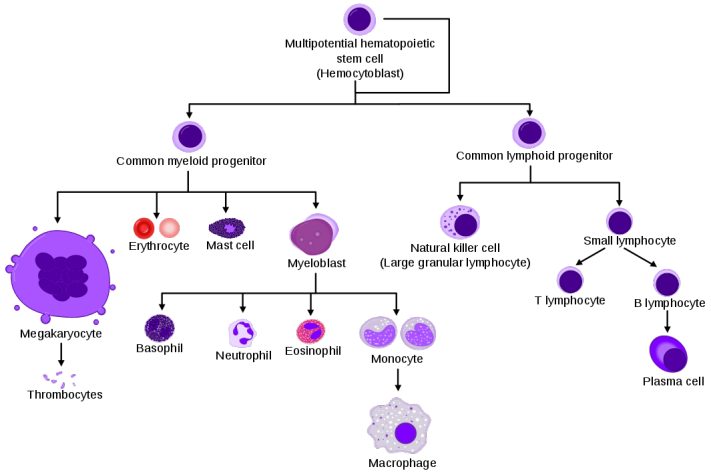
2. Hierarchical, multi-label prediction for automated cell type identification

A common type of data generated in the field of bioinformatics, is so-called "gene expression data", where activity levels of all genes in a particular organism are measured. While the DNA in all cells of an organism is the same, the current set of genes that is active in a cell paints a good picture of what type of cell it is (e.g. cancer cell, blood cell, …), and what is going on in this cell.

A recent technological advance in the field is that due to miniaturization scientists are now able to perform these gene activity measurements for single-cells, and this for many thousands of cells present in a sample. While some cell types are well characterized, other cells are far more unknown, and scientists still discover new cell types, novel subtypes of cells, or novel cell states on a daily basis. For a biologist interpreting such a dataset of thousands of cells, with each cell being described by several thousands of genes, is a nightmare, and thus machine learning techniques provide an excellent way of automating the process of discriminating all these cell types.

In this project, you will develop a machine learning based classifier to predict cell types from single-cell gene expression data. While traditional classification models would treat all the classes (in this case the different cell types) as independent classes, the biological reality is more complex, as often cell types are hierarchically structured (see Figure below). It would thus be interesting to take into account such hierarchical relationships that exist between the different classes, bringing us to the domain of hierarchical multi-label classification.



[1] Abdelaal, T., Michielsen, L., Cats, D. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20, 194 (2019). https://doi.org/10.1186/s13059-019-1795-z

[2] Barros, R. C., Cerri, R., Freitas, A. A., and Carvalho, A. C. P. L. F. Probabilistic Clustering for Hierarchical Multi- Label Classification of Protein Functions. In European Conference on Machine Learning (ECML/PKDD 2013), pp. 385–400. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40991-2.

[3] Bi, Wei and Kwok, James. Multi-Label Classification on Tree- and DAG-Structured Hierarchies. In International Conference on Machine Learning, ICML'11, pp. 17–24. ACM, 2011.