Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-26 (2016) 071–076

A Constrained NMF Approach to Analyze Quantitative Metagenomic Data [⋆]

Sébastien Raguideau [*] Béatrice Laroche [*] Marion Leclerc [**]
Sandra Plancade [*]

[*] *MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas*
[**] *Micalis, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas,*
France (e-mail: sraguideau, blaroche, mleclerc, splancade +
@jouy.inra.fr)

Abstract: In this paper, we propose a new method for inferring the metabolic potential of microbial ecosystems based on gene frequencies generated from shotgun metagenomic data. Our approach is based on Non-Negative Matrix Factorization with constraints accounting for prior biological knowledge of bacterial metabolism. The problem is solved using efficient accelerated projected gradient methods. The approach is illustrated on a toy model and on real data on fiber metabolism by the gut microbiota in humans. We show how this approach leads to the inference of biologically relevant gene clusters.

Keywords: Data reduction, Non-Negative Matrix Factorization, Metagenomics, Microbial ecosystems, Gut, Supervised Machine Learning.

1. INTRODUCTION

The human gut microbiota is a complex community whose interaction with the human host have been extensively studied. The microbiota influences various critical functions such as immune system maturation, human physiology, metabolism and nutrition (Guinane and Cotter (2013)). Its bacterial species composition varies over time and individuals, however metabolic processes at the ecosystem level are ubiquitous and much more stable in time. Since most of micro-organisms present in the distal gut cannot be isolated, observing and studying this ecosystem remained an unreachable goal until recently. The development of high-throughput molecular techniques such as shotgun metagenomics has now allowed to study metabolic processes at the ecosystem level. Large-scale sequencing project (MetaHit (Li et al. (2014)) and HMP (Methé et al. (2012)) consortia) have produced gene catalogues of the human microbiome from which it is now possible to generate genes abundances with functional annotation.
Shotgun metagenomics generate huge amounts of data whose interpretation is still a challenge. In this paper, our goal is to propose a new method for inferring the metabolic potential of microbial ecosystems based on these data. Different approaches have been proposed so far, such as pathway abundance counting (Abubucker et al. (2012)) or gene and samples co-clustering (Jiang et al. (2012)). We focus in the second approach which is based on Non-Negative Matrix Factorization (NMF) techniques. We improve it by taking into account prior biological knowledge of bacterial metabolism, thus obtaining the decomposition of samples over biologically relevant gene clusters.
The paper is organized as follow. The first section focuses on the model presentation and biological assumptions. The second section is devoted to the formulation of the constraints. Section 3 and 4 deal with the optimization problem formulation and its implementation. The method performance and results are evaluated on a toy example in section 5, followed by an illustration on real data in section 6.

2. MODEL PRESENTATION

2.1 Functional Profile Modeling

In a biological sample, genes are associated within bacterial genomes, which are further grouped through micro-organism associations at the ecosystem level, resulting in gene association patterns. These association patterns are indeed the aggregation of many complex co-occurrences or ecological interactions, for instance trophic chains, involving thousands of different bacterial species. They are related to important metabolic processes carried out by the ecosystem under consideration, such as Short Chain Fatty Acid (SFCA) production from dietary fibre degradation for the gut ecosystem.
Our objective is to decipher the patterns resulting from this organisation. Our first modeling hypothesis is that functions are selected regardless of species that fulfill them. This is why we adopt in all that follows a full functional approach and aggregate gene abundances derived from shotgun metagenomics into functional marker abundances by using the Kegg Othology (KO) annotation. Our second hypothesis is that the patterns we are looking for are shared by all samples, because they were selected by specific environmental constraints pertaining to the ecosystem under consideration (e.g. anaerobic conditions, pH and temperature for the gut).
In the sequel we assume that we are interested in an

[⋆] Sebastien Raguideau is funded by a phD grant of the Meta-omics and Microbial Ecosystems (MME) program of INRA.

2405-8963 © 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.12.105

ecosystem level metabolic function, and that our data consist in the abundance of relevant functional markers in several samples. Abundances in each sample are modeled as a mixture of the patterns, in variable proportions across samples. These proportions are influenced by environmental factors and local specific conditions (e. g. host diet, digestion residue composition, age or health status for the gut).

We characterize each pattern by a line vector of functional marker frequencies, which we define to be an Aggregated Functional Traits (AFT). We therefore model functional marker abundances in each sample as a mixture of latent AFT.

The formulation of this model is the following. Let A the $n \times r$ data matrix formed with the observed abundances of r markers (columns) in n samples (lines). We introduce H, the $k \times r$ trait matrix. Each of the k lines of H corresponds to an AFT, described by its composition in terms of proportions of the r markers. We also define W, the $r \times k$ formed with the weights of the k AFT (lines) in the n samples (columns).

For all sample $i = 1, \dots, n$, and marker $j = 1, \dots, r$ the observed frequency A_{ij} of marker j in sample i is the sum of contributions of all AFT leading to

$$A_{ij} \simeq \sum_{l=1}^{k} W_{il} H_{lj} \tag{1}$$

2.2 Imposing Biological Relevant Constraints on Profiles

A first obvious constraint is that the weights in W and proportions in H must be non negative

$$W \geq 0, \quad H \geq 0. \tag{2}$$

Moreover, we would like each trait h to characterize a viable sub-population in the ecosystem. To this end, we assume that a topological network representing all the known metabolic pathways underlying the functions we are interested in is available, such as illustrated in Fig. 1 on the toy example of section 5, where the m nodes are the metabolites and the r edges are the reactions.

Metabolites can be parted into extracellular metabolites, meaning that they can be imported or exported in the environment outside the bacterial cells, and intracellular metabolites which are intermediate products of the biochemical reactions within the cells. In bacterial genomes, we expect some coherence in the genetic structure, that is if genes involved in the production of an intracellular compound are present, genes involved in its utilization should be there, and conversely. As AFTs result from the aggregation of many bacterial genomes, they should obey the same rules. This is why we assume that the knowledge available about the specific metabolic process under consideration and known bacteria in the ecosystem allow to formulate constraints on the AFT structure in the following way.

Let $h \geq 0$ be an AFT, and j the index of an intracellular metabolite. Denote S_{Pj} (resp. S_{Cj}) the sum of all functional markers frequencies in h corresponding to reactions that produce (resp. consume) j, then we assume that the following constraints are satisfied:

$$S_{Pj} \leq \delta_j^- S_{Cj}, \text{ and } S_{Cj} \leq \delta_j^+ S_{Pj}.$$

The choice in the conditions and the values of the positive constants δ_j^- and δ_j^+ is given by biological prior knowledge.

The constraints impose that the non-zero coordinates in h define a union of weighted, connected paths in the graph starting and ending on extracellular metabolites. Moreover, in situations where it would be possible to impose both constraint for all the intracellular metabolites with $\delta_j^- = \delta_j^+ = \delta$, the ratios S_{Cj}/S_{Pj} along these paths would lie in $[1/\delta, \delta]$.

These constraints guarantee that a population with trait h is consistent with bacterial metabolism. The overall set of constraints can be expressed as

$$F_\delta H^T \leq 0 \tag{3}$$

3. NONNEGATIVE MATRIX FACTORIZATION PROBLEM FORMULATION

The determination of W and H by solving (1), (2) and (3) amounts to solving a constrained Nonnegative Matrix Factorization problem (NMF).

NMF was first proposed by Paatero and Tapper (1994), it is a convenient unsupervised model for dimension reduction which can be used either to find a limited number of sources explaining observed data or as a clustering algorithm. The non-negativity constraint allow for a realistic modeling of various problem which explains its high popularity in image processing (Hoyer (2004)), signal processing (Kameoka et al. (2009)), document mining (Lee and Seung (1999)) or metagenomics (Carr et al. (2013); Jiang et al. (2012)). In order to formulate the factorization problem, we first need to define a distance between the data A and the decomposition WH. Then we regularize the distance minimization problem both to handle the ill-posedness of the factorization problem and filter noisy data and obtain our final constrained optimization problem formulation.

3.1 Error model

Equation (1) is translated as a model on the error between A and WH. Many choices are possible, in the absence of any prior information, we chose to use a quadratic error model for its simplicity. In order to account for discrepancies in the mean abundances of the markers, we introduce the diagonal matrix D whose diagonal entries are the L_2 norm of each column of A, and define the normalized data matrix $\tilde{A} = AD^{-1}$ as well as the transformed trait matrix $\tilde{H} = HD^{-1}$. The distance between A and WH is defined as

$$d(A, WH) = \|(A - WH)D^{-1}\|_F = \|\tilde{A} - W\tilde{H}\|_F.$$

Constraints (2,3) can be expressed in terms of W and $\tilde{H}$ by defining the set of constraints $\mathcal{C}$ as the closed, positively invariant convex set

$$\mathcal{C} = \{(W, \tilde{H}) \in \mathbb{R}^{n,k} \times \mathbb{R}^{k,r} \ / \ W \geq 0, \ \tilde{H} \geq 0, \\ \tilde{F}_\delta \tilde{H}^t \leq 0\},$$

with $\tilde{F}_\delta = F_\delta D$.

From now on, we will work with the normalized matrices and drop the tilde on A and H in all that follows.