

Idea for an image extension for Cell2Location

Thomas Walter

August 18, 2024

1 Introduction

Our objective is to propose a deconvolution algorithm for sequence-based spatial transcriptomics (ST) data that uses both the ST data (RNA counts) and the corresponding image data (H&E).

One of the ideas that we want to develop is to use Cell2Location (C2L) [KSD⁺22], a popular method for cell deconvolution and use its output to guide the image classifier applied on H&E. This will provide us with a first version. The limitation is that we do not use the full potential of the complementarity between image and sequencing data, as we do not use the two data jointly to find a common representation more powerful than the representations of each, but the approach still has the benefit to be directly applicable on top of an existing and popular framework. And also: it's a start!

2 Cell classification: learning from label proportion

C2L provides us with an estimation of proportions for each spot in the ST-matrix. Here, we identify the spot S with a set of cells s_i : $S = \{s_i\}_{i=1..N_S}$.

$$C2L : S \rightarrow p \in [0, 1]^C \quad (1)$$

where p is the vector of proportions estimated by CL, C the number of cell types.

We are looking for a classifier that assigns to each cell s_i represented by the image x . As we do not have instance labels, we can apply the LLP (learning from label proportions) paradigm, which has gained some attention recently [DZC⁺19, TL, ZWS23]. According to [DZC⁺19], typical losses used for this are the L_1 , L_2 or KL loss.

This is very close to what Loïc has been doing, but we actually need the cell types that are found by C2L.

3 Adapting the prior probability

In the output of C2L (as defined in 1), we have $\sum_{c=1}^C p_c = 1$ (they are proportions), and we can therefore understand p as a localized prior probability of cell type occurrence.

From section 2, we obtain for every instance (every cell) $p(c|x)$, where $c \in \{1, \dots, C\}$ is the cell type and x the cell image. The Bayes theorem gives us:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \quad (2)$$

Here, all probabilities are estimated from the training set, and we note that in a given spot $p(c)$ is very different from $p(c)$ in the training set, and that we can use the C2L estimation of $p(c)$ instead. We thus want to "correct" the posterior probability to take this changed prior probability into account. For this corrected probability, we get:

$$\tilde{p}(c|x) = \frac{\tilde{p}(x|c)\tilde{p}(c)}{\tilde{p}(x)} \quad (3)$$

Now, we assume that in these two equations, the class conditioned data densities are actually equal:

$$p(x|c) = \tilde{p}(x|c) \quad (4)$$

which allows us to find an expression for the corrected posterior probability:

$$\begin{aligned}\tilde{p}(c|x) &= p(c|x) \frac{p(x)}{\tilde{p}(x)} \frac{\tilde{p}(c)}{p(c)} \\ &= p(c|x) \alpha(x) \frac{\tilde{p}(c)}{p(c)}\end{aligned}\quad (5)$$

Of note, both the data densities $p(x)$, $\tilde{p}(x)$ and the prior probabilities $\tilde{p}(c)$, $p(c)$ are different between the training set and the local spot. The fraction $\alpha(x) = \frac{p(x)}{\tilde{p}(x)}$ can be easily determined by:

$$\sum_{c=1}^C \tilde{p}(c|x) = \alpha(x) \sum_{c=1}^C p(c|x) \alpha(x) \frac{\tilde{p}(c)}{p(c)} = 1 \quad (6)$$

which gives:

$$\alpha(x) = \frac{1}{\sum_c p(c|x) \frac{\tilde{p}(c)}{p(c)}} \quad (7)$$

Indeed, $\alpha(x)$ is just a normalization factor ensuring that the corrected posterior probabilities sum to 1.

References

- [DZC⁺19] Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. Deep multi-class learning from label proportions, June 2019.
- [KSD⁺22] Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaitė, Artem Lomakin, Veronika Kedlian, Adam Gayoso, Mika Sarkin Jain, Jun Sung Park, Lauma Ramona, Elizabeth Tuck, Anna Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, January 2022.
- [TL] Kuen-Han Tsai and Hsuan-Tien Lin. Learning from Label Proportions with Consistency Regularization.
- [ZWS23] Jianxin Zhang, Yutong Wang, and Clayton Scott. Learning from Label Proportions by Learning with Label Noise, September 2023.