Generative MERFISH isoform model

Darren Liu, Rob Bierman, Jordi Abante

August 2021

1 Introduction

Our goal is to create a model that estimates exon presence from MERFISH data. Currently, MERFISH is meant to identify transcripts, but we wish to introduce a model where we could distinguish between different RNA isoforms of the same transcript. In MERFISH, every unique gene of interest is given a binary code and multiple rounds of imaging are preformed where, for each round, a different set of fluorescence probes are used to bind to all transcripts and all fluorescent intensities for all pixels are recorded. A pixel is then given a code depending on which rounds a pixel lit up in and that code is used to identified which gene that pixel is.

The fluorescence intensity for a given round is a function of which exons are present and which probes are bound to the present exons. For every identified transcript, we would like to use their respective fluorescence intensity values to help identify which RNA isoform that transcript could be. Our known values for each gene are:

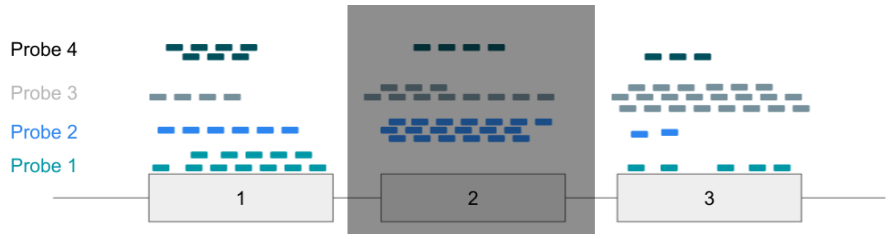
- Fluoresce intensity readings for each imaging round
  - Presented in the form of pixel values. In our case, the images are in uint16, with pixel values ranging from 0 to 65535
- The number of fluorescent probes in each exon in each round

We would like to be able to quantify how likely transcript is a certain isoform given their fluorescence.

2 Model intuition

We will create a different model for each gene, before we try to generalize we'll make a model specific for a fake example gene.

Let's assume we have a gene with 3 exons and is supposed to light up in 4 rounds of imaging. This specific transcript has exons 1 and 3, but exon 2 has been spliced out (indicated by the gray box).



The probe matrix, t, for this example gene is shown below. The rows are the 4 different imaging rounds, the columns are the 3 different exons, and the values are the number of probes in that exon/round combination. We can break t into column vectors to get $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$. Each column vector represents how much each exon is meant to contribute to the overall fluorescence of a given round. t is a known value for each gene, it is not a random variable or a parameter we need to estimate.

$$t = \begin{bmatrix} 12 & 0 & 5 \\ 6 & 19 & 2 \\ 4 & 11 & 22 \\ 7 & 4 & 3 \end{bmatrix}, \mathbf{t_1} = \begin{bmatrix} 12 \\ 6 \\ 4 \\ 7 \end{bmatrix}, \mathbf{t_2} = \begin{bmatrix} 0 \\ 19 \\ 11 \\ 4 \end{bmatrix}, \mathbf{t_3} = \begin{bmatrix} 5 \\ 2 \\ 22 \\ 3 \end{bmatrix}$$

We get to observe the total fluorescence from each of the 4 rounds Y which might be:

$$\mathbf{Y} = \begin{bmatrix} 34 \\ 54 \\ 74 \\ 28 \end{bmatrix}$$

We have chosen to model the presence of each exon in the transcript as an independent Bernoulli random variable, Z $\{Z_n \in \{0,1\} : n \in \{1,...N\}\}$. Z_n will be 1 if exon n is present and 0 otherwise. For this example gene we have Z_1 and Z_3 are equal to 1 since exons 1 and 3 are present. The assumption of independence is maybe questionable since some pairs of exons are going to be found spliced in or out more frequently than other pairs based on frequently used isoforms.

We have created a generative model for the fluorescence, $\mathbf{Y}$, from the Z random variables and the probes per exon matrix t

$$\mathbf{Y} \sim N(\mu(Z,t), \sigma^2)$$

$$\mu = \beta \sum_n Z_n \mathbf{t_n}$$

The intuition for this model is that the observed fluorescence, $\mathbf{Y}$, in a round of imaging is expected to be the sum of fluorescence from each present probe multiplied by the intensity of the probe β.

To show this, let's pretend $\sigma = 0$ so that:

$$\mathbf{Y} \sim N(\mu(Z,t), 0) = \mu(Z,t) = \beta \sum_n Z_n \mathbf{t_n}$$

$$\mathbf{Y} = \beta Z_1 \mathbf{t_1} + \beta Z_2 \mathbf{t_2} + \beta Z_3 \mathbf{t_3}$$

$$\begin{bmatrix} 34 \\ 54 \\ 74 \\ 28 \end{bmatrix} = \beta(1) \begin{bmatrix} 12 \\ 6 \\ 4 \\ 7 \end{bmatrix} + \beta(0) \begin{bmatrix} 0 \\ 19 \\ 11 \\ 4 \end{bmatrix} + \beta(1) \begin{bmatrix} 5 \\ 2 \\ 22 \\ 3 \end{bmatrix}$$

$$\begin{bmatrix} 34 \\ 54 \\ 74 \\ 28 \end{bmatrix} = \beta \begin{bmatrix} 17 \\ 8 \\ 26 \\ 10 \end{bmatrix}$$

$$\beta = 2$$

3 Model definition

3.1 Single-molecule model

A gene has N exons $Z_1, Z_2, ..., Z_N$ which are independently included in the final transcript with unknown probabilities $\alpha_1, \alpha_2, ..., \alpha_N$:

$$\begin{gathered} Z_n \sim Bernoulli(\alpha_n) \ , \quad n = 1, 2, \ldots, N \\ Z_i \perp\!\!\!\perp Z_j \ , \quad i \neq j \end{gathered} \tag{1}$$

Given the included exons, as well as the probe configuration, the fluorescence measurement from the k-th imaging round is assumed to follow a Poisson distribution with a rate λ_k that depends on these two quantities. Letting $\mathbf{Z} = [Z_1, \ldots, Z_N]^T$ be the vector containing the inclusion state of the N exons, and letting the probe configuration be given by $\mathbf{t}_k$, the fluorescent measurement Y_k corresponding to the k-th imaging round is distributed according to

$$Y_k \mid \mathbf{Z} \sim Poisson(\lambda_k(\mathbf{Z}, \mathbf{t}_k)) \ , \quad k = 1, 2, \ldots, K \tag{2}$$

The mean of Y_k, given by $\lambda(\mathbf{Z}, \mathbf{t}_k)$, consists of the sum of the number of probes per exon multiplied by κ, a parameter that linearly relates the number of probes to fluorescence intensity. As a result,

$$\lambda(\mathbf{Z}, \mathbf{t}_k) = \kappa \sum_{n=1}^{N} Z_n t_{k,n} = \kappa \langle \mathbf{Z}, \mathbf{t}_k \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Furthermore, we assume

$$Y_i \perp\!\!\!\perp Y_j \mid \mathbf{Z} \ , \quad i \neq j \tag{4}$$

That is, given the inclusion state of the exons, the k-th fluorescent measurement is conditionally independent from the rest of measurements. This follows

from the fact that, given the exons present, the rest of measurements do not contribute with any further information since each one is done independently. As a result, the joint probability of the (latent) exon presences and he observed fluorescent measurements is given by

$$p(\mathbf{z}, \mathbf{y}) = p(\mathbf{y} \mid \mathbf{z})p(\mathbf{z}) \tag{5}$$

3.1.1 EM approach to solving the model

Derivation of EM algorithm goes here.

3.1.2 Variational inference

As part of the EM algorithm a posterior distribution $p(\mathbf{z} \mid \mathbf{y})$ needs to be computed during inference. This pmf is required in order to find the parameter update equations and can be intractable in some cases. The posterior is

$$p(\mathbf{z} \mid \mathbf{y}) = \frac{p(\mathbf{z}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{z}, \mathbf{y})}{\sum_{\mathbf{z}} p(\mathbf{z}, \mathbf{y})} \tag{6}$$

Here we use an alternative approach that seeks to perform inference through an approximation $q(\mathbf{z} \mid \mathbf{y})$ of this posterior distribution. In particular, we use the mean-field approximation

$$q(\mathbf{z} \mid \mathbf{y}) = \prod_{n} q(z_n \mid \mathbf{y}) \tag{7}$$

and then we take $q(z_n \mid \mathbf{y})$ to be given by

$$q(z_n \mid \mathbf{y}) = Bern(\alpha_{q,n}(\mathbf{y})) \tag{8}$$

where $\alpha_{q,n}(\mathbf{y})$ is such that

$$\text{logit}(\alpha_{q,n}(\mathbf{y})) = \beta_0 + \beta_1 \langle \tilde{t}_n, \mathbf{y} \rangle \tag{9}$$

where $\tilde{t}_n$ is... Thus, our variational parameters will be $\{\beta_0, \beta_1\}$. Once we have estimated these parameters, then we get our guess for $\mathbf{z}$

$$\hat{\mathbf{z}}_{MAP} = \arg\max_{z'} q(\mathbf{z}'|\mathbf{y}) \tag{10}$$

3.2 Cell-level model

Under the assumption that, within each cell, all the Z vectors follow a common underlying distribution, the previous model could be expanded to include all the molecules of a given gene. This does not imply that we expect all the isoforms to be the same, rather we assume there is an underlying distribution of isoforms specific to each cell. In addition, this will provide more power to the method to properly estimate the inclusion probabilities α and to produce better estimates of the exon presence $\hat{\mathbf{z}}$.

Assume that, in a given cell, we observe M molecules that correspond to a given gene. In that case, we are interested in the set of M latent vectors $\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M\}$, where $\mathbf{Z}_m = [Z_{m,1}, Z_{m,2}, \ldots, Z_{m,N}]^T$, where $Z_{m,n}$ indicates the presence of the n-th exon in the m-th molecule. We can assume that, within the cell at hand, the set of indicators for the n-th exon, i.e., $Z_{1,n}, Z_{2,n}, \ldots, Z_{M,n}$, are independently and identically distributed according to $Bernoulli(\alpha_n)$. However, as in the single-molecule model, we only observe a set of M fluorescence vectors $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_M\}$.

Given the included exons, as well as the probe configuration, the fluorescence measurement for the m-th RNA molecule from the k-th imaging round is assumed to follow a Poisson distribution with a rate $\lambda_{m,k}$ that depends on these two quantities. Letting the probe configuration be given by $\mathbf{t}_{m,k}$, the fluorescent measurement $Y_{m,k}$ corresponding to the k-th imaging round of the m-th mRNA molecule is distributed according to

$$Y_{m,k} \mid \mathbf{Z}_m \sim Poisson(\lambda_{m,k}) \ , \quad m = 1, 2, \ldots, M \ , \quad k = 1, 2, \ldots, K \tag{11}$$

The mean of $Y_{m,k}$, given by $\lambda_{m,k} \equiv \lambda(\mathbf{Z}_m, \mathbf{t}_{m,k})$, consists of the sum of the number of probes per exon multiplied by κ, a parameter that linearly relates the number of probes to fluorescence intensity. As a result,

$$\lambda(\mathbf{Z}_m, \mathbf{t}_{m,k}) = \kappa \sum_{n=1}^{N} Z_{m,n} t_{m,k,n} = \kappa \langle \mathbf{Z}_m, \mathbf{t}_{m,k} \rangle \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Furthermore, we assume

$$Y_{m,i} \perp\!\!\!\perp Y_{m,j} \mid \mathbf{Z} \ , \quad i \neq j \tag{13}$$

That is, given the inclusion state of the exons for the m-th molecule, the k-th fluorescent measurement is conditionally independent from the rest of measurements. This follows from the fact that, given the exons present, the rest of measurements do not contribute with any further information since each one is done independently. As a result, the joint probability of the (latent) exon presences and he observed fluorescent measurements is given by

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_M, \mathbf{y}_1, \ldots, \mathbf{y}_M) = \prod_{m=1}^{M} p(\mathbf{y}_m \mid \mathbf{z}_m) p(\mathbf{z}_m) \tag{14}$$

3.2.1 EM algorithm

4 Potential improvements

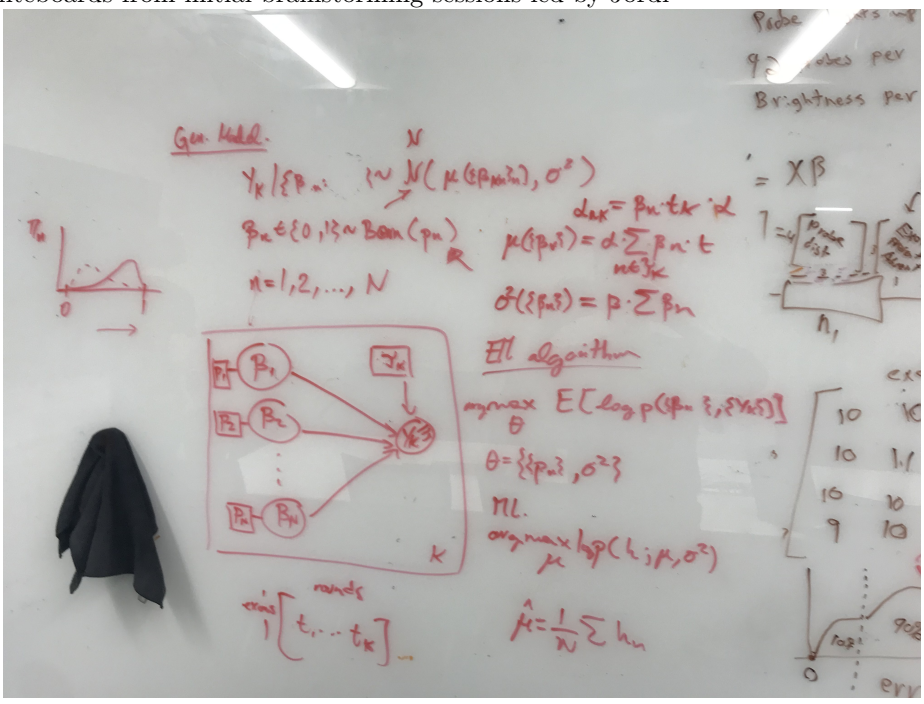
- Incorporate prior knowledge of isoforms. Out of all the possible combination of exons maybe only a few are actually known to exist. Can and should we incorporate this information at the Z level? On one hand, considering all possible exon groupings could illuminate new isoforms, but we

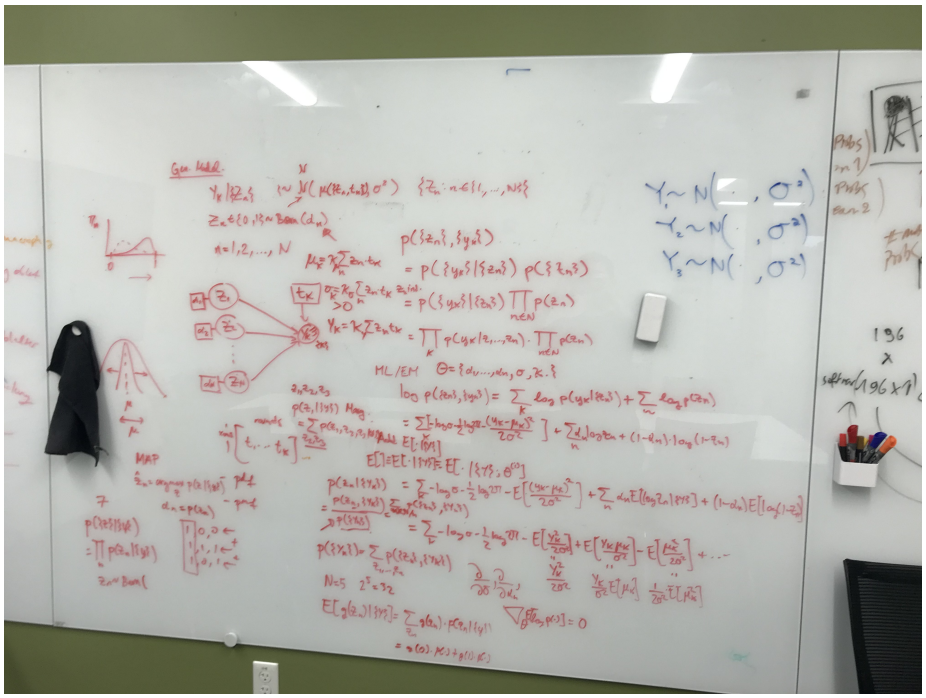
could find an 'optimal' exon grouping that does not exist. In contrast, limiting our search to all known annotated isoforms would lend more confidence in that our results actually represent something biological, but remove our ability to identify exciting new exon groupings.

- Possibility to include priors on α parameters that account for known biology?
- Incorporate information about binding affinity as a function of the exon? For example, RNA secondary structure could shield the certain exons from probes. Thus, some exons could be less likely to bind probes compared to other exons in the same transcript. On the one hand, we are providing Y with randomness that could capture the differences in affinity between exons. On the other hand, Elisabeth brought up a good point about the possibility of the probes being designed to avoid binding affinity biases. For instance, the fluorescent probes could be intentionally designed in a way to avoid RNA secondary structure (such design considerations are known to already take place in PCR)

5 Images

Whiteboards from initial brainstorming sessions led by Jordi




Gen. Model
ML/EM
MAP
196
x
196 x 1