Estimation de la réponse de biosenseurs par l'analyse des signaux spectraux multivariés

Fabrice CALAND[1, 2], Sebastian MIRON[1], David BRIE[1], Christian MUSTIN[2]

[1]CRAN, Nancy-Université,CNRS, Boulevard des Aiguillettes BP 70239
54506 Vandœuvre-lès-Nancy, France

[2]LIMOS, Nancy-Université, CNRS, Boulevard des Aiguillettes BP 70239
54506 Vandœuvre-lès-Nancy, France

fabrice.caland@cran.uhp-nancy.fr

Résumé – Un biosenseur est une bactérie génétiquement modifiée afin d'émettre un signal de fluorescence en présence d'un polluant. Dans ce papier, nous étudions la possibilité d'améliorer la précision des courbes de réponse de gènes dont l'expression dépend de la présence de cadmium métallique. La méthode est fondée sur un protocole expérimental permettant d'obtenir des tableaux tridimensionnels de données. L'extraction de la réponse du biosenseur en fonction de la concentration en polluants sera réalisée par Candecomp/Parafac. Une courbe de réponse du biosenseur plus précise est obtenue, du fait de la suppression de l'autofluorescence parasite.

Abstract – A biosensor is a bacterium that is genetically modified to produce fluorescent signals when exposed to environmental pollutants. In this paper, we investigate the possibility of enhancing the curves of expression of genes sensitive to cadmium.. To do that, we propose an experimental protocol to obtain three-way fluorescence data sets. The extraction of biosensor response to changes in environmental pollutant concentrations will be achieved by Candecomp/Parafac method. This approach provide a more accurate response curve of the biosensor to cadmium by eliminating spurious autofluorescence.

1 Introduction

Un biosenseur est composé d'un système biologique de détection, couplé à un système de transduction transformant l'événement détecté (par exemple, la présence d'un polluant) en un signal physique mesurable (luminescence, fluorescence) [14]. Dans notre étude, un gène rapporteur, codant la synthèse d'une protèine fluorescente, est fusionné avec l'élément génétique d'une cellule bactérienne reconnaissant une molécule ou un stress. La réponse luminescente de la bactérie génétiquement modifiée est détectée ensuite par un spectromètre. Les signaux de fluorescence mesurés sur une population de cellules dépendent de la quantité de polluant présent.

Les principaux avantages des biosenseurs sur les méthodes d'analyses chimiques et physiques sont les meilleures performances en terme de spécificité et de sensibilité. De plus, l'élément sensible étant un système biologique complet, les biosenseurs fournissent des informations complémentaires sur la biodisponibilité ou la toxicité de certains polluants dans l'environnement. Ces informations qualitatives et quantitatives sur les polluants sont utiles pour l'évaluation des risques ou le suivi de la décontamination (eaux,sols).

Aujourd'hui, une des grandes limitations dans l'utilisation des biosenseurs est le manque de précision de la mesure spectrale. L'étalement des spectres fluorescents occasionnent un recouvrement partiel des signaux qui vient détériorer la mesure. De plus, l'existence de fluorescences parasites (autofluorescence spontanée), limite l'exploitation et l'interprétation des données spectrales dans des systèmes naturels ou des Systèmes Minéraux-Bactérie (SMB) complexes. Les parades adoptées consistent à limiter l'autofluorescence au maximum ou à n'utiliser que des systèmes rapporteurs spectralement distincts.

L'approche que nous proposons dans le projet HÆSPRI [1] [13] est différente puisqu'elle consiste à exploiter la diversité des réponses des biosenseurs afin d'extraire, à l'aide d'une méthode de séparation de sources, les signaux effectifs liés à l'expression des systémes rapporteurs. Dans cette communication nous proposons une méthode, fondée sur une factorisation de type Candecomp/Parafac (CP) des tableaux tridimensionnels de données de fluorescence, permettant d'estimer les réponses des biosenseurs aux divers polluants et leur signaux d'autofluorescence. L'autofluorescence est particuliérement génante lorsque la concentration du polluant à détecter est faible, puisque le signal d'autofluorescence spontanée risque de masquer complétement la réponse du biosenseur.

Le modèle Candecomp/Parafac a été introduit de manière indépendante par Carroll et Chang [5] et Harshman [8] dans les années '70. Ils ont également proposé les premiers algorithmes de décomposition et les premières applications respectivement en psychométrie et phonétique. Dans les deux dernières décen-

1. Ce travail a bénéficié du support financier de l'ANR HÆSPRI (ANR-09-BLAN-0336-04).

nies, la décomposition CP, grâce à sa polyvalence et à ses intéressantes propriétés d'unicité, a connu un succès important dans des domaines variés tels que la chimiométrie, les télécommunications, le traitement d'antenne ou l'imagerie médicale. Pour une revue plus complète des différentes applications du CP le lecteur est renvoyé à [1, 11].

2 Modélisation de la réponse multivariée des biosenseurs

2.1 Préliminaires

Avant d'introduire le modèle CP des données de fluorescence nous présentons brièvement quelques notions de base d'algèbre trilinéaire.

Nous appellerons tenseur d'ordre trois tout tableau de données à trois dimensions. Un tenseur d'ordre trois est de rang un s'il peut s'écrire sous la forme du produit tensoriel de 3 vecteurs : $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$. La décomposition CP de rang F d'un tableau tridimensionnel de données $\mathcal{X}$ consiste alors à trouver F tenseurs de rang un dont la somme approche au mieux les données, *i.e.*

$$\mathcal{X} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \tag{1}$$

L'équation (1) peut s'écrire de façon équivalente en utilisant la notation indicielle :

$$x_{i,j,k} = \sum_{f=1}^{F} a_{i,f} b_{j,f} c_{k,f} \tag{2}$$

où $x_{i,j,k}$ est l'élément de $\mathcal{X}$ situé à la place (i, j, k).

L'avantage majeur des décompositions trilinéaires par rapport aux décompositions bilinéaires classiques, est leur unicité sous des faibles contraintes. Pour les modéles trilinéaires, la condition d'unicité la plus connue est due à Kruskal [12] et garantit que les 3 matrices, regroupant respectivement les vecteurs $\mathbf{a}_f, \mathbf{b}_f$ et $\mathbf{c}_f$, peuvent être déterminées de façon unique à partir des données $\mathcal{X}$ si la somme de leurs *rangs de Kruskal*[2] est strictement supérieure à $2F + 1$. Cette condition est vérifiée dans la plus part des applications pratiques. Dans les cas où la condition de Kruskal n'est pas respectée, il existe des conditions plus faible d'unicité garantissant que des sous-familles de vecteurs peuvent être estimées de façon unique [7].

2.2 Le modèle des données de fluorescence

Afin d'introduire le modèle mathématique des données, nous étudierons dans la suite la réponse des biosenseurs et de leurs signaux d'autofluorescence en fonction de trois paramètres/diversités : longueur d'onde, concentration du polluant et temps. Au niveau des données, la réponse du biosenseur au polluant et l'autofluorescence des bactéries représentent deux sources différentes. La méthode est fondée sur l'hypothése que les sources présentent des comportements différents par rapport aux paramétres étudiés. En effet, la réponse du biosenseur au polluant dépend principalement de la protéine fluorescente utilisée alors que l'autofluorescence est essentiellement liée au type de bactérie et dépend de nombreux paramètres (phase de croissance, type de bactérie, etc.).

Afin de donner un caractére plus général à cette présentation, nous considérerons par la suite qu'un nombre F de sources est présent dans le mélange. On notera $s_f(\lambda)$ la signature spectrale de la $f$iéme source, $a_f(x)$ sa réponse à la concentration x en polluant et $c_f(t)$ l'évolution temporelle de son signal de fluorescence. Le signal de fluorescence des F sources peut s'exprimer alors de la façon suivante

$$\mathcal{D}(x, t, \lambda) = \sum_{f=1}^{F} a_f(x) c_f(t) s_f(\lambda). \tag{3}$$

L'équation (3) exprime clairement un modéle trilinéaire du mélange, de type (CP) dans lequel, le temps, en mixant les vitesses de croissance bactérienne et de maturation des différentes protéines [14], apportera la diversité nécessaire pour identifier les différentes sources et leur réponses aux polluants. En considérant N_x valeurs de concentration, N_t instants de temps et N_λ valeurs spectrales, le modéle de mélange donné par (3) peut s'écrire sous la forme

$$\mathcal{D} = [\![\mathbf{A}|\mathbf{C}|\mathbf{S}]\!] \tag{4}$$

où les matrices $\mathbf{A}$, $\mathbf{C}$ et $\mathbf{S}$, de dimensions respectives $(N_x \times F)$, $(N_x \times F)$ et $(N_x \times F)$, contiennent sur leur colonnes, les variations des F sources suivant les trois modes/diversités.

Une des techniques les plus utilisées pour estimer les trois matrices du modèle CP est la méthode des moindre carrés alternés (*Alternating Least Squares* ou *ALS* en anglais), qui consiste à estimer de manière itérative une matrice en fixant les deux autres. Une version améliorée de cet algorithme, permettant d'imposer différents types de contraintes sur les matrices à estimer, peut être trouvée dans la *N-Way Toolbox* developpée par Anderson et Bro [2].

Dans la section suivante nous donnons un exemple de traitement de données réelles, permettant de séparer le signal d'autofluorescence de la réponse des biosenseurs.

3 Résultats

L'expérience réalisée consiste à suivre l'évolution temporelle (4 pas de temps) des spectres d'émission de fluorescence (80 points) mesurés dans les 8 puits d'une microplaque contenant une quantité fixe de biosenseurs. Dans chaque puit, on ajoute des quantités variables de polluant (ici un métal toxique, le cadmium Cd dont les concentrations varient entre 0 et 1mMol). Le biosenseur a été élaboré pour une production de fluorescence (GFP) dose-dépendente, via la fusion d'un gène rapporteur au gène $PPcadA_2$ codant pour une pompe de

2. Le *rang de Kruskal* d'une matrice est le nombre maximum k tel que toute sous-ensemble de k colonnes de cette matrice forme une famille libre de vecteurs.

flux membranaire et inductible par le cadmium. Les spectres de fluorescence ont été acquis à l'aide d'un spectrofluorimètre FLX-Xenius®SAFAS. Les spectres sont relevés de 440 à 600nm avec un pas de 2nm pour une excitation à 390nm. On obtient ainsi un tableau tridimensionnel de données, de dimension $80 \times 8 \times 4$, dont le premier mode est la longueur d'onde du spectre de fluorescence, le second mode la concentration de Cd ajouté et le troisième mode le temps (Fig. 1).

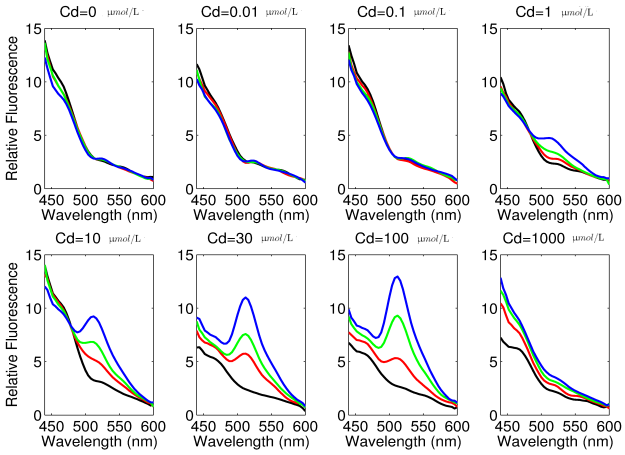

FIGURE 1: Données d'émission de fluorescence. Chaque courbe de chacun des graphiques représente le spectre de fluorescence émis dans un puit à un instant t.

La Fig. 2 montre le résultat de la décomposition CP des données. On peut y voir l'évolution de l'intensité d'une source spectrale (Fig. 2a) en fonction de la concentration en Cd (Fig. 2b) et du temps (Fig. 2c). La décomposition montre l'existence d'une source spectrale S1 dont le maximum d'émission se situe à 440nm et dont l'intensité croit au cours du temps. Cette source peut être associée à l'autofluorescence des bactéries, car son évolution au cours du temps suit la densité bactérienne, qui peut être estimée par la densité optique (absorbance à 600nm). Le biosenseur est représenté par son spectre estimé S2, conforme à celui attendu pour une GFP. La concentration en GFP dans la bactérie augmente avec la concentration de Cd jusqu'au seuil toxique pour la bactérie (entre 0,1 et 1mMol, Fig. 2b). Ces deux phénomènes sont soulignés par la courbe de réponse *en cloche* du gène rapporteur (GFP), estimée par la décomposition CP (en rouge sur la figure 2b). Sur la figure 2c, on peut voir l'évolution croissante de la fluorescence des sources S_1 et S_2 en fonction du temps. Ceci s'explique par la croissance cellulaire au cours du temps. Elle se manifeste par un accroissement du volume cellulaire qui a pour effet d'augmenter la quantité de fluorochromes dans un même volume d'excitation. On observe par conséquent une augmentation de la fluorescence relative de chacun des fluorochromes au cours du temps. L'augmentation plus rapide du fluorochrome GFP s'explique par un phénomène d'accumulation du fluorochrome à l'intérieur de la bactérie qui vient s'ajouter à l'augmentation du nombre de bactéries.

La qualité de ces résultats a été validée en comparant les spectres et réponses estimés aux spectres de fluorescence réalisés *ex vivo* [15] et la courbe d'expression du gène $PPcadA_2$ [3].

La figure 3 montre l'évolution, en fonction de la concentration en Cd, de la fluorescence associée à la GFP, par une méthode standard [10] (ligne pointillée) et la décomposition CP (trait plein). La méthode standard consiste, pour la mesure de l'expression de la GFP, à une mesure de la fluorescence émise à $\lambda = 515\text{nm} \pm 10\text{nm}$ après une excitation à 490nm et à un temps donné (15h après mise en contact avec le polluant). On observe une différence significative de l'estimation pour de faibles concentrations de Cd. Cette différence s'explique par le fait que pour les faibles concentrations de cadmium, la fluorescence de la GFP est minime et qu'elle est cachée par l'autofluorescence de la bactérie. La méthode standard, ignorant l'existence d'autofluorescence, surestime la fluorescence de la GFP. Au contraire, la méthode CP extrait chaque composante et ne fournit que les variations de fluorescence du rapporteur en fonction de la teneur en Cd.

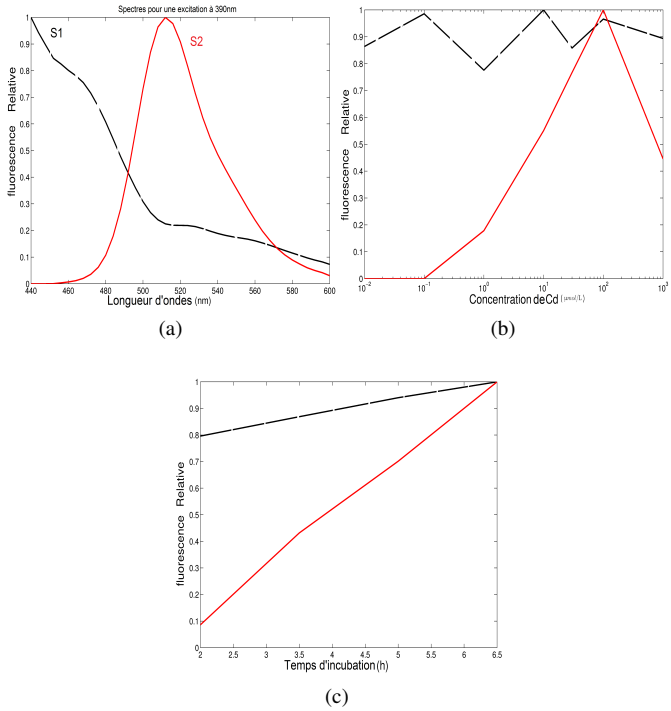

FIGURE 2: Résultats de la décomposition CP des données. Source S_1 (autofluorescence)-Source S_2 (GFP). (a) Spectres de fluorescence de S1 et S2. (b) Réponse en fonction du [Cd] de S_1 et S_2. (c) Évolution en fonction du temps.

4 Discussion

Les modèles de décomposition donnent des résultats qui concordent avec le comportement attendu du biosenseur. Dans le cadre de cette expérience, les paramètres temps et concen-

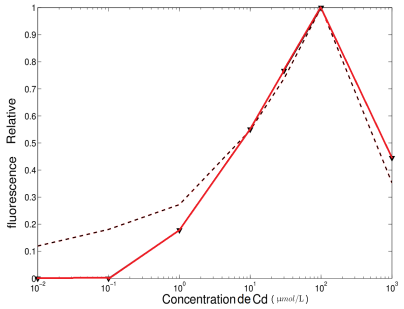

FIGURE 3: Courbe de réponse du biosenseur fluorescent. Mesure à une seule longueur d'onde (en pointillés). Estimation par décomposition trilinéaire (en trait plein).

tration de Cd ont été choisis car ils sont facilement mesurables, toutefois on peut regretter le peu de diversité de comportements qu'ils induisent sur les fluorochromes. Par exemple, l'augmentation du nombre de bactéries au cours du temps induit la même augmentation de fluorescence pour chaque fluorochrome et l'augmentation de la concentration de cadmium n'agit pas sur la première source. Le biosenseur utilisé est particulièrement adapté à la méthode utilisée actuellement par les microbiologistes. La méthode actuelle consiste à choisir des fluorochromes avec des spectres d'émission le plus éloigné possible et d'effectuer des acquisitions à la longueur d'ondes maximum de chacun des fluorochromes. Ce qui permet d'extraire la réponse de chacunes des composantes mais limite le nombre de fluorochromes dans un biosenseur. Dans le cadre de notre approche, le recouvrement spectral n'est pas un problème. Grâce à la séparation des sources, il est possible de considérer un plus grand nombre de fluorochromes et ainsi augmenter le nombre de paramètres étudiés.

5 Conclusion

Les résultats présentés dans cette communication montrent que l'utilisation de plusieurs diversités dans les spectres d'émission des protéines fluorescentes permet de s'affranchir des difficultés majeures (recouvrement spectral, autofluorescence, temps de mesure) limitant l'utilisation des biosenseurs. Des résultats allant dans ce sens ont été obtenu sur d'autres données acquissent sur d'autres biosenseurs et colorants fluorescents. L'amélioration de la qualité des courbes de comportement des biosenseurs, en éliminant la fluorescence propre à la bactérie, ouvre la perspective d'utiliser les biosenseurs en milieu complexe. La possibilité de séparer la contribution de chacun des biosenseurs permet d'envisager la construction de nouveaux biosenseurs pour identifier et quantifier des métaux de transition présents dans un échantillon liquide.

Références

[1] E. Acar et B. Yener, *Unsupervised Multiway Data Analysis : A Literature Survey*. IEEE Transactions on Knowledge and Data Engineering, 2009.

[2] C.A Andersson et R. Bro. *The N-way Toolbox for MATLAB*. Chemometrics and Intelligent Laboratory Systems, 2000.

[3] P. Billard, C. Mustin, T. Béguiristain et C. Leyval. *Approche innovante pour l'étude de la disponibilité des éléments métalliques dans les sols*. Rapport BQR UHP-Région Lorraine, 2006.

[4] R. Bro. *Multi-way analysis in the food industry : Models, algorithms and applications*. PhD Thesis, 1998.

[5] J.D. Caroll et J. Chang. *Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition*. Psychometrika, 1970.

[6] R.B. Cattell. *Parallel proportional profiles and other principles for determining the choice of factors by rotation*. Psychometrika, 1944.

[7] X. Guo, S. Miron, D. Brie et A. Stegeman *Identifiabilité partielle de mélanges trilinéaires de sources linéairement dépendantes*. Proc. GRETSI 2011, Bordeaux, France, Sept. 5-8.

[8] R.A. Harshman. *Foundations of the parafac procedure : Models and conditions for an explanatory multi-modal factor analysis*. UCLA Working Papers in Phonetics, 1970.

[9] F.L. Hitchcock. *The expression of a tensor or a polyadic as a sum of products*. Journal of Mathematics and Physics, 1927.

[10] H. Huot. *Utilisation de biosenseurs bactériens fluorescents pour évaluer la biodisponibilité des éléments métalliques dans les sols*. Rapport Master FAGE Parcours SGEUI, 2009.

[11] T.G. Kolda et B.W. Bader. *Tensor Decompositions and Applications*. SIAM Reviews, 2009.

[12] J.B. Kruskal. *Three-way arrays : Rank and uniqueness of trilinear decomposition, with application to arithmetric complexity and statistics*. Linear algebra and its applications, 1977.

[13] C. Mustin. *Hyperspectral analysis and enhanced surface probing of representative bacteria-mineral interaction.*. Technical report, Programme Blanc, Agence Nationale de la Recherche (ANR), 2009.

[14] R. Tecon et J.R. van der Meer. *Bacterial biosensors for measuring availability of environmental pollutants*. Sensors, 2008.

[15] R.Y. Tsien. *The Green Fluorescent Protein*. Annual Reviews Biochemistry, 1998.