Supplemental Material

A Graph Coarsening Algorithm for Compressing Representations of Single-Cell Data with Clinical or Experimental Attributes

Chi-Jane Chen[†], Emma Crawford[⅄], and Natalie Stanley[‡]

Department of Computer Science and Computational Medicine Program
The University of North Carolina at Chapel Hill,
Chapel Hill, NC, 27599, USA
{[†]*chijane@cs.unc.edu*,[⅄] *emmabc@email.unc.edu*,[‡] *natalies@cs.unc.edu*}

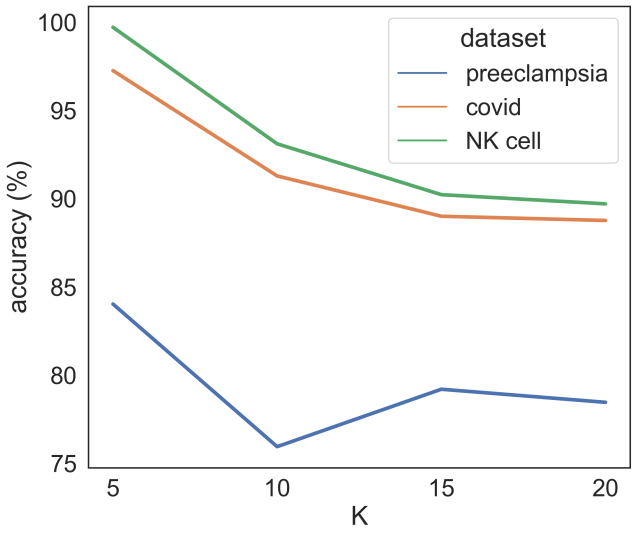

Fig. 1. **Accuracy as a function of k for the KNN graph.** To determine the optimal number of nearest neighbors, K for cytocoarsening, we ran it on three datasets, keeping all other parameters as their default values and varying K between 5 and 20. The accuracy measures the percentage of nodes for which the original attribute labels and coarsened attribute labels align. The default parameter values are $\alpha = 26$, Instead of fixing the number of passes, we ran the algorithm until the coarse graph had less than 75% of the number of nodes of the original graph. For each dataset, we ran the algorithm 30 times on random subsamples of 1000 cells per patient and averaged the accuracy over all 30 runs. We found that the accuracy is highest when K=5 for all three datasets, and the accuracy generally decreases with increasing K. Note that the preeclampsia dataset has continuous labels (gestational age) whereas the COVID and NK cell datasets have discrete labels (disease severity and positive/negative for CMV, respectively). Because the continuous labels are harder to differentiate (ex. 10 weeks and 11 weeks of gestational age would not be markedly different, while positive for CMV and negative for CMV would be), cytocoarsening was less accurate on the preeclampsia dataset. The accuracy of cytocoarsening on the preeclampsia dataset is the lowest at K=10, but the general trend is still consistent with the others - the accuracy decreases as K increases.

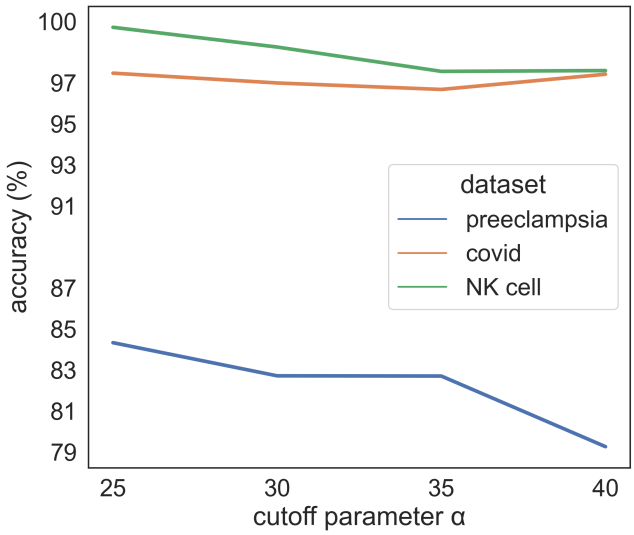

Fig. 2. **Accuracy as a function of cutoff percentile thresholds for the coarsening graph.** We ran cytocoarsening on each of the three datasets for different cutoff percentiles α, ranging between 25 and 40. We fixed the number of neighbors K at the default (5), and ran the algorithm until the number of coarse nodes was less than 75% the number of nodes in the original graph. When plotting accuracy as a function of α, we found the accuracy was higher when the cutoff parameter was lower. For the values of α we tested, our results show the highest accuracy appears when α is set at the 25^{th} percentile. The accuracy of cytocoarsening on the preeclampsia dataset decreased consistently with increasing α. The accuracy for both covid and NK cell datasets achieves its minimum at $\alpha = 35$, the 35th percentile. However, the trend of both two datasets generally shows the accuracy declines as the cutoff parameter increases.

Acknowledgments

Partial support for Emma Crawford is gratefully acknowledged from the National Science Foundation, award NSF-DMS-1929298 to the Statistical and Applied Mathematical Sciences Institute.