Conformal inference for cell type prediction leveraging the cell ontology

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

daniela.corbetta@phd.unipd.it

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

EuroBioconductor 2024 - Sept 4, 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	...	Gene K	Cell type
Cell 1	1	...	5	B cell
Cell 2	0	...	5	B cell
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Cell m	3	...	0	T (CD4+)

	Gene 1	...	Gene K	Cell type
New cell	4	...	0	**?**

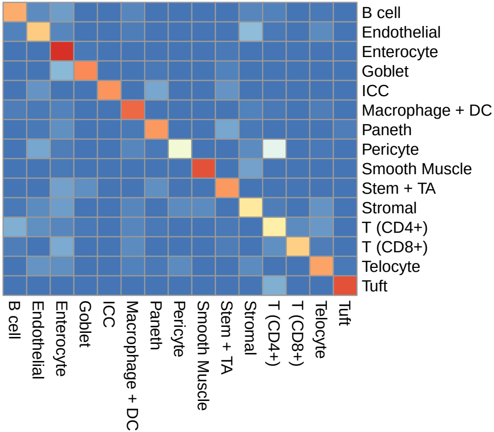
Example

Data: 5163 cells from the mouse ileum sequenced with Merfish (Petukhov et al., 2022)

- 500 cells for the training set
- 4663 cells for the test set
- 15 different cell types

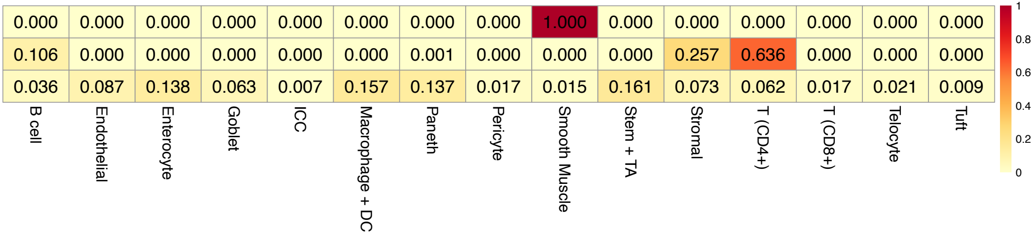
Model: Multinomial model with the 50 HVGs

Results: Accuracy=0.77



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

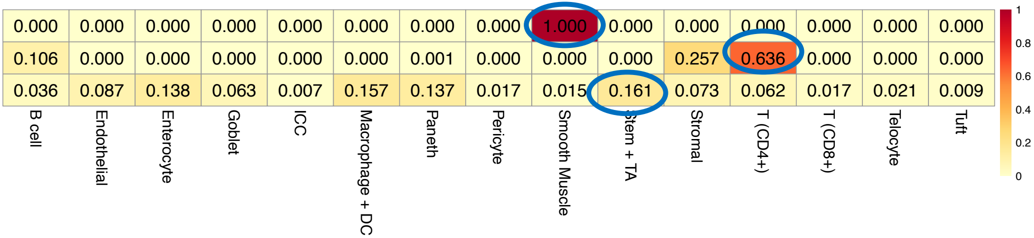


B cell	Endothelial	Enterocyte	Goblet	ICC	Macrophage + DC	Paneth	Pericyte	Smooth Muscle	Stem + TA	Stromal	T (CD4+)	T (CD8+)	Telocyte	Tuft
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
0.106	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.257	0.636	0.000	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

1
0.8
0.6
0.4
0.2
0

Should we rely on point predictions?

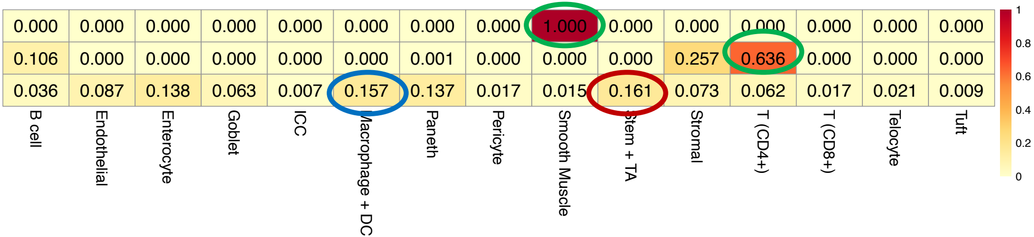
- The model does not provide only a label, but also estimated probabilities for each class
- These probabilities encode how sure the model is of the prediction



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[1]

[1] main reference Vovk et al. (2005), very easy and nice tutorial in Angelopoulos & Bates (2021)

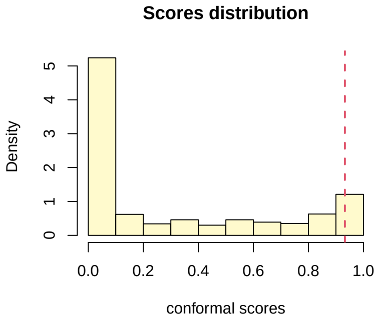
Conformal inference

- 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
  - **test set**: possibly non-annotated data on which we want to do predictions. Need to be exchangeable with the calibration data

Conformal inference - algorithm

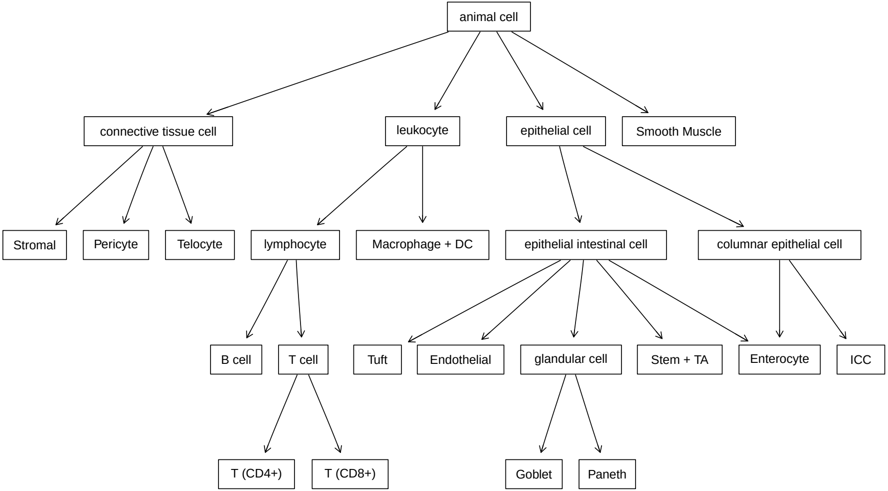
Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the data in the calibration set.

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
4. Obtain predictions for the data in the test set and form prediction sets by including all the classes that have predicted probabilities $\geq 1 - \hat{q}$



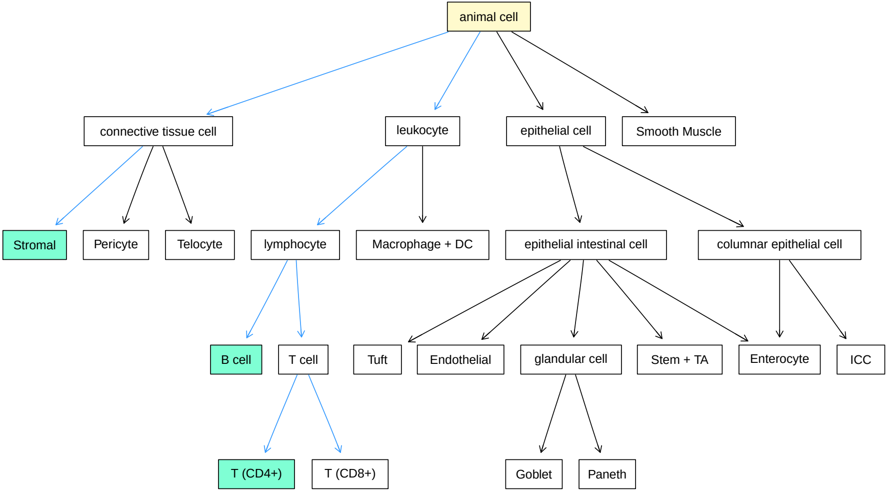
Cell ontology

Cell types are organized into a graph structure



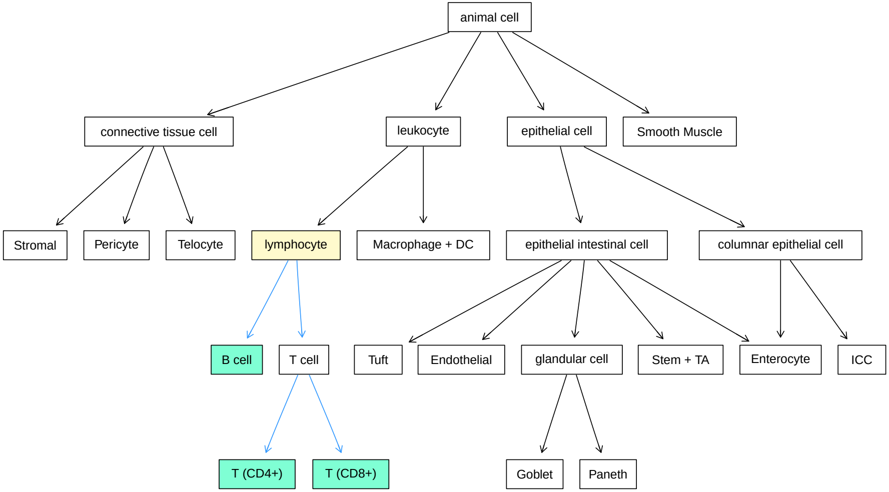
Cell ontology

Cell types are organized into a graph structure



Cell ontology

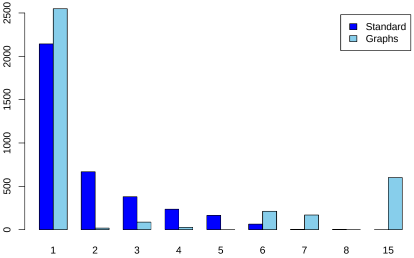
Cell types are organized into a graph structure → select an ancestor of the prediction



Application and methods' comparison

Random split:

- Training (model fit): 500 cells.
  Model: multinomial logit, 50 genes with the highest biological variance. Accuracy is 0.772.
- Calibration: 1000 cells. Used to compute the parameters for split conformal and graph-structured method.
- Test: 3663 cells.



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

scConform R package[2]

- Try our method with the scConform R package, available on Github but soon to be submitted to Bioconductor
- Check out the vignette for package functionalities

[2] https://github.com/ccb-hms/scConform

References

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

Petukhov, V., Xu, R. J., Soldatov, R. A., Cadinu, P., Khodosevich, K., Moffitt, J. R., & Kharchenko, P. V. (2022). Cell segmentation in imaging-based spatial transcriptomics. *Nature biotechnology*, 40(3), 345–354.

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