A CANDECOMP/PARAFAC APPROACH TO THE ESTIMATION OF ENVIRONMENTAL POLLUTANT CONCENTRATIONS USING BIOSENSORS

F. Caland[*,†], *S. Miron*[*], *D. Brie*[*], *C. Mustin*[†],

[*]CRAN, Nancy-Université, CNRS
[†]LIMOS, Nancy-Université, CNRS
Boulevard des Aiguilettes BP 70239, 54506 Vandoeuvre-lès-Nancy, France

ABSTRACT

A biosensor is a bacterium that is genetically modified to produce fluorescent proteins when exposed to environmental pollutants. In this paper, we investigate the possibility of estimating the concentration of environmental pollutants in a sample by using non-specific biosensors. To do that, we first propose an experimental procedure allowing to obtain three-way fluorescence data sets. The extraction of bio-sensor response to changes in environmental pollutant concentrations will be achieved by Candecomp/Parafac. This allows the estimation of multiple environmental pollutant concentrations using calibration curves of biosensors.

Index Terms— multiway data, Candecomp/Parafac, spectrofluorimetry, biosensor, environmental pollution

1. INTRODUCTION

In recent years, the bio-sensors have known a significant development, because of the increasing awareness on the environmental issues [1, 2]. Nowadays, there is a need to improve risk assessment and/or evaluation of remediation technologies by including qualitative and quantitative consideration on the bio-availability of contaminants. For heavy metals (such as cadmium), uncertainty in the relationship between total metal concentrations and those available for uptake by possible biological receptors may lead to situations where risk is over- or underestimated. Thus, qualitative and quantitative considerations on the bio-availability of environmental pollutants versus ecological receptors are required. This motivates the development of bacterial biosensors or whole-cell bio-reporters designed for the detection of environmental (chemical, physical or biological) conditions. The design of a bacterial biosensor consists in adding a reporter gene to the bacterium DNA so that it synthesizes fluorescent proteins in response to the presence of a stressing element. The biologists discovered that bacteria that are resistant to some toxins have developed a toxin detection mechanism. Thus, if the presence of the toxin is detected, they activate their protection mechanism and the genetic modifications enforce the production of fluorescent proteins by the bacteria. The fluorescent elements are produced according to the toxin quantity detected by the bacterium. Figure 1 shows a schematic diagram of a bacterial biosensor and its components interacting to provide information on toxicity level. Biosensors have been developed to react to a number of environmental pollutants. However, one of the main limitation to their use results from the lack of promoters specific to a particular environmental factor. This is why current approaches, favored by microbiologists, aim at developing specific promoters. The approach that we are proposing in the HAESPRI project [3] is quite different since it consists in fully exploiting the diversity of biosensor responses from which the relevant information can be recovered through signal processing methods. In this paper, we

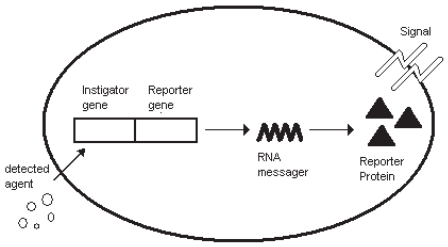

Fig. 1. Schematic diagram of bacterial biosensor

specifically address the problem of estimating the concentration of environmental pollutants in a sample by using non-specific[1] biosensors. Let us mention that this is a problem which has not received a satisfying solution since available methods only allow the concentration estimation of a single pollutant, thus implicitly assuming that no other pollutant is present in the sample. As far as we know, this paper provides the first attempt to solve the problem of jointly estimating the concentrations of multiple environmental pollu-

This work has been supported by the French ANR program through grant ANR-09-BLAN-0336-04.

[1]A non-specific biosensor is a biosensor responding to several environmental factors.

tants present in a sample. The reminder of the paper is organized as follows. In section 2, we present the experimental approach for obtaining data allowing to estimate the pollutant responses. Basically, it consists in performing a metered addition for each pollutant possibly present in the analyzed sample. Combining these metered additions with a controlled concentration profile of the different biosensors yields three way data arrays (one for each pollutant) admitting a Candecomp/Parafac (CP) decomposition. In section 3, some basic notions related to the CP decomposition are recalled and the uniqueness of CP decomposition of these data arrays is addressed. Performing the CP decomposition yields the three factors representing respectively the fluorescence spectra of each bacterium, the concentration of bacteria and the response of each bacteria to the studied pollutant. Remarking that the first value of the pollutant response corresponds to the pollutant concentration, we state in section 4 the pollutant concentration estimation as a minimization problem. The proposed method is illustrated in section 5 by a numerical simulation.

2. GENERATING THREE-WAY FLUORESCENCE DATA BY METERED ADDITIONS AND BIOSENSOR CONTROLLED CONCENTRATIONS

A biosensor is producing a fluorescence signal $s(\lambda)$ when exposed to a given pollutant. For the sake of simplicity let us consider the case of studying the response to two pollutants only. It is worth noticing that the method can be extended straightforwardly to a larger number of pollutants. We denote by $a(x)$ and $b(y)$ the responses of a biosensor to two different pollutants, where x and y are representing the pollutant concentrations. Throughout the paper, we will assume that these individual responses, referred to as calibration curves are available. In practice, they can be obtained through calibration experiments such as those reported in [1]. Suppose now that a solution is containing these two pollutants with respective concentrations x_0 and y_0. Following [1], the fluorescence is known to be proportional to the amount of bacteria denoted $c(z_0)$, where the variable z_0 is introduced for notational homogeneity and whose meaning will be explained in the sequel. Thus, the fluorescence signal emitted by the biosensor is proportional to $(a(x_0) + b(y_0))c(z_0)s(\lambda)$. Supposing that there are a number F of biosensor types marked with different fluorophores and responding differently to these two pollutants, the measured fluorescence signal can be written as:

$$\mathcal{D}(x_0, y_0, z_0, \lambda) = \sum_{f=1}^{F} (a_f(x_0) + b_f(y_0))c_f(z_0)s_f(\lambda). \quad (1)$$

In principle, the more biosensors are used, the better the estimation will be. In the simulation presented, we use two biosensors (F =2) which is the minimum possible value. In practice, this number is limited to $F \leq 4$. It should be noted that, as the model is additive with respect to the response of the biosensors to the two pollutants, building a multidimensional array by performing a metered addition of the two pollutants will not produce an array having a multilinear CP structure.To tackle this problem we propose to perform successive metered additions of a single pollutant combined with a controlled concentration variation of the different types of biosensors. By performing a metered addition of the first pollutant, we obtain a three-way data array model:

$$\mathcal{D}_1(x, z, \lambda) = \sum_{f=1}^{F} \alpha_f(x)c_f(z)s_f(\lambda) \quad (2)$$

where $\alpha_f(x) = a_f(x_0 + x) + b_f(y_0)$. The quantity x is representing the concentration variation of the first pollutant, starting from the initial concentration x_0. The variable z is representing the number of different biosensor concentrations considered in the experiment, z_0 indexing the first concentration. Similarly, for a metered addition of the second pollutant, we get:

$$\mathcal{D}_2(y, z, \lambda) = \sum_{f=1}^{F} \beta_f(y)c_f(z)s_f(\lambda) \quad (3)$$

with $\beta_f(y) = a_f(x_0) + b_f(y_0 + y)$. Clearly, equations (2) and (3) express a CP decomposition of datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ from which the functions $\alpha_f(x)$ and $\beta_f(y)$ can be recovered.

3. CP DECOMPOSITION OF THE DATA

Using tensor decomposition as a data analysis tool traces back to psychometrics [4, 5]. Let us mention in particular the work of Harshman [6] in the 70's who proposed the ALS algorithm for estimating the CP decomposition and who was the first to give an uniqueness result. The most general result regarding the CP decomposition uniqueness is due to Kruskal [7] who showed that uniqueness is achieved under mild conditions which are often met in practice. This is the reason explaining the growing interest of this decomposition in the fields of chemometrics [8] and signal processing [9]. Data mining and neuroscience applications are also reported in [10]. However, the idea of using the CP decomposition to analyze the response of bacterial biosensors, as proposed in [11], seems novel. The reason for considering three-way data rather than two-way data is mainly to uniqueness property of the CP decomposition. Indeed, as mentioned in [12], an ICA method applied tthese non negative data is likely to yield negative values. Similarly, the uniqueness of the NMF results cannot be ensured. On the contrary, introducing a third dimension together with the non negativity of the dat result is unique non diverging CP decomposition [13]. The CP decomposition of a three-way array can be written as:

$$\mathcal{D} = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \circ \mathbf{c}_f \circ \mathbf{s}_f = [\![\mathbf{A}|\mathbf{C}|\mathbf{S}]\!] \quad (4)$$

In (4), $\mathbf{A}$, $\mathbf{C}$ and $\mathbf{S}$ are the matrices formed by the column gathering of vectors $\boldsymbol{\alpha}_f$, $\mathbf{c}_f$, $\mathbf{s}_f$, $f = 1, \ldots, F$, respectively. Denoting by $k_{\mathbf{A}}$, the Kruskal-rank [2] of the matrix $\mathbf{A}$, the condition of Kruskal [7] states that the model is identifiable if $k_{\mathbf{A}} + k_{\mathbf{C}} + k_{\mathbf{S}} \geq 2F + 2$. Assuming that the biosensors are producing different fluorophores results in a matrix $\mathbf{S}$ having full column rank. Similarly, provided that the number of different biosensor concentrations is greater than the number of biosensors, $\mathbf{C}$ can be assumed to be full column rank. As a consequence, identifiability is achieved if $k_{\mathbf{A}} \geq 2$, that is, over the considered pollutant concentration range, the biosensor responses to a given pollutant are not collinear.

4. A MINIMISATION APPROACH TO CONCENTRATION ESTIMATION

After performing the CP decompositions of the two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, estimates of α_f and β_f are obtained. Ideally:

$$\hat{\alpha}_f(x) = a_f(x_0 + x) + b_f(y_0) \tag{5}$$

$$\hat{\beta}_f(y) = a_f(x_0) + b_f(y_0 + y) \tag{6}$$

where a_f and b_f correspond to the responses of the f-th biosensor to the two different pollutants. Let us recall that these responses are known through a calibration procedure. The estimation of the initial concentration x_0 and y_0 can be formulated as an optimization problem aiming at minimizing the following least-squares criterion:

$$\mathcal{J}(x_0, y_0) = \sum_f \Bigg(\sum_x (\hat{\alpha}_f(x) - k_{f,x}(a_f(x_0 + x) + b_f(y_0)))^2 + \sum_y (\hat{\beta}_f(y) - k_{f,y}(a_f(x_0) + b_f(y_0 + y)))^2 \Bigg). \tag{7}$$

The minimization of (7) can also be thought as a maximization of the intercorrelation between the estimated and calibration functions over intervals $x \in [x_0, x_0 + X]$ and $y \in [y_0, y_0 + Y]$, X an Y representing the maximum concentration values of the metered additions. In defining the criterion (7), it was necessary to introduce the weights $k_{f,x}$ and $k_{f,y}$ to take into account the fact that both the calibration and estimated curves are known up to a scale factor (because of the scale indeterminacy). However, when the values of x_0 and y_0 are fixed, the criterion is quadratic with respect to $k_{f,x}$ and $k_{f,y}$ and the values of $k_{f,x}$ and $k_{f,y}$ minimizing $\mathcal{J}$ admit an explicit expression depending on x_0 and y_0. Thus the dependence of $\mathcal{J}$ with respect to $k_{f,x}$ and $k_{f,y}$ can be removed. Let us now turn our attention to the minimization procedure of (7). As evidenced by numerical simulation in section 5 the criterion (7) is neither convex nor unimodal. This is what led us to adopt an exhaustive search approach over a grid, allowing to describe the possible variations of (x_0, y_0). Finally, it should be mentioned that some attention has to be paid to the practical implementation of such an approach, since it requires to interpolate the calibration functions for each possible value of (x_0, y_0) over grids which are compatible with the range of variations of $\hat{\alpha}_f(x)$ and $\hat{\beta}_f(y)$.

[2] The Kruskal-rank of a matrix $\mathbf{A}$ is the maximum number ℓ such that every ℓ columns of $\mathbf{A}$ are linearly independent.

5. RESULTS

For the experiments presented in this section, to compute the CP decomposition of the data, we used an optimized non-negative ALS algorithm which can be found in the Matlab N-way toolbox developed by Bro and Anderson [14]. This section presents the results of the proposed approach when applied to a numerical simulation aiming at reproducing the three-way data as presented in section 2. Figure 2 shows the simulated calibration curves resembling those presented in [15] for biosensors responding to Cadmium (Cd) and Zinc (Zn). These curves are defined on a high resolution grid. In practice, they can be obtained by determining the response of the biosensor for a limiting number of values of the pollutant concentration and then interpolating the response on a high resolution grid. Another possibility would consist in interpolating the calibration curve on a grid starting at x_0 and having the same sampling sequence as the estimated signal $\alpha(x)$. The synthetic data used in this example were corrupted by an additive Gaussian noise to reach a SNR of 20 dB. Figure 3 shows the results of the CP decomposition of the dataset $\mathcal{D}_1$. The first mode, shown on the first plot corresponds to the estimated values of $\alpha_1(x)$ and $\alpha_2(x)$. The other two modes correspond to the estimated biosensor concentration profiles and spectra. Similar results are obtained when decomposing $\mathcal{D}_2$ yielding estimates of $\beta_1(y)$ and $\beta_2(y)$. Figure 4 represents the criterion $\mathcal{J}(x_0, y_0)$, which has a number of local minima while the global minimum is located at coordinates $(196, 431)$ mMol, yielding the results of figure 5. The obtained results show a good estimation of the metal concentrations since the actual values are $(195, 430)$ mMol. Obviously, the quality of the estimation is influenced by the estimation accuracy of the α's and β's which are themselves depending on the noise level.

6. CONCLUSION

In this paper, an approach has been proposed for estimating the concentration of multiple pollutants by using non-specific biosensors. To overcome the problem coming from the additivity of the biosensor responses, we have proposed to use metered additions coupled with varying concentration mixtures of the biosensors. A three-way data array has to be generated for each pollutant. By performing the CP decomposition of the data coupled with a minimization procedure, it is possible to estimate the pollutant concentrations. In its present form, the algorithm assumes that the exact number and types of pollutants in the sample are known. Future works will be directed at studying how the procedure can be modified to

cope with the partial knowledge of the pollutant types in the sample. Currently, biosensors reacting to different environmental pollutants are under design. In a nearby future, it is expected to have operating biosensors available and to report the results of the data processing corresponding to real biosensor responding to environmental solicitations.

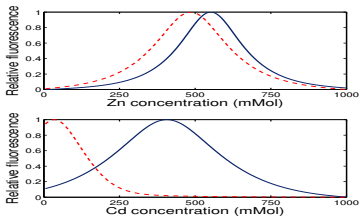

Fig. 2. Calibration curves

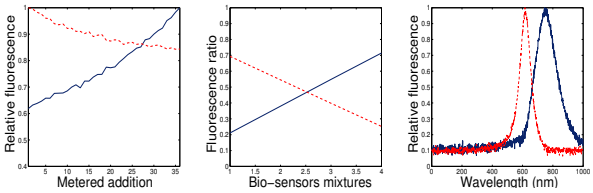

Fig. 3. Results of CP decomposition of the dataset $\mathcal{D}_1$

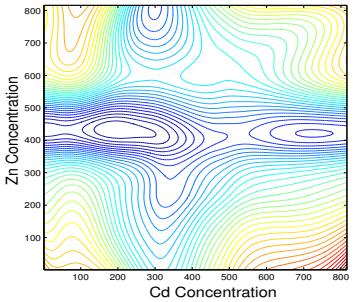

Fig. 4. Criterion $\mathcal{J}(x_0, y_0)$

7. REFERENCES

[1] R. Tecon and J. R. Van der Meer, "Information from single-cell bacterial biosensors: what is it good for?," *Current Opinion in Microbiology*, vol. 17, no. 1, pp. 4–10, Feb. 2006.

[2] E. Larrainzar, F. O'Gara, and John P. Morrissey, "Applications of autofluorescent proteins for *in situ* studies in microbial ecology," *Annual Review of Microbiology*, vol. 59, pp. 257–277, Oct. 2005.

[3] C. Mustin, "Hyperspectral analysis and enhanced surface probing of representative bacteria-mineral interaction," Tech. Rep., Programme Blanc, Agence Nationale de la Recherche (ANR), 2009.

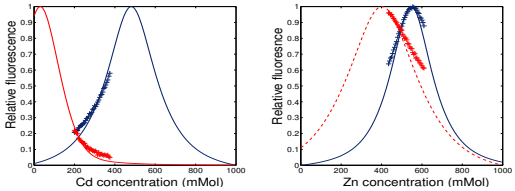

Fig. 5. Result of the concentration estimation

[4] R.B. Cattell, "Parallel proportional profiles and other principles for determining the choice of factors by rotation," *Psychometrika*, vol. 9, pp. 283, 1944.

[5] J.D. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, pp. 283–319, 1970.

[6] R.A. Harshman, "Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[7] J.B. Kruskal, "Three-way arrays: Rank and uniqueness of of trilinear decomposition, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, vol. 18, pp. 95–138, 1977.

[8] R. Bro, *Multi-way analysis in the food industry: Models, algorithms and applications*, Ph.D. thesis, University of Amsterdam, 1998.

[9] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, "Parallel factor analysis in sensor array processing," *IEEE Transactions on Signal Processing*, vol. 48, no. 8, pp. 2377–2388, 2000.

[10] Tamara G. Kolda and Brett W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[11] D. Brie, S. Miron, F. Caland, and C. Mustin, "An uniqueness condition for the 4-way Candecomp/Parafac model with collinear loadings in three modes," in *Proc. ICASSP*, 2011 (to appear).

[12] S. Moussaoui, D. Brie, A. Mohammad Djafari, and C. Carteret, "Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4133–4145, 2006.

[13] L. H. Lim and P Comon, "Nonnegative approximations of nonnegative tensors," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 432–441, 2009.

[14] C. A. Andersson and R. Bro, "The N-way toolbox for MATLAB," *Chemometrics and Intelligent Laboratory Systems*, vol. 52, pp. 1–4, 2000.

[15] K. B. Riether, M. A. Dollard, and P. Billard, "Assessment of heavy metal bioavailability using escherichia coli zntap::lux and copap::lux-based biosensors," *Appl. Microbiol. Biotechnol*, vol. 57, pp. 712–716, 2001.