1 Background and Objectives

We assume there are S single cells in our dataset. We do not known which cell cycle phase they come from and have only recorded the gene expression of these cells. Not all of these genes bear cell cycle specific information, so we constrain our focus to only the cell cycle genes. Before we try to model the data we have, lets have a glimpse at what we have and what we want to achieve.

- Objective: We have some cells, we do not know which phase they are in. We want to find out which phase they are in and the relative order
- We know a set of genes are expressed in these phases. Is that the full set is not known. Lengths of the phases - not known. Also very likely not all of the genes in the list are actually expressive or oscillatory. Should we do some preprocessing to filter them (aka Macosko approach of correlation thresholding). May be look at qtlCurves plot to get a sense.
- Researchers assume sinusoidal wave function for modeling the oscillatory genes. Is that a good model? What does qtlCurves say? Can we do Fourier or wavelet fitting instead.
- We want a hierarchical ordering. First we want to group the single cells into the higher level order of cell cycle phases, namely G1.S, S, G2.M, M and M.G1. But then once we have done that, we also want to reorder the cells inside each of these phases. That is a little extra than just finding the ordering of the cells, because that does not tell us which phase it is from and does not fully draw from phase specific cell cycle genes.

2 Model

Let the vector of time orders for the S cells is given by t_S. Usually S could be a pretty big number and that could lead to lots of parameters to estimate. So, we trim down the time points into time classes $\mathcal{T}$ which is a set of uniformly spaced time points on $(0, 2\pi)$. We can choose the time spacings depending on our computational resources and how much fine tuning we want on the ordering. We follow an iterative scheme, where for each cell s, we pick a time class $t_s^{(0)}$ randomly from the $|\mathcal{T}|$ time classes. For any n, starting from 0, we fit the following model for gene g and cell s.

$$Y_{sg} = \alpha_g sin(t_s^{(n)} + \phi_g) + \epsilon_{sg} \qquad\qquad \epsilon_{sg} \sim N(0, \sigma_g^2)$$

Note that the frequency is 1 because the reference frame is the cell cycle which is a circle and the period is thus assumed to be 2π. This model is assumed for all g which are sinusoidal. Without loss of generality, we assume that all the genes are such. So, we can write down the model for cell s as

$$\mathcal{L}_s \propto \prod_{g=1}^{G} N\left(\alpha_g sin(t_s^{(n)} + \phi_g), \sigma_g^2\right)$$

The full model over all the S cells is given by

$$\mathcal{L} \propto \prod_{s=1}^{S} \prod_{g=1}^{G} N\left(\alpha_g sin(t_s^{(n)} + \phi_g), \sigma_g^2\right)$$

Here t_s are the cell specific parameters and α_g, ϕ_g and σ_g^2 are gene speciifc parameters. I do not think this model is identifiable.

Given the vector t_s, we can write

$$\begin{aligned} Y_{sg} &= \alpha_g cos(\phi_g) sin(t_s^{(n)}) + \alpha_g sin(\phi_g) cos(t_s^{(n)}) + \epsilon_{sg} \\ &= \beta_{1g} sin(t_s^{(n)}) + \beta_{2g} cos(t_s^{(n)}) + \epsilon_{sg} \end{aligned}$$

There is a bijective mapping from (α, ϕ) to (β_1, β_2) and so it is enough to find the ML estimates of β_1 and β_2 instead of α and ϕ. The bijective map is given by

$$\alpha_g = \sqrt{\beta_{1g}^2 + \beta_{2g}^2}$$

$$\phi_g = tan^{-1}\left(\frac{\beta_{2g}}{\beta_{1g}}\right)$$

We can thus write down the model in matrix notation as

$$Y_{S\times 1}^g = M_{S\times 2}^s \beta_{2\times 1}^g + \epsilon_{S\times 1}^g$$

2.1 Classical model

Fit a classical linear model for each g and get estimates $\hat{\beta}_{1g}$, $\hat{\beta}_{2g}$ and $\hat{\sigma}_g$.

2.2 Bayesian model

Assume flat prior for β^g

$$P(\beta^g) \propto 1$$

Then the posterior for β^g is given by

$$P(\beta^g | \sigma_g^2, t_S, Y) \propto N(\beta^g | \hat{\beta}^g, \sigma^2 (M^T M)^{-1})$$

We assume the prior for σ_g^2 to be

$$P(\sigma_g^2) \propto \frac{1}{\sigma_g^2}$$

Then the posterior is given by

$$P(\sigma_g^2|\beta^g, t_S, Y) \propto \frac{1}{\sigma_g^2} N\left(Y^g|M\beta^g, \sigma_g\right)$$
$$\propto InvGamma\left(\sigma_g^2|\frac{S}{2}, \frac{1}{2}(Y^g - M^s\beta^g)^T(Y^g - M^s\beta^g)\right)$$

2.3 Updating the time classes

Next we need to find the posterior

$$P(t_S|Y, \sigma^{(n)}, \alpha^{(n)}, \phi^{(n)})$$

. Since for each s, t_s can take $|\mathcal{T}|$ values in the range $0, \frac{2\pi}{T-1}, \frac{4\pi}{T-1}, \cdots$. For each of the $|\mathcal{T}|$ values, we calculate

$$p_s(c|Y_s, \theta^{(n)}) \stackrel{def}{=} P\left(t_s = c|Y_s, \sigma^{(n)}, \alpha^{(n)}, \phi^{(n)}\right)$$
$$\propto P(Y_s|t_s = c, \sigma^{(n)}, \alpha^{(n)}, \phi^{(n)})$$
$$\propto \prod_{g=1}^{G} N\left(\alpha_g^{(n)} sin(c + \phi_g^{(n)}), \sigma_g^{(n)}\right)$$

where $\theta^{(n)} = \left(\sigma^{(n)}, \alpha^{(n)}, \phi^{(n)}\right)$.

We calculate this for each c and then generate a sample from the multinomial distribution $Mult\left(1, p_s(.|Y_s, \theta^{(n)})\right)$ and assign the cell s to that time class. We repeat this procedure for all the single cells s. Note there that the finer the set of time classes $\mathcal{T}$, the higher would be the resolution of the order, but the greater would be the computational expense as well. So, there is a trade-off in how refined we want $\mathcal{T}$ to be.

2.4 A continuous time order

There may be multiple cells that get assigned to the same time class using the method above. One may seek a more continuous time order pattern, more so if $|\mathcal{T}|$ is not so large. We tend to follow a similar approach to Macosko 2012 paper, but in a slightly different way. Our algorithm to find the total cell re-order is as follows. Suppose at the end of the above algorithm, once we have observed convergence, $t_s = c^\star$ with the parameters $\theta^\star$. . Define

$$c_{pre} = |\mathcal{T}| \qquad c^\star = 1$$
$$= c^\star - 1 \qquad c^\star \neq 1$$

$$\begin{aligned} c_{post} &= 1 && c^\star = |\mathcal{T}| \\ &= c^\star + 1 && c^\star \neq |\mathcal{T}| \end{aligned}$$

Then we define

$$r_s = \frac{\frac{p_s(c_{pre}|Y,\theta^\star)}{p_s(c^\star|Y,\theta^\star)}}{\frac{p_s(c_{pre}|Y,\theta^\star)}{p_s(c^\star|Y,\theta^\star)} + \frac{p_s(c_{post}|Y,\theta^\star)}{p_s(c^\star|Y,\theta^\star)}}$$

Then the fully recovered time order T_s is given by

$$T_s = c_{pre} + r_s \times \frac{4\pi}{(|\mathcal{T}| - 1)}$$

We repeat the above mechanism for all s and we will get a more continuous ordering. This is more like a piecewise linear statistical interpolation scheme applied on the $p_s\left(.|Y, \theta^\star\right)$ data.

3 Simulation Results

I coded up the above mechanism and then used a simulation model to check whether the code is giving back the true ordering of the cells or not. For this simulation, I did not use any phase specific information of any of the genes or any of the cells. I started with 400 single cells and 500 genes (in the Yoav data, there were 578 single cells and 543 cell cycle genes- so pretty close) and then generated the signal from the model discussed above with

$$\begin{aligned} \alpha_g^{true} &= 10 && \forall g \\ \sigma_g^{true} &\sim \chi^2(4) \\ \phi_g^{true} &\sim \mathcal{U}(0, 2\pi) \end{aligned}$$

The cell cycle times were generated for the 400 cells at uniform spacing from 0 to 2π. Then we fitted the classical model mentioned above with number of time classes $\mathcal{T} = 100$ and grouped the cells in these time classes. We ran around 30 iterations (a total time of around 15 minutes) and although the convergence tolerance was not met, the log posterior increase was of the order of $0.4 - 0.5$ when we stopped (we started from log posterior increase of the order of 1000s) . The estimates we observed were compared with the true values of the simulation. The plots are presented in Fig 1. Next, we observed the patterns of estimated cell cycle time classes and the true cell cycle times and observed there indeed seems to be a rotation of the circle and when we adjust for this rotation, the two plots indeed look pretty same which kind of suggests that the model is working modulo identifiability (Fig 2).

4 Discussion

4.1 Why sinusoidal and not Fourier

One suggestion we were discussing recently in the context that the gene expression curved did not look very sinusoidal was to use a sum of sinusoids (which is basically Fourier) or a mixture of sinusoids (which I think will be analogous to sum of sinusoids as we will not be able to separate out the mixing proportions and the amplitudes of the individual components due to lack of identifiability). I think this will not really make a difference because in that case, we shall assume

$$Y_{sg} = \sum_{l=1}^{L} \alpha_{lg} sin(t_s + \phi_{lg}) + \epsilon_{sg}, \qquad \epsilon_{sg} \sim N(0, \sigma_j^2)$$

Note that we keep the frequency of each sinusoid 1, because firstly, I guess we want them to have period 2π as we are looking at the state space which is a circle for the cell cycle. The other reason is computational. If we have a frequency for each sinusoid, say ω_{lg}, then we have serious lack of identifiability. We can write $\omega_{lg}^{'} = \omega_{lg} \times 100$ for all l and g and $t_s^{'} = t_s/100$ for all s and the model remains the same. I do not know of any good prior that can handle such a scenario.

In that case, we can write the total loglikelihood as

$$\mathcal{L} \sim\propto \prod_{s=1}^{S} \prod_{g=1}^{G} N(\sum_{l=1}^{L} \alpha_{lg} sin(t_s + \phi_{lg}), \sigma_g^2)$$

or we can write

$$\begin{aligned} Y_{sg} &= \sum_{l=1}^{L} \alpha_{lg} sin(t_s + \phi_{lg}) + \epsilon_{sg} \\ &= sin(t_s) \sum_{l=1}^{L} \alpha_{lg} cos(\phi_{lg}) + cos(t_s) \sum_{l=1}^{L} \alpha_{lg} sin(\phi_{lg}) + \epsilon_{sg} \\ &= \beta_{1g} sin(t_s) + \beta_{2g} cos(t_s) + \epsilon_{sg} \\ &= \lambda_g sin(t_s + \nu_g) + \epsilon_{sg} \end{aligned}$$

where

$$\lambda_g = \sqrt{\beta_{1g}^2 + \beta_{2g}^2}$$

$$\nu_g = tan^{-1}\left(\frac{\beta_{2g}}{\beta_{1g}}\right)$$

So basically with the frequencies across the sinusoids remaining the same, we are ultimately getting a sinusoid only it seems (if my calculations are correct). That is taking us back to the previous assumption we had.

4.2 Cell phase length

One of the important considerations has been how should we move from the time classes to actual cell cycle phases - G1, S, G2.M, M and M.G1. Originally the idea was to split up the cell cycle appropriately and assign the partitions to these cell phases. However, the task is pretty difficult from biological standpoint. For a typical rapidly proliferating human cell with a total cycle time of 24 hours, the G1 phase lasts around 11 hours, S phase lasts about 8 hours, G2 about 4 hours and M about 1 hour. So, the partitioning is definitely not uniform across the cell phases. To further complicate matters in embryonic stem cells which are very rapidly proliferating, cell cycles are about 30 minutes long with just the S phase and the M phase (minimal growth phases observed). There is also the possibility that a cell in a G1 phase may hit an energy barrier and stop growing due to lack of nutrients or extracellular impulse and enter a quiescent phase called G0, and stay there until it gets activated by enzymes to overcome the barrier. So, overall, the moral of the story is that it is nearly impossible to deduce the cell phases merely looking at the relative ordering of the cells or their position on the cycle. This is where, it is vital that we pool in information from the cell cycle phase specific genes in order to assign these cells into different cell phases. Also, it makes it very difficult to assign a particular phase value for these phase specific genes, because assigning a phase value will require one to have information about the phase lengths of each phase, which is difficult to track down.

4.3 Problem of identifiability

If we do not constrain the phases, the model no longer remains identifiable. One can just take

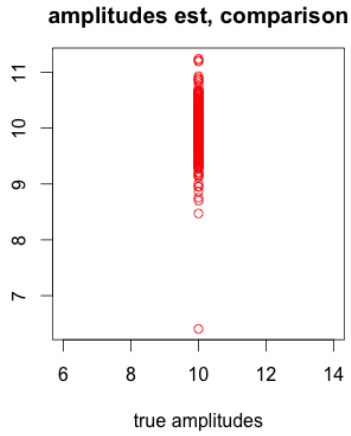
$$t_s^{'} = t_s + \epsilon \qquad \forall s$$
$$\phi_g^{'} = \phi_g - \epsilon \qquad \forall g$$

As a result of this, we may get estimated phases and estimated cell times that are shifted. If we do not have prior information about relative order of the phases or the relative order of the cell times, then it will be difficult to extract the correct estimates of t and ϕ and what we would end up estimating is a rotation of the actual time points. To solve this problem, again we may need to use the phase specific genes expression to drive the knowledge of a broad order among the cells, which can give us the suitable anti-rotation to get the back estimates of the original times.

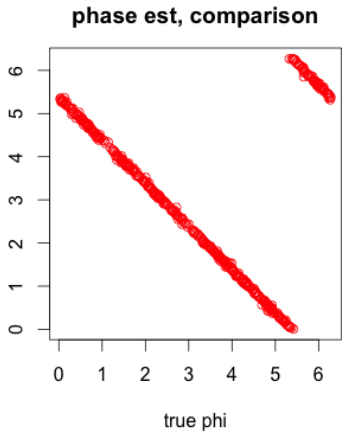
4.4 Road ahead

- To fix the identifiability issue in a meaningful way, possibly pooling in information from phase specific cell cycle genes. Would probably be worthwhile to think about whether we can fix this without relying on cell cycle genes so much, as the identification of these phase specific genes may not be foolproof.
- To implement the Bayesian version, which should not be very difficult. We may think about what priors we want to set though. I am more interested in having a meaningful prior for phases that takes into account the cell cycle gene information, thereby fixing identifiability.

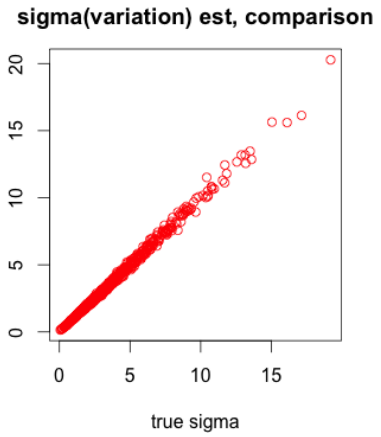
- To implement the method (both classical and Bayesian versions) on Yoav's single cell data and figure out a way to compare our results with Macosko's results and measures to check which method is doing better. The big question: How much relevant would be the assumption of all genes being sinusoidal. More importantly, do we want a filtering scheme to filter out the sinusoidal genes. That would be a difficult task as there is a lot of noise, seems like signal to noise ratio (SNR) will not be very high.
- Whether we should move to wavelets from sinusoidal. I do not think there is much gain to be had in doing so probably, but personally, my knowledge of wavelets is pretty limited right now. Will be able to give a better call on this after 1 month.



(a) amplitude

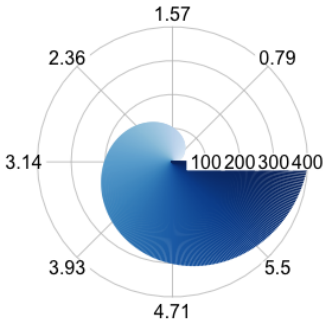


(b) phase

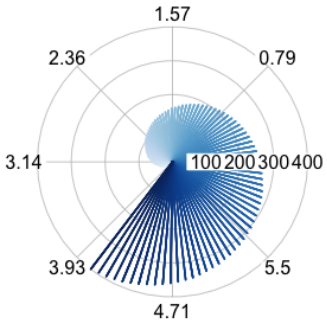


(c) sigma

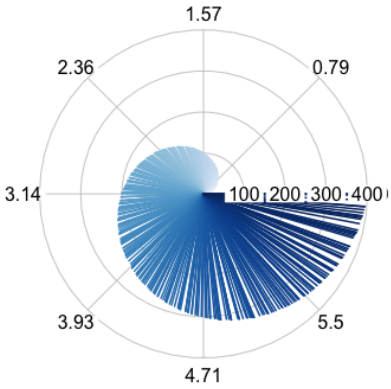
Figure 1. The plots to compare the estimated amplitudes, phase and error variance of the genes with the true values. Ideally we would want to match the estimated values with the true values. While that seems to be fairly the case for the amplitudes and the sigma (the gene variance), it seems the results for the phase are not matching up, and looking at the linear trend in which the estimated phases are associated with the true phases, it seems the identifiability issue is making its presence felt.



(a) True radial plot



(b) Est. radial plot - time classes



(c) Est radial plot- time, phase adjusted

Figure 2. We present the true and the estimated radial plots, where we put the angles in radians on the circle and then colored them in a continuous pattern from light blue to deep blue based on the order of the observations. To make it more interpretable, we also considered the length of the angle to be proportional to the position of the cell in the full list of 400 cells considered for simulation. The first plot (*top,left*) shows the true cell cycle times of the cells. The second plot (*top,right*) shows the estimated time classes with $\mathcal{T} = 100$ for the cells and one can see easily that this looks more like a rotation of the first plot. The third plot (*bottom*) shows the phase adjusted fully recovered estimated cell cycle times of the cells, which looks pretty similar to the true cell cycle time plot.