Conformal inference for cell type prediction leveraging the cell ontology

Daniela Corbetta[1], Ludwig Geistlinger[2],
Livio Finos[1], Davide Risso[1]

[1]Department of Statistical Sciences, University of Padova
[2]Center for Computational Biomedicine, Harvard Medical School

Ascona Workshop: Spatial and temporal statistical modeling in molecular biology - Sept 12, 2024

Motivating Application: cell type annotation

Aim: Starting from a set of already annotated cells (reference set), predict the cell type of a new, unknown cell

How:

- Choose a model
- Fit the model on the reference set
- Obtain predictions for the new cell

	Gene 1	$\dots$	Gene K	Cell type
Cell 1	1	$\dots$	5	B cell
Cell 2	0	$\dots$	5	B cell
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Cell m	3	$\dots$	0	T (CD4+)

	Gene 1	$\dots$	Gene K	Cell type
New cell	4	$\dots$	0	**?**

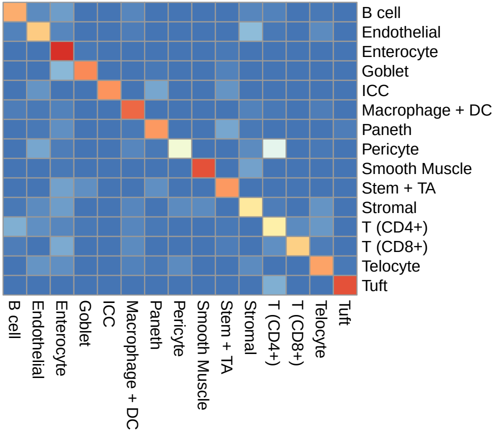
Example

Data: 5163 cells from the mouse ileum sequenced with Merfish[a]

- 500 cells as reference
- 4663 cells as query
- 15 different cell types

Model: Multinomial logit model with the 50 HVGs

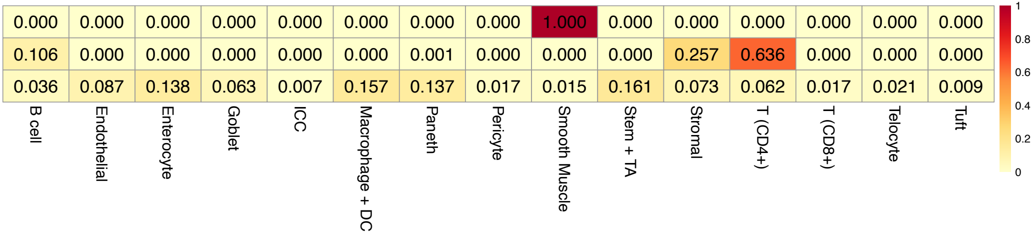
Results: Accuracy=0.77



[a] Petukhov, V., et al. (2022). Cell segmentation in imaging-based spatial transcriptomics. Nature biotechnology, 40(3), 345–354.

Should we rely on point predictions?

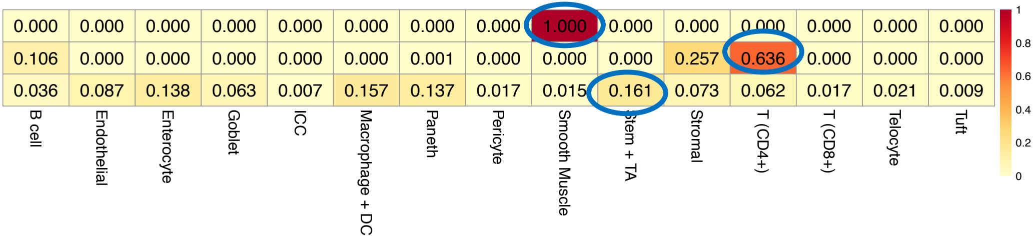
- The model does not provide only a label, but also estimated probabilities for each class
- These probabilities encode how sure the model is of the prediction



B cell	Endothelial	Enterocyte	Goblet	ICC	Macrophage + DC	Paneth	Pericyte	Smooth Muscle	Stem + TA	Stromal	T (CD4+)	T (CD8+)	Telocyte	Tuft
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
0.106	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.257	0.636	0.000	0.000	0.000
0.036	0.087	0.138	0.063	0.007	0.157	0.137	0.017	0.015	0.161	0.073	0.062	0.017	0.021	0.009

Should we rely on point predictions?

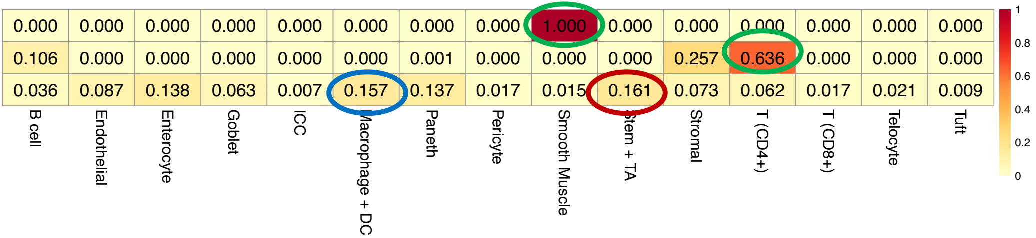
- The model does not provide only a label, but also estimated probabilities for each class
- These probabilities encode how sure the model is of the prediction



B cell	Endothelial	Enterocyte	Goblet	ICC	Macrophage + DC	Paneth	Pericyte	Smooth Muscle	Stem + TA	Stromal	T (CD4+)	T (CD8+)	Telocyte	Tuft
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
0.106	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.257	0.636	0.000	0.000	0.000
0.036	0.087	0.138	0.063	0.007	0.157	0.137	0.017	0.015	0.161	0.073	0.062	0.017	0.021	0.009

Should we rely on point predictions?

- The model does not provide only a label, but also estimated probabilities for each class
- These probabilities encode how sure the model is of the prediction



B cell	Endothelial	Enterocyte	Goblet	ICC	Macrophage + DC	Paneth	Pericyte	Smooth Muscle	Stem + TA	Stromal	T (CD4+)	T (CD8+)	Telocyte	Tuft
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
0.106	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.257	0.636	0.000	0.000	0.000
0.036	0.087	0.138	0.063	0.007	0.157	0.137	0.017	0.015	0.161	0.073	0.062	0.017	0.021	0.009

How can we translate the level of confidence?

- Instead of a point prediction, return a prediction set: a set of different labels that we think our new cell might be
- Intuitively, the prediction set has to include more labels when we are less sure of the point prediction
- Let Y_{new} be the true label of the new cell and $C(X_{new})$ be the prediction set. We define a level α and we want the set to be valid at a level $1-\alpha$:

$$P(Y_{new} \in C(X_{new})) \geq 1 - \alpha.$$

$\rightarrow$ Conformal inference

Conformal inference

- Proposed by Vovk[1], very easy and nice tutorial in Angelopoulos & Bates[2]
- Provides prediction sets that satisfy $P(Y_{new} \in C(X_{new})) \geq 1 - \alpha$, it's distribution-free and works with every model (even terrible ones)
- Based on data splitting:
  - **training set**: annotated data used to fit the model
  - **calibration set**: annotated data that we need to calibrate the prediction sets construction
  - **query set**: new data on which we want to do predictions. Need to be exchangeable with the calibration data
- Algorithm: calibration step and prediction step

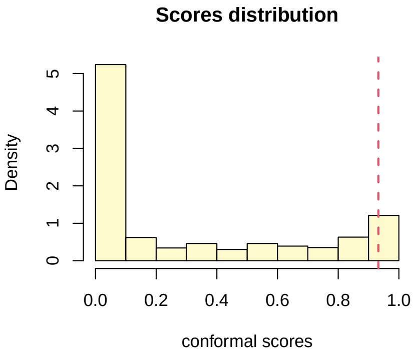
[1] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world, volume 29. Springer

[2] Angelopoulos, A. N. & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

Calibration step

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the data in the calibration set.

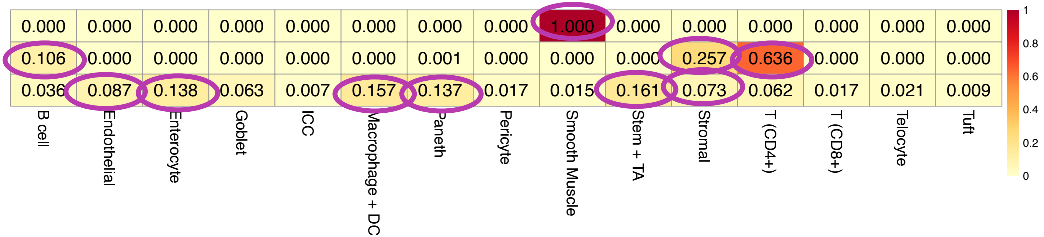
1. Compute predictions for the data in the calibration set
2. Obtain the conformal score: $s_i = 1 - \hat{p}(X_i)_{Y_i}, \; i = 1, \ldots, n$ (i.e. 1 - the predicted probabilities for the true class)
3. Compute $\hat{q}$, the $\lceil (n+1)(1-\alpha) \rceil / n$ empirical quantile of the conformal scores



Prediction step

1. Obtain predictions for the data in the query set
2. Form prediction sets by including all the classes that have predicted probabilities $\geq 1 - \hat{q}$

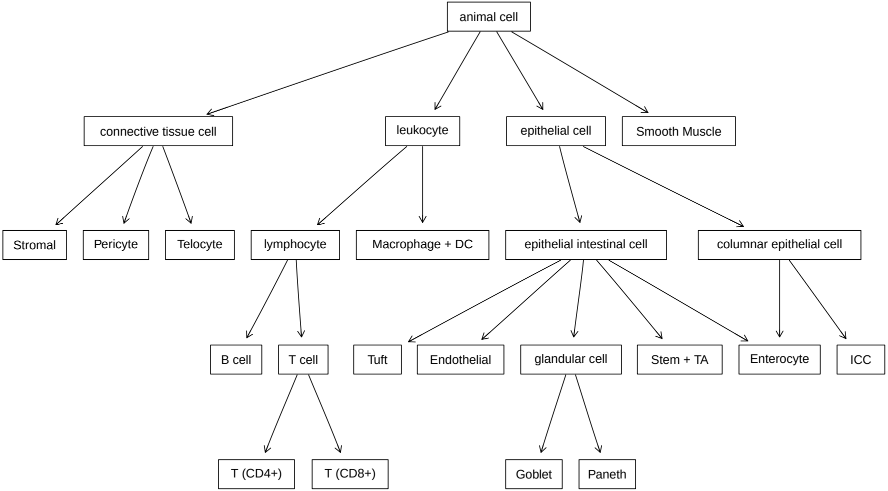
Back to the example: $1 - \hat{q} = 0.068$



B cell	Endothelial	Enterocyte	Goblet	ICC	Macrophage + DC	Paneth	Pericyte	Smooth Muscle	Stem + TA	Stromal	T (CD4+)	T (CD8+)	Telocyte	Tuft
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
0.106	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.257	0.636	0.000	0.000	0.000
0.036	0.087	0.138	0.063	0.007	0.157	0.137	0.017	0.015	0.161	0.073	0.062	0.017	0.021	0.009

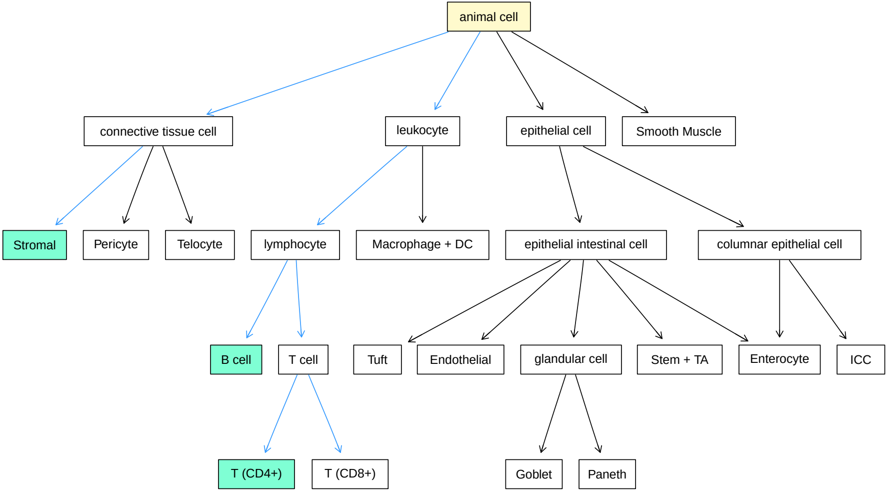
Additional information

Cell types are organized into a graph structure:



Additional information

Back to the example 2:



Additional information

- Question: is there a way to exploit this information when we build prediction set?
- Desired result: instead of returning a set of potentially unrelated labels, return an ancestor of the predicted class.

→ Conformal risk control[3]

[3] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2022). Conformal risk control. arXiv preprint arXiv:2208.02814.

Conformal risk control

- Split the reference data into train set and calibration set.
- Choose an algorithm to build prediction sets. This algorithm must depend on a parameter λ that controls how big the prediction sets are. The only requirement is that the prediction sets are nested when λ increases.
- Choose a loss function $L_i(\lambda) \rightarrow$ miscoverage

$$L_i(\lambda) = \begin{cases} 1 & \text{if } y_i \notin C_\lambda(x_i) \\ 0 & \text{if } y_i \in C_\lambda(x_i) \end{cases}$$

- Choose λ based on the data in the calibration set as

$$\hat{\lambda} = \inf \left\{ \lambda : \frac{n}{n+1} \hat{R}_n(\lambda) + \frac{1}{n+1} \leq \alpha \right\},$$

where $\hat{R}_n(\lambda)$ is the empirical risk for observations in the calibration set, to ensure, for a new observation in the query set,

$$E\left[L_{new}(\hat{\lambda})\right] \leq \alpha \overset{\text{with miscoverage}}{\Longleftrightarrow} P(Y_{new} \in C_\lambda(X_{new})) \leq \alpha$$

How do we build the prediction sets?

- Exploit the graph structure.
- Define for each node v a score $g(v)$ as the sum of the predicted probabilities of the leaf nodes that are descendants of v.
- Start from the predicted class $\hat{y}(x)$. Let $\mathcal{P}(v)$ and $\mathcal{A}(v)$ be the set on descendant nodes that are leaves of the graph and ancestor nodes of a node v, respectively.
- Choose λ and build a prediction set as

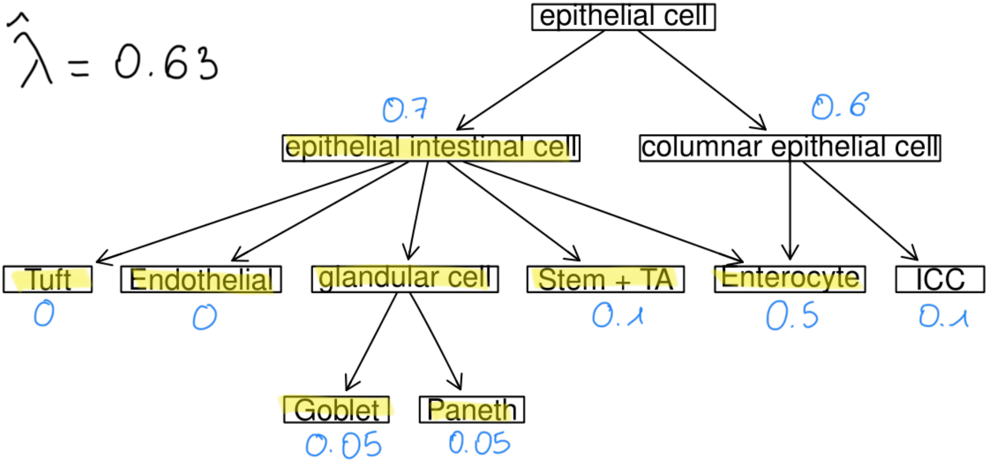
$$\mathcal{P}(v) \cup \{\mathcal{P}(a) : a \in \mathcal{A}(\hat{y}(x)) : g(a, x) \leq \lambda\},$$

  where $v : v \in \mathcal{A}(\hat{y}(x)),\ g(v, x) \geq \lambda,\ v = \arg\min_{u : g(u,x) \geq \lambda} g(u, x)$.

- In words, we start from the predicted class and we go up in the graph until we find an ancestor of $\hat{y}(x)$ that has a score that is at least λ and include in the prediction sets all its descendants. To ensure that the sets are nested, to this subgraph we add all the other ones that contain $\hat{y}(x)$ for which the score is less than λ.

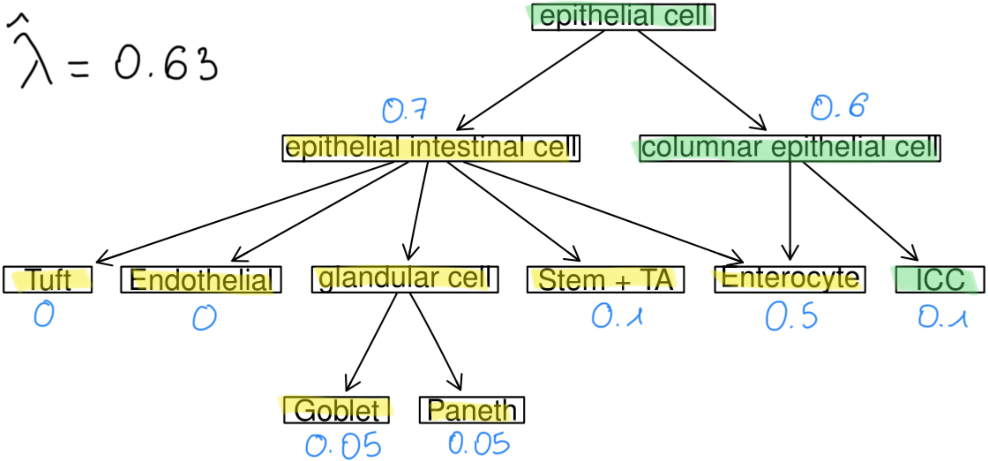
A simple example

$\hat{\lambda} = 0.63$

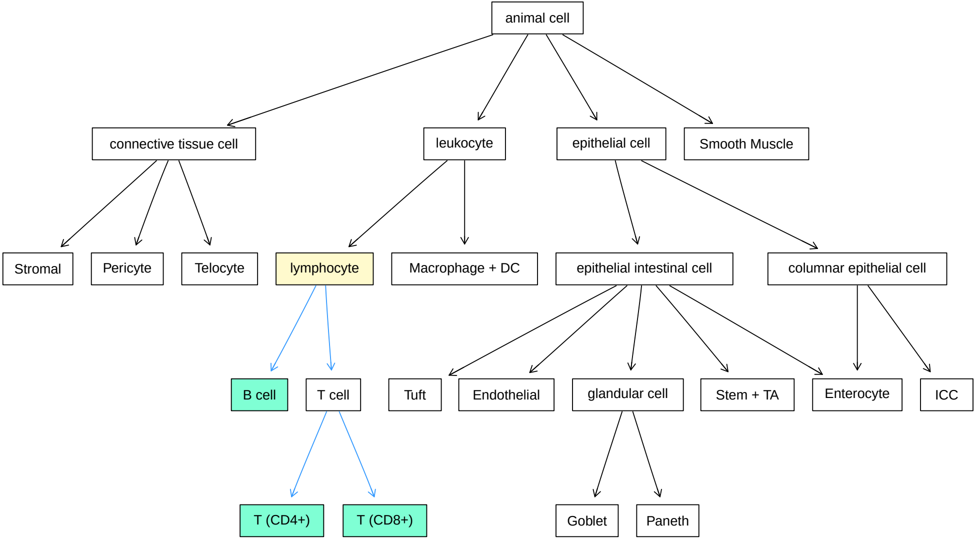


A simple example

$\hat{\lambda} = 0.63$



Back to example 2



Application

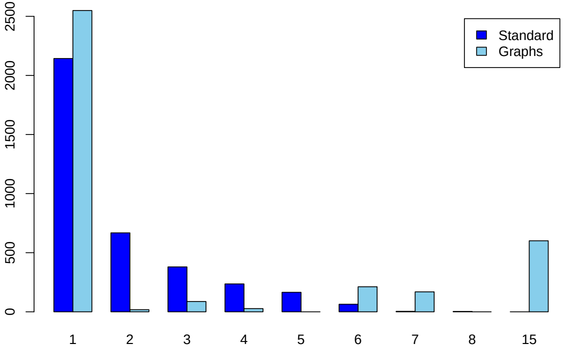
Random split:

- Training (model fit): 500 obs.

  Model: multinomial logit, 50 genes with the highest biological variance of the log-expression. Accuracy is 0.772.
- Calibration: 1000 obs. Used to compute the quantiles for split conformal and graph-structured
- Query: 3663 obs.

Comparison

Method	Coverage	Avg. Size	Avg. Dist.
(Standard) Conformal	0.901	1.842	1.564
Graph Conformal	0.903	3.577	1.003



Open problems

- Size of the calibration set
  It affects the precision of the coverage. Standard results (i.e. Beta distribution) does not apply in the Graph-structured procedure.
- Exchangeability of calibration data and query data is assumed, but in practice there are different sources of distribution shift:
  1. different technologies
  2. batch effects
  3. different proportions of cell types in calibration and query set

Acknowledgements

University of Padova

- Livio Finos
- Davide Risso

Harvard Medical School

- Tram Nguyen
- Anthony Christidis
- Ludwig Geistlinger
- Robert Gentleman

Thank you for your attention!

daniela.corbetta@phd.unipd.it

https://github.com/ccb-hms/scConform