Deconvolution approaches for automatically detecting novel secretion systems

Sophie Abby
TIMC, CNRS

Nelle Varoquaux
TIMC, CNRS

November 9, 2022

Internship subject

Secretion systems are crucial for bacterial 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.

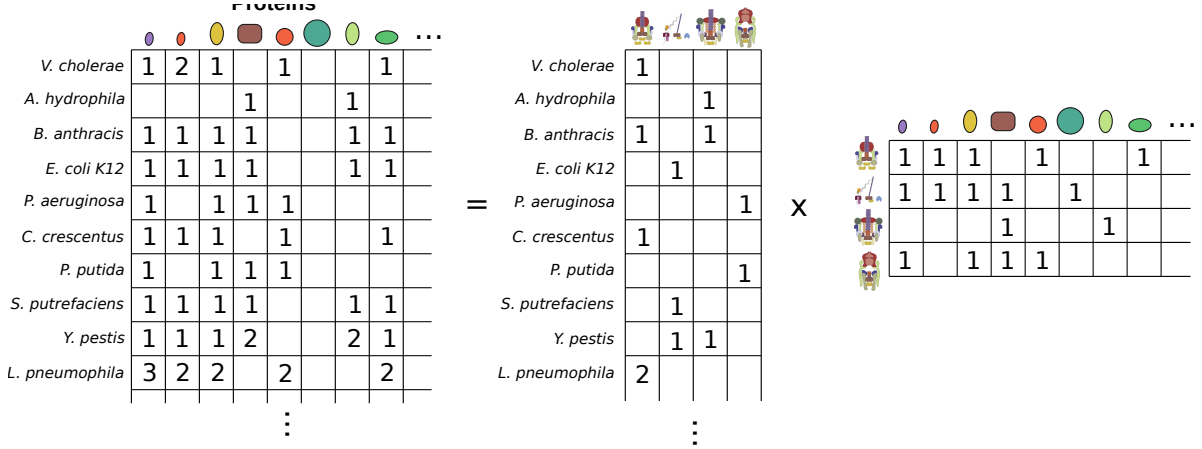

Figure 1: bf A schema representing an NMF approach to tackle the problem at hand. The goal is to decompose a matrix X (where each row corresponds to an organism, each column a protein known to be involved in a secretion system, and each entry the number of times this protein is found in this genome) into two matrix V and U to automatically detect the number of each time of secretion systems in each genome.

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, denoted by V, corresponding to which homologs are found in which type of secretion system, the other, denoted by U corresponding to the number of secretion system in each organism.

A challenge when using such non-supervised approaches is to tune the hyper-parameters of the model (*e.g.*, k, the number of secretion systems). We propose to investigate two strategies. First, in this particular case, some secretion systems are well-known and well studied. We can thus leverage "classical" autoML strategies to tune the hyperparameters of the model through cross-validation. We will thus perform the deconvolution

for a large number of k on a subpart of the data and check whether (1) the V matrix corresponding to the match between homolog and secretion system type is correct or not; (2) whether the matrix U on the training or the test data is correct or not using classification metric of successes (*e.g.*, balanced accuracy, precision). Second, to ensure such an approach can be applied on systems for which there is no prior knowledge, we will investigate whether stability analysis can be used in this particular case. Several approaches can be investigated here: (1) bootstrapping the input data and assessing whether the matrix V is robustly estimated between bootstrapped samples for a particular k (akin to stability estimation of the number of clusters; [2]; (2) separating the data into a training and testing set and estimating the accuracy of the reconstruction on the testing dataset.

The goal of the internship is to investigate how non-negative matrix factorization (NMF) 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, 3].
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.

Candidate profile

The candidate should be a master (1st or 2nd year) student or equivalent (engineering school) in applied mathematics, computer science, or bioinformatics, with a interest in biological applications. The candidate should have strong communication skills and be capable of interacting with researchers from other fields (*e.g.* medical doctors, bioinformaticians, biologists). Opportunities for

The application file should contain:

- Cover letter
- CV

Applications

Applications are to be addressed to Nelle Varoquaux: `nelle.varoquaux@univ-grenoble-alpes.fr`

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] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 6–17, 2001.

[3] 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.