AMSI VACATION RESEARCH SCHOLARSHIPS 2020–21

Get a Thirst for Research this Summer

What kind of random walk are these biological cells doing?

Rebecca Rasmussen

Supervised by Prof. Barry Hughes

The University of Melbourne

Vacation Research Scholarships are funded jointly by the Department of Education, Skills and Employment and the Australian Mathematical Sciences Institute.

AMSI Australian Mathematical Sciences Institute

Australian Government
Department of Education, Skills and Employment

Contents

1 Abstract

Observing cellular motion on artificial, planar substrates has potential for distinguishing phenotypes and ultimately assisting in the study and diagnosis of disease. Practical limitations ensure that position measurements can only be made at relatively low frequencies and precision is limited. To investigate how reliably such data can be used to describe the true motion of the cell, we use synthetic data generated from two known stochastic models to see if we can distinguish the models and recover their parameters. In this context, the cells of interest typically display some directional persistence. Correspondingly we used a lattice-based persistent random walk (PRW) and a continuum run and tumble (R&T) to generate the data. First we fit the data to a probability density function (pdf) describing the displacements of run and tumble motions. The models were found to be distinct under this analysis but could both have been generated by a run and tumble mechanism. Next, we computed a number of measures relating to the speed and persistence of a cell. Again we found that the estimated parameters are not meaningful in the context of the underlying mechanism but they do allow us to differentiate the models. Only when positions were sampled at a rate similar to the cell turning rate, could we gain meaningful insight into the movement mechanisms. So, when information is lost from snapshot data due to a relatively large time between measurements, it is unlikely estimates of movement are correct and meaningful. However, we may still be able to use these data as a means of diagnosis and cell sorting.

2 Introduction

Cellular motion is a key aspect of embryonic growth, homeostasis and disease research. Different cells function in different ways, resulting in various type-specific movements [1, 2]. Similarly, if function varies within a organism, as between diseased and healthy cells, movement can vary too [3, 4]. This has potentially huge applications for the detection and diagnosis of diseases. When morphological differences are too difficult to observe, we may be able to use a detectable difference in movement as a diagnostic tool.

It is possible to observe isolated cells migrating on artificial substrates, removed from any interactions that may complicate the data we are trying to model [5, 6]. While the observed cell movements are therefore unnatural, the same underlying difference in mechanism causing the movement change is still present. Even when a cell is isolated, the collection of movement data is not a simple task. The collected data often comprise a set of position snapshots of a cell at regular time intervals, watched over a number of hours or days [7]. There is an upper and lower limit of the frequency with which we can make observations. There is an unavoidable limit to the accuracy of position measurements due to equipment and tracking limitations, and the shape-change of the cell. The tracked nucleus may move while the cell itself remains stationary, so measuring position too frequently can result in very noisy data [8]. To combat this, we can effectively bin the data to a lattice (where positions are integer multiples of the same length scale) where cells may occasionally be interpreted as stationary, but more often move a few to a hundred lattice spacings between measurements. There is also the

matter of accumulating sufficient amounts of data. To do so it is often necessary to observe a large number of cells simultaneously from a given tissue sample, necessarily increasing the time between samples of a given cell. Measuring too infrequently however, generates potential to lose information. We cannot be certain of a cell's movements between observations, so important information about turning events and speed may be lost. When the cell motion can be conjectured to contain an element of directional persistence with occasional changes in direction, another issue arises. A very broad class of stochastic processes approach Brownian motion over long time intervals [9]. In such case, if we wait too long between measurements, we may see none of the underlying directional persistence, the legacy of which is concealed within the apparent diffusion constant and cannot be extracted from it in any meaningful, model independent way

To explore this kind of data, where information has potentially been lost due to relatively large sampling intervals, we employ random walk models. Random walks have been used to model many systems in biology and have proven to be a useful tool in the analysis of movement mechanisms [10, 11, 12, 13]. Many types of random walks exist but one of the simplest gives an agent the option to jump a discrete distance per discrete time step in any of the cardinal directions on a lattice, a simple random walk. If the agent has a higher-than-random probability to move in the direction of the previous step, we have a persistent random walk (PRW). This kind of walk simulates the tendency for cells to keep moving in straight lines, at least over short periods of time. Other models have been suggested for cells that seem to deviate from the predictions of the PRW [1, 2]. One such model is the run and tumble (R&T). This model takes place in continuous space and time, and has an agent move in one direction for an exponentially distributed random time, turn in a random direction and continue. Data generated from random walk models have a known mechanism and thus known movement parameters. This provides a useful basis to analyse how informative snapshot data can be.

To investigate the feasibility of obtaining useful information from such data, we used synthetic data generated by known stochastic models; a persistent random walk on a two-dimensional square lattice and a run and tumble model. We used pdf fitting and movement measures to investigate whether limited sets of synthetic cell migration data can be used to distinguish between the two models. Most movement measures pertain to either cell speed or directional persistence—the tendency for the cell to move in the same direction. The R&T had faster agents than the PRW, and the PRW had more directionally persistent agents (straighter trajectories). Thus, we expected the estimated speed (and associated parameters) to be higher for the R&T than for the PRW. Additionally, we expected the estimated persistence (and associated parameters), to be larger for the PRW than for the R&T.

Statement of Authorship

Barry Hughes conceived the main conceptual ideas and the project outline, supervised the work and proofread the report. Yining Ding (another UoM undergraduate) developed the initial PRW code in Python and produced

some of the data. Rebecca Rasmussen developed the R&T code in MATLAB, produced the majority of the data/results, reported and interpreted the results and wrote the report.

3 Methods

3.1 Models

All modelling and pdf fitting was done in MATLAB (R2018b).

Persistent Random Walk

The persistent random walk (PRW) takes place on an infinite two-dimensional lattice with an associated integer coordinate at the centre of each cell. An agent undertaking this walk can move one step up, down, left or right per time unit, or not move at all. The agent is initialised at the origin, $(0, 0)$. To decide the initial direction, we generate a random number $rand \overset{d}{\sim} R(0, 1)$. If;

- $rand \in (0, 0.25]$, agent moves in the positive x direction
- $rand \in (0.25, 0.5]$, agent moves in the negative x direction
- $rand \in (0.5, 0.75]$, agent moves in the positive y direction
- $rand \in (0.75, 1)$, agent moves in the negative y direction

After the initial step, the direction of subsequent steps are decided:

- With probability a, the agent moves forward (in the direction of the previous step)
- With probability b, the agent reverses
- With probability c, the agent does not move
- With an equal chance of $d = (1 - a - b - c)/2$, the agent moves left or right

If $a > b, c, d$, the agent is most likely to continue in the same direction at each step, giving the agent persistence—this was the case for each instance of this model used in this study. If the agent does not move at a given time, the direction of the next step will be randomly selected as per the mechanism of the initial step. Since the agent moves unit distance in unit time with probability $1 - c$ and pauses with probability c, the average speed at the individual step scale is $1 - c$. Over longer length scales, this speed would only be maintained if the walker underwent no direction changes (the case $b = d = 0$) so that in general the measured speed will drop as we increase the time interval between successive position measurements. For those random processes that converge to diffusion in the long-time limit, the root-mean-square displacement grows as the square root of the time between observations and the inferred speed decays to zero as the time between observations grows.

Run and Tumble

An agent moving per the run and tumble model moves in continuous 2-dimensional space and continuous time. This agent also begins at the origin, $(0, 0)$. The initial direction is generated from $\overset{d}{\sim} R(0, 2\pi)$. The agent moves for an exponentially distributed length of time, from $\overset{d}{\sim} Exp(\lambda)$ with mean $1/\lambda$, at constant speed v. Then, a new random direction is generated from $\overset{d}{\sim} R(0, 2\pi)$ and the agent continues by this mechanism.

Recording Synthetic Data

Both models were run over N time units. To simulate the recording of real-world snapshot data at regular time intervals, the positions of each agent was recorded every $n \ll N$ time units. This leaves us with $\lfloor (N/n) \rfloor$ coordinates for each simulated trajectory. The run and tumble model data was binned to a lattice (with *lattice spacing* $= 1$) to match the effective binning of real cell motion data.

3.2 Generating Data

We generated 100 trajectories for each model and 100 coordinates were generated for each path (i.e. $N = 100n$). To synchronise the length scale of the models, we chose model parameters so that trajectories rendered a mean displacement of ≈ 6 units per 15 units of time. This occurred when the PRW had parameters: $a = 0.6, b = 0.1$ and $c = 0.1$, and the R&T had parameters: $\lambda = 14$ and $v = 3$.

At this point, we note that the speed of the R&T ($v = 3$) is greater than the PRW ($v \approx 1$). On average, the PRW agent will make a turn 35.25% of time steps, while the run and tumble agent will turn 14 times per time unit. So we have relatively low persistence in the PRW but even lower persistence in the R&T. Thus, the walks are suitably distinct from each other for the purposes of this comparison. We also have a low persistence for each model and so if the data has potential to lose persistence information, we should be able to see this at relatively short measurement intervals (convenient for modelling).

In the first part of this study, we compared snapshot data from the persistent random walk model to snapshot data from the run and tumble model for sampling interval $n = 15$. For the second part, we decreased the measurement interval n from 15, to see the effects on the data.

3.3 Analysing Data

Part 1 - Preliminary comparison: $n = 15$

pdf fitting:

The first analysis we performed involved fitting the data from each model to a pdf. The pdf describes the probability density of displacement R at time t resulting from a run and tumble model with *turning rate* $= \lambda$ and *speed* $= v$ (adapted from [14], Eqn. 1.3/1.4).

$$f_R(r) = \frac{r\lambda}{v^2} \exp\left[\lambda\left(t^2 - \frac{r^2}{v^2}\right)^{1/2} - \lambda t\right]\left(t^2 - \frac{r^2}{v^2}\right)^{-1/2} + \frac{2r}{t^2 v^2}\exp(-\lambda t), \qquad 0 \leq r \leq vt \tag{1}$$

The displacements

$$r = \sqrt{(x_2^2 - x_1^2) + (y_2^2 - y_1^2)} \tag{2}$$

from coordinates 0 to 100, 50 to 100, 75 to 100, 90 to 100, 95 to 100 and 99 to 100 (that is, over time intervals of 1500, 750, 375, 150, 75 and 15 units) from each model were fitted to the pdf (Eqn. 1) with MATLAB's maximum likelihood estimation function *mle*. The initial values of v and λ were 2 and 10 respectively (chosen to be somewhat close to the real parameters of the R&T). The upper and lower bound set for the parameters (same for both v and λ) were 30 and 0.1 respectively—the parameters are positive and no larger than 30. The maximum number of iterations and function evaluations for *mle* were set to 10,000, which was adequate for this analysis.

The parameters λ and v were estimated for the set of displacements and empirical distributions were plotted. The estimated parameters were qualitatively compared between models.

Movement measures:
Next we compared a series of parameters relating to motion, estimated from position snapshot data. These measures were *speed*, *directionality ratio*, *mean square displacement* (MSD) and the α *value*—the exponent for the MSD.

To calculate these values, we used an open source program, DiPer, developed by researchers Gorelik and Gautreau [15]. It was run as a series of Microsoft Excel macros through the built-in application 'Visual Basic Editor'. DiPer takes position snapshot data and returns the above movement measures, among others. This program was developed for cellular motion data collected often enough to not miss small cell displacements. However, in this study we test whether it is possible to acquire meaningful values from sparse data through the same analysis.

The *speed* calculated by DiPer was the estimated average speed between measured positions for a population of C cells (Eqn. 3; Fig. 1).

$$speed_i = \langle \frac{d_i}{n} \rangle_C \tag{3}$$

where d_i denotes the straight-line distance between positions i and $(i-1)$ and n is the time between measurements. It should be noted that this is only an effective speed as the agent does not necessarily move in a straight line between consecutive positions.

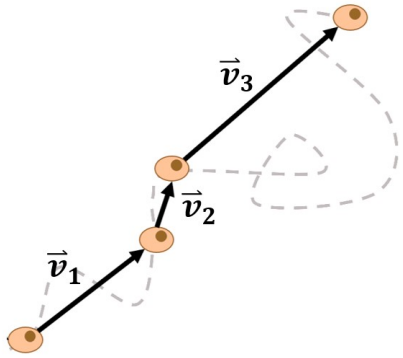

Figure 1: Illustration of the straight-line distance between measured positions (black arrows) in contrast to the actual path taken (grey dashed line). The speed is calculated using the straight-line distances

The *directionality ratio* is the average ratio of displacement over accumulated distance at time t over a population of C cells (Eqn. 4; Fig. 2).

$$Directionality\ ratio_t = \langle \frac{d_t}{D_t} \rangle_C \tag{4}$$

where d_t is the straight-line distance between the initial agent position and the agent position at time t and D_t is the sum of the distance between adjacent points up to time t. This value gives a measure of directional persistence ranging from 0 to 1, where 1 indicates a perfectly straight path.

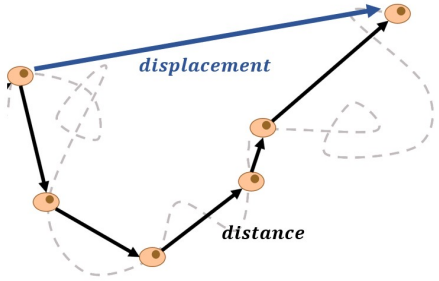

Figure 2: Illustration of the different distance measures: the straight-line distance between times 0 and t (blue arrow), the straight-line distance between adjacent points (black arrows) and the actual path taken by the cell (grey dashed line).

The *mean square displacement* is calculated (with DiPer) using overlapping time intervals to increase the number of points for each trajectory (Eqn. 5; see [15] for discussion).

$$MSD(s) = \frac{1}{S-s+1}\sum_{i=0}^{S-s}[(x_{(i+s)n} - x_{in})^2 + (y_{(i+s)n} - y_{in})^2] \tag{5}$$

Equation 5 gives mean square displacement of an agent over a moving window of s steps. S is the total number of steps in the trajectory and n is the time between steps. This equation is applied to each trajectory and then we take an average (for each time step). The MSD provides a good measure of the surface area covered by the agents over time.

The MSD is often expressed as a log-log plot over time. The slope of the linear fit to this plot is the α value, which gives another measure of directional persistence. We have a straight path when $\alpha = 2$ and diffusion (randomly moving agents) when $\alpha = 1$.

The trajectory snapshot data generated from both models were run through the DiPer program. The differences between average speeds and between average α values were analysed with a t-test. A paired t-test was used to analyse MSD and directionality ratio differences across each time point. All statistical analysis was carried out in RStudio (version 3.6.3, R Core Team 2020).

Part 2 - Changing n

For the second part of the study, we decreased the time between measurements: we tested $n = 1, 3, 5$ and 10. Again, these snapshot data were run through the DiPer program and speeds and directionality ratios were compared between models (the same way as in part 1).

4 Results

4.1 Part 1 - Preliminary comparison: $n = 15$

pdf fitting:

The maximum likelihood function for this pdf has multiple local maxima and is not concave enough to directly optimise the parameters. As a result, the estimated parameters $\hat{v}$ and $\hat{\lambda}$ for the R&T model were at least an order of magnitude less than the actual parameters (Table 1). All estimations for the PRW were substantially larger than those for the R&T model. The estimated parameters for the PRW were themselves a poor match for the known mechanism.

Table 1: Parameters estimated by mle pdf fitting for the R&T and the PRW. Data was fit over coordinates 0 to 100, 50 to 100, 75 to 100, 90 to 100, 95 to 100 and 99 to 100, corresponding to time intervals of 1500, 750, 375, 150, 75 and 15 units.

	Run and Tumble	Persistent Random Walk
$t = 1500$	$\hat{v} = 0.58$	$\hat{v} = 6.22$
	$\hat{\lambda} = 0.10$	$\hat{\lambda} = 30.0$
$t = 750$	$\hat{v} = 0.41$	$\hat{v} = 6.80$
	$\hat{\lambda} = 0.10$	$\hat{\lambda} = 30.0$
$t = 375$	$\hat{v} = 0.29$	$\hat{v} = 6.18$
	$\hat{\lambda} = 0.10$	$\hat{\lambda} = 30.0$
$t = 150$	$\hat{v} = 0.18$	$\hat{v} = 6.10$
	$\hat{\lambda} = 0.10$	$\hat{\lambda} = 30.0$
$t = 75$	$\hat{v} = 0.13$	$\hat{v} = 6.36$
	$\hat{\lambda} = 0.10$	$\hat{\lambda} = 30.0$
$t = 15$	$\hat{v} = 0.10$	$\hat{v} = 0.98$
	$\hat{\lambda} = 0.23$	$\hat{\lambda} = 0.73$

The analysis did distinguish between the two models over most time intervals (Fig. 3). The empirical distributions only looked similar over a time interval of 750. The estimated parameters remained consistent over the larger time intervals (Table 1) but the analysis broke down at lower times ($t \leq 75$).

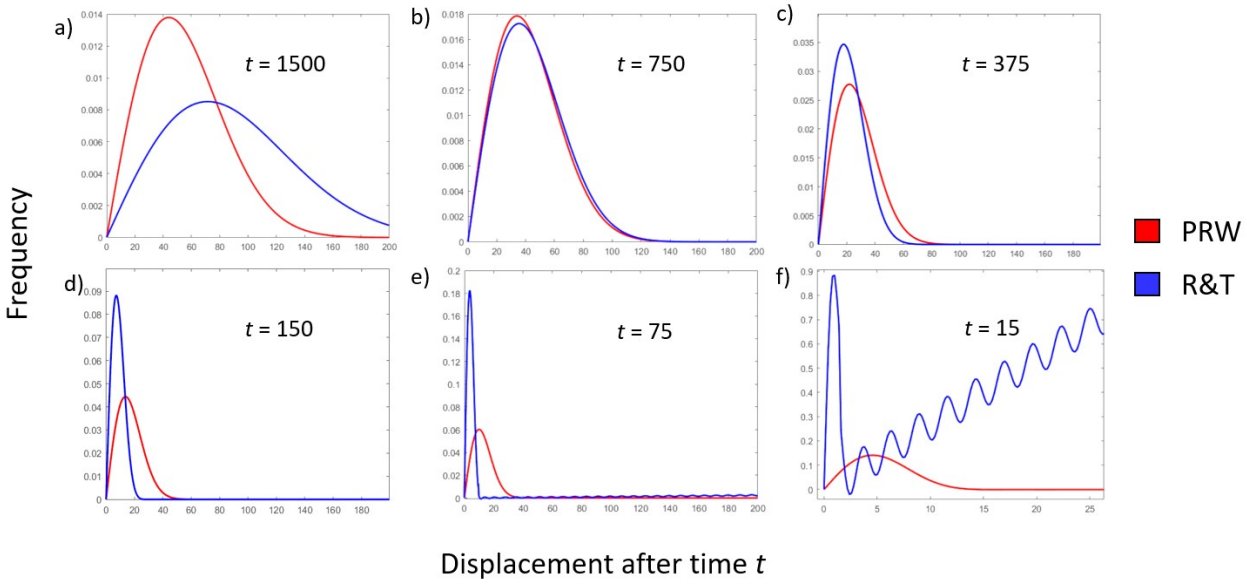

Figure 3: Empirical pdf's from the mle parameters for the PRW and the R&T. Data was fit over coordinates a) 0 to 100, b) 50 to 100, c) 75 to 100, d) 90 to 100, e) 95 to 100 and f) 99 to 100.

Movement measures:

From our movement measure comparison we found that the estimated speed was significantly larger ($t_1 62 = 42.3$, p-value< 0.001) for the PRW than for the run and tumble model ($mean \pm SE$: PRW $= 0.365 \pm 0.002$; R&T $= 0.256 \pm 0.001$; Fig. 4a). Additionally, the MSD was significantly larger for the PRW than for the run and tumble model over time ($t_4 7 = 11.8$, p-value < 0.001; Fig. 4b).

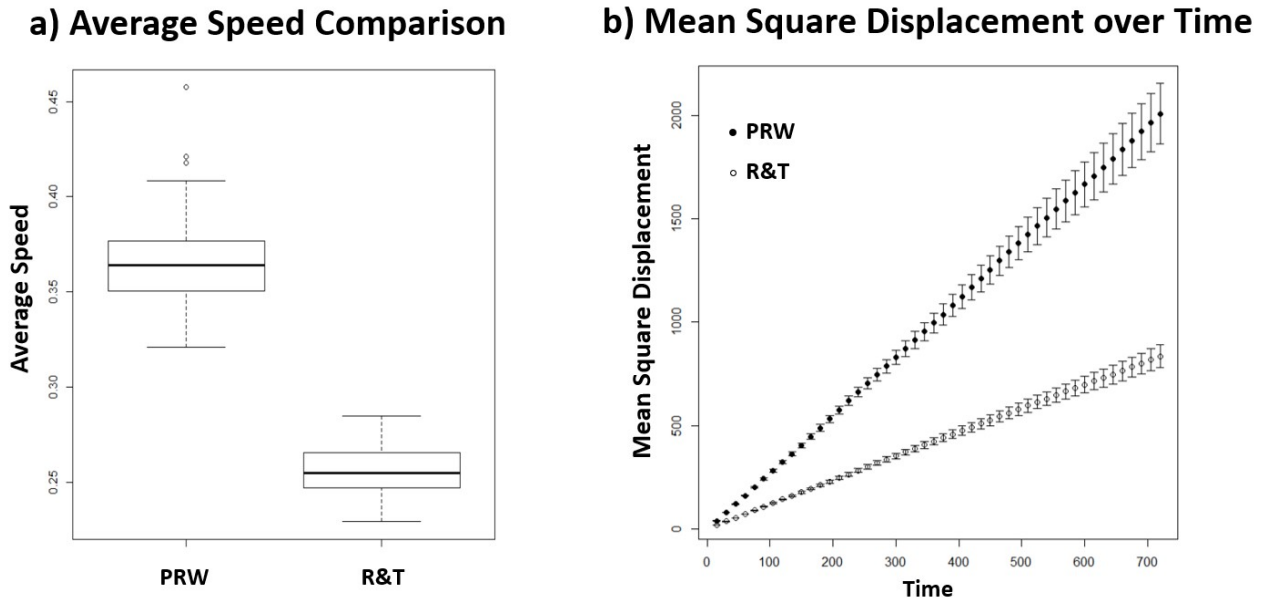

Figure 4: Comparison of estimated movement measures for the PRW and R&T data. a) Box plot showing average estimated speed; bold lines represent medians, boxes represent the IQR's, lower/upper whiskers represent first/third quantile ± (1.5*IQR) and data points represent outliers. b) Time series plot showing average mean square displacement over time; Error bars represent mean ± SE.

The directionality ratio was not significantly different between the two models over time ($t_9 8 = -0.9$, p-value= 0.38; Fig. 5a). Likewise, the α value was not significantly different ($t_1 98 = -1.0$, p-value = 0.31) between the two models ($mean \pm SE$: PRW $= 0.98 \pm 0.02$; R&T $= 1.01 \pm 0.02$; Fig. 5b).

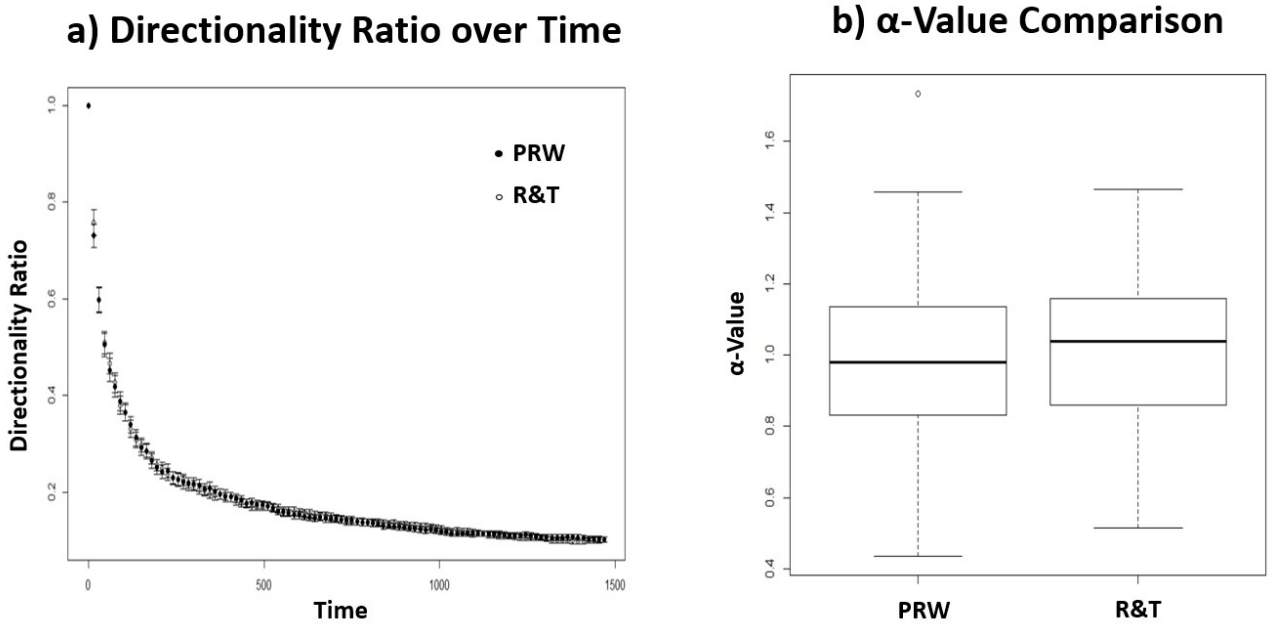

Figure 5: Comparison of estimated movement measures for the PRW and R&T data. a) Time series plot showing average directionality ratio over time; Error bars represent mean ± SE. b) Box plot showing average α-value; bold lines represent medians, boxes represent the IQR's, lower/upper whiskers represent first/third quantile ± (1.5*IQR) and data points represent outliers.

4.2 Part 2 - Changing n

As we decrease n from 15, the estimated average speed increases at similar rates for both models, slowly approaching the real speeds (Fig. 6a). For $n > 1.5$, the estimated speeds are still opposite in magnitude to the real model speeds. When $n < 1.5$, the estimated speed for the R&T is finally greater than that for the PRW. At $n = 1$ (the minimum for the DiPer program), the estimated speed for the PRW is close to the real PRW speed but the R&T speed is still greatly underestimated (Table 2).

Table 2: Estimated average speed from PRW and R&T data over decreasing measurement intervals n. Bold statistics: speeds that are significantly different and in the correct order magnitude-wise.

n	Run and Tumble (mean ± SE)	Persistent Random Walk (mean ± SE)
1	**0.978 ± 0.009**	**0.909 ± 0.003**
3	0.577 ± 0.004	0.691 ± 0.003
5	0.444 ± 0.003	0.576 ± 0.003
10	0.315 ± 0.002	0.434 ± 0.002
15	0.256 ± 0.001	0.365 ± 0.002

The directionality ratios for the two models are only significantly different for approximately $n < 2.5$ (Fig. 6b; Table 3). At this point, the directionality ratio (at the last point in the trajectories) for the PRW becomes greater than for the R&T.

Table 3: Estimated average directionality ratio between first and last coordinates of PRW and R&T data over decreasing measurement intervals n. Bold statistics: directionality ratios that are significantly different and in the correct order magnitude-wise.

n	Run and Tumble (mean ± SE)	Persistent Random Walk (mean ± SE)
1	**0.099 ± 0.006**	**0.153 ± 0.008**
3	0.094 ± 0.005	0.121 ± 0.006
5	0.096 ± 0.005	0.110 ± 0.006
10	0.101 ± 0.005	0.109 ± 0.006
15	0.100 ± 0.005	0.101 ± 0.006

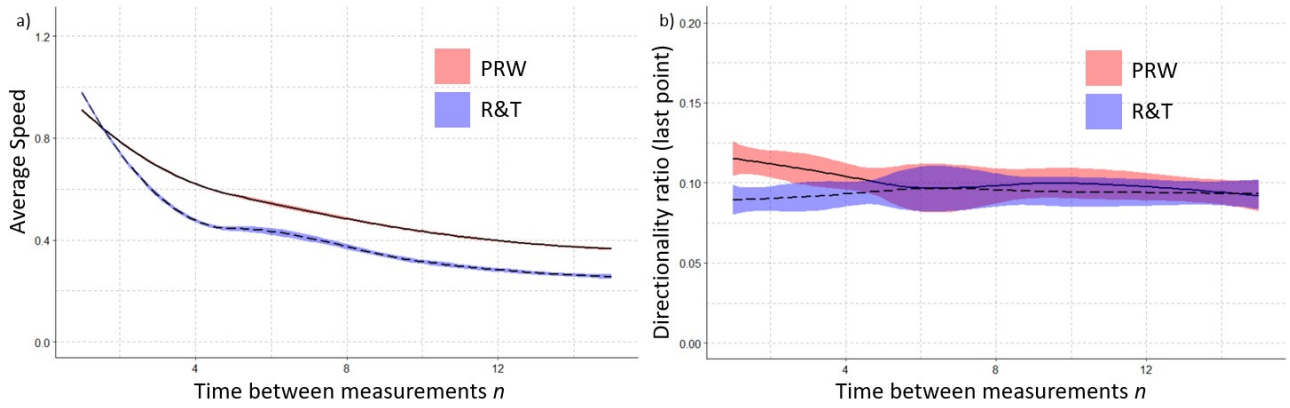

Figure 6: Local polynomial regression curves of a) average speed and b) directionality ratio (between first and last coordinate) over decreasing measurement intervals n, for PRW and R&T data. Shading = 95% confidence intervals.

5 Discussion

In this study, we generated limited cellular position data from two different models: the persistent random walk (PRW) and the run and tumble (R&T) model. First, we compared the fit of each model to the run and tumble pdf. The analysis did not indicate that the PRW data originated from a mechanism other than a run and tumble. Additionally, the estimated speed $\hat{v}$ and turning rate $\hat{\lambda}$ did not accurately describe the underlying mechanisms - we expected $\hat{v}$ and $\hat{\lambda}$ to be lower for the PRW data than for the R&T data, but the results showed the opposite. While this analysis did not return the correct parameters for either model, the models appeared distinct most of the time. At all times the estimated parameters differed between the models and at most times the empirical pdf's were different. However, when fitting displacements over 750 time units, the empirical pdf's looked quite similar. The time interval is specific to this model comparison but these results demonstrate that displacement frequencies may not always distinguish different mechanisms. Furthermore, the likelihood function is quit flat and so the estimated parameters are not necessarily always going to be distinct, as they were in this study. So while different cell phenotypes may not necessarily be differentiated by this analysis, those that are can reliably be sorted. We cannot however, make quantitative conclusions about the cell motion based on the maximum likelihood estimators.

Next, various movement measures calculated by the DiPer program were analysed for differences between the models for $n = 15$. The estimated speed and MSD were significantly higher for the PRW than for the R&T model (opposite to what we expected) indicating that we have likely lost information during the infrequent sampling. The data fed into DiPer do not describe the full cell trajectories, they merely capture the positions of the cells (at relatively large intervals in this case). The PRW model has a higher directional persistence than the run and tumble model (lower turning rate: $PRW = 0.35$; $R\&T = 14$). This means the PRW agents will tend to move directionally for longer periods of time and will have straighter paths on average, while the run

and tumble agents will move in a much less direct manner (Fig. 7a,b). Since we measured position relatively infrequently, it is likely that many turns in the run and tumble paths were missed. This has the effect of slowing down the agents from the data's perspective (Fig. 7c, d), hence the underestimated speed and MSD for the R&T. It should be noted that the measures for both the PRW and R&T have been underestimated (speed of PRW is actually closer to 1) so all estimations cannot be directly interpreted.

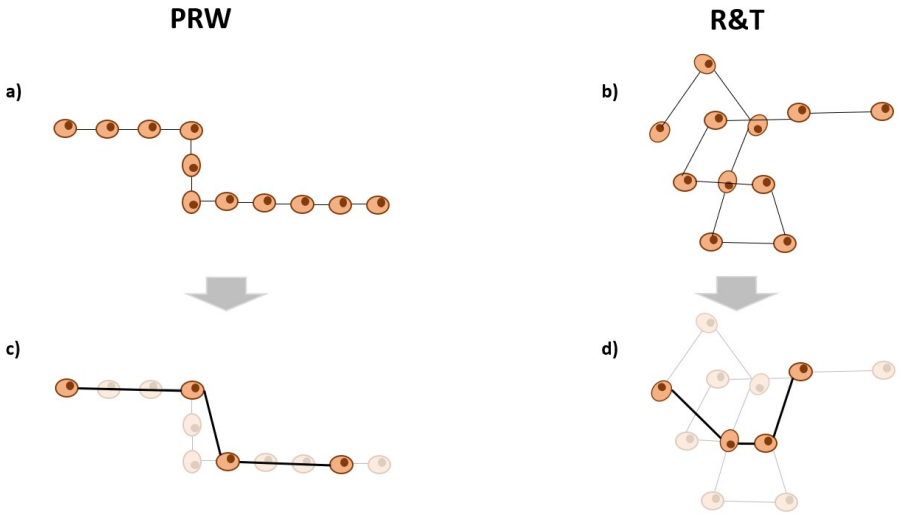

Figure 7: Depiction of the effect of infrequent sampling on speed and MSD. In the figure, the PRW measures would only be slightly underestimated while the much less directional R&T would be greatly underestimated.

Neither the estimated directionality ratio nor the α value differed significantly between the PRW and the R&T models. Additionally, the α values are not significantly different from 1, indicating a completely random walk for both the PRW and R&T. A large class of stochastic models approach diffusion over long time intervals. One such model is the PRW: an agent undertaking a persistent random walk moves directionally only over short time intervals and loses persistence over longer times [16]. It is likely that the infrequent position measurements did not capture the persistence of the PRW and so both models appear to describe equally random walks.

This first part of our study demonstrates that it is not reasonable to assume movement measures arising from limited cell position data are correct and meaningful. If we pretended not to know the source of the two data sets, we would conclude the following: both mechanisms are completely random with no persistence and the PRW has faster moving agents than the R&T. Both of these conclusions are incorrect. The estimated speed, MSD and directional persistence measures do not mean anything in the context of the models. The analysis did however, manage to distinguish between the two models. The simulated cells appeared to move at different rates and cover a different amount of area in a given time. It is possible that different cell phenotypes, moving in different ways, would also produce different estimates through similar movement analysis. So we may be able to use limited snapshot data as a diagnostic tool, even when the estimated parameters are incorrect. It may not be true however, that cells with identical movement measure estimates have identical phenotypes. Since

the estimates are not meaningful, there is a chance we could miss movement information in such a way that the cells look identical through DiPer.

In the second part of the study, we changed the time between measurements n to probe how often we must sample cell positions for movement measures to be meaningful. Both the comparison between speeds and directionality ratios only became meaningful at around $n = 1$. At this point, the output correctly indicated that the R&T agents were faster than the PRW agents and the PRW agents moved more directionally. At $n = 1$, the estimated speed for the PRW is approximately equal to the real speed, however the speed for the R&T was still underestimated (average speed $= 0.978$ vs $v = 3$). We can associate a natural time scale to the models: the average time between turns. For the PRW, this time is ≈ 2.8 and for the R&T ≈ 0.7. Thus, at $n = 1$ we are measuring faster than the PRW agent turns on average but not quite as fast as the R&T agent turns. This could account for the accurate estimate of speed for the PRW but not for the R&T. The estimated directionality ratios probably follow along similar lines although the true directionality ratio is hard to define.

In conclusion, it appears that we must sample cell positions at smaller time intervals than the average time between turning events in order for estimated movement measures to have any meaning. However, if we merely want to differentiate between cell lines, larger sampling intervals or displacement frequencies can be utilized. It is likely this rule-of-thumb extends past our PRW and R&T to other stochastic mechanisms as information will be lost whenever we miss cell turns, regardless of specifics. It would be pertinent for future studies to explore the effect of the lattice on the accuracy of results. The mean displacement (between adjacent points) of 6 at $n = 15$ is quite small relative to *lattice spacing* $= 1$. Thus, the results presented in this study are likely a worst case scenario. It is possible that with mean displacements much larger than the length scales arising from position uncertainty, resultant movement measures may be more accurate. However, caution should still be taken when drawing conclusions about underlying cellular mechanisms when observation frequencies are low.

6 Acknowledgements

I would like to express my gratitude to my supervisor, Professor Barry Hughes, who guided me for the duration of the project. I would also like to thank Yining Ding, a fellow University of Melbourne student, who worked alongside me for a few weeks and provided me with the initial ($n = 15$) PRW data and the PRW code.

References

[1] Li, L, Norrelykke, SF & Cox, EC 2008, Persistent cell motion in the absence of external signals: a search strategy for eukaryotic cells. PLoS ONE 3, e2093.

[2] Webre, DJ, Wolanin, PM & Stock, JB 2003, Bacterial Chemotaxis. Curr. Biol. 13, 47-49.

[3] Fan, Y, Abrahamsen, G, Mills, R, Calderon, CC, Tee, JY, Leyton, L. & Mackay-Sim, A 2013, Focal adhesion dynamics are altered in schizophrenia. Biol Psychiatry, 74(6), 418-426.

[4] Li, X, Vlahovska, PM & Karniadakis, GM 2013, Continuum- and particle- based modelling of shaped and dynamics of red blood cells in health and disease. Soft Matter. 2013 January 7; 9(1): 28–37.

[5] Simpson, MJ, Landman, KA & Hughes, BD 2009, Distinguishing between directed and undirected cell motility within an invading cell population. Bull. Math. Biol. 71, 781–799.

[6] Tremel, A, Cai, A, Tirtaatmadja, N, Hughes, BD, Stevens, GW, Landman, .A &. O'Connor, AJ 2009, Cell migr-ation and proliferation during monolayer formation and wound healing. Chem. Eng. Sci. 64, 247–253.

[7] Diao, W, Tong, X, Yang, C et al. 2019, Behaviors of Glioblastoma Cells in in Vitro Microenvironments. Sci Rep 9, 85:

[8] Dickinson, RB & Tranquillo, RT 1993, Optimal estimation of cell movement indices from the statistical analysis of cell tracking data. Bioengineering, Food and Natural Products. 39(12), 1995-2010.

[9] Billingsley, P 1995. Probability and Measure. Wiley, New York. 3rd ed.

[10] Codling, EA, Plank, MJ & Benhamou, S 2008, Random walk models in biology, J. R. Soc. Interface, 5, 813–834.

[11] Kaiser, D 2003, Coupling cell movement to multicellular development in myxobacteria. Nat. Rev. Microbiol., 1, 45–54.

[12] Kasyap, TV, Koch DL & Wu, M 2014, Hydrodynamic tracer diffusion in suspensions of swimming bacteria. Physics of Fluids, 26, 081901.

[13] Popkin, G 2016, From flocking birds to swarming molecules, physicists are seeking to understand 'active matter' — and looking for a fundamental theory of the living world. Nature, 529, 16–18.

[14] Stadje, W 1987, The Exact Probability Distribution of a Two-Dimensional Random Walk. Journal of Statistical Physics. 46(1/2) 207-216

[15] Gorelik, R & Gautreau, A 2014, Quantitative and unbiased analysis of directional persistence in cell migration. Nat Protoc, 9(8), 1931-1943.

[16] Dunn, GA 1983, Characterising a kinesis response: time averaged measures of cell speed and directional persistence. Agents Actions Suppl. 12, 14–33.