Deconvolution approaches for automatically detecting novel secretion systems

Sophie Abby
TIMC, CNRS

Nelle Varoquaux
TIMC, CNRS

September 8, 2022

Secretion systems are crucial for prokaryotic organisms to interact with their environment, such as acquiring nutriments, setting up biotic defenses, as well as delivering virulence factors. There are currently 12 bacterial secretion systems known varying in size (1 to 15 proteins involved). Some organisms can have several times the same secretion system, and sometimes a single protein or group of proteins is involved in several secretion systems. Thus, detecting which protein belongs to which secretion system is a difficult task.

If we assume we know all homologs (a protein family with a common ancestor) involved in all secretion systems, can we find which homolog is involved in which secretion system? We propose to formulate this task as a **deconvolution approach** (*e.g.*, non-negative matrix factorization [1]): given a matrix $\mathbf{X}$ where each row corresponds to an organism, each column to a homolog, and each entry the number of times a homolog is found in an organism, factorize $\mathbf{X}$ into two matrices, one corresponding to which homologs are found in which type of secretion system, the other corresponding to the number of secretion system in each organism.

The goal of the internship is to investigate how non-negative matrix factorization approaches can be used to perform this task.

1. We will first perform a literature review on this topic, specifically for NMF approaches on count data [1, 2].
2. Then, we will test a couple of algorithms on this particular task.
3. We will investigate how to tune the hyperparameters of the model using prior knowledge and/or stability analysis.
4. Finally, we will investigate better feature engineering to consider each protein's genomic context (*e.g.*, the gene organization along the genome): Taking inspiration from nature language processing, we will extract n-grams of proteins and investigate how to add structured penalization to the deconvolution.

References

[1] Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Adv. Neural Inf. Process. Syst.*, volume 13. MIT Press, 2001.

[2] Cédric Févotte and A. Taylan Cemgil. Nonnegative matrix factorizations as probabilistic inference in composite models. In *In In Proc. 17th European Signal Processing Conference (EUSIPCO-09)*, 2009.