Talanta 160 (2016) 148–156

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

A modified weighted mixture model for the interpretation of spatial and temporal changes in the microbial communities in drinking water reservoirs using compositional phospholipid fatty acid data

I. Stanimirova [a,*], A. Woznica [b], T. Plociniczak [b], M. Kwasniewski [b], J. Karczewski [b]

[a] Institute of Chemistry, University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] Faculty of Biology and Environmental Protection, University of Silesia, 28 Jagiellonska Street, 40-032 Katowice, Poland

ARTICLE INFO

Article history:
Received 18 May 2016
Received in revised form
30 June 2016
Accepted 3 July 2016
Available online 4 July 2016

Keywords:
Chemometrics
Compositional data
Multivariate curve resolution for compositional data
Missing elements
Bray-Curtis similarity
Beta diversity

ABSTRACT

The aim of this work was to check whether a methodology based on the analysis of data that contain the entire phospholipid fatty acid, PLFA, compositions of water samples can be successfully used to interpret spatial and temporal changes in the microbial communities in water reservoirs. The proposed methodology consists of the construction of a modified weighted multivariate mixture model for the PLFA profiles of the water samples collected in a given monitoring campaign and the identification of latent PLFA components through a comparison with the known PLFA profiles of some cultivated or non-cultivated microbial communities. A 16S rDNA analysis of some of the selected water samples in the monitoring campaign was performed in order to verify the results of the PLFA analysis.

The results showed that the proposed methodology can be useful for a dynamic and sensitive evaluation of changes in the microbial quality of water before and after flash flooding and can help in taking a decision regarding further risk assessment.

1. Introduction

The assessment and control of bacterial diversity in the natural water ecosystems used as drinking water suppliers is a top priority for any water production company. A successful strategy for the management of the microbial safety of water in the catchment area requires information about possible sources of contamination; knowledge about spatial and seasonal microbiological variations that are influenced by various physico-chemical factors such as the temperature and pH of the water, salinity, turbidity, the amount of solids, etc.; regular monitoring at carefully selected monitoring points; and a fast and cost-effective methodology for the identification and elimination of pathogens in order to prevent serious waterborne diseases. An attractive but challenging idea for microbial safety management is the creation a dynamic model of the source water reservoir through the regular monitoring of microbial diversity and measuring some selected physicochemical parameters. Such a model should allow an interpretation of any spatial changes in the microbial communities and the identification of specific pathogens at the pollution source.

Nowadays, high resolution metabarcoding approaches [1,2], which are based on nucleic acid extraction and particularly gene coding for ribosomal 16S RNA, are preferred over the analysis of individual phospholipid fatty acids, PLFA, for a direct target independent microbial identification [3–5]. PLFA analysis is now considered to be an unspecific method [4,5] and this is not surprising since the source tracking of some coliforms or the identification of some strains of bacteria is done using single quantified phospholipid fatty acids and ignores the fact that the same individual fatty acids occur in different amounts in many species depending on the growth conditions. The use of the ratios of selected fatty acids or the conversion of the PLFA data to so-called diversity indices has also been shown to lead to misleading conclusions when the influence of physico-chemical parameters [6] such as pH are not considered.

Even though the cost of the metabarcoding analysis has been reduced over the last decade through new technologies, it is still not the method of choice for the routine identification of microbial communities in a large number of environmental samples. Moreover, metabarcoding approaches provide little information about the activity of the microorganisms and their phenotype, although they do offer the possibility to identify unknown pathogens.

Taking into account the advantages and disadvantages of both methods, a successful risk assessment strategy for controlling the microbial quality of source water might be to reduce the risk of the false identification of potential sources of pollution using the

* Corresponding author.
E-mail address: istanimi@us.edu.pl (I. Stanimirova).

http://dx.doi.org/10.1016/j.talanta.2016.07.006
0039-9140/

analysis of single quantified PLFA and to perform metabarcoding analysis only for selected samples in order to assess the actual risk of pollution.

Specifically, since most of the taxonomic groups have characteristic PLFA compositions, but only some of them have unique chemotaxonomic markers, the identification of potential sources can be performed by analyzing the data in which each water sample from a given monitoring campaign is described by its entire quantitative PLFA composition. The amount of each phospholipid fatty acid is expressed as a part of the total sample phospholipid fatty acids content. Such data represent mixtures of phospholipid fatty acids, which may originate from different species in varying amounts depending on the growth conditions at the source, but for which the mixing mechanism is unknown. Variations of pH and temperature in the growth environment lead to quantitative modifications in the PLFA composition of the bacterial cell membrane which may further complicate the interpretation of the PLFA profiles for water samples.

The goal of this work was to propose a methodology for the analysis of data that contain the entire PLFA content of every water sample in order to enhance the level of information that can be obtained. We also wanted to check whether the relation of the results of the PLFA data analysis with those of the environmental metabarcoding data analysis can serve as a successful strategy for the interpretation of spatial and temporal changes in the microbial communities in water reservoirs.

In general, in order to interpret the PLFA mixture data, one should find the so-called pure contributions due to the latent PLFA components, which may be characteristic for individual microbial communities. In geology, this type of modeling is called 'end-member analysis', while in the social sciences it is known as 'latent budget analysis' [7,8]. There are two approaches that can be used to obtain the latent components ('end-members' or 'latent budgets') from compositional data. In the first approach [9], the original compositional data are log-ratio transformed and classical factor analysis is then used to obtain the latent mixture components that are further rotated and normalized to ensure the non-negativity of its elements and finally transformed back in order to be interpreted as compositions. If some of the original data elements are missing (taking into account their mechanism of missingness), they should be replaced before the transformation. In contrast to this multi-step approach, the aim of the second approach is to obtain the latent non-negative PLFA compositions and their physically meaningful non-negative contributions directly. Only the inclusion of non-negativity and closure constraints within the steps of the estimation of latent components and their contributions can guarantee that their final estimations approximate the original compositional data in a least-squares sense optimally [10]. The latter approach was adopted in the weighted multivariate curve resolution, MCR, algorithm that has been proposed for compositional data by Mooijaart et al. [10]. This algorithm was specifically designed for compositional data analysis and the weighting scheme differs from the weighting scheme used in the multivariate curve resolution-weighted alternating least squares method, MCR-WALS, [11,12] or positive matrix factorization, PMF, [13] which incorporate a measurement uncertainty estimation. We further modified a version of this algorithm for compositional data in order to handle any type of missing elements.

The interpretation of the latent PLFA mixture components in terms of the profiles of a specific community of microorganisms requires a comparison with the known PLFA profiles of selected cultivated and non-cultivated microbial communities. There are several important issues to be considered in this comparison. The elements of each profile sum to one (closure constraint) and every estimated latent profile and the known mixture profile contain 'rounded zeros', e.g. elements the values of which are below the reporting limit. The correlation coefficient seems to be a straightforward quantitative measure of similarity for each pair of profiles. However, the closure constraint for these profiles leads to the estimation of 'spurious correlations' [8]. One way to avoid this problem is to consider the squared Aitchison distance. However, the estimation of the Aitchison distance on data that contain 'rounded zero' elements is problematic. Several approaches have been proposed in the literature and they will be discussed further in the text.

The verification of the results that are obtained from a PLFA data analysis and the assessment of possible risks may be performed through the analysis of the 16S rDNA data obtained from selected water samples. Here, established methods such as hierarchical clustering using the Bray-Curtis [14,15] or UniFrac metrics [16] are usually adopted for a good visualization and an adequate interpretation in terms of the similarity between pairs of samples.

The methodology is illustrated on the PLFA and 16S rDNA-based metabarcoding data that were obtained from water samples that were collected within a one-year monitoring campaign at nine sampling points that are located at the Goczalkowice Reservoir. The Goczalkowice Reservoir is one of the main sources of drinking water for the Upper Silesian region of Southern Poland and is owned by the Upper Silesian Waterworks Plc. Company. This company is one of the biggest in Europe and includes several surface and deep water treatment plants. The research work that is presented in this article is part of the European Union Innovative Economy Operational Programme project, abbreviated as ZiZOZap, for creating an integrated support system for the management and protection of a dam reservoir.

2. Experimental section

The water data set that was analyzed in this work included the content of 97 phospholipid fatty acids which were determined in 54 water samples collected between June 2010 to September 2011 at nine sampling points (abbreviated as z1, z2, …, z9) that are located at the Goczalkowice water reservoir. The sampling points are situated on the Vistula River, which is one of the main tributaries, as well as above and below the tributary and in front of the pipe section into the water treatment plant. Each sample was described by its phospholipid fatty acids profile. Additionally, 16S rDNA metabarcoding analysis was performed for eight samples that were collected within the sampling period.

2.1. Analysis of phospholipid fatty acids

PLFAs were isolated from the biomass that was collected on the surface of a cellulose filter (pore size $=0.45\ \mu m$) after filtration of 5 L of water. The analytical procedure was described by Pennanen et al. [17]. A mixture of chloroform, methanol, and a citrate buffer (v:v:v=1:2:0.8) was added to each sample. The solution was mixed in a horizontal shaker for 12 h at room temperature and then a mixture of 2.5 mL chloroform and 2.5 mL buffer was added. The organic phase, which contained the lipids, was separated from the water phase and after drying with nitrogen, the sample was stored at $-20\ ^{\circ}C$ prior to further analysis. A chromatographic preparation on silicic acid columns was performed using 5 mL chloroform, 10 mL acetone and 5 mL methanol, respectively, in order to separate neutral, glycolipids and polar lipid fatty acids. Each sample was dissolved in a 1 mL mixture of toluene: methanol (v:v=1:1) and then subjected to a mild alkaline methanolysis using 1 mL 0.2 M KOH in methanol. Methyl nonadecanoic acid (19:0) was used as the internal standard in order to calculate the concentration of individual fatty acids. The samples that contained