Genetic Tool to Cell Type Bayesian Mapper (GT2CT Bayesian Mapper)

Nikhil Karthik and Chaitali Khan

May 29, 2025

Genetic Tool to Cell Type Bayesian Mapper (GT2CT Bayesian Mapper) is a python based toolkit for finding the likely fractions of brain cell-types (class, subclass, supercluster or cluster) present in the GFP/RFP positive cells tagged by genetic tools.

In the note below, we will use "VISp" to specify a brain volume, but it is just a placeholder for any brain volume parcellation. We will use the term "subclass" to specity cell-type, but it is a placeholder for cell-type at class, subclass, supercluster or cluster. This makes the discussion focus on the challenge.

1 Mathematical background

1.1 Notation

Let $\mathbf{x} = (x, y, z)$ be the three-dimensional CCF coordinates in the VISp area. The brain slices are labelled by constant z values. Let S be the set of subclasses present in VISp area, and let N be the size of this set of subclasses. That is, for VISp, $S = \{$L6 CT CTX Glut, Pvalb Gaba,...$\}$ witb 28 other subclasses in the set. Note that this information on cell-type is infered from the MERFISH cell data.

Next, we have the GFP/RFP positive cells tagged by a gene tool. Let K be the target specificity of the gene tool. Either one knows how many K cell types can be tagged by the genetic tool, or one knows a prior distribution over K for gene tools used. For now, we assume that the value of K is known, and later we will use prior knowledge that K is small. Given the value of K, there are $\binom{N}{K}$ subclass combinations possible. Let $\mathcal{K}$ is the set of these $\binom{N}{K}$ subclasse K-plets, and let its members be k. These K-plets k are, for example, such as $k_1 =$ (Oligo, Astro, Sst Gaba), $k_2 =$(L5 ET CTX Glut, Sncg Gaba, Microglia NN), etc., for $K = 3$.

1.2 Probability distributions

Let $P(s|\mathbf{x})$ be the probability to find a cell of type $s \in S$ at position $\mathbf{x}$. The net probability to find any of the subclasses at a location $\mathbf{x}$ should be one; that is

$$\sum_{s \in S} P(s|\mathbf{x}) = 1. \tag{1}$$

Let $P'(k|x)$ be the probability to find a subclass K-plet $k \in \mathcal{K}$ at CCF location x. This is related to the distribution of individual subclass $P(s|\mathbf{x})$ as

$$P'(k|\mathbf{x}) \propto \sum_{s \in k} P(s|\mathbf{x}), \tag{2}$$

with the proportionality constant that correctly normalizes $P'(k|\mathbf{x})$. We also want the spatial distribution of a K-plet k given by $q(\mathbf{x}|k)$. This is given by the Bayes theorem to be

$$q(\mathbf{x}|k) = \frac{P'(k|\mathbf{x})}{\sum_{\mathbf{x}'} P'(k|\mathbf{x}')}, \tag{3}$$

assuming a uniform prior distribution over position. For example, let us say that we know that a genetic tool specifically targets $k =$(L5 ET CTX Glut, Sncg Gaba, Microglia NN). Then, the distribution $q(\mathbf{x}|k)$ gives the spatial distribution to find one of the members of k to be at various x.

Let $Q_{\rm GFP}(\mathbf{x})$ be the spatial distribution of GFP/RFP cells. If one had a really high resolution image of GFP/RFP cells, and one knew with 100% confidence which were GFP/RFP cells versus the background, then we can still describe the cells through a distribution $Q_{\rm GFP}(\mathbf{x})$ that are non-zero only at exact locations of the GFP cells. In general, this is not the case, and we take the approach to assign only a continous spatial distribution $Q_{\rm GFP}(\mathbf{x})$ that there is a GFP/RFP cell at various $\mathbf{x}$. We will define our procedure to assign $Q_{\rm GFP}(\mathbf{x})$ based on intensity distribution later in this note.

1.3 Mathematical statement of the algorithm

The basis of our algorithm is the following. Given $Q_{\rm GFP}(\mathbf{x})$ of the GFP/RFP cells via the STPT images of the brain, given a spatially smooth distribution $P(s|x)$ of the celltypes inferred from MERFISH data, and given thet the gene tool can tag only certain cell types from K-plets k all over VISp, we find the single best K-plet k whose $q(\mathbf{x}|k)$ is the most similar to $Q_{\rm GFP}(\mathbf{x})$.

Such a measure of similarity between two probability distributions p and q is given by the cross-entropy $\mathcal{H}(p,q)$ which is a minimum only if $p=q$. For our case, we want to find k such that

$$\mathcal{H}(k) = -\sum_{\mathbf{x}\in \rm VISp} Q_{\rm GFP}(\mathbf{x}) \log\left[q(\mathbf{x}|k)\right], \tag{4}$$

is minimized among all $k \in \mathcal{K}$. Let $k=k^*$ be the optimal K-plet. Then, given the assumption/model that the gene tool has specificity of K cell types, the expected fractions f_s of subclasses $s \in k^*$ present all over VISp is

$$f_s = \sum_{\mathbf{x}\in \rm VISp} Q_{\rm GFP}(\mathbf{x}) \frac{P(s|\mathbf{x})}{\sum_{s'\in k^*} P(s'|\mathbf{x})}, \quad \text{for} \quad s \in k^*. \tag{5}$$

For $s \notin k^*$, $f_s = 0$. The assumptions of specificity K and the optimal k^* from $\mathcal{K}$ are implicit, and it is to be understood that f_s is the above equation is actually $f_s\left(k^*(K)\right)$.

1.4 Model averaging

The above step concludes the algorithm. But, one could go further to alleviate the model dependence on the choice of gene tool specificity K by a weighted average of $f_s(k^*)$ for different K. Studies are needed on how best to perform the weighted average, but for this challenge, we do a simple model averaging as

$$f_s = \frac{1}{K_{\rm max}} \sum_{K=1}^{K_{\rm max}} f_s\left(k^*(K)\right), \tag{6}$$

for $K_{\rm max} = 4$.