable from single time points. To ensure the timescales of the observed experimental data and simulated data match, we use the model with production of mRNA at rate γ and infer this parameter also.

Applying inference with ABC to these data using a Euclidean distance measure and binned spatial histograms in the x direction as summary statistics, we obtain posteriors for the model parameters as shown in Figure 10. Since simulating from the model with production over longer timescales is more computationally intensive than simulating from the original model, as in Section 3.2, we must sample fewer parameter values, taking $N = 500$ and $\alpha = 0.1$ in the ABC rejection sampling algorithm. The results of the ABC inference show a very tight distribution for γ, despite the low value of N used in these simulations. This suggests that the output of the production model depends most strongly on γ when summarised by our chosen summary statistics. Little extra information beyond the prior is gained for the other parameters, based on these results with $N = 500$.

4 Conclusion

Through use of a model-based approach for explaining mRNA transport and localization, we have been able to represent the dynamics of this process and reproduce behaviour observed experimentally, showing that a biased network of microtubules can maintain localization of mRNA. We constructed a velocity jump process with two phases of

./Auto_hist_w_production2-eps-converted-to.pdf

(a) Histogram for each parameter.

./Auto_heatmap_w_production2-eps-converted-to.pdf

(b) Heatmap for pairwise parameters.

Figure 10: Posterior for each parameter in the full model with production approximated via ABC rejection sampling, using $N = 500$, $\alpha = 0.1$, based on real data from one single image frame. In 10a, histograms for each parameter are shown, with a

motion to represent directed movement via active transport on microtubules and random motion in free diffusion. Since simulations of this model were accessible but the likelihood was intractable, we employed ABC methods to infer the parameters from our model from data collected via live *in vitro* imaging and compared the results of this with values obtained from direct measurement.

The timescales for MFPTs based on simulations from the model are on the order of tens of minutes, rather than hours, for the inferred parameters, suggesting that the transport step is not the rate-limiting stage of the mRNA localization process. Based on the modelling work presented here, we postulate that the production of mRNA in the nucleus of the nurse cells is instead the rate-limiting stage.

By incorporating production at a constant rate into our model, we recover transport on timescales of the order of hours seen during oogenesis. We are also able to simulate spatial distributions of mRNA particles similar to what is observed experimentally.

4.1 Further work

The subcellular environment is in some ways drastically different to that assumed in our spatially homogeneous two-dimensional model. In reality, cells exist in three dimensions and are crowded inside with a heterogeneous environment. One possible method of representing heterogeneity in the cell would be by use of a potential to account for volume exclusion effects, as used by (**?**) in the context of nuclear export of mRNAs. The three-dimensional nature of the environment presents further challenges for tracking, as RNPs move between frames. Although some progress has already been made in this area (**?**), we hope that incorporation of modelling into a tracking framework could enable tracking of particles over longer timescales at increased resolution. Motivated by a Kalman filter type approach (**?**), we would like to investigate further using our model combined with regular inputs of data over time to assist particle tracking algorithms in the linkage step of tracking. Therefore in future the model should be extended to enable us to capture behaviour in three dimensions.

Furthermore, it would be beneficial to examine quantitatively the agreement between our simulated results using the model with production and experimental data to provide model validation. If further time were available, we would also seek to parametrise this full model from experimental data via ABC.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

A Analytics

For many models, analytical techniques can offer more general insight than simulating from the model with specific parameters. We show here what can be achieved with these techniques and give some indicative results. First, we must make some simplifying assumptions to ensure the model is analytically tractable. We will consider only one spatial distribution on the interval $[0, L_x]$, parallel to the x axis. Furthermore, we assume $\nu_1 \gg \nu_2$ so that it is reasonabe to ignore the diffusive phase of motion, leaving only reorientations with the active transport phase at rate ω_1.

Suppose we have a velocity jump process with one mode of motion in one dimension where the speed in the x direction varies as $\nu_1 \sin(\theta)$ when the velocity is oriented in direction θ in two dimensions. Let $p(x, \theta, t | y, \alpha, s)dt$ be the probability that the particle is at position x moving in direction θ in the time interval $[t, t + dt]$ having started previously at time s at position y moving in direction α. We will now suppress the conditioning and write only $p(x, \theta, t)$.

The evolution of the probability is governed (**?**) by :

$$\frac{\partial p}{\partial t} + \nu_1 \sin\theta \frac{\partial p}{\partial x} = -\omega_1 p + \omega_1 \int_0^{2\pi} T(\theta) p(x, \theta', t) \mathrm{d}\theta', \quad (2)$$

where the turning kernel $T(\theta)$ gvies the probability of reorienting into a new angle θ and we assume this does not depend on the previous angle. We define $\langle f(\theta) \rangle = \int_0^{2\pi} f(\theta) \mathrm{d}\theta$ and $\langle\langle f(x, \theta) \rangle\rangle = \int_0^{2\pi} \int_0^{2\pi} f(\theta) \mathrm{d}\theta \mathrm{d}x$.

The orientation of the velocity, θ, is updated via the reorientation kernel $T(\theta)$ after every transition. It depends only on the initial condition and the kernel $T(\theta)$. After a sufficiently long time, the initial condition will have been forgotten and we should be able to decompose the probability p such that:

$$p(x, \theta, t) = T(\theta) q(x, t), \quad (3)$$

since whenever the angle θ is updated, this is independent of time or position. For a particle to forget the initial condition, it must simply execute at least one jump. After a time $t_{\min} = 1/\omega_1 \log(100)$, then 99% of the particles should have executed a jump, since this is the 99th percentile of an exponential distribution with parameter ω_1. Assuming a value of $\omega_1 = 0.42 \mathrm{s}^{-1}$, then $t_{\min} = 11\,\mathrm{s}$ which is much less than our timescales of interest.

Now integrating (2) over θ, we obtain:

$$\frac{\partial}{\partial t} \langle p \rangle + \frac{\partial}{\partial x} \langle \nu_1 \sin(\theta) p \rangle = 0, \quad (4)$$

which becomes via our assumption (3):

$$\frac{\partial q}{\partial t} + \nu_1 \langle T(\theta) \sin(\theta) \rangle \frac{\partial q}{\partial x} = 0. \quad (5)$$

From multiplying by x and considering the first spatial moment, we have:

$$\frac{\partial}{\partial t}(\langle\langle xp\rangle\rangle) = \nu_1 \langle\langle \sin\theta p\rangle\rangle,$$

$$\frac{\partial}{\partial t}(\langle\langle xT(\theta)q\rangle\rangle) = \nu_1 \langle\langle \sin\theta T(\theta)q\rangle\rangle,$$

$$\frac{\partial}{\partial t}\left(\int_0^L xq\,\mathrm{d}x\right) = \nu_1 \langle\sin\theta T(\theta)\rangle \int_0^L q\,\mathrm{d}x.$$

Suppose we let $\mu^{(1)}(t) = \int_0^L xq(x,t)\mathrm{d}x$. Then we can write:

$$\frac{\mathrm{d}\mu^{(1)}}{\mathrm{d}t} = \nu_1 \langle\sin(\theta)T(\theta)\rangle \int_0^L q\,\mathrm{d}x,$$

$$\mu^{(1)}(t) = \mu^{(1)}(0) + \nu_1 \langle\sin(\theta)T(\theta)\rangle \int_0^t \int_0^L q\,\mathrm{d}x\mathrm{d}t. \quad (6)$$

We can interpret $\int_0^L q(x,t)\,\mathrm{d}x$ as a survival probability of a particle at time t. After very short times, when the particle cannot physically have exited the system yet, this will equal one, leading to a linear form for the average position, $\mu^{(1)}(t)$. At later times, the particle may have exited the system and the survival probability will decrease to 0. If we release an ensemble of particles, we have the average position of this ensemble changes according to the mean net velocity, $\nu_1 \langle\sin(\theta)T(\theta)\rangle$, until particles begin to be absorbed. We note that is mean net velocity depends on the microtubule bias. For our specified choice of $T(\theta)$, we have $\nu_1 \langle\sin(\theta)T(\theta)\rangle = \nu_1(4\phi - 2)/\pi$. Using our estimated parameters, $\phi = 0.58$ and $\nu_1 = 1.16\mu\mathrm{ms}^{-1}$, we find an estimate of the mean net velocity of $0.12\mu\mathrm{ms}^{-1}$.

By similar arguments, expressions for higher spatial moments and angular moments can be obtained. For the jth spatial moment, we have

$$\frac{\partial\mu^{(j)}}{\partial t} = j\nu_1 \langle\sin(\theta)T(\theta)\rangle \mu^{(j-1)}, \quad (7)$$

leading to a hierarchy of equations. Extensions of these arguments to the full model with two phases of motion are also possible.