Unsupervised Barcode Demultiplexing

Adrián Goga, Broňa Brejová, Tomáš Vinař

Computational Biology Research Group, Faculty of Mathematics, Physics and Informatics,
Comenius University in Bratislava, Mlynská dolina, 842 48 Bratislava, Slovakia
http://compbio.fmph.uniba.sk/

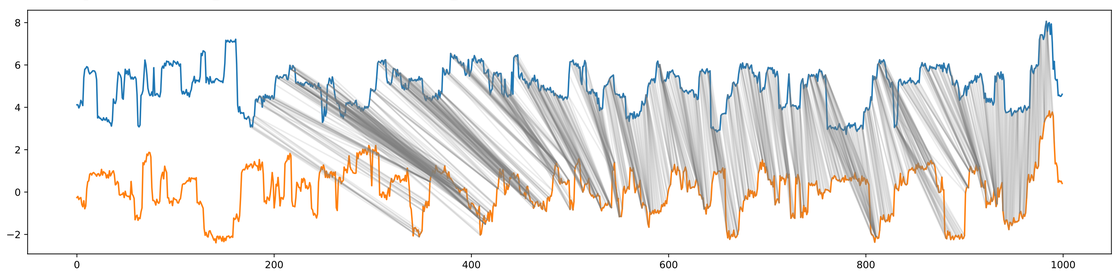
Introduction

The current approaches to demultiplexing of barcoded reads typically use base called sequences and tend to render up to 20% of the reads unusable due to base calling errors. In contrast, Deepbinner [Wick et al., 2018] works with the raw signal by employing a convolutional neural network and loses only $\approx 5\%$ of reads, while retaining the precision of $\approx 98\%$. We present a novel approach that also operates in the signal space, but is based on unsupervised learning.

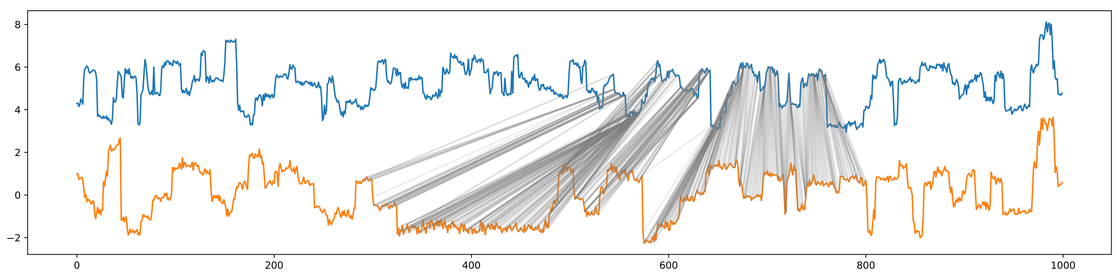
Local Dynamic Time Warping

Idea: compare squiggles by the similarity of their barcode-containing regions, which are typically starts of the reads (in some kits also their ends). Our similarity function combines the concepts of Dynamic Time Warping (DTW) and local alignment, resulting in local DTW (LDTW). The LDTW score highlights the most similar subsequences of two squiggles (regions of length about 1000 signal observations typically containing the barcode). If they share the same barcode, the LDTW should yield the alignments of barcode subsequences.

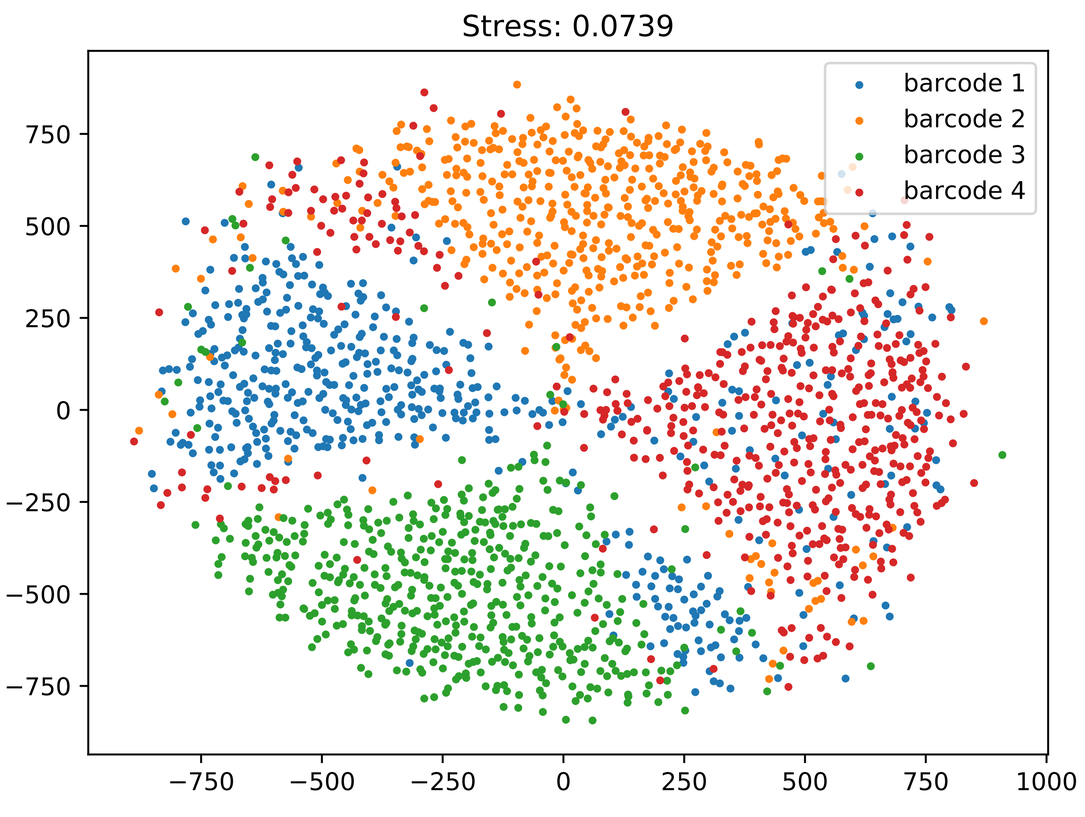
Example of regions containing the same barcodes:



Example of regions containing different barcodes: (much shorter alignment, likely corresponding to the barcode flanking sequences)



Euclidean space embedding based on LDTW distances using multidimensional scaling:



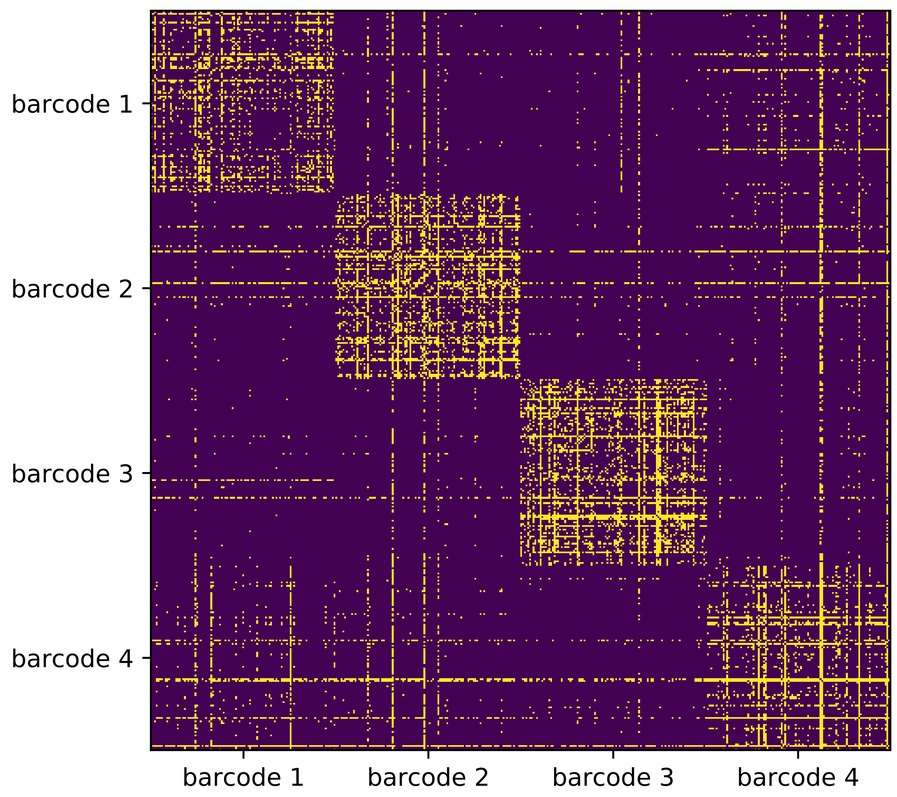
Acknowledgements

This research was supported in part by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 872539, and by grants from the Slovak Research and Development Agency (APVV-18-0239) and VEGA (1/0458/18 to TV and 1/0463/20 to BB).

Clustering of Reads Based on Similarity

1. Randomly sample a small **discovery** set of reads $\mathcal{A}$
2. **Discovery step:** Compute the all-pairs similarities in $\mathcal{A}$ and run **spectral clustering** algorithm [Von Luxburg, 2007]
3. Select r representatives from each barcode class
4. **Binning step:** Align remaining reads to the representatives and assign the class based on the alignment scores, or conclude that the read is ambiguous and discard it.

Example of an adjacency matrix used in the discovery step:



Experimental Results

Experiment	Barcodes	Precision(%)	Recall(%)	Binned reads(%)
1	5, 6, 7	98.77	93.43	94.59
2	5, 6, 7, 8	88.69	59.36	66.82
3	5, 7, 9, 12	97.41	89.53	91.91
4	8, 9	98.95	92.29	93.26
5	5, 6, 11, 12	98.60	93.85	95.18
6	11, 12	98.50	91.41	92.81

(experiments are based on Deepbinner data set)

Our approach appears to have problems when barcode classes are imbalanced (experiment 2). The less frequent barcodes may get either omitted or polluted by dominant barcodes.

We also tested our approach in supervised setting. While training on a much smaller data set than Deepbinner, our method achieved comparable results. Moreover, our approach leads to an interpretable model.

	Precision (%)	Recall (%)	Binned reads (%)
our method	98.39	91.35	92.84
Deepbinner	98.41	93.33	94.84

Side note: Not all pairs of barcodes are equally dissimilar in the signal space, and therefore not all subsets of barcodes are equally suitable for use in a single experiment. For example, barcode 1 of the NBK is often very close to barcode 4 (see the MDS visualization).

References

[Von Luxburg, 2007] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[Wick et al., 2018] Wick, R. R., Judd, L. M., Holt, K. E., and Pertea, M. (2018). Deepbinner: Demultiplexing barcoded oxford nanopore reads with deep convolutional neural networks. *PLOS Computational Biology*, 14(11).