Markov Chains

Summary

This document explains the use of discrete-time Markov chains to assess whether two quantities changing over time are independent or not. We used the method here to study the accumulation of a fluorescent protein within the tips of protrusions (filopodia) of neuronal growth cones, and the dynamic behaviour (extension or retraction) of these tips. The analysis documented here makes use of the R package 'markovchain' (developed by Giorgio Alfredo Spedicato, Tae Seung Kang, Sai Bhargav Yalamanchi, Mildenberger Thoralf and Deepak Yadav; https://CRAN.R-project.org/package=markovchain (https://CRAN.R-project.org/package=markovchain)).

The question to be answered is: does the level of fluorescence of TOCA within filopodial tips at a given timepoint affect their likelihood to grow or shrink? Available data describes TOCA fluorescence levels within tips over time (up to 121 timepoints) for 21 filopodia, and corresponding tip extension/retraction rates. Both quantities exhibit auto-correlated behaviour within a time series. Therefore it is not possible to use statistical tools that assume independence of measurements; however, they can be modelled as discrete time Markov chains. **The null hypothesis to be tested is: 'TOCA fluorescence and tip movement are fully independent of each other'**.

In the first section ('A worked example') we develop and demonstrate this approach on one example filopodium. The second section ('Application to the dataset') demonstrates the application of this method to a complete dataset of 21 filopodia.

Required data and packages:

Dataset: http://link-to-data-here (http://link-to-data-here) (data not publicly available yet).

```
load('~/Documents/Postdoc/ANALYSIS_local-files/ANALYSIS LOGS/2016-11_CCFs_Improvements/LastWorkspace_CCF_TOCA.Rdata')
```

Installing and importing the required R package:

```
#install.packages('markovchain', dependencies=TRUE, repos='http://cran.rstudio.com/')
library(markovchain)
```

```
## Package:  markovchain
## Version:  0.6.5.1
## Date:     2016-09-10
## BugReport: http://github.com/spedygiorgio/markovchain/issues
```

Introduction to the data:

Data for all filopodia are stored in the workspace as variables called 'all.move' (tip movement) and 'tip.f' (tip fluorescence).

Showing data for the chosen example filopodium:

```
##   Time.in.s Movement Fluorescence
## 1         0       NA     1.104360
## 2         2       NA     1.164520
## 3         4   0.0124     1.162039
## 4         6   0.0248     1.058733
## 5         8   0.0124     1.055989
## 6        10   0.0124     1.108486
```

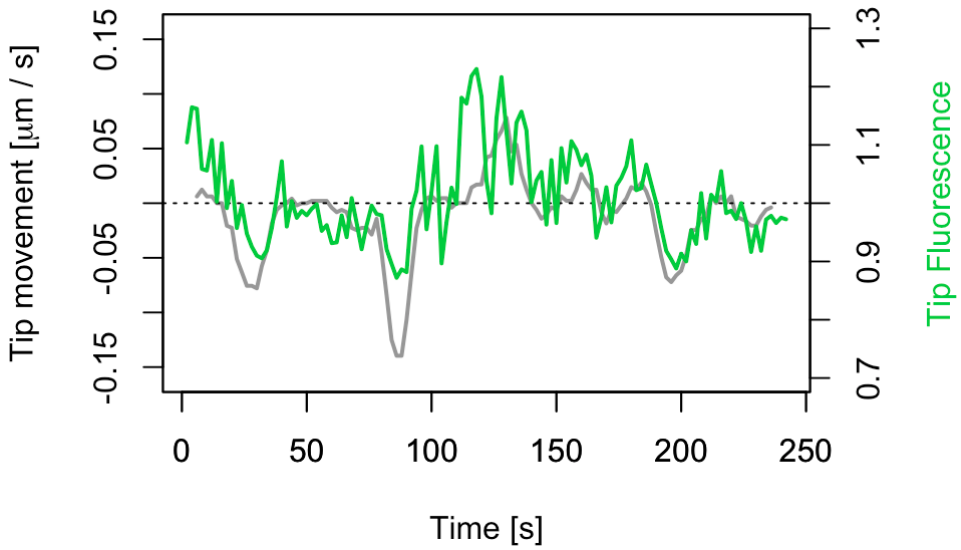

Fig. 1. Measured TOCA tip fluorescence and tip movement for an example filopodium from our dataset.

Fluorescence and movement for the example filopodium appear to be positively correlated:

```
## Warning in cor.test.default(as.numeric(move.0), as.numeric(tip.f.0), method
## = "spearman"): Cannot compute exact p-value with ties
```

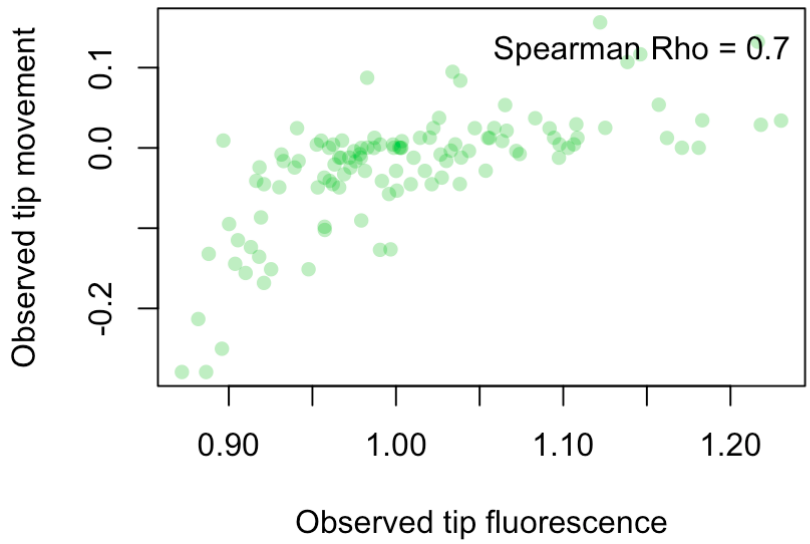

Fig. 2. Relationship between measured TOCA tip fluorescence and tip movement for the example filopodium in Fig.1.

Is this correlation above what can be expected to occur at random in two unrelated time series? Time series that are autocorrelated may be liable to producing spurious correlations that are not truly meaningful. If that is the case, we would expect the same extent of correlation to also occur in simulated datasets modelled to fit

the same ‘rules’ (i.e. transition probabilities between states). In other words, if the null hypothesis is true and the two time series are independent of each other, modelling the two processes as independent Markov chains is likely to recapitulate at least in a significant proportion of simulated cases.

An assessment of this hypothesis on the given example is provided in the following section.

Implementing the Markov chains approach: A worked example

For implementing the Markov chain approach, the data given above first need to be discretised (binned):

```
states.move <- 1:9
states.tip.f <- 1:9
move.0d <- cut(move.0, 9, labels = states.move, na.omit = TRUE)  # 'd' for discrete
move.0d
```

```
##   [1] <NA> <NA> 7    7    7    7    6    6    5    5    4    4    3    3
##  [15] 3    4    4    6    6    6    6    6    6    6    6    6    6    6
##  [29] 6    6    6    6    6    5    5    5    5    5    6    4    3    1
##  [43] 1    1    2    4    5    6    6    7    6    6    6    6    6    6
##  [57] 6    7    7    7    8    8    9    9    9    8    8    7    7    6
##  [71] 6    6    6    6    6    7    6    6    7    7    7    7    7    6
##  [85] 6    6    6    6    6    7    7    7    7    6    5    4    3    3
##  [99] 4    4    4    5    5    6    6    7    6    7    6    7    6    6
## [113] 6    5    5    6    6    6    <NA> <NA> <NA>
## Levels: 1 2 3 4 5 6 7 8 9
```

```
tip.0d <- cut(tip.f.0, 9, labels = states.tip.f, na.omit = TRUE)
tip.0d
```

```
##   [1] 6 8 8 5 5 6 4 6 4 5 3 4 2 2 1 1 2 3 4 6 3 4 3 3 3 3 4 3 3 2 2 3 2 4 3
##  [36] 2 3 4 3 3 2 1 1 1 1 3 4 6 3 4 6 1 3 4 4 8 8 9 9 8 5 3 7 9 7 5 7 8 7 4
##  [71] 5 5 3 6 3 6 5 6 6 5 6 5 2 3 4 3 4 5 5 6 4 4 5 5 4 3 2 1 1 2 1 3 2 4 2
## [106] 4 4 5 3 3 3 4 3 2 3 2 3 3 3 3 3
## Levels: 1 2 3 4 5 6 7 8 9
```

The following table shows the intervals according to which the data is assigned to bin number (“category”, or “state” in Markov chain terminology):

```
##        Movement Category
## 1 (0.872,0.912]        1
## 2 (0.912,0.952]        2
## 3 (0.952,0.991]        3
## 4  (0.991,1.03]        4
## 5   (1.03,1.07]        5
## 6   (1.07,1.11]        6
## 7   (1.11,1.15]        7
## 8   (1.15,1.19]        8
## 9   (1.19,1.23]        9
```

Table 1. Intervals describing the tip movement states for the example filopodium.

Assuming the null hypothesis is true, we describe the fluorescence and movement as separate, completely independent discrete time Markov chains (using functions provided by the *‘markovchain’* package).

We first need to **calculate transition probabilities** between states for the each of the two time series:

```
tMatrixMove <-  markovchainFit(move.0d)$estimate@transitionMatrix
tMatrixTipF     <- markovchainFit(tip.0d)$estimate@transitionMatrix

print(tMatrixMove)
```

```
##              1         2   3          4         5         6         7
## 1    0.6666667 0.3333333 0.0 0.00000000 0.0000000 0.0000000 0.0000000
## 2    0.0000000 0.0000000 0.0 1.00000000 0.0000000 0.0000000 0.0000000
## 3    0.1666667 0.0000000 0.5 0.33333333 0.0000000 0.0000000 0.0000000
## 4    0.0000000 0.0000000 0.3 0.40000000 0.2000000 0.1000000 0.0000000
## 5    0.0000000 0.0000000 0.0 0.15384615 0.5384615 0.3076923 0.0000000
## 6    0.0000000 0.0000000 0.0 0.01886792 0.0754717 0.7358491 0.1509434
## 7    0.0000000 0.0000000 0.0 0.00000000 0.0000000 0.3913043 0.5652174
## 8    0.0000000 0.0000000 0.0 0.00000000 0.0000000 0.0000000 0.2500000
## 9    0.0000000 0.0000000 0.0 0.00000000 0.0000000 0.0000000 0.0000000
## <NA> 0.0000000 0.0000000 0.0 0.00000000 0.0000000 0.0000000 0.2500000
##               8         9       <NA>
## 1    0.00000000 0.0000000 0.00000000
## 2    0.00000000 0.0000000 0.00000000
## 3    0.00000000 0.0000000 0.00000000
## 4    0.00000000 0.0000000 0.00000000
## 5    0.00000000 0.0000000 0.00000000
## 6    0.00000000 0.0000000 0.01886792
## 7    0.04347826 0.0000000 0.00000000
## 8    0.50000000 0.2500000 0.00000000
## 9    0.33333333 0.6666667 0.00000000
## <NA> 0.00000000 0.0000000 0.75000000
```

```
# Now use these transition matrices to generate Markov chain objects

mcMove <- new("markovchain",
    states = as.character(colnames(tMatrixMove)),
    byrow = TRUE,
    transitionMatrix = tMatrixMove,
    name = "Movement")

mcTip <- new("markovchain",
    states = as.character(colnames(tMatrixTipF)),
    byrow = TRUE,
    transitionMatrix = tMatrixTipF,
    name = "Fluorescence")
```

Having calculated Markov chain transition probabilities for movement and fluorescence of the given filopodium, we can **generate Markov chain simulations** that follow the same transition probabilities. Here is an example of a single pair of simulated time series created in this way:

```
move.sim1 <- as.integer(rmarkovchain(121, object = mcMove, t0 = move.0d[3]))
move.sim1
```

```
##   [1]  7  6  6  6  6  6  6  7  7  7  6  6  6  7  6  7  7  6  6  7  6  7  7
##  [24]  7  6  6  6  6  6  6  7  7  7  7  7  7  7  6  6  5  5  5  5  6  6  6
##  [47]  7  7  6  6  7  7  7  7  6  7  7  6  5  5  6  6  6  7  7  6  6  6  6
##  [70]  6  6  6  7  6  7  7  7  6  6  6  6  6 NA NA NA NA NA  7  6  4  5  5
##  [93]  5  5  5  6  6  6  6 NA NA NA NA  7  7  7  7  7  6  6  6  6  6  5  6
## [116]  6  6 NA NA NA NA
```

```
tip.sim1 <- as.integer(rmarkovchain(121, object = mcTip, t0 = tip.0d[1]))
tip.sim1
```

```
##   [1] 3 3 2 1 1 3 3 2 1 3 3 4 5 3 4 3 3 3 3 2 1 3 2 1 1 1 3 4 6 6 6 4 5 4 6
##  [36] 5 6 6 4 3 2 1 2 3 2 4 2 1 3 4 3 6 4 6 3 2 3 2 3 3 4 3 3 4 6 3 3 2 2 3
##  [71] 2 2 4 3 4 2 2 3 3 3 3 4 2 1 1 1 1 2 3 3 4 6 4 6 5 6 4 5 5 6 6 3 4 3 4
## [106] 3 4 2 4 2 1 3 4 3 3 4 5 7 8 5 5
```

Illustrating this visually:

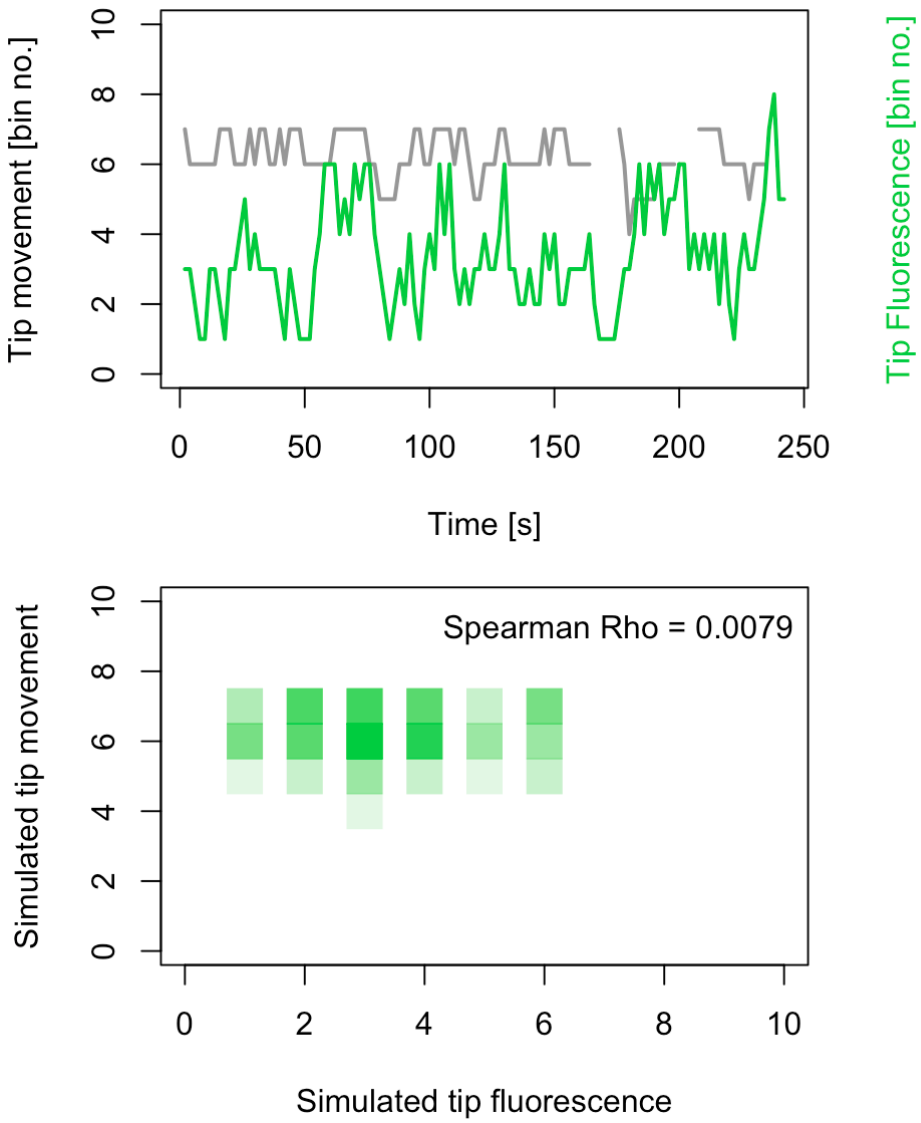

Fig. 3. Fluorescence and movement values over time (upper panel) and in relation to each other (lower panel) in one instance of a Markov chain simulation. The simulation is based on transition probabilities found in real data for the single example filopodium. The two time series are assumed to be independent of each other.

Now, the question is: if we generate a large number of such simulations, in what proportion of the simulations do we observe the extent of correlation that matches the correlation in the real (observed) dataset? To answer this question, we create 10,000 simulated objects here (all based on transition probabilities from the real dataset), and record the correlation coefficient of each. With this we can make a rough estimate the likelihood of the correlation observed in the real-world dataset having occured by chance.

```
# Number of simulations per filopodium
n.sim.per.filo <- 10000
set.seed <- 0.1

# Creating the variables used in loop below:
sim.rho <- c()
sim.p <- c()
sim.move <- data.frame(matrix(NA, ncol = n.sim.per.filo, nrow = length(move.0d)-bb))
sim.tip  <- data.frame(matrix(NA, ncol = n.sim.per.filo, nrow = length(tip.0d)-bb))

# Correlation in the original (observed) dataset:
rho.0 <- cor.test(as.numeric(move.0d), as.numeric(tip.0d))$estimate
p.0 <- cor.test(as.numeric(move.0d), as.numeric(tip.0d))$p.value

# Loop to run through 10,000 simulations for the same filopodium:
for (i in 1:n.sim.per.filo) {

    move.sim.i <- as.integer(rmarkovchain(121, object = mcMove, t0 = move.0d[3]))
    tip.sim.i <- as.integer(rmarkovchain(121, object = mcTip, t0 = tip.0d[1]))

    sim.move[, i] <- move.sim.i
    sim.tip[, i] <- tip.sim.i

    # Record the correlation for each simulation (at each iteration of the loop):
    p.i <- cor.test(as.numeric(move.sim.i), as.numeric(tip.sim.i), method = "spearma
n")$p.value
    rho.i <- cor.test(as.numeric(move.sim.i), as.numeric(tip.sim.i), method = "spearm
an")$estimate

    sim.p[i] <- p.i
    sim.rho[i] <- rho.i
}
```

Which of these simulations has the best correlation between fluorescence and movement?

```
sim.max <- which.max(sim.rho)
print(sim.max)
```

```
## [1] 3402
```

Visualise the most positively correlated simulation:

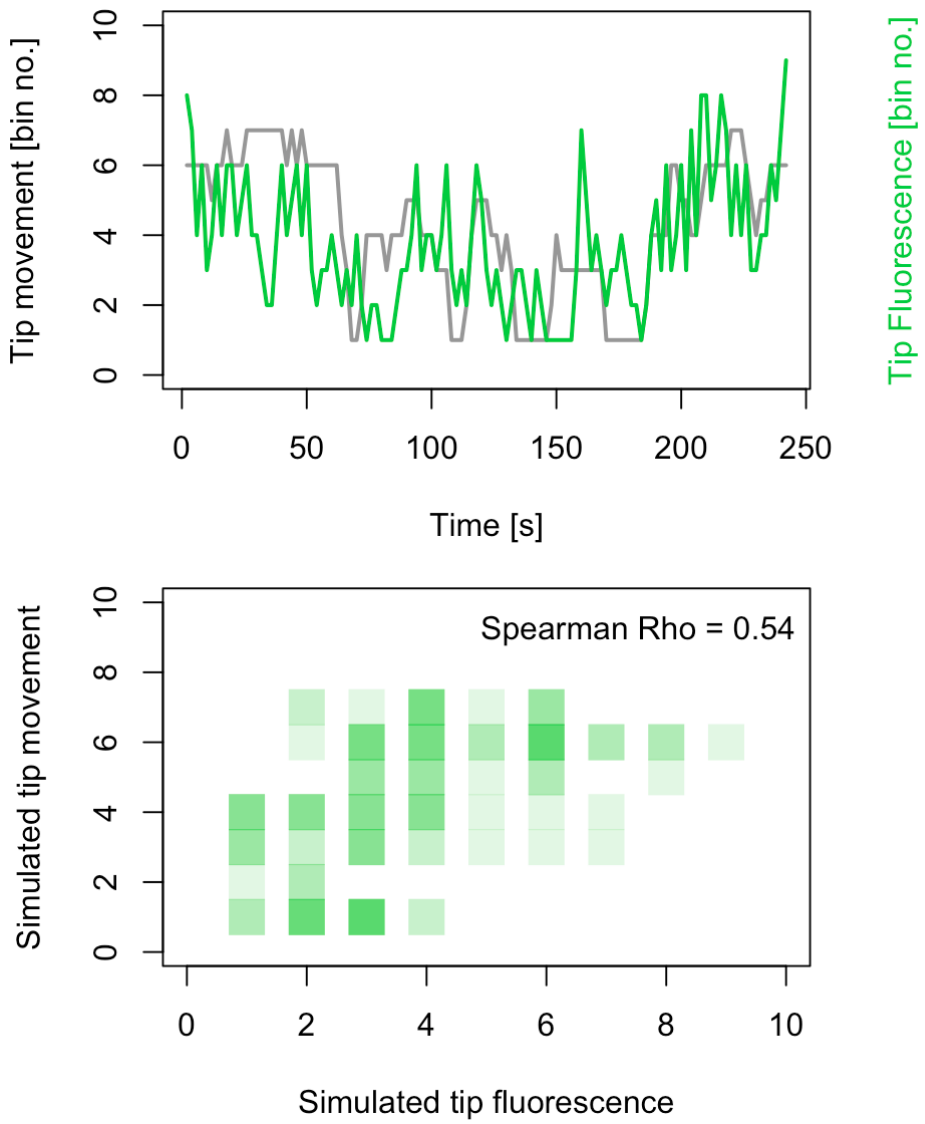

Fig. 4. Simulated fluorescence and movement in the simulation with the highest correlation between all 10,000 generated simulations.

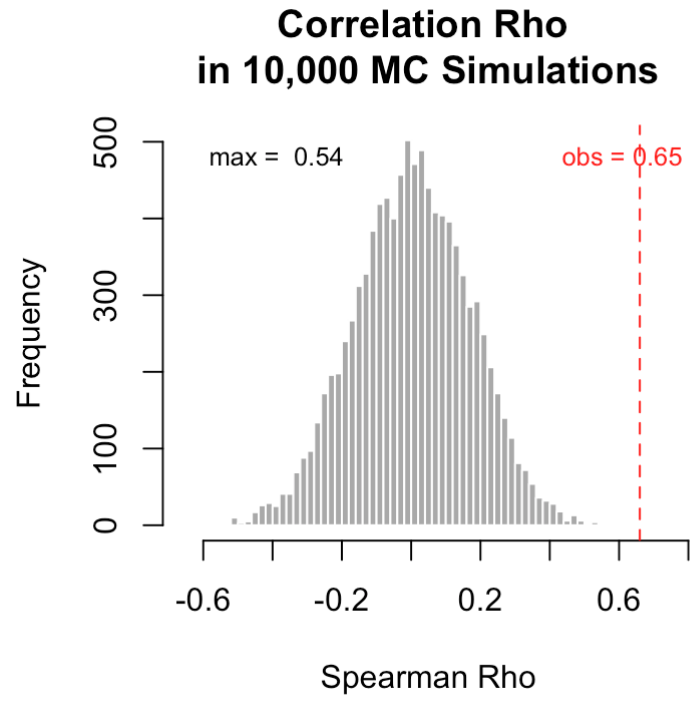

Fig. 5. Histogram of correlation coefficients for all 10,000 simulations. In the case of this example filopodium, the correlation seen in real data (red line, Rho = 0.65) exceeds the correlation in any of the 10,000 simulations (grey, max Rho = 0.55).

To save a visualisation of each of these simulations, see original markdown script file at this position (not executed here).

In conclusion, none of the 10,000 simulations recapitulate the correlations bewteen fluorescence and movement observed in the real measured dataset. The simulations were generated under the assumption that the two modelled quantitites (fluorescence and movement) behave independently of each other. We can thus conclude that in a given filopodium with transition probabilities that match those of the example filopodium, the probability of observed correlation arising by chance is less than 1 in 10,000, i.e. < 0.0001. This can be interpreted as evidence that in the measured example filopodium, fluorescence and movement are not independent of each other (unless the example was chosen as the best example in a very large dataset).

Application to the dataset:

In order to apply the above method to a larger dataset encompassing multiple filopodia, each with its own transition probabilities for fluorescence and movement, we need a new function which performs the following tasks for each filopodium given its data on fluorescence and movement:

Function *SimulateFilo*: 1. extracts Markov chain transition probabilities for its fluorescence measurements 2. likewise, extracts Markov chain transition probabilities for its movement measurements 3. performs a specified number of simulations with these transition probabilities for both fluorescence and movement 4. for each pair of simulations for fluorescence and movement, calculates a correlation metric (Spearman's Rho) 5. (output) counts the number of simulations in which the recorded correlation metric is higher than in the measured dataset 6. (output) calculates the distance (in standard deviation) of the observed Rho in the measured dataset from the population of Rho values from the simulated datasets

See the associated markdown file for code.

Looping the above function (*SimulateFilo*) through all filopodia:

```
## [1] 0
## [1] 417
## [1] 1059
## [1] 2531
## [1] 449
## [1] 30
## [1] 26
## [1] 1606
## [1] 160
## [1] 2846
## [1] 1781
## [1] 392
## [1] 5181
## [1] 3934
## [1] 7067
## [1] 13
## [1] 4539
## [1] 0
## [1] 0
## [1] 6048
## [1] 0
```

Here is a summary of results for each of the 21 filopodia in our dataset:

```
##          Name Sim.in.10000 Sim.proportion
## 1    DCTM (0)            0         0.0000
## 18  DCTM (10)            0         0.0000
## 19 DCTM (0).2            0         0.0000
## 21 DCTM (2).2            0         0.0000
## 16 DCTM (7).1           13         0.0013
## 7    DCTM (6)           26         0.0026
## 6    DCTM (5)           30         0.0030
## 9  DCTM (0).1          160         0.0160
## 12 DCTM (3).1          392         0.0392
## 2    DCTM (1)          417         0.0417
## 5    DCTM (4)          449         0.0449
## 3    DCTM (2)         1059         0.1059
## 8    DCTM (7)         1606         0.1606
## 11 DCTM (2).1         1781         0.1781
## 4    DCTM (3)         2531         0.2531
## 10 DCTM (1).1         2846         0.2846
## 14 DCTM (5).1         3934         0.3934
## 17 DCTM (9).1         4539         0.4539
## 13 DCTM (4).1         5181         0.5181
## 20 DCTM (1).2         6048         0.6048
## 15 DCTM (6).1         7067         0.7067
```

In conclusion: for four filopodia in the dataset (n = 21, 19%) there is a less than 1 in 10,000 chance ($P < 0.0001$) that the observed correlation could have arisen by chance, given the transition matrices describing the dynamic behaviour of their tip fluorescence and tip movement. For an additional 3 filopodia (14%; total 7/21, 33%), there is a less than 1 in 1,000 chance ($P < 0.001$) of the observed correlation between tip fluorescence and movement having occured by chance. This can be interpreted as evidence against the null hypothesis, indicating instead that at least in a third of the filopodia in the measured dataset, there is a positive correlation between their tip fluorescence and tip movement.