Décomposition tensorielle de signaux luminescents émis par des biosenseurs bactériens pour l'identification de Systèmes Métaux-Bactéries

Fabrice Caland

École doctorale Ressources Procédés Produits Environnement (RP2E)
Collégium Sciences et technologies

Décomposition tensorielle de signaux luminescents émis par des biosenseurs bactériens pour l'identification de Systèmes Métaux-Bactéries

THÈSE

présentée et soutenue publiquement le 17 septembre 2013
pour l'obtention du

Doctorat de l'Université de Lorraine

spécialité Géosciences

par

Fabrice CALAND

Composition du jury

Rapporteurs :	Nadège Thirion-Moreau, Professeur des Universités, Université du Sud Toulon-Var Gérald Thouand, Professeur des Universités, Université de Nantes
Examinateurs :	Bernard Humbert, Professeur des Universités, Université de Nantes Valeriu Vrabie, Maître de Conférences, Université de Reims Champagne-Ardennes
Directeurs :	Christian Mustin, Directeur de Recherche, CNRS, Université de Lorraine David Brie, Professeur des Universités, Université Lorraine Sebastian Miron, Maître de conférences, Université Lorraine

Laboratoire Interdisciplinaire des Environnements Continentaux UMR 7360
Centre de Recherche en Automatique de Nancy UMR 7039
Campus Aiguillettes - 54506 Vandœuvre Lès Nancy

Mis en page avec la classe thloria.

Remerciements

Je tiens tout d'abord à remercier mes directeurs de thèse, M. Mustin Christian, M. Brie David et M. Miron Sebastian pour leur encadrement et pour m'avoir donné l'occasion de travailler sur cette thèse. Ainsi que Mme Corinne Leyval (directrice du LIMOS/LIEC) et M. Alain Richard (directeur du CRAN) pour m'avoir accueilli au sein de leurs laboratoires.

Je remercie M. Jean-Christophe Ponsart pour l'expérience d'enseignement, ainsi que mes collègues et mes étudiants de la licence Sciences Pour l'Ingénieur (EEAPR) de la Faculté des Sciences de Nancy.

Je remercie les rapporteurs de ma thèse Mme Nadège Thirion-Moreau et M. Gérald Thouand, ainsi que les examinateurs M. Bernard Humbert et M. Valeriu Vrabie, pour avoir accepté d'évaluer mon travail ainsi que pour leurs remarques qui m'ont permis d'améliorer la qualité de ce manuscrit.

Merci ensuite à mes amis et mes collègues du LIMOS et du CRAN pour leur soutien, leur bonne humeur et pour tout ce que nous avons partagé au cours de ces années de thèse. Je tiens à remercier plus particulièrement ma deuxième famille Thierry, Anne-Marie, Mathéo et leur autre fils adoptif Olivier pour leur accueil, les défis sportifs que nous avons réalisés et les séances de récupération qui ont suivi en espérant que de nombreuses autres suiveront encore, JB et Eric qui je l'espère ne regréterons bientôt plus d'avoir suivi mes conseils et pour finir Fabien, Romain, Alexis, Wahiba pour avoir été présents.

Enfin, je remercie ma famille, en particulier Stéphanie et mes parents pour ce qu'ils m'ont apporté et ce qu'ils m'apportent encore.

Table des matières

Table des figures

Notations

a, A	scalaire
$\mathbf{a}$	vecteur
$\mathbf{A}$	matrice
$\boldsymbol{\mathcal{X}}$	tableau de données de dimension supérieure à trois (tenseur)
$[\![.]\!]$	Notation alternative d'un tenseur
$\mathbf{X}_{(k)}$	matrice du tenseur $\boldsymbol{\mathcal{X}}$ déplié suivant le mode k
$\mathbf{X}_k$	k^e tranche du tenseur$\boldsymbol{\mathcal{X}}$
$\otimes$	Produit de Kronecker
$\odot$	Produit de Khatri-Rao
$\circ$	Produit tensoriel
$\dagger$	Pseudo inverse
CP	Candecomp/Parafac
DCP	Décomposition Candecomp/Parafac
promoteur :: *rapporteur*	Gène rapporteur induit par un gène promoteur
$\#\mathcal{B}_{2G}$	Quantité du biosenseur $\mathcal{B}_{2G}$

Abréviations

ACP	Analyse en Composantes Principales
ALS	Alternating Least Squares
BPSS	Bayesian positive source separation
CFP	Cyan Fluorescent Protein
CorConDia	Core Consistency Diagnostic
CP	Candecomp/Parafac
DCAA	Milieu de culture carencé en fer
DsRed	Discosoma Red (protéine fluorescente rouge)
EYFP	Enhanced Yellow Fluorescent Protein
GFP	Green Fluorescent Protein
$ICP - MS$	Inductively coupled plasma mass spectrometry
LASSO	Least Absolute Shrinkage and Selection Operator
LB	Milieu de culture Lysogeny broth
mCherry	Protéine fluorescente rouge
MEEF	Matrice d'émission-excitation de fluorescence
NMF	Non-negative matrix factorization
*OG*514	Oregon Green 514
PARALIND	Parallel profiles with linear dependencies
ppb	part per billion
PPcadA	cadmium-translocating P-type ATPase
R6G	Rhodamine 6G
RB	Rhodamnie B
RPS	Résonance des Plasmons de Surface
SVD	Singular Value Decomposition

Chapitre 1

Biosenseurs et spectrométrie de fluorescence

Sommaire

1.1 Introduction

Afin de satisfaire des demandes croissantes en matières premières et en ressources énergétiques, les activités humaines ont occasionné une perte substantielle et parfois irréversible de la qualité des sols et des milieux aquatiques. Depuis la fin du XIX^{e} siècle, les développements économiques et technologiques ont engendré, au fil du temps, de profondes modifications physiques, chimiques et biologiques entraînant une détérioration de la qualité écologique de ces milieux et perturbant leur biocénose [1]. Par ignorance sur les conséquences environnementales des techniques d'exploitation (*e.g.* extraction minière, procédés sidérurgiques, *etc.*) ou de l'usage régulier de produits chimiques nocifs, ces activités humaines ont conduit à l'implantation progressive d'écosystèmes fortement pollués et perturbés. Aujourd'hui, rares sont les milieux exempts de polluants [56]. Qu'il s'agisse de rejets industriels ou miniers, d'agriculture intensive ou d'activités domestiques, ces sources de contaminations ont amené dans tous les milieux des métaux, des pesticides ou des fertilisants et divers polluants organiques [50].

Le déclin industriel de ces dernières décennies a mis à jour l'existence de millions d'hectares de friches industrielles et de milliers de kilomètres de cours d'eaux pollués. En Europe, 3,5 millions de sites dégradés sont recensés dont 500 000 sont contaminés par divers polluants. Avec 6000 ha de sols contaminés (recensés en 1995), soit 9% des sites et sols pollués français, la région Lorraine porte encore les traces de son lourd passé industriel. En particulier, les substances et produits ne se dégradant ou ne se volatilisant pas naturellement, comme les métaux, se sont accumulés dans les sols, les eaux souterraines et les sédiments de rivières.

Omniprésents dans l'environnement, les métaux lourds (*e.g.* plomb, chrome, mercure, cadmium, cuivre, *etc.*) sont actuellement la cause des plus graves problèmes de pollution et de toxicité [13]. Même à de faibles concentrations, ils sont une menace pour l'environnement et la santé humaine, car le taux d'exposition des personnes ne cesse de croître [101]. En revanche, d'autres métaux (Fe, Co, Zn, Mo, *etc.*) sont indispensables aux fonctionnements des organismes vivants et des écosystèmes. Par exemple, le fer est un oligo-élément essentiel, dont l'absence ou l'excès peuvent avoir des conséquences néfastes sur la physiologie des organismes ou la fertilité des sols. Son statut chimique (spéciation) constitue un excellent indicateur des processus d'oxydo-réduction chimique ou biologique dans les sols [36]. En particulier, la précipitation ou la dissolution d'oxydes de fer sont considérés comme des processus majeurs intervenant dans la dégradation de la matière organique et la mobilisation ou immobilisation des éléments essentiels ou des contaminants organiques ou inorganiques présents dans les sols [57].

Inverser la tendance de dégradation des écosystèmes sans nuire à leur usage futur n'est pas une chose facile et implique des changements significatifs aux niveaux scientifique et métrologique. La mise en place de stratégies de gestion des écosystèmes visant le maintien ou la restauration de leur qualité impose le développement de systèmes d'observation de l'état des milieux et le développement de dispositifs de détection efficaces et robustes des polluants. L'apparition

1. Communauté d'espèces animales ou végétales vivant dans un milieu donné (biotope).

dans l'environnement d'un nombre croissant de produits potentiellement nocifs (métaux, nanoparticules, perturbateurs endocriniens, *etc.*) et le déploiement intense de réseaux de mesures et de systèmes de surveillance exigent des techniques d'analyses rapides, peu coûteuses et efficaces [9, 10, 24]. Or, les exigences techniques pour l'application des méthodes traditionnelles d'analyses des métaux constituent souvent un obstacle important à leur déploiement *in situ*.

Dans ce contexte et sur le plan métrologique, l'usage de capteurs environnementaux robustes et pertinents constitue un levier essentiel pour la mise en œuvre de stratégies de reconstruction ou de préservation des ressources naturelles. En particulier, en parallèle des capteurs conventionnels et classiques (sondes pH, conductimétrique, hygrométrique, *etc.*), des travaux visent au développement de senseurs biologiques ou de bio-essais capables d'identifier et doser, voire localiser, directement les métaux *in situ* [73]. Par exemple, des biosenseurs ont été mis au point pour la détection de l'arsenic dans l'eau potable, dans une gamme de concentration (10 à 50 µg/L) compatible avec le seuil de toxicité de l'élément [89]. Cette technique utilise un biosenseur bactérien, dont le principe repose sur un mécanisme de résistance et de détoxification de l'arsenic(III) connus chez les bactéries. L'élément sensible du biosenseur est composé de cellules bactériennes génétiquement modifiées afin de produire une coloration bleue en présence de faibles concentrations d'arsenic. Quelques gouttes de cultures bactériennes sont déposées sur des bandelettes, semblables aux papiers indicateurs utilisés pour mesurer le pH. Il suffit de plonger ensuite la bandelette dans quelques millilitres d'échantillon d'eau et d'évaluer l'intensité de la coloration au bout d'une demi-heure. La compacité et la portabilité de ce biosenseur s'avère compatible avec des analyses de terrain. D'après les concepteurs, ce test, peu onéreux, est suffisamment sensible pour mesurer les variations saisonnières de la concentration en arsenic dans des stations de distribution d'eau comme les millions de puits d'eau potable à usage privé ou collectif répartis sur le territoire du Bangladesh [44]. Il remplace avantageusement les techniques traditionnelles et portatives utilisant des indicateurs colorés qui s'avèrent inefficaces pour des concentrations en arsenic inférieures à 100 µg/L.

Le développement de ce type de méthodes de diagnostic fondées sur des indicateurs simples ou des capteurs *in situ* de la qualité physique et chimique des milieux autorisant des mesures spatiale et temporelle quasi-continues est donc essentiel pour développer des systèmes de surveillance adaptés aux problématiques environnementales actuelles.

1.2 Analyse environnementale et méthodes de détection *in situ* des métaux

Classiquement, la détection et l'identification des métaux est réalisée *ex situ* (en laboratoire), sur des solutions, par des analyseurs couplant la spectromètrie de masse à une génération de plasma (ICP-MS). Ces instruments sont très sensibles : plusieurs métaux peuvent être identifiés simultanément et dosés avec des limites de détection largement inférieures aux ppb (*i.e* µg/L), soit des concentrations inférieures à une dizaine de nanomoles (nM) pour la plupart des métaux.

En revanche, ces techniques sont lourdes à mettre en œuvre et présentent un coût élevé. Les échantillons liquides doivent être filtrés et acidifiés, les échantillons solides doivent être dissouts par fusion alcaline à haute température ou digestion thermique sous pression en milieu oxydant [103]. Certains éléments (Hg, As, Se, *etc.*), pour être dosés correctement et avec sensibilité, nécessitent des techniques de préparation spécifiques (génération d'hydrure) et une certaine expertise [9]. L'utilisation d'une ICP-MS *in situ* nécessiterait le transport de bouteilles d'argon, du matériel de préparation des échantillons et d'une source d'énergie. A l'inverse des techniques plus « portables » utilisant la spectrométrie de fluorescence X, la complexométrie (indicateurs colorés) et la spectrométrie d'absorbance, ne sont pas aussi précises que l'ICP-MS en raison des interférences liés aux effets de matrice (*i.e.* effets du milieu contenant l'élément à doser).

Ces analyses physico-chimiques peuvent évidemment indiquer avec précision les quantités totales de métaux toxiques présents dans des échantillons d'eau, de sols ou de sédiments mais elles ne donnent pas d'indication directe sur leur spéciation (*i.e.* leur statut chimique) ou leur biodisponibilité. Il est alors très difficile, à partir de ces analyses totales, de déterminer les effets sur les écosystèmes que peuvent engendrer des combinaisons de plusieurs métaux [62]. Ce besoin récurrent de systèmes et outils pour des applications environnementales, en particulier en écotoxicologie, a motivé le développement de biotechnologies et méthodologies plus appropriées, comme les bio-essais ou les biosenseurs simples, utilisés pour la surveillance des concentrations de métaux dans l'environnement et l'évaluation de leur toxicité [33, 65, 103].

Très employé en écotoxicologie, les bio-essais et bio-tests permettent d'évaluer de façon qualitative ou quantitative l'effet toxique d'un échantillon sur divers micro-organismes et organismes vivants : bactéries, algues, plantes, mollusques, insectes ou poissons. En intégrant les effets multiples (antagonistes, synergiques) de l'ensemble des contaminants présents dans l'environnement, ces batteries de tests fournissent un des indicateurs de la qualité biologique des milieux cibles [69]. Ce sont des outils complémentaires des analyses physico-chimiques classiques, dans l'évaluation de la toxicité d'un milieu et la gestion des risques associés [102]. De même, les biosenseurs microbiens, comme celui à l'arsenic présenté précédemment, apparaissent comme une alternative intéressante pour le dosage des métaux dans les milieux aquatiques. Les principaux avantages des biosenseurs sur les méthodes d'analyses chimiques et physiques sont la dynamique de leur réponse, la portabilité, la miniaturisation et la sensibilité qui permet la détection de polluants y compris dans des milieux complexes ou en présence de substances interférentes. Cependant, la bonne utilisation de ce type d'outils nécessite une meilleure compréhension de la dynamique des interactions entre le biosenseur et son environnement. Leur application aux milieux continentaux passe par la modélisation de la réponse fonctionnelle des systèmes bactériens et par l'étude de la robutesse de la mesure dans des milieux poreux plus organisés comme les sols, en tenant compte de l'accessibilité des polluants aux biosenseurs ou l'hétérogénéité des échantillons.

Néanmoins, il est difficilement concevable qu'un simple biosenseur soit sensible à tous les métaux potentiellement présents dans un échantillon. De plus, il est peu probable de trouver des biosenseurs spécifiques pour tous les métaux potentiellement dangereux et compatibles avec leur

limite de toxicité. Le choix arbitraire de biosenseurs spécifiques d'un seul contaminant métallique peut conduire, comme pour les analyses physico-chimiques, à une mauvaise appréciation de la qualité du milieu. À l'opposé, un senseur sensible à une variété de contaminants permettra de vérifier la toxicité ou la salubrité d'un environnement sans identifier nécessairement les contaminants impliqués. C'est pourquoi l'usage d'une série de capteurs ou d'indicateurs est souvent recommandé afin d'augmenter la robustesse de la mesure et de l'interprétation [80, 35, 70]. Ces études servent ensuite à la constitution de guides d'interprétations des indications qualitatives et des mesures quantitatives afin d'établir un diagnostic.

1.3 Biosenseurs

Les biosenseurs sont des dispositifs analytiques souvent miniaturisés utilisant pour la détection un système sensible d'origine biologique, doué de capacité de reconnaissance moléculaire. L'élément biologique immobilisé, sensibles aux composés ou molécules d'intérêt, est en contact intime avec un transducteur qui convertit un signal biochimique en un signal physique mesurable (électrique, optique, magnétique). Les biosenseurs sont présents dans de nombreux domaines (agriculture, médecine, *etc.*) et applications allant des contrôles qualité de procédés biotechnologiques au monitoring environnemental, en passant par le diagnostic thérapeutique. Ils peuvent être considérés comme un sous-groupe des capteurs chimiques dans lequel le système de détection utilise un mécanisme biochimique. Ils sont utilisés dans différents domaines (médical, militaire, agronomique) et applications, comme le contrôle de procédés biologiques ou de la qualité des aliments. Un biosenseur est composé d'un système biologique de détection, couplé à un système de transduction transformant l'évènement détecté en un signal physique mesurable (luminescence, fluorescence) [93]. Les biosenseurs sont habituellement classés en fonction du système de détection ou selon les transducteurs physico-chimiques utilisés : électrochimique, optique, piézoélectrique ou thermique (fig. 1.1). Les micro-organismes et les enzymes constituent les éléments sensibles principalement utilisés dans les biosenseurs dédiés aux mesures environnementales et surtout à la détection des métaux. Les transducteurs électrochimiques sont le plus souvent utilisés car ils offrent des perspectives attrayantes en terme de portabilité, d'interfaçage et de facilité d'utilisation, même si aujourd'hui grâce aux progrès récents en opto-électronique, les détections optiques sont en plein essor.

1.3.1 Biosenseurs moléculaires

Les capteurs biochimiques sont couramment utilisés dans la reconnaissance de molécules ou d'ions d'intérêt en solution et l'estimation de leur concentration. Leur principe repose sur la détection d'une réaction biochimique entre le ligand et un récepteur protéique ou enzymatique. En effet, certains métaux se lient ou interagissent spécifiquement à des protéines et enzymes. Le senseur réagit aux interactions récepteur-métal et produit un signal électrique, généralement proportionnel à la quantité de métal fixé. Par exemple l'activité catalytique de certaines enzymes

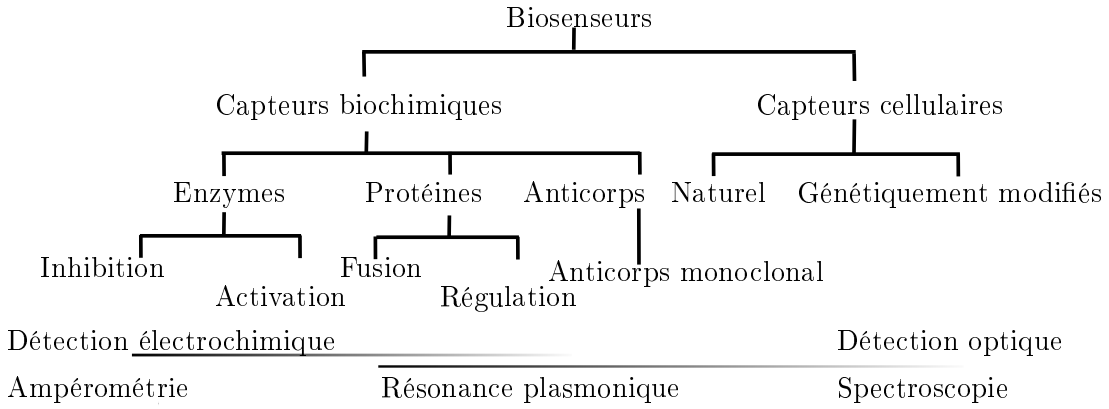


FIGURE *1.1 – Classification des biosenseurs utilisés en fonction de l'élément sensible utilisé.*

est inhibée de façon très sélective par la présence à faibles concentrations d'ions métalliques [3]. Ainsi, Domínguez-Renedo et *al.* [31] ont développé des biocapteurs ampérométriques enzymatiques pour la mesure du mercure. Durrieu et Tran-Minh [34] ont développé un biocapteur optique pour détecter le plomb et le cadmium par inhibition de la phosphatase alcaline présente dans la paroi d'une microalgue. Ces biosenseurs développés à base d'algues [2] ont été aussi validés pour déterminer la présence du cuivre (II) dans des échantillons aqueux et les concentrations estimées coïncidaient avec celles obtenues par spectrophotométrie d'absorption atomique. Aujourd'hui les recherches s'orientent vers de nouveaux types de biosenseurs moléculaires polyvalents à haute performance. Ces senseurs bénéficient d'une sensibilité et sélectivité accrue grâce à l'utilisation de nouveaux systèmes de transduction optique utilisant des sondes à résonance plasmonique de surface (RPS). La surface du transducteur optique est fonctionnalisée afin d'assurer la détection sélective d'interactions récepteur-molécules. Grâce à cette technologie récemment commercialisée, les biosenseurs à transducteur RPS sont devenus un outil central dans la recherche biomédicale pour caractériser et mesurer des interactions biomoléculaires sur des surfaces : anticorps-ligand-métaux, protéine-ADN, enzymes-substrat. La RPS est couramment employée pour analyser des biopuces ou micro-arrays à protéines, anticorps, ADN, ANR. Ces outils miniaturisés de dosage de molécules organiques simples ou complexes ou de métaux sont utilisés dans des domaines variés tels que la santé, l'environnement, la police scientifique [81]. Cependant, la sensibilité et la sélectivité des biosenseurs-RPS sont très dépendantes de la distribution et l'orientation des biomolécules sur les supports transducteurs. Le tableau 1.1 regroupe, de manière non-exhaustive, quelques exemples de biocapteurs biochimiques développés pour la détermination des concentrations en métaux.

1.3.2 Biosenseurs à cellules entières

D'autres biosenseurs, dits à cellules entières, utilisent comme éléments sensibles des cellules vivantes (*e.g* bactéries) dont les éléments génétiques régulant la production d'enzymes ou de protéines spécifiques sont connus. Des promoteurs sont activés spécifiquement en réponse à des substances chimiques cibles, détectés à l'extérieur ou à l'intérieur de la cellule. Ces systèmes de

Enzymes	Métal	Limite de détection	Référence
Alcaline phosphatase	Zinc	10 µM	Satoh 1991 [82]
L-lactate déhydrogénase	Mercure	0.002 µM	Gayet et *al.* 1993 [37]
	Argent	0.02 µM	
	Plomb	0.2 µM	
	Cuivre	0.5 µM	
	Zinc	5.0 µM	
Peroxydase	Mercure	0.02 µM	Shekhovtsova et *al.* 1993 [84]
Protéines			
Glutathione-S-transferase-SmtA	Zinc	0.1 pM	Bontidean et *al.* 2003 [12]
	Cadmium	0.1 pM	
	Cuivre	0.1 pM	
	Mercure	0.1 pM	
Anhydrase carbonique	Cuivre	0.1 pM	Zeng et *al.* 2003 [106]
Sondes			
DsRed	Cuivre	10 nM	Sumner et *al.* 2006 [90]
Engineered GFP		5 µM	Richmond et *al.* 2000 [79]
Anticorps			
Anticorps monoclonaux	Cadmium	0.25 nM	Blake et *al.* 2001 [11]
	Cobalt	10 nM	
	Plomb	6 nM	
Micro-organismes			
Staphylococcus aureus	Cadmium	10 nM	Tauriainen et *al.* 1998 [92]
	Plomb	33 nM	
Escherichia coli	Cuivre	1.0 µM	Corbisier et *al.* 1999 [28]
	Chrome	1.0 µM	
	Plomb	0.5 µM	
	Cadmium	25 nM	Biran et *al.* 2000 [8]
	Mercure	10 nM	Jouanneau et *al.* 2011 [48]
	Arsenic III	28 µM	
	Cuivre	16 µM	
	Plomb	2 µM	
Saccharomyces cerevisiae	Cuivre	0.1 mM	Lehmann et *al.* 1999 [55]
Ralstonia eutropha	Nickel	1.0 mM	Tibazarwa et *al.* 1999 [95]
	Cobalt	9.0 mM	

TABLE *1.1 – Tableau des biosenseurs moléculaires et à cellules entières (micro-organismes) sensibles aux métaux lourds. D'après Verma et Singh (2005) [103].*

reconnaissance assez spécifiques sont très utiles pour qualifier ou quantifier la toxicité, la biodisponibilité, voire dans certains cas la bio-accessibilité de certains contaminants dans l'environnement : métaux lourds, hydrocarbures, composés aromatiques, dérivés chlorés, pesticides, *etc.* Cette réponse cellulaire d'ordre physiologique peut se traduire à l'échelle de la cellule par changement de forme, de taille, de couleur, de mobilité [6] ou à l'échelle moléculaire par l'induction de gènes promoteurs et la production de systèmes protéiques. Si ces éléments génétiques capables de détecter des interactions contaminant-cellule sont répertoriés, des manipulations génétiques permettent de transformer ces systèmes de reconnaissance afin que les cellules génèrent des signaux luminescents ou fluorescents. Dans ce type de capteurs, le système de détection est plus robuste car la cellule est un système souple, énergétiquement autonome et auto-poïètique [2]. Il s'avère moins sensibles aux perturbations ou aux interférences extérieures qu'un capteur usuel. De plus, les progrès réalisés dans le domaine de l'imagerie et de la spectroscopie de fluorescence permettent d'étudier la réponse de ces biosenseurs à l'échelle d'une population de cellules par spectrofluorimétrie ou d'un individu par microscopie de fluorescence. Différents systèmes d'imagerie sont utilisés pour visualiser et quantifier les réactions de bioluminescence et de chimioluminescence.

1.4 Systèmes rapporteurs luminescents

Pour éviter les dommages cellulaires provoqués par la pénétration des métaux et maintenir l'homéostasie, les bactéries ont développés au cours du temps différents mécanismes de résistance actifs ou passifs. En particulier, les concentrations cytoplasmiques en métaux sont régulées très finement par :

- la séquestration intracytoplasmique des métaux par des protéines soufrées, riches en cystéine. Ces molécules chélatantes piègent et inactivent les métaux les plus chalcophiles (Cd^{2+}, Cu^{2+} et Zn^{2+}) au sein du cytoplasme [98, 20, 40].
- l'oxydation ou la réduction enzymatique intracellulaire des métaux potentiellement toxiques (*e.g.* As, Sb) afin d'accroître leur innocuité.
- le contrôle passif de la diffusion pariétale, qui permet d'exclure certains métaux par absorption des ions dans les couches polymériques périphériques des cellules [83]. Ce mécanisme passif assez efficace réduit le flux d'internalisation des métaux et le risque toxique pour les systèmes vitaux des bactéries.
- le contrôle actif de la concentration intracellulaire des ions par des systèmes protéiques membranaires : les pompes d'efflux. Ces mécanismes de transport actif sont couramment utilisés par les bactéries pour exporter les métaux toxiques depuis leur cytoplasme jusqu'au milieu extracellulaire [20].

Parce qu'il confère une résistance efficace et contrôlée des métaux et qu'il s'agit du mécanisme de résistance le plus utilisé par les micro-organismes, le fonctionnement des pompes d'efflux fournit les principes de base à la construction de biosenseurs bactériens sensibles aux métaux [77].

2. Propriété d'un système à se produire lui-même (*e.g.* division cellulaire).

1.4.1 Construction génétique des biosenseurs

Les pompes d'efflux membranaires pour les éléments métalliques divalents (Zn^{2+}, Cd^{2+} et Pb^{2+}) sont très répandues chez les bactéries Gram-positives et Gram-négatives [76]. Ce sont souvent des systèmes de transport actifs, consommateurs d'énergie de type ATP (pompes P type ATPase). Ces transporteurs protéiques intégrés dans les parois forment des canaux et des systèmes de transport [78, 100, 67].

Des études physiologiques et génétiques sur des souches bactériennes résistantes à un certain nombre de métaux comme le zinc, le cuivre, l'étain, l'argent, le mercure et le cobalt ont permis d'isoler des systèmes récepteurs et promoteurs de réponses biologiques spécifiques aux métaux [61, 74, 75]. A partir des données génétiques, il est possible de repérer les séquences du génome impliqués dans la fabrication de ces pompes d'efflux. En particulier, les sites promoteurs qui régulent l'expression des gènes adjacents peuvent être instrumentés pour réguler la production de molécules fluorescentes. La construction de biosenseurs cellulaires pour la détection de métaux à partir de la fusion des gènes codant ces systèmes d'efflux avec les gènes codants pour des protéines bioluminescentes, par exemple la luciférine. Dans ce cas, la production de la lumière, qui peut être mesurée par luminomètres et photomètres, indique la présence d'un métal dans l'environnement cellulaire et est généralement proportionnelle à la quantité présente [103, 74].

Bien que de nombreuses protéines et enzymes différents puissent servir en tant que rapporteurs, les biosenseurs bactériens intégrent des gènes codant pour des protéines luminescentes (*e.g.* luciférase) ou fluorescentes (*e.g.* GFP ou DsRed) en raison de leur stabilité chimique et de leur rendement lumineux.

1.4.2 Fonctionnement d'un biosenseur cellulaire

Le déclenchement de ces promoteurs est assuré par des protéines régulatrices qui changent de conformation et de comportement lorsqu'elles s'associent à un métal. Elles jouent le rôle d'un interrupteur dans un circuit électrique. Par exemple, en l'absence de métal, la protéine est fixée sur un élément spécifique de l'ADN, le promoteur, et bloque la transcription des gènes codant pour la synthèse des systèmes protéiques de défense. Le circuit est ouvert. Cette répression n'est jamais totale et une quantité minimale de pompes des flux ou de protéines de séquestration est toujours présente. Quand le métal pénètre dans le cytoplasme et interagit avec la protéine, celle-ci perd son affinité pour le site promoteur dont elle se détache. Le circuit se ferme. Le mécanisme de défense n'est plus inhibé et la bactérie produit alors une quantité de pompes d'efflux ou de protéines permettant d'expulser ou de neutraliser le métal.

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. Ces gènes spécifiques, responsables de la résistance bactérienne à cet élément, ont été isolés sur des souches bactériennes ayant montrées une résistance à un certain nombre de métaux tels que le zinc, le cuivre, l'étain, l'argent, le mercure ou le cobalt [61]. Le gène de résistance

est induit uniquement lorsque l'élément atteint le cytoplasme de la bactérie. La spécificité de l'activation de ce mécanisme de résistance contribue à la construction de biosenseurs pour la détection de métaux grâce à la fusion de ce gène résistant avec un gène codant la production de protéine fluorescente ou luminescente. La réponse fluorescente ou luminescente de la bactérie génétiquement modifiée est détectée ensuite par un spectromètre. L'intensité des signaux mesurés sur une population de cellules dépend alors, en général, de la quantité de polluants présent (Voir figure 1.2).

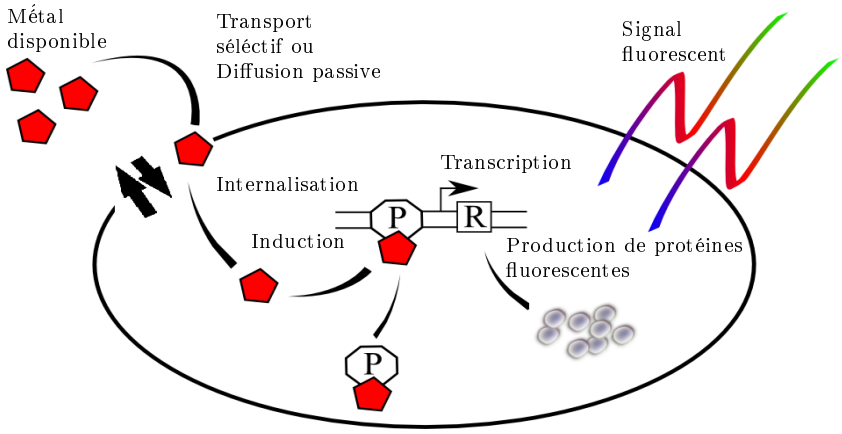


FIGURE 1.2 – *Principe de fonctionnement d'un biosenseur bactérien luminescent. Le gène de promoteur (P) est induit lorsque l'élément atteint le cytoplasme de la bactérie. Le gène rapporteur (R) codant pour la protéine fluorescente étant fusionné au gène de promoteur, le signal de fluorescence est corrélé à la concentration intracellulaire du métal.*

Ce niveau d'expression du gène rapporteur, reflété par l'intensité de luminescence, peut être mesuré périodiquement ou de façon continue et calibré par rapport à une gamme étalon.

Toutefois, la faible spécificité des biosenseurs bactériens est une limitation pour parvenir à un niveau de sensibilité exploitable. Les biosenseurs vont souvent répondre à des composés chimiques structurellement semblables, parce qu'ils détectent aussi les variations des conditions physiologiques induites par la présence de composés toxiques dans la cellule. Par exemple, des gènes qui sont capables de réagir avec le benzène vont généralement réagir au toluène, éthylbenzène et xylène, tandis que ceux qui sont capables de détecter du cadmium vont également détecter d'autres métaux lourds, tels que le mercure ou le zinc.

Une des pistes d'amélioration de la sensibilité consiste à améliorer le rapport signal à bruit ou à combiner différents biosenseurs. En outre, des réactions secondaires indésirables causées par des enzymes dans les cellules doivent être minimisées (par exemple, en prenant en compte les interactions dans la modélisation). La prise en compte des temps de maturation différents pour chaque protéine fluorescente a également une grande importance. Dans le cas où l'instant de mesure est mal choisi, le retard, variable d'une protéine à l'autre, entre l'induction du gène et le moment où la fluorescence est mesurable peut biaiser la mesure de la réponse. Cependant, dans le cadre d'une étude temporelle des réponses des biosenseurs, la variété engendrée par les différents temps de maturation peut être utilisée comme source supplémentaire de diversité.

Contrairement aux méthodes d'analyse chimique ou biochimique, qui déterminent la concentration totale de métaux présents dans l'échantillon ou solubles (*i.e.* libres), ces systèmes biologiques répondent uniquement à la fraction de métaux pouvant diffuser à travers les parois cellulaires.

Ainsi, dans le cas des métaux lourds, c'est la spéciation et la disponibilité des métaux résultant des différents processus de transformations dans l'environnement plutôt que leur concentration totale qui détermine les effets physiologiques et toxiques sur ces systèmes biologiques et par conséquent la réponse du biosenseur [77]. L'écart entre la fraction totale et la fraction de métal biodisponible peut être importante, surtout dans le cas des contaminants avec de faibles constantes de dissociation ou de faible solubilité aqueuse [93].

La section suivante présente les différentes méthodes d'acquisition de spectres de fluorescence dont nous disposons. Ces méthodes d'acquisition sont particulèrement adaptées pour l'étude multiparamétrique.

1.5 Mesures spectroscopiques des signaux émis par des biosenseurs fluorescents

Contrairement à la bioluminescence, où l'on ne mesure que l'intensité lumineuse, l'acquisition de spectres de fluorescence permet d'identifier la source d'émission. Le spectre de fluorescence reflète la distribution électronique de la molécule émettrice et la caractérise de manière unique, ce qui permet de détecter les signaux fluorescents émis par des mélanges de biosenseurs.

1.5.1 Spectrométrie de fluorescence (émission/excitation, MEEF)

Le phénomène de luminescence caractérise l'émission de lumière résultant de l'excitation électronique de la matière. La distinction entre les différents types de luminescence se fait par le mode d'excitation. On parle d'électroluminescence lorsqu'un courant électrique va engendrer l'émission de photons en traversant la matière ou de chimioluminescence ou bioluminescence lorsque des réactions chimiques ou biochimiques sont respectivement à l'origine de la dissipation d'énergie sous forme de lumière. Le phénomène de fluorescence résulte de l'absorption de l'énergie d'un photon par une molécule. Le photon est un quanta d'énergie électromagnétique que l'on peut définir par son énergie avec la relation suivante :

$$E = h\nu = \frac{hc}{\lambda} \tag{1.1}$$

avec h la constante de Planck ($\approx 4,13.10^{-15} eV.s$), c la vitesse de la lumière, ν la fréquence, λ la longueur d'onde. Ce processus d'activation se produit en un temps très court (10^{-15} secondes) durant lequel la molécule n'a pas le temps de changer de configuration. Il place la molécule dans un état électronique excité instable dont le retour à l'état fondamental se fait en émettant un photon dans un laps de temps très court (10^{-9} à 10^{-6} secondes). Une molécule est dite

fluorescente lorsqu'elle possède cette propriété d'absorber de l'énergie lumineuse et de la restituer sous forme de lumière ; elle est alors appelée fluorochrome ou fluorophore. Ce phénomène étudié au cours du XIX^{e} siècle par Becquerel et Stokes, est accompagné par un déplacement en longueur d'onde entre la lumière incidente et la fluorescence. Le photon émis est d'énergie inférieure à celle du photon absorbé suite à une perte d'énergie par vibration. La distance séparant le maximum d'absorption du maximum d'émission est appelé déplacement de Stokes.

L'intensité de lumière émise à une longueur d'onde j, par un fluorochrome excité à une longueur d'onde i, peut s'exprimer par l'équation suivante :

$$x_{ij} = \zeta \varphi \epsilon_i s_j \tag{1.2}$$

où s_j indique la fraction de fluorescence émise à la longueur d'onde j, ϵ_i est un coefficient reflétant la quantité de lumière absorbée à une longueur d'onde i, φ est un rendement de fluorescence, puisqu'une partie de l'énergie absorbée peut être dissipée par transfert non radiatif [85] et ζ correspond à la concentration de fluorochrome. Le tableau à deux entrées décrivant, pour tout i et j, l'équation (1.2) est appelée matrice d'excitation-émission. Les lignes et les colonnes de la matrice d'excitation-émission correspondent, respectivement, aux spectres d'émission et d'excitation. Un spectre d'émission (resp. d'excitation) de fluorescence représente l'intensité de fluorescence émise (resp. absorbée) par les molécules en fonction de la longueur d'onde. En fixant la longueur d'onde d'excitation λ_{exc}, on obtient le spectre d'émission de fluorescence et l'équation (1.2) devient

$$x_j = \zeta \varphi \epsilon_{\lambda_{exc}} s_j \tag{1.3}$$

Inversement, en fixant la longueur d'onde d'émission λ_{em}, on obtient le spectre d'excitation et l'équation (1.2) devient

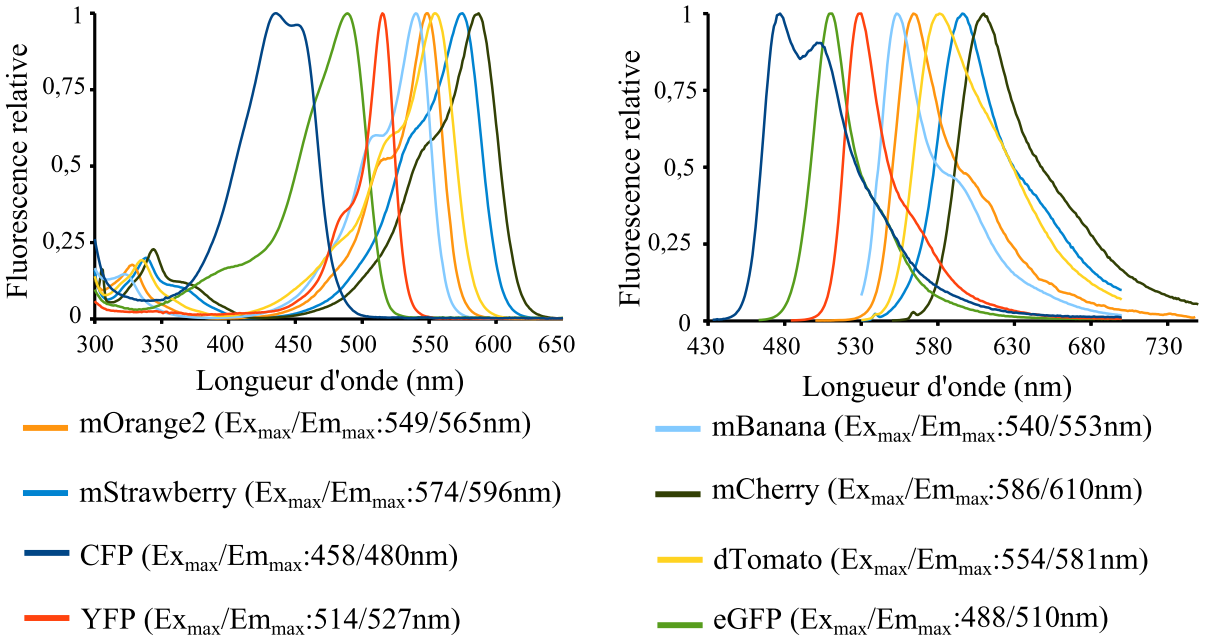
$$x_i = \zeta \varphi \epsilon_i s_{\lambda_{em}} \tag{1.4}$$

La figure 1.3 représente les spectres d'émission et d'excitation des molécules fluorescentes utilisables dans la conception de systèmes rapporteurs des biosenseurs. Chaque molécule possède des spectres d'excitation et d'émission spécifiques permettant de les différencier. Toutefois, les spectres de fluorescence présentent des recouvrements spectraux importants qui peuvent se combiner si différents biosenseurs sont mélangés.

Si l'on considère le cas d'un échantillon hétérogène contenant plusieurs biosenseurs, la fluorescence des F différents rapporteurs se combine, et l'intensité d'émission de fluorescence globale de l'échantillon à la longueur d'onde j peut s'écrire comme une combinaison linéaire des émissions des F molécules (sources) :

$$x_j = \sum_{f=1}^{F} \alpha_f s_{jf}, \tag{1.5}$$

avec f représentant la variable identifiant le fluorochrome et un coefficient α_f regroupant la



(a) Spectres d'excitation (b) Spectres d'émission

FIGURE *1.3 – Spectres d'excitation et d'émission de différentes protéines fluorescentes, obtenus en collectant les photons émis au voisinage du maximum d'émission ou d'excitation respectivement.*

concentration, la fraction de lumière absorbée par le fluorochrome f.

Si l'on connaît les J valeurs de chacun des F spectres et si $J \geq F$, on peut estimer les coefficients α_f par une méthode de moindres carrées :

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_F \end{bmatrix} = \begin{bmatrix} \mathbf{s}_{11} & \mathbf{s}_{12} & \dots & \mathbf{s}_{1F} \\ \mathbf{s}_{21} & \mathbf{s}_{22} & \dots & \mathbf{s}_{2F} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{s}_{J1} & \mathbf{s}_{J2} & \dots & \mathbf{s}_{JF} \end{bmatrix}^{\dagger} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_J \end{bmatrix}. \tag{1.6}$$

.

Cependant, cette méthode nécessite de connaître les spectres purs de toutes les sources de fluorescence présentes dans le milieu. Or, les sources inconnues ou inattendues sont possibles et rien ne garantit que les sources spectrales ne seront pas modifiées par une interaction avec le milieu. De plus, la qualité d'estimation des sources ayant un très faible rapport signal à bruit est grandement dégradée. Une solution à ce problème consiste à utiliser uniquement des combinaisons de molécules ayant un minimum de recouvrement spectral ou de faire appel à des méthodes de séparation de sources. Les spectres individuels peuvent être déterminés s'il existe des plages de longueurs d'onde dans lesquelles l'un des spectres reste non nul, tandis que les autres s'annulent [43]. Cette méthode est rapidement limitée par le nombre de combinaisons possibles, par le choix des longueurs d'onde d'excitation ou encore par le compromis entre l'absorption et le rendement quantique de chacune des protéines fluorescentes. Prenons l'exemple d'un mélange de

deux protéines (GFP et mCherry) dont les spectres d'émission optimaux (a et b) sont représentés sur la figure 1.4. Les maxima d'absorbance et d'émission sont respectivement à 488/509 nm pour la GFP et 587/610 nm pour la mCherry. Comme le montre la figure 1.4, une excitation à 488 nm favorise l'excitation de la GFP au détriment de la mCherry dont le rendement de fluorescence devient insuffisant pour séparer son signal de fluorescence du bruit. Une excitation à 587 nm permet d'augmenter la flurorescence de la mCherry mais la fluorescence de la GFP est invisible. Une excitation médiane à 520 nm permet d'avoir un niveau de fluorescence équivalent pour les deux protéines mais réduit fortement l'intensité globale du signal de fluorescence et nécessite d'augmenter le gain, le temps d'acquisition ou la puissance de la lumière excitatrice. Ceci accroit le risque de photoblanchiment [3], qui diminue l'intensité des signaux de fluorescence et rend difficile les études temporelles.

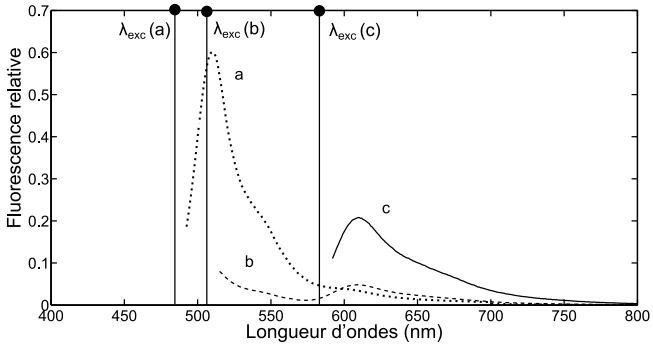


FIGURE 1.4 – *Spectres d'émission de fluorescence d'un mélange de deux protéines fluorescente (GFP et mCherry). La courbe* **a** *(pointillé) correspond au spectre d'émission pour une longueur d'onde d'excitation de* 488 nm*, soit la longueur d'onde du maximum d'absorbance de la GFP. La courbe* **c** *(trait plein) correspond au spectre obtenu avec une excitation à* 587 nm*, soit la longueur d'onde du maximum d'absorbance de la mCherry. La courbe* **b** *(tiret) correspond au spectre obtenu avec une excitation à* 520 nm.

Ces problèmes inhérents à l'utilisation de mélanges de molécules fluorescentes peuvent être en partie résolus en générant des matrices d'excitation-émission de fluorescence (MEEF). En utilisant successivement les deux monochromateurs (excitation et émission) d'un spectrofluorimètre, il est possible de mesurer les spectres d'émission pour une large gamme de longueurs d'ondes d'excitation. On obtient ainsi une MEEF. Cependant, l'acquisition de cette matrice est lente car elle nécessite l'acquisition successives de nombreux spectres. La figure 1.5 représente la matrice d'excitation-émission d'un mélange de deux molécules fluorescentes (GFP et mCherry). La zone de valeurs nulles (en noir) dans la partie supérieure gauche de la matrice est due au déplacement de Stokes [4]. Une ligne représente un spectre d'émission et une colonne un spectre d'excitation. Grâce à cette méthode, on distingue très nettement les spots d'émission des deux molécules et leur contribution relative au signal de fluorescence.

3. Dégradation de la molécule fluorescente entraînant la perte de la propriété de fluorescence. C'est généralement la conséquence d'une réaction photochimique avec l'oxygène.

4. Dissipation d'énergie qui rend l'énergie émise plus faible que l'énergie excitatrice

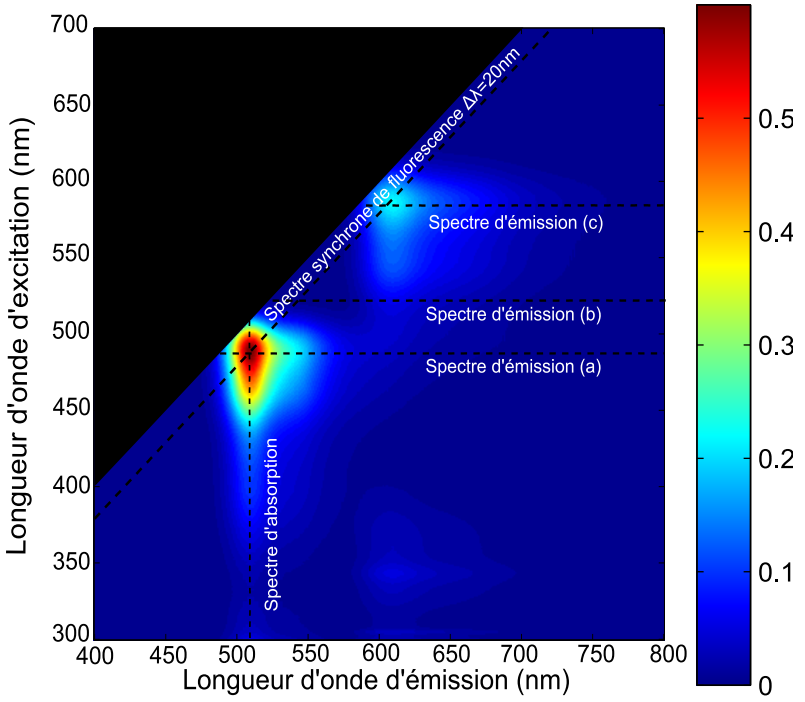


FIGURE *1.5 – Matrice d'excitation-émission du mélange GFP/mCherry. Les lignes correspondent aux spectres d'émission (ex : $\lambda_{exc} = 488\,\text{nm}$). Les colonnes correspondent aux spectres d'émission (ex :$\lambda_{emi} = 510\,\text{nm}$).*

1.5.2 Spectrométrie synchrone

L'acquisition du spectre d'émission d'un mélange de protéines fluorescentes est souvent rendu difficile par le choix de la longueur d'onde d'excitation. Chaque protéine ayant des spectres d'excitation, d'émission et des rendements quantiques différents, il est difficile de trouver une longueur d'onde d'excitation permettant d'obtenir une fluorescence suffisante de chacune d'elles. La spectrométrie synchrone a l'avantage de mieux exciter les différentes sources en utilisant différentes longueurs d'onde d'excitation. Dans cette méthode, les longueurs d'onde d'émission et d'excitation varient simultanément, tout en conservant entre elles un décalage constant ($\Delta\lambda$). Sur la figure 1.5, la diagonale représente un spectre synchrone de fluorescence. Pratiquement, on choisit un décalage correspondant au déplacement de Stokes de la molécule ayant le rendement quantique le plus faible ou dont la concentration est la plus faible. Les deux monochromateurs sont synchrones et les longueurs d'onde d'émission et d'excitation sont balayées simultanément. Le spectre obtenu dépend de l'absorbance à la longueur d'onde λ_{exc} et de l'émission à la longueur d'onde $\lambda_{exc} + \Delta\lambda$.

Les spectres synchrones donnent une information plus complète sur les mélanges de molécules fluorescentes que les spectres classiques. La figure 1.6 montre les spectres synchrones de fluorescence ($\Delta\lambda$=20 nm et $\Delta\lambda$=30 nm) et le spectre d'émission pour une longueur d'onde d'excitation de 475 nm. L'intensité de fluorescence des spectres synchrones est plus importante que l'intensité du spectre d'émission.

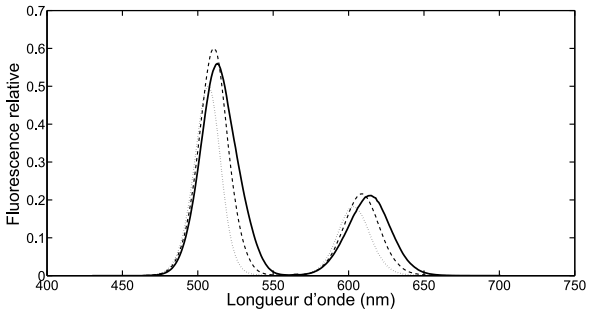


FIGURE *1.6 – Spectres synchrones d'un mélange de deux protéines fluorescentes (GFP et mCherry). La courbe en trait plein correspond au spectre pour* $\Delta\lambda = 30$ nm*, soit un* $\Delta\lambda$ *favorisant la mCherry. La courbe discontinue (tiret) correspond à* $\Delta\lambda = 20$ nm*, soit un* $\Delta\lambda$ *favorisant la GFP. La courbe en pointillé correspond au spectre d'émission de fluorescence pour une longueur d'onde d'excitation de* 475 nm.

Cette technique est un bon compromis entre le spectre de fluorescence et la MEEF. Car l'excitation de chacune des molécules du mélange est optimisée. L'acquisition d'un spectre synchrone est aussi plus rapide que l'acquisition de la matrice d'excitation-émission et diminue les risques de photoblanchiment.

1.5.3 Réduction des phénomènes non linéaires

Lors de la mesure, divers phénomènes physiques ou chimiques peuvent interférer et perturber la mesure des spectres. On qualifiera de non linéaire tout phénomène influençant la mesure du spectre de fluorescence et qui ne peut être modélisé par l'ajout d'un ou plusieurs composants supplémentaires dans l'approximation linéaire de la loi de Beer-Lambert(Eq. 1.5). Lorsque la concentration en composés fluorescents devient trop importante, des effets de filtre et de masquage réduisent l'intensité de fluorescence. En particulier, les photons émis à des longueurs d'ondes situées dans la zone de recouvrement entre les spectres d'excitation et d'émission de deux molécules différentes, peuvent être réabsorbés (transfert radiatif) [58]. Lorsque les molécules fluorescentes sont physiquement proches, on peut également observer un transfert d'énergie par résonance de type Förster (transfert non radiatif). La probabilité d'un transfert non radiatif entre deux molécules énergétiquement compatibles est inversement proportionnelle à la distance les séparant. La désexcitation de la molécule par transfert non radiatif aura pour effet de diminuer l'énergie à dissiper par la fluorescence. La fluorescence peut être également diminuée par le phénomène d'écrantage. Si la concentration moléculaire est très importante, la probabilité d'activation des molécules est diminuée par l'absorption des molécules avoisinantes. Pour rester dans l'approximation de la loi de Beer-Lambert, il faut donc que l'absorbance de l'échantillon reste faible.

Toutefois, l'utilisation de biosenseurs bactériens apportent de nombreux avantages. Les problèmes de non linéarité liés aux variations de force ionique et de température sont moins marqués. En effet, l'homéostasie cellulaire tend à réguler la composition chimique interne de la cellule et

limite les sur-concentrations moléculaires. De plus, les sondes fluorescentes internes ne sont pas affectées par les paramètres externes comme la température ou le pH. Enfin, la production interne de sondes à haut rendement de fluorescence permet de réduire la densité cellulaire et par conséquent réduire les problèmes d'écrantage et de quenching. Dans la plupart des cas, on peut donc considérer que la suspension de biosenseurs est suffisamment diluée et qu'il n'y a pas d'interaction entre les biosenseurs [52, 47, 60, 59].

La section suivante présente succintement l'approche proposée pour la séparation des spectres de fluorescence. Les méthodes de séparation de sources étudiées et modifiées pour mieux s'adapter aux problèmes de l'étude des biosenseurs bactériens sont présentées en détail dans le chapitre 2.

1.6 Orientation et objectifs de l'étude

Comme nous l'avons vu précédement, les défis techniques et scientifiques majeurs dans le domaine de l'environnement consiste à améliorer la compréhension des interactions entre les microorganismes et les minéraux, les métaux et les minéraux, dans des environnements naturels ou fortement détériorés et perturbés par les activités humaines.

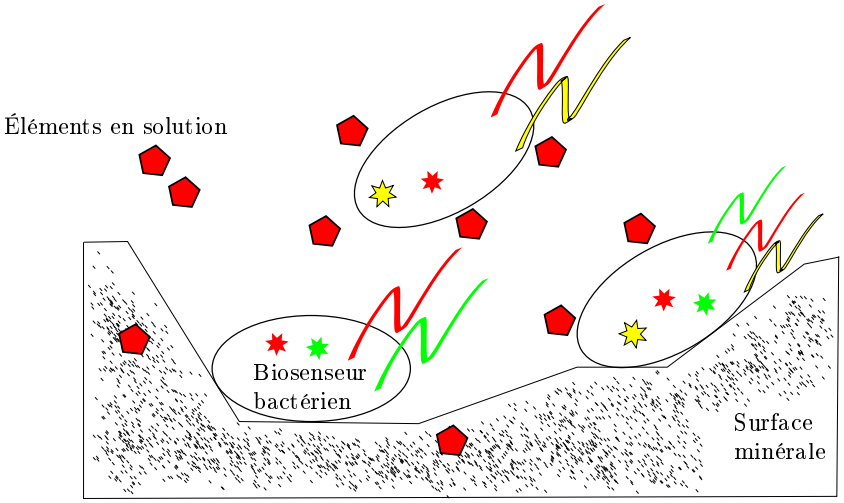


FIGURE *1.7 – Étude des interactions bactéries-minéraux à l'aide de biosenseurs multicolores ou des mélanges de biosenseurs, dont les réponses varient en fonction de l'environnement immédiat de la cellule (présence de molécules en solution, particules, état physiologique du biosenseur, etc.). Cas d'un biosenseur tricolore équipé d'un marqueur constitutif rouge et deux marqueurs jaune et vert dont les productions sont respectivement dépendantes de la présence d'un métal en solution et de la proximité d'une surface minérale, contenant ce métal.*

Pour vivre dans ces environnements accueillants ou hostiles, les bactéries ont développées des mécanismes moléculaires assurant la fixation, la séquestration ou l'expulsion sélective et contrôlée des métaux en fonction de leur caractère essentiel ou toxique. Une connaissance approfondie du fonctionnement de ces systèmes *metal-sensitive* et leur transformation en senseurs biologiques serait très utile pour qualifier ou quantifier la biodisponibilité, voire dans certains cas la bioaccessibilité de certains métaux dans l'environnement ou d'estimer leur toxicité. Ainsi, à côté des

capteurs conventionnels et classiques (sondes pH, conductimétrique, hygrométrique, *etc.*), les chercheurs développent des biosenseurs bactériens plus ou moins spécifiques capable d'identifier et doser, voire localiser, directement les métaux in situ. En insérant au droit des gènes promoteurs des systèmes rapporteurs luminescents, ces micro-organismes génétiquement modifiés produisent un signal dont l'intensité est dose-dépendante.

Comme il a été montré ces biosenseurs répondent uniquement à la fraction métallique pouvant diffuser à travers leurs parois cellulaires. Ainsi, l'écart entre la fraction totale et la fraction biodisponible du métal peut être important, surtout pour les métaux hydrolysables, oxydo-réductibles ou peu solubles à pH neutre, comme le fer. De plus, comme pour toute cellule vivante, cette réponse biologique est contingente et intégrative ; elle s'avère sensible à de nombreuses grandeurs d'influence, comme l'ensemble des facteurs physico-chimiques et biologiques externes ou internes affectant la physiologie ou le cycle de vie des cellules.

Pour améliorer la robustesse de l'estimation et conférer à la réponse des biosenseurs un caractère univoque, deux approches sont aujourd'hui possibles.

La première vise à améliorer nos connaissances des réseaux de régulation génétique des systèmes de résistance aux métaux chez les bactéries et à rechercher des systèmes promoteurs spécifiques par des approches réductionnistes. Néanmoins, une des plus grandes limitations dans l'utilisation des biosenseurs est le caractère non univoque de la réponse de la plupart des senseurs et les promoteurs très spécifiques répondant à un seul métal sont rares. Les modifications réalisées pour la construction de biosenseurs très spécifiques peut alors modifier le fonctionnement du système. Par exemple, la réduction de l'autofluorescence bactérienne passe par la désactivation de mécanismes génétiques dont les effets ne sont pas totalement déterminées. Au vu des modifications génétiques, parfois importantes, réalisées sur les bactéries, on peut raisonnablement se demander si le comportement du système n'a pas été modifié et si la réponse à la sollicitation étudiée reste inchangée.

La seconde approche suivie dans le présent travail, consiste plutôt à utiliser la diversité de comportement des différents biosenseurs pour élaborer des combinaisons de biosenseurs ou de gènes permettant d'identifier les interactions bactérie-métaux (biodisponibilité). Cette approche de type traitement du signal amène à considérer le problème de modélisation du comportement de biosenseurs métal-sensitive comme un problème de séparation de sources. Toutefois, bien que les mécanismes de résistance et de transports chez les bactéries aient été largement étudiés, on ne sait toujours pas comment ces différents promoteurs et mécanismes cellulaires sont intégrés et gérés au sein d'un réseau d'homéostasie/résistance efficace. L'existence de réponses univoques pour la séquestration ou l'expulsion de métaux lourds est très contreversée. Une approche multivariée s'impose donc. La modularité des biosenseurs va permettre de multiplier les sources de fluorescence en limitant les modifications génétiques réalisées aux marquages des gènes promoteurs étudiés. Les différentes sources de fluorescence se mélangeront entre elles et avec la fluorescence naturelle du milieu pour donner la fluorescence globale. L'utilisation de méthodes de séparation de sources permettront d'extraire les signaux d'intérêt et de réduire les interférences des autres

sources de fluorescence.

La question posée est la suivante : peut-on identifier avec un minimum d'*a priori* les interactions métaux-bactérie à partir des signaux spectraux émis par des biosenseurs bactériens ? Par exemple, une variation en qualité ou en quantité des métaux présents en solution aura pour conséquence de modifier les réponses de biosenseurs fluorescents et par conséquent de faire varier le coefficient α_f de l'équation (1.5) en fonction de paramètres comme par exemple le pH, la concentration en éléments dissouts ou la température.

En ajoutant une nouvelle diversité conservant les conditions nécessaire à l'approximation de Beer-Lambert, le modèle des signaux de fluorescence devient le suivant :

$$x_{jk} = \sum_{f=1}^{F} \alpha_{kf} s_{jf} \tag{1.7}$$

où k représente la variable identifiant les variations du paramètre associées à la première diversité. La résolution unique de ce problème, par une méthode de séparation aveugle de sources, ne peut être assurée sans l'application de contraintes fortes. L'ajout d'un diversité supplémentaire permet alors de réduire les degrés de liberté du modèle et permet d'utiliser des méthodes de décomposition trilinéaire dont l'unicité peut être assurée sous de faibles contraintes. Avec l'ajout d'une troisième diversité le modèle des signaux de fluorescence devient alors le suivant :

$$x_{jkl} = \sum_{f=1}^{F} \alpha_{fkl} s_{jf} \tag{1.8}$$

Le choix du troisième paramètre est alors fait de manière à fournir une diversité permettant d'obtenir un modèle de fluorescence compatible avec le modèle CP, ce qui permet d'écrire $\alpha_{fkl} = \alpha'_{kf}\beta_{lf}$. Dans notre étude, cette diversité est liée à la quantité de biosenseurs. La quantité de biosenseurs pourra varier en fonction du paramètre temporel, par l'intermédiaire de la croissance bactérienne, ou en fonction du ratio de biosenseurs dans un mélange.

Le modèle CP des signaux de fluorescence devient alors le suivant :

$$x_{jkl} = \sum_{f=1}^{F} \alpha'_{kf}\beta_{lf} s_{jf} \tag{1.9}$$

L'objectif est de retrouver les contributions α'_f, β_f et s_f de chacune des F sources de fluorescence et d'estimer les variations de fluorescence en fonction des paramètres étudiés.

L'algèbre multilinéaire constitue un cadre mathématique bien adapté pour la résolution de ce type de problèmes, car elle permet d'exploiter les diversités multiples des réponses des biosenseurs afin d'extraire les signaux effectifs liés à l'expression des systèmes rapporteurs. On s'affranchit ainsi des fluorescences parasites (autofluorescence spontanée) limitant l'exploitation et l'interprétation des données spectrales dans des systèmes naturels.

Les méthodes dévelopées dans ce travail de thèse, sont inspirées des méthodes CP [15], PA-

RALIND [17] et Block Component Decomposition [27, 29]. L'avantage de ce type d'approches, par rapport aux approches classiques, est l'identification unique des sources sous des faibles contraintes. Cependant, les algorithmes de type moindres carrés alternés (ALS) utilisés pour les décompositions multilinéaires (de type CP) présentent plusieurs inconvénients tels que la vitesse de convergence et la détermination de l'ordre de la décomposition (nombre des sources), fixé *a priori* d'une manière empirique. L'un des objectifs de la thèse est donc de développer des algorithmes efficaces de séparation de sources pour les signaux multidimensionnels acquis à partir de biosenseurs bactériens luminescents sensibles aux variations physico-chimiques du système (solution et minéral). Le point original de la thèse est la prise en compte des contraintes liées à la physique des phénomènes analysés tels que la parcimonie des coefficients de mélange ou la positivité des signaux sources, afin de réduire au maximum l'usage d'*a priori* sur la composition du mélange. L'intérêt porte également sur le développement conjoint de nouveaux biosenseurs, de plans d'expériences et de méthodes d'analyses des données.

La démarche envisagée comprend trois grandes orientations. La première est l'identification avec un minimum d'*a priori* des sources fluorescentes et leur comportement. Le but est de mieux comprendre le fonctionnement des biosenseurs en étudiant la réponse fonctionnelle du biosenseur à des concentrations croissantes en métaux plutôt qu'en élaborant une courbe d'étalonnage. Dans un premier temps, nous étudierons la réponse temporelle d'un biosenseur répondant au cadmium, en fonction de la concentration en métal. Ensuite, nous complexifierons le problème en estimant conjointement la réponse d'un mélange de deux biosenseurs antagonistes réagissant respectivement à la carence et à l'excès de fer. Ces deux études permettront de valider l'approche multiparamétrique pour l'analyse de signaux de biosenseurs.

Pour le second aspect, orienté capteur, nous développerons de nouvelles méthodes afin d'utiliser des biosenseurs et des algorithmes pour mettre en évidence la capacité des biosenseurs à fournir une information quantitative sur la concentration des métaux biodisponibles dans un système bactérie-solution.

Enfin pour la troisième orientation méthodologique, nous développerons de nouvelles méthodes d'expérimentation afin de fournir des données suffisament diversifiées pour mener à bien les deux aspects précédents. Nous nous intéresserons en particulier au développement de nouvelles méthodes de décomposition de tenseurs capables de gérer les interactions entre les différentes sources.

Chapitre 2

Méthodes multilinéaires pour la séparation de sources en spectroscopie de fluorescence

Sommaire

Dans le chapitre précédent, nous avons montré l'intérêt d'une meilleure compréhension du fonctionnement des biosenseurs bactériens par l'étude de leurs réponses fonctionnelles dans des environnements complexes. Nous avons également présenté le fonctionnement des biosenseurs bactériens ainsi que les méthodes de mesure de la fluorescence associée. Néanmoins, les signaux mesurés ne sont pas toujours exploitables en l'état. Les traitements numériques proposés auront pour buts principaux :

- l'identification des sources spectrales ;
- l'extraction des réponses des gènes ;
- la calibration des biocapteurs ;
- l'estimation de la concentration.

Les méthodes traditionnelles de mesures de fluorescence sont sensibles aux problèmes de recouvrement spectral. Le recouvrement peut provenir de l'autofluorescence du biosenseur, d'une source spectrale inconnue issue du milieu ou de la bactérie ou encore des différentes sources nécessaires à l'analyse multiparamétrique. De part les interactions possibles entre les éléments composant le milieu complexe, l'étude de la fluorescence des biosenseurs dans un tel milieu nécessite l'utilisation de méthodes d'analyses multiparamétriques. Nous proposons alors d'extraire les réponses du biosenseur aux différents paramètres étudiés à l'aide de méthodes multilinéaires de séparation de sources.

Avant de s'intéresser en détail aux méthodes de séparation aveugles de sources, la première partie de ce chapitre présente les méthodes de production de données de fluorescence proposées. De l'étude du modèle de fluorescence, nous tirons différents types de plans d'expériences produisant des données uni- ou multi-variées. L'utilisation de microplaques rend aisée la production de données multi-paramétriques en mesurant les spectres de fluorescence dans chaque puits et en associant les lignes et les colonnes à des paramètres différents. De plus, l'utilisation d'un robot pipeteur pour le remplissage des microplaques assure une grande précision et permet de produire des combinaisons variées difficiles à produire par un expérimentateur humain.

Nous étudions ensuite les méthodes de factorisation d'un tableau bidimensionnel de type décomposition en valeurs singulières et factorisation en matrices non-négatives. Nous rappelons leur principe et illustrons leur mise en œuvre sur un exemple simple d'analyse d'une simulation de mélange de spectres. Cette simulation met en évidence l'intérêt d'ajouter une diversité supplémentaire pour utiliser des méthodes de factorisation des tableaux tridimensionnels ou tenseurs. Nous étudions, ensuite, la décomposition CP afin d'en déterminer les conditions d'identifiabilité et de validité. Puis, nous validons l'intérêt de cette méthode en comparant les résultats obtenus, par les différentes méthodes étudiées, dans l'étude d'un exemple de signaux de fluorescence réels mesurés sur un mélange de biosenseurs bactériens.

Nous nous intéresserons ensuite au cas particulier de la gestion des dépendances linéaires dans un mode. Pour cela, nous étudierons le modèle PARALIND et la décomposition CP quadrilinéaire.

2.1 Production de données de fluorescence

Le modèle des signaux de fluorescence des biosenseurs bactériens fait intervenir trois paramètres : la longueur d'onde, la concentration en élément stressant et la quantité de bactéries. On distingue alors trois types de données :

Expériences uni-paramétriques. Les plans d'expériences uni-paramétriques produisent des données uni-variées. En analyse spectrale, seule la longueur d'onde varie et on considère indépendamment chacun des spectres mesurés. Pour ce type de données, on ne dispose que d'un seul spectre (émission, absorption ou synchrone) mesuré sur un seul mélange. Ce cas est le plus défavorable. Pour retrouver les paramètres du mélange, il faut alors avoir connaissance des sources spectrales ou du mélange si l'on cherche les sources spectrales.

Le vecteur de données $\mathbf{x}$ contient le spectre mesuré. En pratique, la mesure du spectre se fait après avoir mis en contact un mélange de biosenseurs avec un élément stressant dans un seul puit de la microplaque. Avec $\mathbf{s}_f$ le spectre du composant f du mélange et $\mathbf{a}_f$ son poids dans le mélange, le modèle est le suivant :

$$\mathbf{x} = \sum_{f=1}^{F} \mathbf{a}_f \mathbf{s}_f$$

avec F le nombre de sources de fluorescence. Dans le cas de l'étude de la réponse de gène, la résolution du problème d'estimation des coefficients de mélanges est impossible car les spectres ne sont pas connus *a priori*. De plus, cette approche ne permet pas d'estimer la réponse d'un gène. Elle peut uniquement être utilisée pour retrouver la proportion d'un élément stressant dans le mélange à condition d'avoir précédemment calibré le biocapteur pour cette quantité de biosenseur.

Expériences bi-paramétriques. Les plans d'expériences bi-paramétriques produisent des données bi-variées. Un des paramètres (proportion d'élément stressant ou quantité de bactéries) varie en plus de la longueur d'onde. Ce type de données est obtenu en mesurant un spectre sur plusieurs échantillons. Un seul des paramètres (proportion d'élément stressant ou quantité de bactéries) varie. En pratique, on fera généralement varier la proportion d'éléments stressant en fixant la quantité de bactéries de manière à estimer la réponse du gène.

On obtient un lot de spectres mesurés sur plusieurs échantillons (puits) que l'on regroupe dans une matrice $\mathbf{X}$ de dimension $(I \times J)$, avec I le nombre de mesures et J le nombre de longueurs d'ondes, dont le modèle est le suivant :

$$\mathbf{X} = \sum_{f=1}^{F} \mathbf{a}_f \mathbf{s}_f^T$$

avec $\mathbf{s}_f$ vecteur, de dimension $(J \times 1)$, représentant le spectre du composant f du mélange et $\mathbf{a}_f$ de dimension $(I \times 1)$, le vecteur des poids respectifs dans les mélanges. L'estimation des spectres et des coefficients de mélanges revient à la résolution d'un problème inverse. La section

suivante présente des méthodes de décompositions bilinéaires permettant de résoudre ce type de problème. Cependant, dans le cadre de l'étude de la fluorescence des biosenseurs les conditions assurant l'unicité sont rarement remplies. Toutefois, en augmentant la diversité des mesures, le modèle bilinéaire passera à un modèle trilinéaire : on pourra alors facilement assurer l'unicité de la décomposition.

Expériences tri-paramétriques. Les plans d'expériences tri-paramétriques produisent des données tri-variées. Ce type de données est obtenu en mesurant un spectre sur plusieurs échantillons. Dans chacun des puits, on fera varier les deux paramètres suivants : proportion d'éléments stressant et quantité de bactéries, suivant les deux dimensions de la microplaque.

En pratique, on fera varier conjointement les deux paramètres sur une microplaque de manière à obtenir un tableau tridimensionnel de données $\boldsymbol{\mathcal{X}}(I \times J \times K)$ dont le modèle paramétrique est le suivant :

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

avec $\circ$ le produit tensoriel de deux vecteurs, $\mathbf{s}_f$ le spectre de la source f, $\mathbf{a}_f$ l'évolution de la fluorescence de la source f en fonction des variations du premier paramètre et $\mathbf{b}_f$ l'évolution de la fluorescence de la même source en fonction des variations du second paramètre. Il est possible (voir section 2.7) de faire varier la fluorescence d'une source par rapport à un troisième paramètre de manière à obtenir des modèles quadrilinéaires.

La section suivante décrit des techniques d'analyses bilinéaires permettant l'estimation des signaux de fluorescence. Nous montrerons que ces méthodes ne sont pas les plus adaptées à l'étude de la réponse de biosenseurs. La section 2.3 présentera des méthodes de décomposition trilinéaire plus adaptées aux données considérées.

2.2 Factorisation d'un tableau de données bidimensionnel

Considérons une matrice de données $\mathbf{X}$ de dimension $(I \times J)$ contenant I spectres (excitation, émission ou synchrone) de longueur J correspondant à I échantillons différents. Pour cette expérience, on choisit donc de faire varier un seul paramètre, soit la proportion en élément stressant, soit la quantité de biosenseurs.

Dans la suite, on suppose que les mélanges sont des combinaisons de F fluorochromes (sources) et que $I > F$ (problème sur-déterminé). Idéalement, F est égal au nombre de gènes instrumentés. En pratique, les sources de fluorescence du milieu et les sources intrinsèques aux bactéries font que le nombre de sources est inconnu.

La séparation des F sources revient alors à résoudre un problème de factorisation de matrice qui correspond au modèle de mélange linéaire instantané donné par l'équation suivante :

$$\mathbf{X} = \mathbf{A}\mathbf{S}^T \tag{2.1}$$

où $\mathbf{X}$ sont les signaux observés, $\mathbf{A}$ $(I \times F)$ contient sur les colonnes la contribution de chacune des F sources dans l'observation et $\mathbf{S}$ $(J \times F)$ contient les spectres des F sources.

La factorisation de matrice est un problème inverse est qualifié de « mal posé » (au sens de Hadamard) car sans informations supplémentaires sur les sources et/ou les coeffcients du mélange, il admet une infinité de solutions. En effet, pour toute matrice $\mathbf{M}(F \times F)$ inversible l'équation (2.1) peut s'écrire :

$$\mathbf{X} = (\mathbf{A}\mathbf{M}^{-1})(\mathbf{M}\mathbf{S}^T) = \mathbf{A}'\mathbf{S}'^T, \tag{2.2}$$

avec $\mathbf{A}'$ et $\mathbf{S}'$ sont deux autres matrices de mélange et de spectres admissibles. Or, l'objectif principal de la séparation de sources est de retrouver de manière unique les sources et coefficients de mélange, à partir des données, *i.e.* l'identifiabilité du modèle. Pour arriver à un modèle identifiable, il est nécessaire de prendre en compte des informations supplémentaires sur le mélange et/ou sur les sources. Les contraintes couramment employées sont l'orthogonalité, l'indépendance, la parcimonie. Dans le domaine de la spectroscopie les contraintes de positivité sont intéressantes car elle correpondent à une réalité physique, les spectres et les proportions sont des quantités positives. Dans cette partie, nous nous intéressons essentiellement à deux types de techniques :

- la décomposition en valeurs singulières ;
- la factorisation en matrice non négatives

Pour ces deux approches nous rappelons leur principe et illustrons leur mise en œuvre en analyse de mélanges spectraux sur un exemple simple.

2.2.1 Décomposition en valeurs singulières (Singular Value Decomposition (SVD))

La décomposition en valeur singulière est une méthode de factorisation matricielle [38] souvent utilisée en séparation de sources [26, 104], pour l'analyse en composantes principales (Principal Component Analysis PCA).

Soit une matrice de données $\mathbf{X}$ de taille $I \times J$. La méthode SVD consiste à trouver deux matrices orthogonales $\mathbf{U}(I \times I)$ et $\mathbf{V}(J \times J)$ et une matrice diagonale positive $\mathbf{\Sigma}(I \times J)$ telles que :

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T. \tag{2.3}$$

La matrice $\mathbf{\Sigma}$ contient sur sa diagonale les valeurs singulières de la matrice $\mathbf{X}$. L'analyse de la décroissance des valeurs singulières peut être utilisée pour déterminer le nombre de sources.

Dans le cas où la matrice $\mathbf{X}$ regroupe des spectres de fluorescence, on peut estimer les coefficients de mélange $\mathbf{A}$ et les sources spectrales $\mathbf{S}$ (voir équation (2.1)) par

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma} \text{ et } \mathbf{S} = \mathbf{V} \text{ ou } \mathbf{A} = \mathbf{U} \text{ et } \mathbf{S} = \mathbf{V}\mathbf{\Sigma}^T$$

Il existe alors une indétermination d'échelle en fonction de l'intégration de la matrice $\mathbf{\Sigma}$ aux sources de fluorescence ou aux coefficients de mélange.

Exemple Dans cette simulation, nous testons la séparation des signaux de fluorescence de biosenseurs bactériens par la méthode SVD. Pour cela, une matrice $\mathbf{X}(5 \times 200)$ est engendrée. Cette matrice simule la mesure des spectres de 5 échantillons de 400 nm à 600 nm (soit 5 spectres de 200 points). Chaque spectre simule un mélange différent de deux protéines fluorescentes. Ils ont été choisis pour simuler le fonctionnement de deux gènes répondant à un même paramètre de manière antagoniste. La fluorescence de la source $\mathbf{s_1}$ augmente avec le paramètre P_1 alors que la fluorescence de la source $\mathbf{s_2}$ diminue. La figure 2.1 montre les spectres des deux protéines du mélange et les coefficients de mélange utilisés.

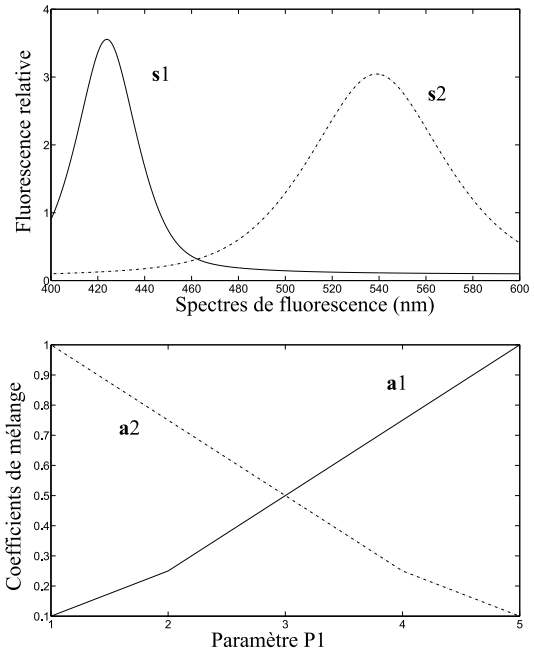


FIGURE *2.1 – Spectres de fluorescence simulés ($\mathbf{s_1}$ et $\mathbf{s_2}$) de deux protéines fluorescentes (longueurs d'ondes comprises entre* 400 nm *et* 600 nm*) et coefficients de mélange utilisés ($\mathbf{a_1}$ et $\mathbf{a_2}$) pour simuler les réponses antagonistes de deux gènes en fonction d'un paramètre P_1.*

En effectuant la SVD sur nos données, nous analysons les P premières valeurs singulières correspondant aux P sources spectrales prépondérantes. Les deux premières colonnes de $\mathbf{S}$ donnent les deux sources les plus importantes (les deux valeurs singulières les plus fortes). Les deux premières colonnes de $\mathbf{A}$ donnent la quantité de chaque spectre dans $\mathbf{X}$. Les deux spectres prépondérants et coefficients de mélange associés sont représentés figure 2.2. On peut voir que les spectres de fluorescence ne sont pas séparés correctement et présentent des valeurs négatives rendant impossible toute identification. De plus, les coefficients de mélange ne représentent pas l'aspect antagoniste des mélanges initiaux.

L'estimation du nombre de sources, dans la méthode SVD, est facilité par l'analyse de la décroissance des valeurs singulières. De plus, l'unicité de la décomposition en valeurs singulières

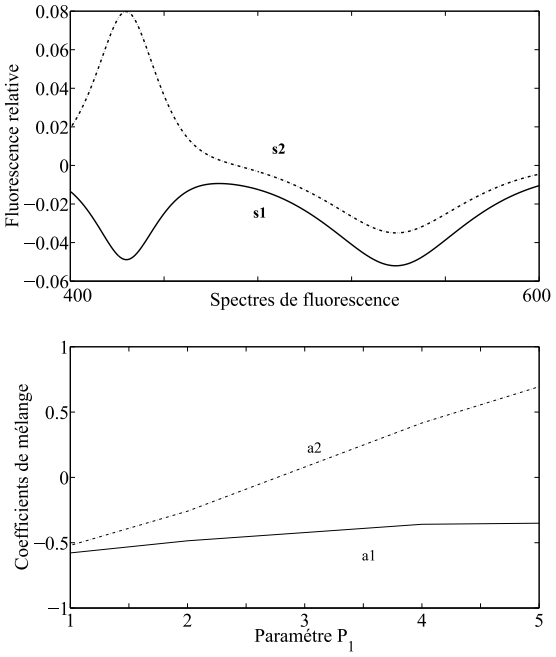


FIGURE *2.2 – Spectres de fluorescence* ($\mathbf{s_1}$ *et* $\mathbf{s_2}$) *et coefficients de mélange* ($\mathbf{a_1}$ *et* $\mathbf{a_2}$) *estimés par la méthode SVD.*

peut être assurée à une indétermination d'échelles près. L'inconvénient de la méthode SVD vient de la contrainte d'orthogonalité imposée entre les colonnes de $\mathbf{A}$ et de $\mathbf{S}$. La décomposition implique une hypothèse d'orthogonalité des sources inadaptée à des sources spectrales de fluorescence. Sous cette contrainte les réponses estimées peuvent être négatives ce qui rend inexploitable l'interprétation physique des résultats.

2.2.2 Factorisation en matrices non-négatives (Non-Negative Matrix Factorization (NMF))

De nombreuses applications réelles impliquent la positivité. C'est le cas en spectroscopie et plus particulièrement, en analyse de mélanges chimiques puisque les sources spectrales, non négatives, représentent les composés chimiques purs et les coefficients de mélange, également non négatifs, représentent les abondances de composés chimiques. La formulation du problème d'analyse de mélange de sources spectrales comme un problème de factorisation de matrices non négatives a été proposée par Lawton et Sylvestre [53] sous le terme de Self Modeling Curve Resolution. Une méthode de moindres carrés alternés a été proposée par Tauler [91]. Une version régularisée de cette approche a été proposé par Paatero et Tapper [5, 68] sous le terme PMF. La terminologie Non Negative Matrix Factorisation a été introduite par Lee et Seung [54] sans référence aux travaux antérieurs. L'approche qu'ils proposent s'apparente à une descente de gradient alternée avec remise à jour des matrices $\mathbf{A}$ et $\mathbf{S}$ selon une règle d'adaptation multiplicative garantissant la non-négativité des matrices.

Avant de tester ces algorithmes, nous nous intéressons à la question de l'unicité de la solution dans le cas particulier de la factorisation de matrices de données spectrales issues de biosenseurs bactériens fluorescents. Une première condition nécessaire et suffisante à l'unicité de la décomposition NMF a été formulé par Chen [25], mais cette condition ne permet pas de déterminer *a priori* l'unicité de la solution. Park et *al.* [71] et Smilde et *al.* [87] ont exprimé des conditions suffisantes pour l'unicité de la solution, mais ces conditions sont trop restrictives pour être appliquables à nos données. La question de l'unicité de la décomposition NMF a été abordée par Donoho et Stodden [32]. Ce travail a permis de dégager une condition suffisante à l'unicité de la solution fournie par l'algorithme NMF.

Selon cette condition, si pour chaque source k, il existe une composante du vecteur $\mathbf{s}_k$ qui le caractérise, c'est-à-dire qui est nulle pour tout autre vecteur $\mathbf{s}'_k$, $k \neq k'$, il y a unicité de la solution. Dans notre cas, cela revient à pouvoir associer une longueur d'onde à chaque composant fluorescent du mélange. La fluorescence à cette longueur d'onde serait uniquement dûe au composant associé et le spectre des autres composants serait nul à cette longueur d'onde.

Cas des signaux de fluorescence.

Nous testons donc la séparabilité des signaux de fluorescence de biosenseurs bactériens par la méthode NMF. Pour cela, une matrice de signaux $\mathbf{X}$ est générée. L'approche NMF revient à factoriser la matrice d'observation $\mathbf{X}$ (2.1) en deux matrices avec contraintes de positivité sur $\mathbf{S}$ et/ou sur $\mathbf{A}$. Dans notre cas, les colonnes de la matrice $\mathbf{S}$ représentent les spectres et les colonnes $\mathbf{A}$ l'évolution de la quantité de protéines fluorescences produites en fonction de la proportion de l'élément stressant. Il est donc naturel d'appliquer des contraintes de positivité sur chacunes des deux matrices. La figure 2.1 montre les spectres des deux protéines fluorescentes du mélange et les coefficients de mélanges utilisés.

On peut voir que les sources spectrales, comme pour des sources spectrales réelles, n'ont pas de valeurs nulles. De plus, les gènes ont des réponses impliquant un mélange des sources pour toutes les valeurs du paramètre P_1. On en déduit que la condition d'unicité n'est pas valide, ce qui implique une infinité de solutions possibles dépendant des valeurs d'initialisation. La figure 2.3 présente une famille de dix solutions admissibles obtenues pour différentes initialisations.

Dans le cas des biosenseurs bactériens, le problème de séparation de sources de fluorescence sous la seule contrainte de non négativité n'admet pas de solution unique, car la condition de séparabilité n'est pas valide. Afin de pallier les problèmes liés à la non unicité de la NMF, il est nécessaire d'introduire des contraintes supplémentaires, c'est le cas de l'approche BPSS, développée par Moussaoui [63, 64], qui s'apparente à la méthode NMF avec cependant une différence essentielle liée à la technique d'optimisation employée.

2.2.3 Bayesian Positive Source Separation (BPSS)

La méthode de séparation BPSS [64] est une méthode de séparation par approche bayésienne. La méthode consiste à utiliser des densités gamma paramétrables comme *a priori* sur la distribution des sources et des coefficients de mélanges. Cela permet d'obtenir une solution particulière

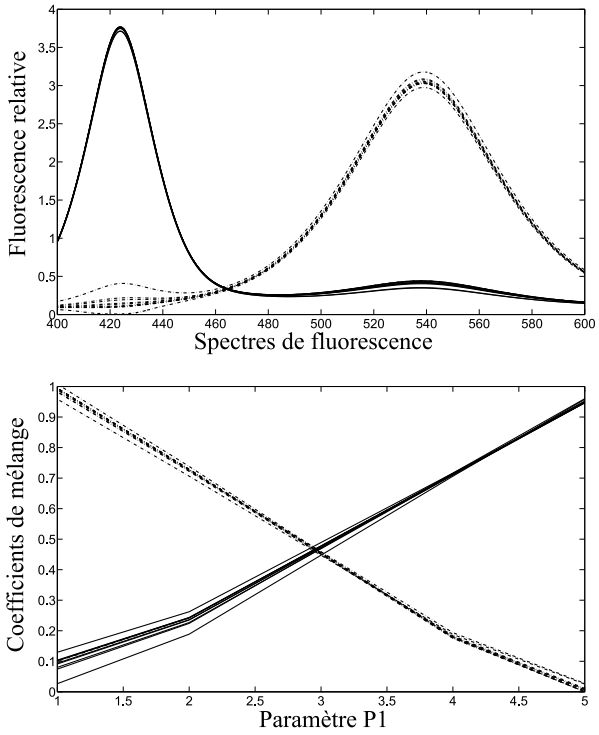


FIGURE *2.3 – La condition d'unicité de la décomposition NMF n'étant pas valide il existe une infinité de solutions admissibles. Nous avons représenté ici quelques unes des solutions admissibles obtenus par l'application de la méthode NMF sur les données générées par le mélange des spectres et coefficients de mélange représentés figure 2.1.*

parmi les solutions admissibles.

La figure 2.4 représente le résultat de la séparation par BPSS. Sur le premier graphique, on peut voir l'estimation des deux sources spectrales. La source $\mathbf{s}_1$ correspond à la source attendue. La source $\mathbf{s}_2$ montre deux pics car il existe encore un mélange entre les deux sources dans l'estimation de la source $\mathbf{s}_2$. L'étude de l'estimation des coefficients de mélange montre une sous estimation des coefficients pour les valeurs de P_1 où la source $\mathbf{s}_1$ est faible et où $\mathbf{s}_2$ est plus forte ce qui s'explique par la présence de la source $\mathbf{s}_1$ dans la source $\mathbf{s}_2$.

La solution trouvée par BPSS est unique mais ne correspond pas aux sources et aux coefficients de mélanges attendus, car les performances de la méthode BPSS dépendent de la capacité de la densité gamma à représenter la distribution des sources recherchées. Or, la densité gamma n'est pas adaptée pour modéliser des évolutions monotones des coefficients de mélanges comme c'est la cas dans les simulations, ce qui explique le piètre résultat de la décomposition BPSS. De plus, la distribution gamma est bien adaptée aux spectres Raman qui sont très piqués mais moins aux spectres de fluorescence qui sont plus mous.

Conclusion

Dans cette section, nous avons abordé le problème de séparation de sources bilinéaires. La contrainte d'orthogonalité imposée par la décomposition en valeurs singulières est inadaptée à

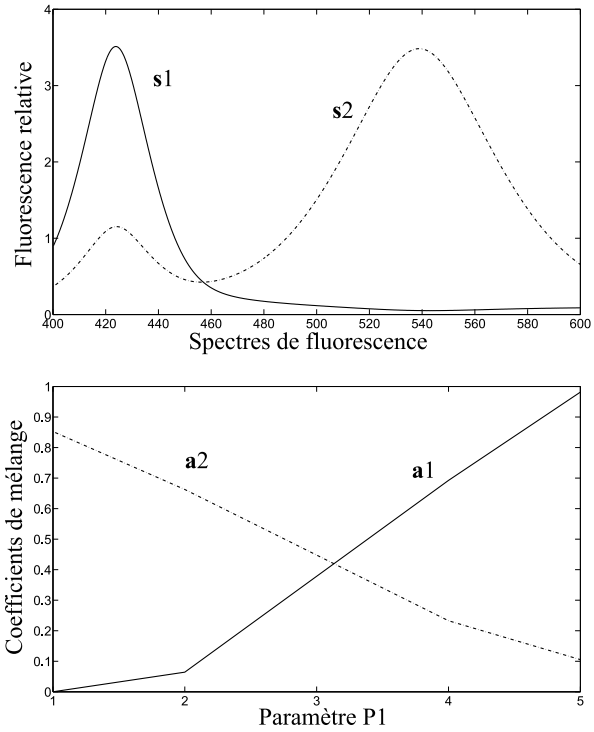


FIGURE 2.4 – *Représentation des résultats de la séparation de sources, par BPSS, des données générées par les mélanges présentés sur la figure 2.1. On retrouve sur le premier graphique l'estimation des sources spectrales ($\mathbf{s}_1$ et $\mathbf{s}_2$) des protéines fluorescentes et les coefficients de mélange associés ($\mathbf{a}_1$ et $\mathbf{a}_2$)*

des sources spectrales de fluorescence. Les estimations sous cette contrainte peuvent prendre des valeurs négatives rendant impossible l'interprétation physique. La méthode NMF utilise une contrainte de positivité pour conserver la réalité physique des signaux de fluorescence. Cette méthode n'est pourtant pas adaptée à l'étude de signaux issues des biosenseurs bactériens, car les réponses des gènes ne permettent pas d'assurer l'unicité de la solution. Nous avons essayé ensuite de résoudre le problème de séparation de sources en utilisant une approche bayésienne (BPSS). La méthode proposée consiste à utiliser une densité gamma comme modèle *a priori* sur la distribution des sources et des coefficients de mélange, ce qui permet d'obtenir une solution particulière parmi les solutions admissibles. Néanmoins, la représentation des sources spectrales par une distribution gamma n'est pas des plus adaptée. Afin de réduire le nombre de contraintes imposées sur les signaux sources et sur les coefficients de mélange, nous avons choisi, dans l'étude effectuée de rajouter une troisième/quatrième diversité aux données acquises. Cela permettra, de résoudre de façon naturelle le problème d'identification du modèle, grâce à une réduction du nombre de degrés de liberté du modèle. De plus, l'aspect multiparamétrique de ces méthodes permet d'analyser le comportement du système dans des milieux plus compliqués.

2.3 Décomposition trilinéaire

Dans la section suivante, nous étudions le cas de la décomposition trilinéaire de données de fluorescence. Les spectres de fluorescence mesurés sont regroupés sous forme de tenseur de données $\boldsymbol{\mathcal{X}}(I \times J \times K)$ contenant JK spectres de taille I correspondant à JK échantillons pour lesquels on aura J différentes valeurs de l'élément stressant et K différentes quantités de biosenseurs.

2.3.1 Modèle Candecomp/Parafac (CP)

Considérons un jeu de données $\boldsymbol{\mathcal{X}}$, contenant les mesures de fluorescence $x_{i,j,k}$ en fonction de trois paramètres i, j, k. Les vecteurs représentent $\mathbf{a}_f, \mathbf{b}_f$ et $\mathbf{s}_f$ l'évolution de la source de fluorescence f en fonction des paramètres i, j, k.

La décomposition CP d'un tenseur $\boldsymbol{\mathcal{X}}$ d'ordre 3 (tableau à 3 modes) en une somme de F tenseurs de rang un s'exprime de la manière suivante :

$$\boldsymbol{\mathcal{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{s}_f \tag{2.4}$$

où $\mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{s}_f$ est un tenseur de rang un pour lequel $\mathbf{a}_f(I \times 1)$, $\mathbf{b}_f(J \times 1)$ et $\mathbf{s}_f(K \times 1)$ sont des vecteurs représentant les composantes de chacun des modes (dimensions). L'entier F désigne le nombre de sources.

Cette décomposition, pour des tableaux à trois dimensions, a été proposée de manière indépendante par Carroll et Chang [23] et par Harshman [41], qui l'ont nommé, respectivement, Candecomp (CANonical DECOMPosition) et Parafac (PARAllel FACtor analysis). De part sa polyvalence et ses propriétés d'identifications intéressantes, la décomposition Candecomp/Parafac (CP) est largement utilisée dans divers domaines, tels que la chimiométrie [86, 14], l'industrie alimentaire, l'analyse numérique [1] et le traitement du signal [49].

Identifiabilité du modèle trilinéaire

La popularité de la décomposition CP est due, essentiellement, à ses propriétés d'identifiabilité. Jennrich est le premier à donner une preuve de l'unicité de la décomposition CP en montrant qu'une solution unique existe lorsque le rang des matrices de chaque mode est égal ou supérieur au nombre de composantes recherchées. Cependant, Harshman suggère par une approche empirique sur des données synthétiques [41] que les conditions requises par le théorème d'unicité de Jennrich sont plus fortes que nécessaires. L'approche empirique proposée consiste à générer des données avec des structures et propriétés mathématiques connues. On répète les décompositions à partir de différents points de départ lorsque le jeu de données converge sur la même solution de tous les points de départ, il est conclu que la solution est unique. Cette approche a permis à Harshman d'énoncer la condition non minimale de l'unicité selon laquelle la solution de CP est

unique lorsque deux des trois matrices sont de rang colonne plein et que la troisième n'a pas de colonne colinéaire.

Un assouplissement de cette condition d'unicité a été donné par Kruskal [51] qui a montré que, même si aucune des matrices de composants n'est de rang plein colonne, une solution peut-être unique. Si la condition

$$rang_k(\mathbf{A}) + rang_k(\mathbf{B}) + rang_k(\mathbf{C}) \geq 2F + 2 \tag{2.5}$$

est respectée, alors la décomposition du tenseur $\boldsymbol{\mathcal{X}}$ d'ordre 3 et de rang F est unique. Dans 2.5, le $rang_k(.)$ dénote le rang de Kruskal d'une matrice. Le $rang_k$ d'une matrice équivaut à la plus grande valeur m telle que toutes les sous-matrices de m colonnes soient de rang plein. La condition de Kruskal est nécessaire et suffisante pour les tenseurs de rang $R = 2$ ou 3. Ten Berge et Sidiropoulos ont montré [94], que pour des tenseurs de rang 4 ou plus l'unicité peut être obtenue, même si la condition de Kruskal n'est pas remplie.

Cependant, une dépendance linéaire entre les vecteurs de décomposition d'un même mode peut constituer une violation des conditions d'unicité. Harshman [41] a rapporté l'existence de cas où la détermination de vecteurs de certains modes est unique alors que certains vecteurs des autres modes sont sujets à l'indétermination par rotation. Ce phénomène a été appelé unicité partielle et des résultats relatifs à ce type d'unicité ont été proposés récemment dans [39].

D'après l'équation 2.4, si on permute les composantes ou si l'on multiplie $\mathbf{a}_f$ par γ_a, $\mathbf{b}_f$ par γ_b et $\mathbf{s}_f$ par γ_s avec $\gamma_a\gamma_b\gamma_s = 1$ cela ne change pas la qualité de la décomposition. Autrement dit, si ($\mathbf{A}$,$\mathbf{B}$,$\mathbf{S}$) sont les solutions de la décomposition CP, alors ($\mathbf{A\Pi\Gamma}_a$,$\mathbf{B\Pi\Gamma}_b$,$\mathbf{S\Pi\Gamma}_s$) sont également solutions, où $\mathbf{\Pi}$ est une matrice de permutation et $\mathbf{\Gamma}_a$,$\mathbf{\Gamma}_b$,$\mathbf{\Gamma}_s$ sont des matrices diagonales telles que $\mathbf{\Gamma}_a\mathbf{\Gamma}_b\mathbf{\Gamma}_s = I_R$. On dit alors que la décomposition CP est unique à une permutation et à un facteur d'échelle près ou essentiellement unique.

Algorithmes pour la décomposition d'un tableau de données tridimensionnel

Dans cette section, une brève présentation des algorithmes de décomposition CP est donnée. L'idée n'est pas de fournir une étude exhaustive des méthodes, ni de comparer tous les algorithmes disponibles mais de fournir les informations nécessaires à la compréhension de la méthode proposée. Certaines limitations de ces méthodes et éléments de réponses sur les choix effectués seront présentés.

Moindres carrées alternés (Alternating Least Squares (ALS)). Dans l'approche ALS, le problème non linéaire de l'ajustement du modèle CP est divisé en sous problèmes de moindres carrés linéaires qui sont résolus de façon itérative. A chaque étape, un sous-ensemble des paramètres du modèle est estimé et une itération est terminée lorsque tous les sous-problèmes ont été résolus. Le plus souvent, les sous-ensembles de paramètres représentent les matrices des modes et le nombre de sous-étapes est égal à l'ordre du tenseur.

La décomposition CP peut s'écrire sous forme matricielle en regroupant les vecteurs $\mathbf{a}_f$ dans

une matrice $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_F]$ et de la même manière les vecteurs $\mathbf{b}_f$ et $\mathbf{s}_f$ dans des matrices $\mathbf{B}$ et $\mathbf{S}$, *i.e.* $\boldsymbol{\mathcal{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{S}]\!]$. Le tenseur $\boldsymbol{\mathcal{X}}$ peut être déplié sous la forme d'une matrice de trois manières différentes :

$$\mathbf{X}_{(1)} = \mathbf{A}(\mathbf{S} \odot \mathbf{B})^T,$$

$$\mathbf{X}_{(2)} = \mathbf{B}(\mathbf{S} \odot \mathbf{A})^T,$$

$$\mathbf{X}_{(3)} = \mathbf{S}(\mathbf{B} \odot \mathbf{A})^T$$

où $\mathbf{X}_{(1)}$ est une matrice de taille $I \times JK$ représentant le tenseur $\boldsymbol{\mathcal{X}}$ déplié suivant le premier mode et $\mathbf{X}_{(2)}$, $\mathbf{X}_{(3)}$ sont dépliés suivant les modes 2 et 3, respectivement. Le symbole $\odot$ désigne le produit de Khatri-Rao (produit de Kronecker colonne par colonne de deux matrices). Une autre façon de représenter le modèle CP pour un tenseur d'ordre 3 est à l'aide des matrices représentant une tranche du tenseur :

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}^{(k)}\mathbf{B}^T, k = 1, \ldots, K$$

où $\mathbf{X}_k$ est la k^e tranche et $\mathbf{D}^{(k)} = diag(\mathbf{S}_k)$ est une matrice diagonale qui prend sur sa diagonale la ligne k de la matrice $\mathbf{S}$. Avec ces notations l'algorithme ALS pour la décomposition d'un tableau tridimensionnel est donné par la table 2.1.

	Entrée : $\boldsymbol{\mathcal{X}}$, λ, F
1:	Initialisation $\mathbf{A}, \mathbf{B}, \mathbf{S}$
2:	$\mathbf{A} = \mathbf{X}_{(1)}(\mathbf{S} \odot \mathbf{B}) \left[(\mathbf{B}^T\mathbf{B})(\mathbf{S}^T\mathbf{S})\right]^{-1}$
3:	$\mathbf{B} = \mathbf{X}_{(2)}(\mathbf{S} \odot \mathbf{A}) \left[(\mathbf{A}^T\mathbf{A})(\mathbf{S}^T\mathbf{S})\right]^{-1}$
4:	$\mathbf{S} = \mathbf{X}_{(3)}(\mathbf{B} \odot \mathbf{A}) \left[(\mathbf{A}^T\mathbf{A})(\mathbf{B}^T\mathbf{B})\right]^{-1}$
5:	Si condition d'arrêt non satisfaite, aller étape 2
	Sortie : Estimation de $\mathbf{A}, \mathbf{B}, \mathbf{S}$

TABLE *2.1 – algorithme ALS-PARAFAC*

La méthode ALS devient moins efficace avec l'accroissement de l'ordre du tenseur, tandis que la complexité de ALS croit linéairement avec le nombre de facteurs. La convergence de la méthode ALS peut être très lente en particulier dans le cas où les facteurs sont fortement cohérents. Plusieurs modifications ont été proposées pour accélérer la convergence de la méthode ALS, la plus utilisée étant la recherche linéaire (Line Search). Elle extrapôle à partir des itérations précédentes l'évolution des paramètres afin d'estimer leurs valeurs après plusieurs itérations. Une extension de cette approche, proposée par Rajih, Comon et Harshman [72], intègre une extrapolation plus sophistiquée utilisant simultanément les tendances de tous les paramètres.

Alternating Slicewise Diagonalization (ASD). La méthode ASD [46] est une alternative à l'approche ALS lorsque le temps de calcul est important. Dans cette méthode, on emploie un

procédé de compression basé sur la décomposition en valeurs singulières permettant ainsi de réduire le nombre d'opérations par itération. Cependant, cette méthode, bien que plus rapide, semble être inférieure aux autres méthodes du point de vue de la qualité des solutions [97].

dGN et PMF3. Les méthodes dGN et PMF3 sont toutes deux basées sur l'optimisation simultanée de toutes les matrices. D'après Tomasi et Bro [97], ces méthodes sont supérieures à ALS en termes de propriétés de convergence mais beaucoup plus coûteuses en termes de mémoire et de temps de calcul.

Contrainte de positivité dans la décomposition trilinéaire

L'application d'une contrainte de positivité pour l'ajustement de modèles de fluorescence peut avoir une importance pratique. En effet, un modèle qui prend en compte la positivité de propriétés physiques, telles que la concentration ou l'absorption, est plus réaliste et facilite l'interprétation des résultats. De plus, la contrainte de positivité réduit l'espace des solutions admissibles et donc, dans certains cas, fournit l'unicité. La contrainte de positivité est une contrainte que l'on peut facilement intégrer dans l'algorithme ALS. Un algorithme rapide pour ALS avec contrainte de positivités a été proposé par Bro [16].

La boîte à outils utilisée est la toolbox N-Way [4] pour MATLAB. Cette boîte à outils fournit un grand nombres d'algorithmes de calcul de décompositions tensorielles. En outre, la plupart des méthodes peuvent gérer les contraintes (*e.g.* positivité).

Nombre de sources

Dans la décomposition CP, le choix du nombre de sources est une étape essentielle, car à la différence de la SVD, les sources d'un modèle d'ordre N sont différentes des sources d'un modèle d'ordre N-1. De plus, le critère de convergence (erreur entre les données vraies et les données reconstruites) ne permet pas de dire si une composante supplémentaire améliore la modélisation du signal de fluorescence ou celle du bruit. Par conséquent, plus le nombre de sources est grand plus le modèle est proche des données. Le meilleur ordre de décomposition est difficile à déterminer. Il existe des méthodes pour aider à la détermination du nombre de sources mais aucune ne résoud définitivement le problème. Nous donnons par la suite deux approches permettant de guider l'utilisateur sur le choix du nombre de sources.

La première méthode consiste à analyser la décroissance des valeurs singulières du tenseur $\boldsymbol{\mathcal{X}}$ déplié suivant un mode. Le nombre de sources recherchées correspond alors aux F premières valeurs singulières non négligeables. Cependant, cette méthode peut être insuffisante dans le cas où une des sources spectrales à une fluorescence beaucoup plus faible que les autres, dans ce cas elle peut être considérée à tort comme négligeable.

La seconde méthode consiste à analyser l'évolution de l'erreur de reconstruction du modèle en fonction du nombre de sources recherchées. Théoriquement, on devrait observer une diminution de la vitesse de décroissance de l'erreur autour du nombre de sources adéquat, car l'erreur diminue rapidement jusqu'au vrai nombre de sources puis diminue plus lentement par la suite car le

signal sera modélisé et les sources supplémentaires chercheront à modéliser le bruit. Cependant, d'un point de vue pratique, cette méthode est coûteuse car elle nécessite de calculer plusieurs décompositions.

Validation du modèle

Dans cette section, nous nous intéressons à quelques unes des méthodes permettant d'analyser la validité de la décomposition obtenue par la méthode CP. Ces méthodes sont :

- Split Half Analysis
- Core Consistency Diagnostic (CorConDia)

La méthode Split Half Analsysis [42] consiste à analyser indépendemment différents sous-ensembles de données. Si l'ordre de décomposition choisi est correct, on obtiendra le même modèle pour chacun des sous ensembles de données (aux permutations et facteurs d'échelle près).

La méthode CorConDia [18] utilise le fait que le modèle CP peut être écrit comme un modèle TUCKER dont les seuls éléments non nuls du tenseur cœur se trouvent sur son hyperdiagonale. Un modèle CP est alors considéré comme approprié si l'ajout d'interactions entre les composants des différents modes (*i.e.* l'ajout des éléments sur l'hyperdiagonale) n'améliorent pas l'ajustement du modèle. Dans le cas contraire, l'amélioration s'expliquerait par le fait que les données doivent être modélisées par un modèle prenant en compte les interactions, ou que l'ordre de la décomposition est trop important et donc qu'une ou plusieurs sources contribuent plus ou moins à toutes les combinaisons du modèle. Dans les deux cas, le modèle CP est inapproprié.

La méthode proposée par Bro et Kiers [18] consiste à calculer un indice de pertinence du modèle CP. On estime le tenseur cœur du modèle de TUCKER à partir des composantes du modèle CP, puis on calcule le rapport de cohérence entre le tenseur cœur et le tenseur théorique hyperdiagonal. Le rapport de cohérence du cœur est calculé par

$$\textit{cohérence du cœur} = 100\left(1 - \frac{\sum^F \sum^F \sum^F (\mathcal{G}_{ijk} - \mathcal{T}_{ijk})^2}{F}\right) \tag{2.6}$$

où $\mathcal{G}_{ijk}$ est un élément du cœur d'interactions estimé et $\mathcal{T}_{ijk}$ le cœur théorique. Si l'on augmente progressivement l'ordre de la décomposition, la cohérence du cœur diminue progressivement parce que l'influence du bruit augmente avec le nombre de composantes. Lorsque l'on dépasse le nombre de sources vraies, la cohérence du cœur décroit plus rapidement.

La simulation suivante consiste à étudier l'indice de cohérence du cœur pour différentes valeurs du nombre de sources recherchées et pour différents niveaux de bruit. Considérons un exemple simple de spectrofluorescence. On mesure les spectres d'émission de fluorescence (81 points) d'un biosenseur pouvant produire trois molécules fluorescentes en réponse à deux éléments (*i.e.* deux métaux) dont les concentrations varient (6 points). Le tenseur de données est de dimension $81 \times 6 \times 6$. Le mélange est ensuite bruité par un bruit additif gaussien de moyenne nulle dont le rapport signal sur bruit est compris entre -20 et 20 dB. Le tableau 2.2 donne l'évolution de l'indice de cohérence du cœur en fonction du nombre de sources recherchées et du rapport signal

à bruit.

SNR (dB) \ Nb sources	1	2	**3**	4	5	6	7
-20	100	100	99	99	99	99	80
-10	100	99	99	99	99	98	20
0	100	99	99	99	99	99	39
10	100	100	99	99	99	98	29
20	100	100	99	99	99	98	42

TABLE 2.2 – *Comparaison des valeurs de l'indice de cohérence du cœur pour la décomposition CP d'un jeu de données bruitées simulant le mélange de trois sources spectrales.*

On constate une diminution de l'indice à partir de sept sources recherchées et ceci quelque soit le niveau de bruit. Aucune évolution significative ne permet de retrouver le nombre de sources utilisées. L'indice CorConDia ne peut pas être utilisé pour déterminer si l'ordre de décomposition choisi est le meilleur mais une valeur proche de 100 est une condition nécessaire pour valider l'adéquation aux données du modèle CP.

En pratique, on cherche à retrouver un modèle physique interprétable. On utilisera alors les connaissances *a priori* sur le système étudié. Les spectres des molécules, la réponse des gènes ou la quantité de biosenseurs peuvent être connues et fournir une information suffisante pour choisir le meilleur ordre de décomposition.

2.4 Exemple sur des signaux réels de fluorescence de biosenseurs bactériens

Après avoir présenté les méthodes dont nous disposons pour effectuer la séparation des signaux de fluorescence, nous présentons un exemple d'application sur une expérience basée sur différents biosenseurs bactérien dont la fluorescence est dite constitutive. Pour cela, quatre biosenseurs ont été conçus pour produire chacun en permanence une protéine fluorescente différente. Les quatre protéines fluorescentes sont GFP, YFP, E2Orange et DsRedExpress2 et leurs spectres respectifs sont notés $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$ et $\mathbf{s}_4$. Nous proposons de simuler les variations de fluorescence suivant deux paramètres P_1 et P_2 en effectuant des mélanges de biosenseurs. Ces deux paramètres pourraient, par exemple, représenter la proportion d'un élément stressant pour P_1 et les variations du ratio de biosenseurs pour P_2.

Le modèle théorique de l'évolution de la fluorescence pour les quatre biosenseurs en fonction de la longueur d'ondes d'émission et des paramètres P_1 et P_2 est représenté sur la figure 2.5.

Cette expérience permet de tester les différentes méthodes sur des données réelles parfaitement maîtrisées. Le plan d'expérience proposé génère 35 mélanges différents de biosenseurs (7 suivant le paramètre P_1 et 5 suivant le paramètre P_2). La fluorescence émise par chaque échantillon est mesurée de 450 à 600 nm avec un pas de 2 nm, soit 76 points. Les données de fluorescence sont regroupées dans un tenseur $\boldsymbol{\mathcal{X}}$ de dimension $76 \times 7 \times 5$. La figure 2.6 représente

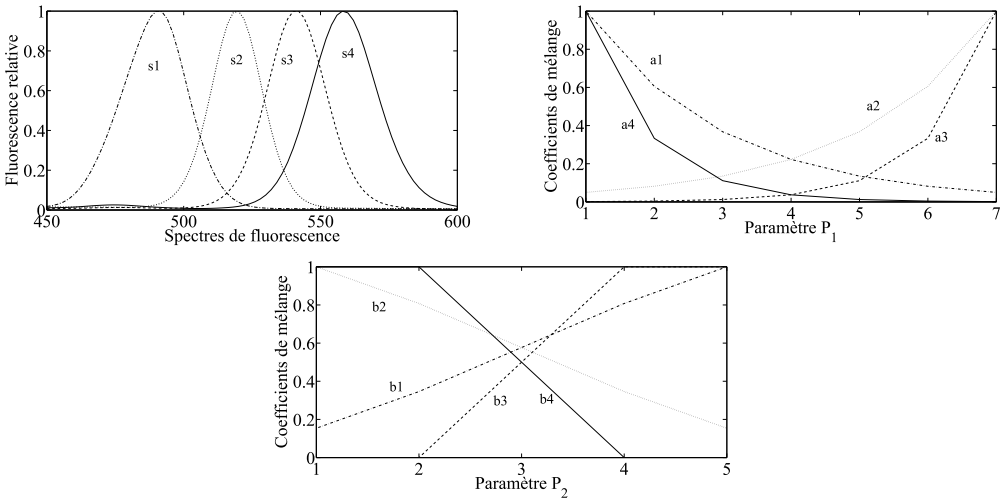


FIGURE *2.5 – Représentation des mélanges effectuées pour simuler les réponses de quatre biosenseurs répondant à un paramètre P_1. Les variations en fonction du paramètre P_2 correspondent à différents ratios des biosenseurs.*

les 35 spectres (de taille 76 points) mesurés. On peut y voir l'évolution de la fluorescence des différentes sources en fonction des deux paramètres P_1 et P_2.

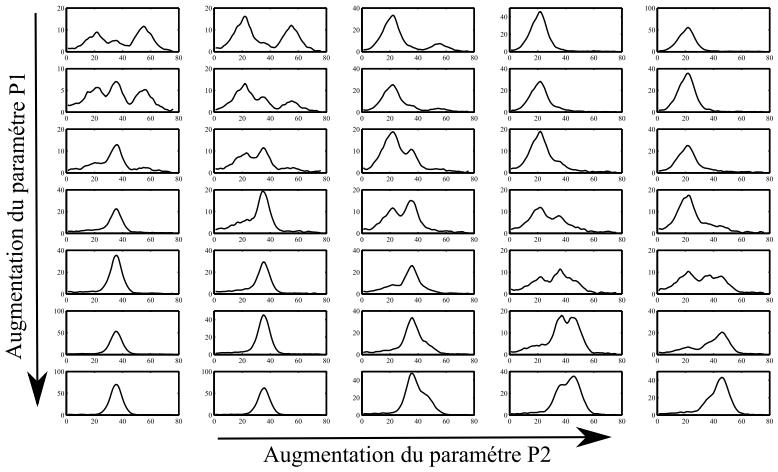


FIGURE *2.6 – Spectres synchrones de fluorescence mesurés (λ de 450 à 600 nm pas de 2 nm) pour différents mélanges de biosenseurs. Le modèle théorique des mélanges est représenté sur la figure 2.5.*

2.4.1 Décomposition bilinéaire

Nous proposons de tester plusieurs méthodes de séparation de sources bilinéaires afin d'illustrer les problèmes pouvant être recontrés et démontrer l'intérêt de passer à des méthodes trilinéaires. Pour appliquer les méthodes bilinéaires, nous représentons le tenseur $\boldsymbol{\mathcal{X}}$ sous sa forme matricielle :

$$\mathbf{X}_{(3)} = \mathbf{S}(\mathbf{B} \odot \mathbf{A})^T$$

où $\mathbf{X}_{(3)}(76 \times 35)$ est la matrice formée par la concaténation des 35 spectres mesurés et $\mathbf{S}$,$\mathbf{A}$ et $\mathbf{B}$ sont les matrices contenant respectivement les spectres et coefficients de mélange en fonction du paramètre P_1 et P_2 des F composants. Nous avons donc un problème bilinéaire standard où $(\mathbf{B} \odot \mathbf{A})$ est une matrice des mélanges et $\mathbf{S}$ la matrice des sources spectrales. Les F colonnes de la matrice de mélanges sont formées par le produit de Khatri-Rao des matrices $\mathbf{B}$ et $\mathbf{A}$. Chaque colonne peut être transformé en une matrice $J \times K$ dont la décomposition bilinéaire permet d'estimer deux vecteurs représentant respectivement une colonne de $\mathbf{A}$ et une colonne de $\mathbf{B}$.

Ce qui donne l'algorithme suivant :

1. Estimation de la matrice des spectres $\mathbf{S}$ et de la matrice des mélanges $(\mathbf{B} \odot \mathbf{A})$ par une méthode de décomposition bilinéaire (SVD, BPSS, *etc.*).
2. Décomposition des F matrices formées par les F colonnes de la matrice des mélanges pour estimer chacune des colonnes de $\mathbf{A}$ et $\mathbf{B}$.

Estimation du nombre de sources

L'analyse de la décroissance des valeurs singulières est souvent utilisée pour déterminer le nombre de sources du milieu à séparer, bien que cette méthode soit discutable car fortement influencée par le niveau de bruit et l'amplitude des sources. Dans le cadre de cette expérience, nous connaissons le nombre de sources à rechercher ce qui nous permet de tester cette méthode.

La figure 2.7 représente les valeurs singulières calculées sur les matrices issues du dépliement du tenseur $\boldsymbol{\mathcal{X}}$ suivant chacun des modes. On peut voir qu'un plateau est rapidement atteint et que trois sources sont au moins nécessaire pour représenter les données. On voit également que l'influence des sources supérieures à la cinquième est négligeable. La décroissance des valeurs singulières ne permet pas de déterminer le nombre exact de sources mais, dans notre cas, elle fournit un intervalle encadrant le bon nombre de sources.

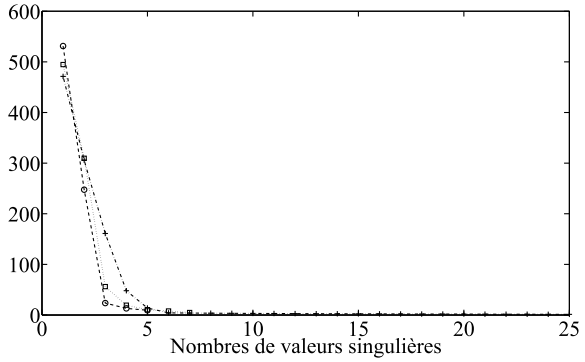


FIGURE *2.7 – Valeurs singulières de la décomposition SVD des matrices* $\mathbf{X}_{(1)}$ *(+),* $\mathbf{X}_{(2)}$ *(◇) et* $\mathbf{X}_{(3)}$ *(○).*

Application de la SVD

En appliquant la SVD sur nos données, nous analysons les 4 premières valeurs singulières

qui donnent une information sur la contribution des quatre différentes sources spectrales prépondérantes. La figure 2.8 représente les quatre sources spectrales et les coefficients de mélange associés en fonction des paramètres P_1 et P_2.

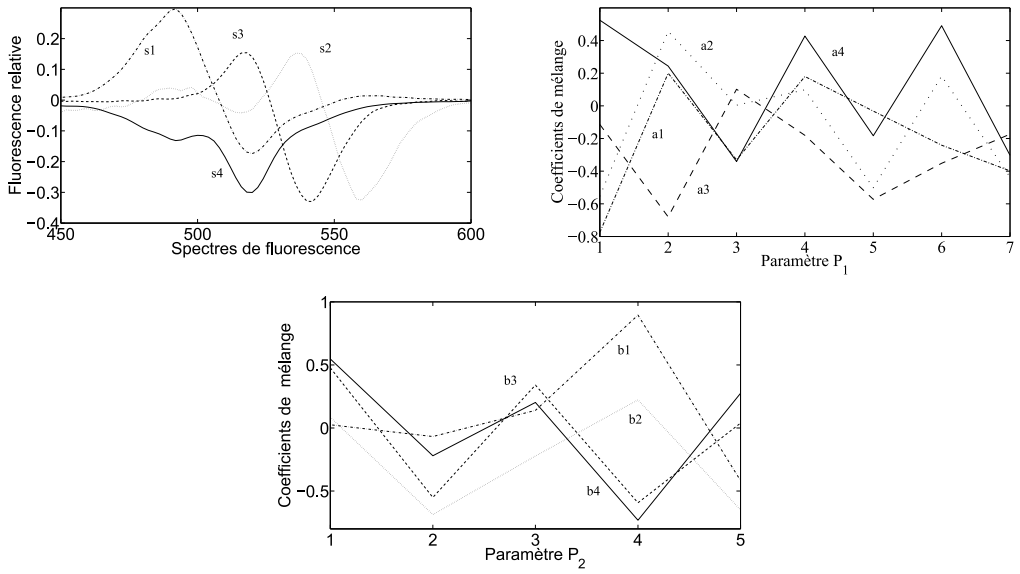


FIGURE *2.8 – Décomposition SVD du tenseur de données d'ordre 3. Deux décompositions successives ont permis d'estimer les sources spectrales et les coefficients de mélange puis la décomposition SVD des coefficients de mélange ont permis l'estimation des coefficients de mélange en fonction des paramètres P_1 et P_2.*

Le premier graphique représente l'évolution de la fluorescence en fonction de la longueur d'onde. Les quatre spectres de fluorescence estimés sont désignée par $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{s}_3$ et $\mathbf{s}_4$. On remarque que les sources spectrales initiales ne sont pas bien estimées et qu'il est difficile d'associer les spectres aux protéines fluorescentes.

Le second graphique représente l'évolution de l'intensité de fluorescence en fonction du paramètre P_1. On peut remarquer que les mélanges sont beaucoup plus chahutés que les mélanges initiaux.

De même, l'évolution des coefficients de mélange en fonction du paramètre P_2 est éloigné de l'évolution attendue.

Pour conclure, on comprend aisément que la séparation par la SVD n'est pas adaptée dans le cas de la séparation de sources de signaux de fluorescence de biosenseurs bactériens. La contrainte d'orthogonalité n'est pas applicable aux sources de fluorescence utilisées, ce qui empêche l'identification des protéines fluorescentes et l'étude du comportement des biosenseurs.

Application de la NMF

Du fait que la contrainte d'orthogonalité n'est pas adaptée aux spectres de fluorescence, l'idée d'appliquer une contrainte de positivité semble plus adaptée car plus proche de la réalité physique des signaux. En appliquant la NMF sur nos données, nous chercherons les quatre sources et les coefficients de mélange associés qui respectent cette contrainte de positivité. La figure 2.9

représente les quatre sources spectrales et les coefficients de mélange associés en fonction des paramètres P_1 et P_2.

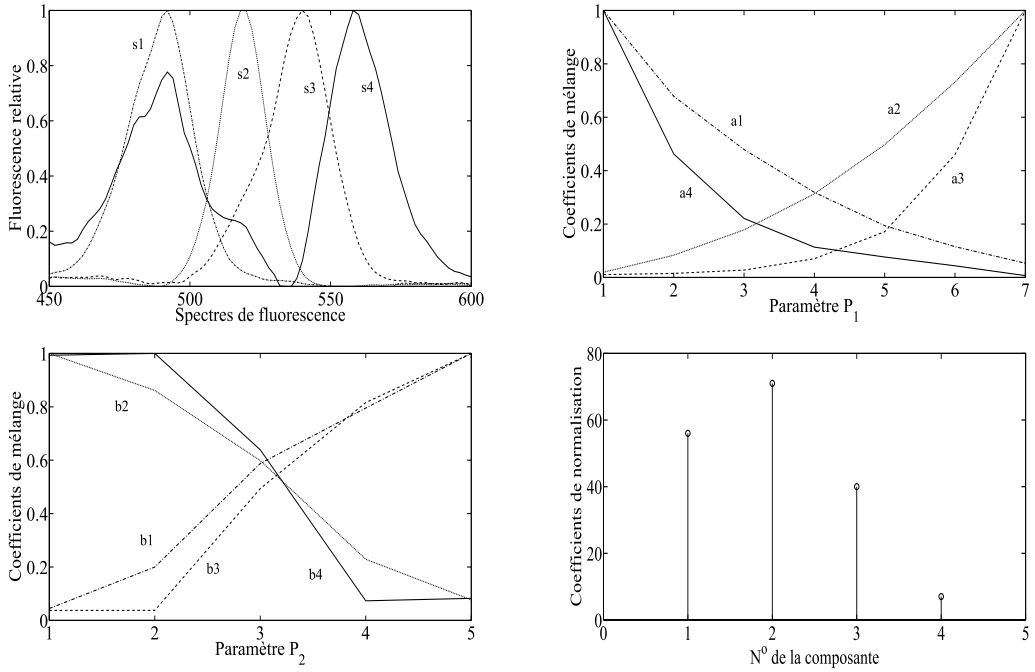


FIGURE *2.9 – Décomposition NMF du tenseur de données d'ordre 3. Les décompositions successives ont permis d'estimer les sources spectrales et les coefficients de mélange puis la décomposition NMF des coefficients de mélange a permis l'estimation des coeffcients de mélange en fonction des paramètres P_1 et P_2. Le quatrième graphique représente les coefficients de normalisation.*

Le premier graphique représente l'évolution de l'intensité de fluorescence en fonction de la longueur d'onde. Les quatre spectres estimés ont respectivement leur maxima à 490, 520, 540 et 558 nm. Les spectres $\mathbf{s}_1$ et $\mathbf{s}_2$ sont proches des sources spectrales initiales. Les spectres $\mathbf{s}_3$ et $\mathbf{s}_4$ montrent qu'il existe encore un mélange entre les spectres ce qui se manifeste par un pic autour de 490 nm dans le spectre $\mathbf{s}_4$ et dans le pied du pic de fluorescence du spectre $\mathbf{s}_3$ correspondant à un mélange avec la spectre $\mathbf{s}_2$.

Le second graphique représente l'évolution de l'intensité de fluorescence en fonction du paramètre P_1. Les estimations des coefficients de mélange des quatre sources sont proches des mélanges réels.

Le troisième graphique représente l'évolution de l'intensité de fluorescence en fonction du paramètre P_2. L'allure des courbes estimées pour les mélanges suivant le paramètre P_2 sont proches de celles attendues mais les valeurs normalement nulles sont surévaluées.

Pour conclure, l'estimation des sources spectrales montre que la séparation des différentes sources n'est pas parfaite et qu'il subsiste un mélange que l'on retrouve dans les différents coefficients de mélange où certains coefficients sont surestimés.

Application de BPSS

La figure 2.10 représente le résultat de la décomposition de nos données par la méthode BPSS.

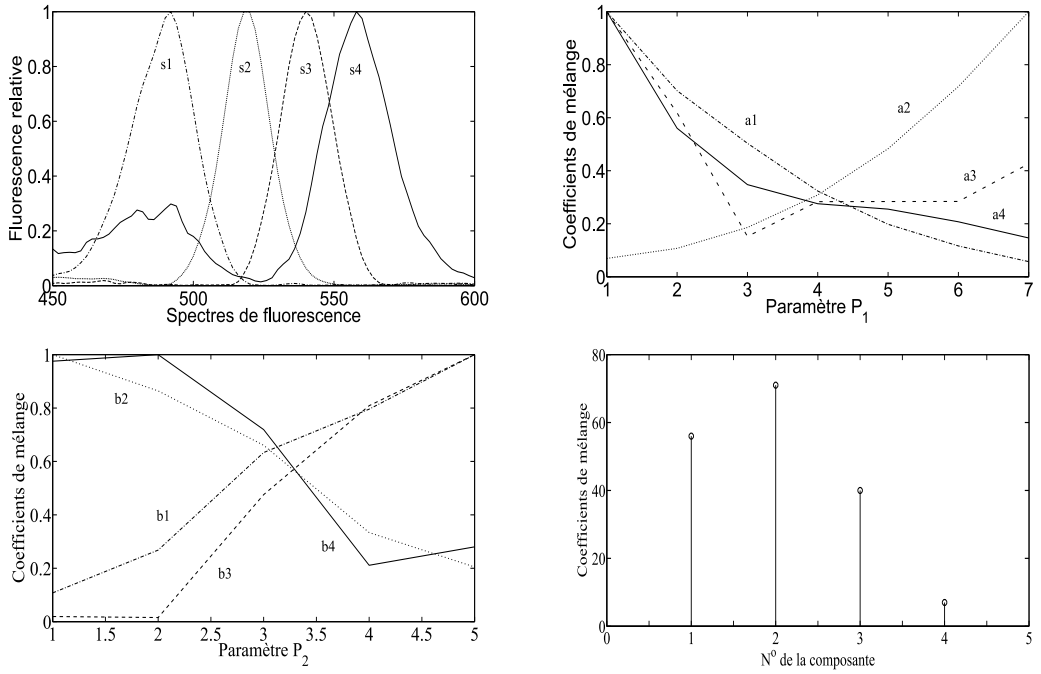


FIGURE *2.10 – Décomposition BPSS du tenseur de données d'ordre 3. Les décompositions successives ont permis d'estimer les sources spectrales et les coefficients de mélange puis la décomposition BPSS des coefficients de mélange ont permis l'estimation des coeffcients de mélange en fonction des paramètres P_1 et P_2. Le quatrième graphique représente les coefficients de normalisation.*

Le premier graphique représente l'évolution de la fluorescence en fonction de la longueur d'onde. Les trois premières sources sont bien estimées et permettent, grâce aux spectres de références, d'associer la source $\#_1$ à la protéine GFP, la source $\#_2$ à la protéine YFP et la source $\#_3$ à la protéine E2Orange.

Le second graphique représente l'évolution de la fluorescence en fonction du paramètre P_1. Les coefficients de mélange $\mathbf{a}_1$ et $\mathbf{a}_2$ sont bien estimés. Les coefficients de mélange $\mathbf{a}_4$ sont surestimés car la source $\#_4$ est mal estimée et contient une partie de la source $\#_1$.

Le troisième graphique représente l'évolution de la fluorescence en fonction du paramètre P_2. Les coefficients de mélange $\mathbf{b}_4$ sont surestimés notamment pour les plus grandes valeurs de P_2 ce qui entraîne une mauvaise estimation des autres coefficients de mélanges.

Les mauvaises estimations des réponses en fonction des paramètres P_1 et P_2 sont dues à l'inadéquation des lois gamma aux profils de réponses utilisées.

2.4.2 Application de la décomposition CP

Ordre

Comme nous l'avons présenté précédement, il n'existe pas de méthode permettant de déterminer, avec précision, *a priori* le nombre de sources à rechercher, hormis en utilisant les

connaissances que nous possédons sur la construction du jeu de données. Dans le cadre de cette étude, il existe au moins 4 sources spectrales (4 biosenseurs produisant chacun une protéine fluorescente différente). Le tracé de l'évolution de l'erreur de reconstruction en fonction du nombre de sources (Voir fig. 2.11) ne permet pas de déterminer le nombre de sources le plus adapté. Le tableau 2.3 regroupe différents indices calculés pour des nombres différents de sources recherchées. Les valeurs CorConDia des décompositions permettent de valider le modèle jusqu'à sept sources recherchées. La fraction d'énergie résiduelle, calculée par le rapport de l'énergie de l'erreur de reconstruction sur l'énergie des données, montre qu'au moins quatre sources sont nécessaires. L'évolution de l'erreur d'estimation montre que l'ajout de source au delà de cinq n'améliore pas significativement la représentation des données.

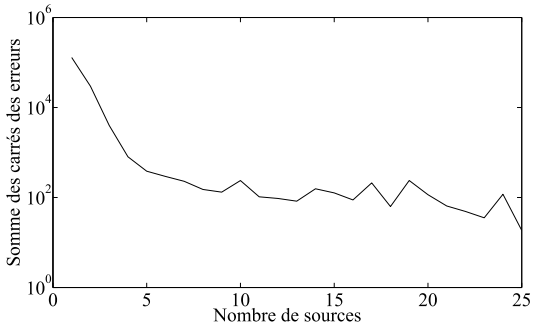


FIGURE 2.11 – *Évolution de l'erreur en fonction du nombre de sources recherchées.*

Nb sources	1	2	3	4	5	6	7	8
CorConDia	100	100	99	99	99	95	98	69
Erreur	129250	29700	7830	783	380	370	310	840
Fraction d'énergie résiduelle	38%	9%	2%	0.2%	0.1%	0.1%	0.1%	0.2%

TABLE 2.3 – *Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie expliquée, rapport d'amélioration de l'erreur par l'ajout d'une source supplémentaire pour la décomposition CP du jeu de données pour des nombres différents de sources recherchées.*

Résultats

Les résultats des deux décompositions (4 et 5 sources recherchées) sont représentés sur les figures 2.12 et 2.13. Chacune des F sources est normalisée par un coefficient qui minimise l'erreur quadratique entre la composante estimée et la composante théorique. Le coefficient de normalisation obtenu suite à la normalisation des trois modes permet de connaitre l'importance des sources.

La figure 2.12 montre le résultat de la décomposition CP des données pour quatre sources. On peut voir l'évolution de l'intensité d'une source spectrale (Fig. 2.12(a)) en fonction du paramètre P_1 (Fig. 2.12(b)) et du paramètre P_2 (Fig. 2.12(c)). La décomposition montre l'existence de quatre sources spectrales distinctes dont l'évolution en fonction des paramètres P_1 et P_2 correspond aux

mélanges effectués. Prenons, par exemple, la source spectrale ayant son maximum à 556 nm (source spectrale la plus à droite). Son évolution en fonction du paramètre P_1 correspond à une dilution en cascade au 1/2 et son évolution en fonction du paramètre P_2 correspond à une diminution à partir de la troisième valeur pour passer de un dans les deux premiers mélanges à zéro dans les deux derniers. Le graphique (Fig. 2.12(d)) montre les coefficients de normalisation des différentes sources.

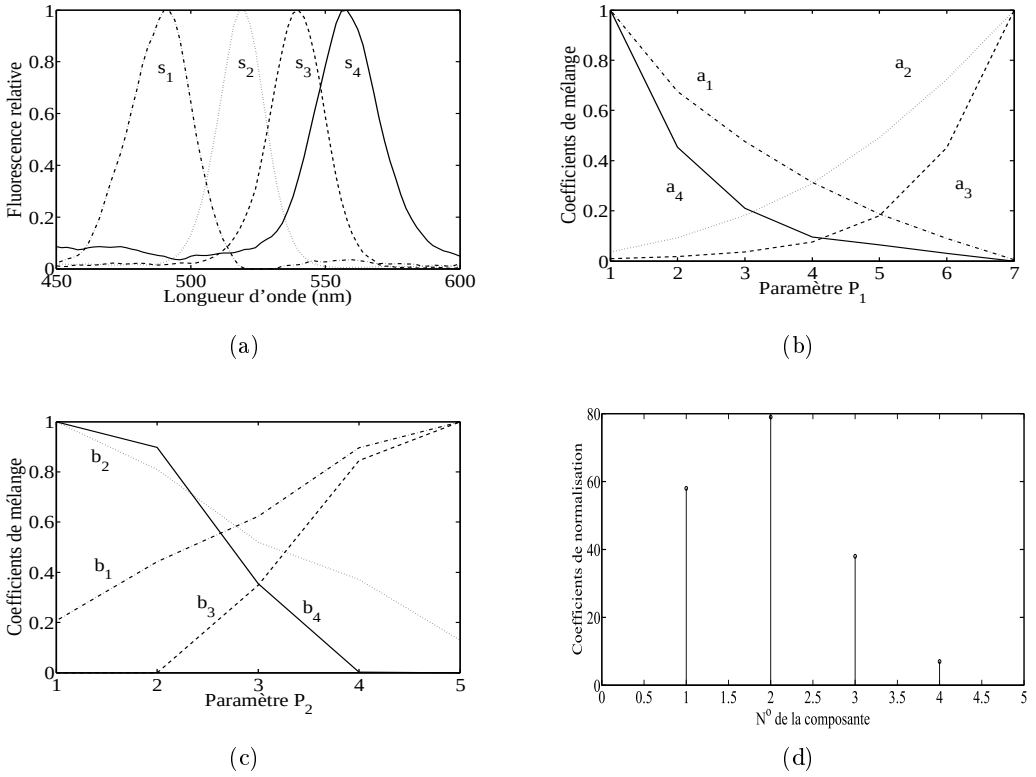


FIGURE *2.12 – Décomposition CP des données pour quatre sources. (a) Spectres de fluorescence des quatre protéines. (b) Réponse en fonction du paramètre P_1. (c) Réponse en fonction du paramètre P_2. (d) Coefficients de normalisations obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

La figure 2.13 montre le résultat de la décomposition CP des données en considérant cinq sources. On voit que les quatre premières sources spectrales sont distinctes et que le spectre $\mathbf{s}_5$ semble être un mélange des spectres $\mathbf{s}_1$ et $\mathbf{s}_2$. Les coefficients de mélange $\mathbf{a}_5$ de la source $\#_5$ sont nuls pour les valeurs extrêmes de P_1 alors que les coefficients $\mathbf{b}_5$ croit sur les premières valeurs de P_2. Les coefficients de normalisation montrent que cette cinquième source est du même ordre d'importance que la DsRed (source $\#_4$).

Aucune méthode ne nous permet de déterminer si la décomposition CP par 4 sources représente mieux ou moins bien les données que la décomposition 5 sources. Dans les deux cas les sources que l'on maîtrise sont bien estimées et correspondent au modèle théorique de mélange.

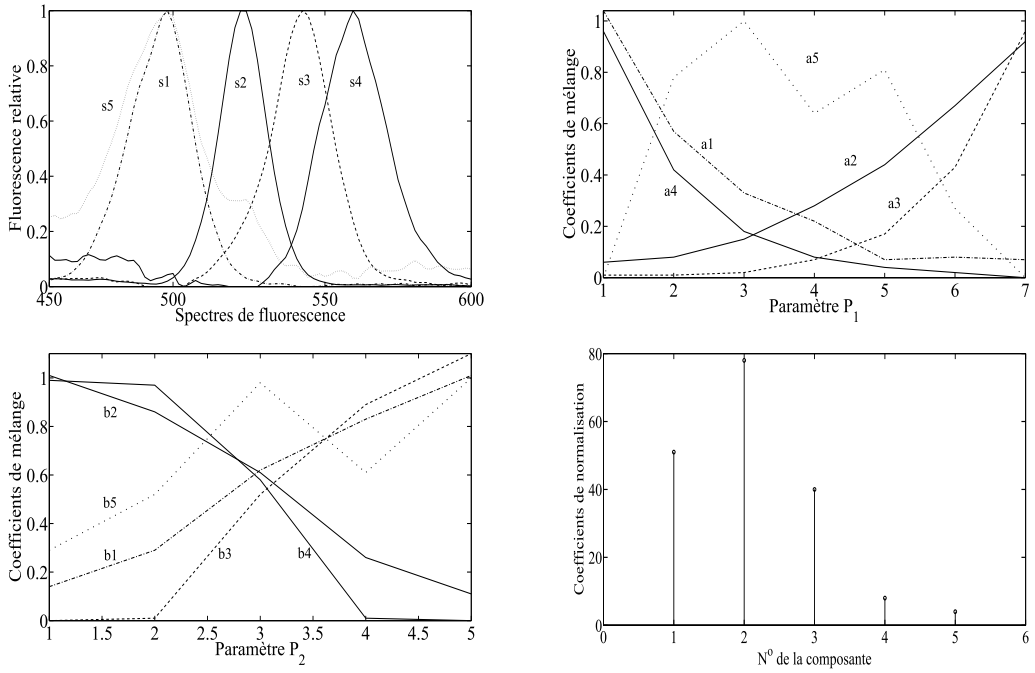


FIGURE *2.13 – Décomposition CP des données pour cinq sources.(a) Spectres de fluorescence. (b) Réponse en fonction du paramètre P_1. (c) Réponse en fonction du paramètre P_2. (d) Coefficients de normalisations obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

La présence de cette cinquième source pourrait s'expliquer par un artefact de mesure ou d'expérimentation qui éloignerait le modèle théorique du modèle réel mesuré. Dans cette expérience, on a fait varier les mélanges de biosenseurs de manière à suivre un modèle théorique offrant une grande diversité mais la quantité de biosenseur n'est pas la même dans chacun des puits. La cinquième source pourrait provenir d'un phénomène d'écrantage dû aux variations de quantité de biosenseurs entre les puits ou d'interactions inattendues.

Conclusion

L'étude de ce problème de fluorescence de biosenseurs bactériens par les différentes méthodes bilinéaires (SVD, NMF et BPSS) et trilinéaire (CP) a permis de mettre en évidence les différents problèmes liés aux méthodes bilinéaires. On a pu voir que la méthode SVD ne fournit pas de solutions interprétable car la contrainte d'orthogonalité n'est pas adaptée aux signaux de fluorescence. Nous avons vu également que la méthode NMF n'a pas permis de séparer totalement les sources spectrales. Un mélange persiste ce qui influence l'estimation des coefficients de mélange, notamment en surestimant les valeurs proches de zéro. L'estimation des coefficients de mélange par la méthode BPSS montre que la loi gamma n'est pas adaptée pour des profils monotones de variations.

La méthode CP permet une bonne estimation rendant possible l'identification des différentes sources et l'étude de leurs variations en fonction de plusieurs paramètres. Cela s'explique par le fait que dans le cas de la décomposition CP, les seules hypothèses faites sont liées à la structure

multilinéaire du mélange, bien qu'il subsiste quand même le problème de la détermination du nombre de sources recherchées.

2.5 Unicité partielle

Dans cette section, nous abordons le problème d'unicité de la décomposition CP lorsqu'il y a dépendance linéaire entre les vecteurs colonnes des matrices des trois modes.

Le problème de non unicité de la décomposition se pose lorsqu'une des matrices $\mathbf{A}$, $\mathbf{B}$ ou $\mathbf{S}$ a deux colonnes proportionnelles, par exemple, lorsque $F = 2$ et $\mathbf{S} = [\mathbf{s}, \gamma\mathbf{s}]$ où γ est un scalaire et $\mathbf{s}$ un vecteur $\mathbf{s} = [s_1, s_2, ..., s_k]^T$. D'après 2.3.1, pour toutes les tranches d'un tenseur $\boldsymbol{\mathcal{X}}$, on a $\mathbf{X}_k = \mathbf{A}diag(\mathbf{S}_k)\mathbf{B}^T = [\mathbf{a_1 s_k a_2}\gamma\mathbf{s_k}]\mathbf{B}^T = \mathbf{s}_k[\mathbf{a}_1\gamma\mathbf{a}_2]\mathbf{B}^T = \mathbf{s}_k[\mathbf{a}_1\gamma\mathbf{a}_2]\mathbf{W}\mathbf{W}^{-1}\mathbf{B}^T$ pour toute matrice $\mathbf{W}$ inversible. Si l'on choisit $\mathbf{W}$ autre que le produit d'une diagonale ou d'une matrice de permutation, on voit qu'il existe des solutions alternatives à $\mathbf{A}$ avec $[\mathbf{a}_1\gamma\mathbf{a}_2]\mathbf{W}$. Cet argument peut être étendu au cas où $F > 2$, si deux colonnes de $\mathbf{S}$ sont proportionnnelles, alors les colonnes composantes de $\mathbf{A}$ et $\mathbf{B}$ peuvent être mélangées sans perte d'ajustement. On considère, par exemple, un biosenseur dont deux gènes sont instrumentés de manière à générer deux protéines fluorescentes différentes. On suppose que les réponses, notées $\mathbf{a}_1$ et $\mathbf{a}_2$, de ces deux gènes à un paramètre P_1, sont très différentes (voir fig. 2.14), alors que les réponses, notées $\mathbf{b}_1$ et $\mathbf{b}_2$, à un paramètre P_2, sont linéairement dépendantes (dans l'exemple $\mathbf{b}_2 = 2\mathbf{b}_1$). Les spectres de fluorescence de ces protéines sont notés $\mathbf{s}_1$ et $\mathbf{s}_2$ et représentés sur la figure 2.14.

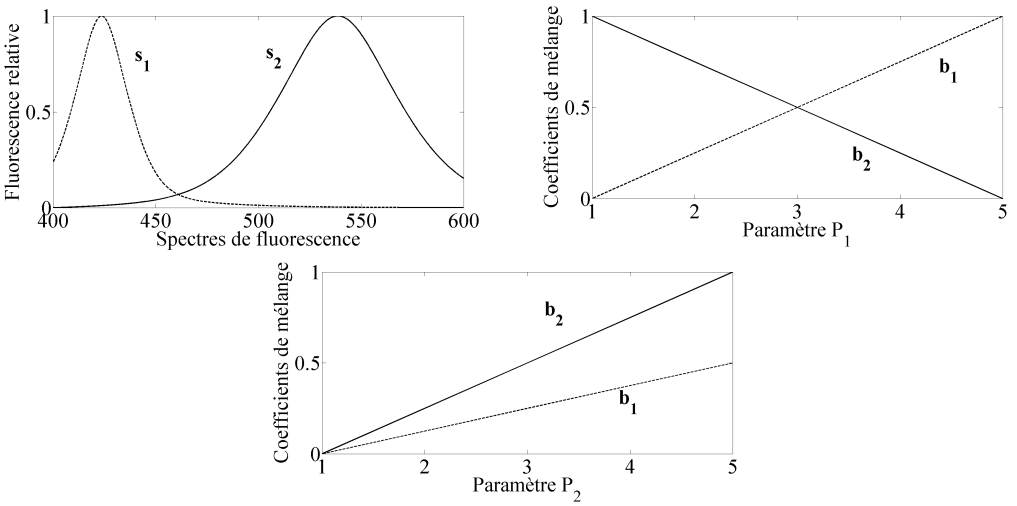


FIGURE *2.14 – Données initiales utilisées pour générer le tenseur $\boldsymbol{\mathcal{X}}$.s_1 et s_2 représentent deux spectres de fluorescence de protéines fluorescentes. a_1 et a_2 représentent les coefficients de mélange des deux sources en fonction d'un paramètre P_1 et b_1, b_2 en fonction d'un paramètre P_2.*

La figure 2.15 représente des décompositions CP du tenseur de données pour une même erreur d'estimation. Ces solutions font partie d'une infinité d'autres solutions admissibles. Cependant une seule de ces solutions correspond au modèle attendu. Les autres solutions proviennent de la transformation des matrices $\mathbf{A}$ et $\mathbf{B}$ par une matrice $\mathbf{W}$ inversible.

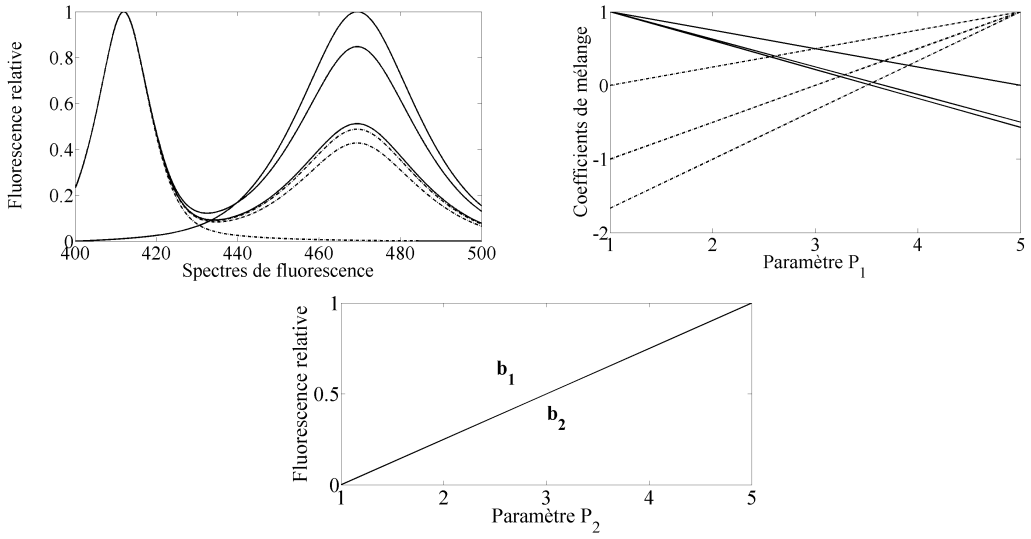


FIGURE *2.15 – Le lien de linéarité entre les réponses des deux sources suivant le troisième mode engendre une infinité de solutions possibles. Cette figure représente quelques décompositions CP possibles du jeu de données simulé suivant les mélanges présentés figure 2.14. Toutes ces solutions conduisent à une même erreur d'estimation.*

Nous avons présenté précédement une condition (la condition de Kruskal) suffisante pour assurer l'unicité de la décomposition CP. Cette condition n'est pas valide lorsqu'il y a une dépendance linéaire (*e.g.* colinéarité) entre les colonnes d'une des matrices de la décomposition. La question de l'unicité du modèle devient alors problématique.

Pour ces cas, il existe des conditions assurant l'identifiabilité soit de certaines colonnes des trois modes (unicité partielle) soit d'une seule des trois matrices (unicité unimodale). Des conditions suffisantes pour ces types d'unicité ont été énoncées dans [39]. Ces conditions sont similaires à la condition de Kruskal à la différence que les sources colinéaires sont autorisées dans un mode. Il est montré que dans le cas où l'une des matrices est identifiable on peut décomposer le tenseur de rang F en sous-tenseurs de rang inférieur et analyser l'unicité pour chacun de ces tenseurs de façon indépendante. Pour la partie non-identifiable du modèle, l'application d'une contrainte de type positivité, indépendance, orthogonalité permet, dans certains cas, d'assurer l'unicité de la décomposition.

Une autre façon de gérer les dépendances linéaires entre les colonnes des trois modes est de les prendre explicitement en compte à l'aide des matrices de contraintes. Cela conduit à un nouveau modèle appelée PARALIND [17], qui est étudié dans la partie suivante.

2.6 Modèle PARALIND

Le modèle PARALIND (PARAllel profiles with LINear Dependences) a été proposé dans [17], dans le but de résoudre les problèmes liés à la dépendance de facteurs dans la décomposition CP.

Le modèle PARALIND introduit dans le modèle CP de nouvelles matrices $\mathbf{\Psi}$, $\mathbf{\Phi}$ et $\mathbf{\Omega}$ appelées

matrices des dépendances. Le nom de matrices d'interactions est également utilisé par analogie au tenseur cœur de la décomposition suivant le modèle de Tucker. Ainsi, au lieu de la décomposition CP de $\boldsymbol{\mathcal{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{S}]\!]$, on a $\boldsymbol{\mathcal{X}} = [\![\tilde{\mathbf{A}}\boldsymbol{\Psi}, \tilde{\mathbf{B}}\boldsymbol{\Phi}, \tilde{\mathbf{S}}\boldsymbol{\Omega}]\!]$ avec $\tilde{\mathbf{A}}$,$\tilde{\mathbf{B}}$ et $\tilde{\mathbf{S}}$ des matrices de rang colonne plein.

Considérons un modèle CP, où les colonnes de la matrice $\mathbf{A}$ s'écrivent $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_1 + \mathbf{a}_2]$. Le modèle PARALIND exprime $\mathbf{A}$ comme étant le produit matriciel $\mathbf{A} = \tilde{\mathbf{A}}\boldsymbol{\Psi}$ où $\tilde{\mathbf{A}} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ et la matrice d'interactions

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}. \tag{2.7}$$

En général, les algorithmes d'estimation du modèle PARALIND supposent que les matrices d'interactions soient connues *a priori*. Dans ce cas, les conditions d'unicité ont été énoncées par Stegeman et De Almeida dans [88]. Cependant, pour les applications visées, notamment l'identification des systèmes Minéral/Bactérie, les matrices d'interactions sont inconnues. Or, il est intéressant d'estimer ces matrices car elles fournissent des informations importantes sur les interactions physiques entre les différents composants de ces systèmes. La matrice de dépendances sur le mode spectral permet d'identifier différentes évolutions d'un même marqueur fluorescent. On peut imaginer deux gènes différents marqués par une même protéine fluorescente. La matrice de dépendances sur l'un des modes $\mathbf{A}$ et $\mathbf{B}$ permettrait donc de mettre en évidence deux gènes ayant le même comportement par rapport aux paramètres P_1 et P_2. Une estimation par moindres carrés alternées de la matrice de contraintes du modèle PARALIND a été proposée par Bro *et al.* dans [17]. L'algorithme d'estimation du modèle PARALIND consiste à estimer tour à tour chacune des matrices et est présenté pour un modèle $[\![\mathbf{A}\boldsymbol{\Psi}, \mathbf{B}, \mathbf{S}]\!]$ dans le tableau 2.4. Après avoir choisi le nombre de sources recherchées, les matrices des modes sont fixées et la matrice des interactions est estimée par la méthode des moindres carrés, puis les matrices des modes 2, 3 et la matrice des interactions sont fixées et la matrice du premier mode est estimée et ainsi de suite jusqu'à validation de la condition d'arrêt.

Pour des raisons algorithmiques, l'estimation de la matrice de contraintes est effectuée directement dans la boucle principale mais cela revient à estimer $\boldsymbol{\Psi}$ et $\tilde{\mathbf{A}}$ à partir de $\mathbf{A}$. Sans contraintes supplémentaires ce problème bilinéaire d'estimation est mal posé et admet une infinité de solutions. Nous proposons alors de limiter le nombre des solutions possibles en appliquant une contrainte de parcimonie lors de l'estimation de la matrice d'interaction [22]. L'application de cette contrainte de parcimonie revient à chercher la matrice d'interactions qui explique les données par le nombre minimum d'interactions entre ses colonnes.

Algorithme S-PARALIND

Dans cette section, nous présentons un algorithme (S-PARALIND) pour estimer le modèle PARALIND avec des contraintes de parcimonie sur la matrice d'interaction. L'algorithme est

	Entrée : $\boldsymbol{\mathcal{X}}$, λ, F, F_1
1:	Initialisation $\tilde{\mathbf{A}}, \mathbf{B}, \mathbf{S}$
2:	$\text{vec}\boldsymbol{\Psi} = \arg\min_{\boldsymbol{\Psi}} \|\text{vec}\mathbf{X}_{(1)} - [(\mathbf{S} \odot \mathbf{B}) \otimes \tilde{\mathbf{A}}]\ \text{vec}\boldsymbol{\Psi}\|_2^2$
3:	$\tilde{\mathbf{A}} = \mathbf{X}_{(1)}(\mathbf{S} \odot \mathbf{B})\boldsymbol{\Psi}^T \left\{\boldsymbol{\Psi}[(\mathbf{B}^T\mathbf{B})(\mathbf{S}^T\mathbf{S})]\boldsymbol{\Psi}^T\right\}^{-1}$
4:	$\mathbf{B} = \mathbf{X}_{(2)}[\mathbf{S} \odot (\tilde{\mathbf{A}}\boldsymbol{\Psi})] \left[(\boldsymbol{\Psi}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\boldsymbol{\Psi})(\mathbf{S}^T\mathbf{S})\right]^{-1}$
5:	$\mathbf{S} = \mathbf{X}_{(3)}[\mathbf{B} \odot (\tilde{\mathbf{A}}\boldsymbol{\Psi})] \left[(\boldsymbol{\Psi}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\boldsymbol{\Psi})(\mathbf{B}^T\mathbf{B})\right]^{-1}$
6:	Si condition d'arrêt non satisfaite, aller étape 2
	Sortie : Estimation de $\tilde{\mathbf{A}}, \boldsymbol{\Psi}, \mathbf{B}, \mathbf{S}$

TABLE *2.4 – algorithme ALS-PARALIND*

présenté pour estimer les dépendances linéaires dans un seul mode (*i.e.* mode-1) mais l'extension à deux ou trois modes est directe.

La représentation parcimonieuse des signaux consiste dans la représentation d'un signal avec un faible nombre de coefficients significatifs. Par définition, un signal est dit parcimonieux lorsque la plupart de ses coefficients sont nuls. La dernière décennie a vue de nombreuses avancées dans le domaine de la parcimonie, quelles soient méthodologiques [19] ou algorithmiques [96, 99]. L'application d'une contrainte de parcimonie sur la matrice d'interaction permet de prendre en compte un faible nombre d'interactions entre les composantes de manière à améliorer la modèlisation des données.

En présence de bruit ou d'erreurs du modèle, l'estimation de $\boldsymbol{\Psi}$ à partir d'une estimation CP de $\mathbf{A}$ n'est pas judicieuse à cause de l'effet d'accumulation d'erreurs. Par conséquent, dans [17] la matrice de contraintes est estimée directement dans la boucle principale de l'algorithme (ALS) lors de la mise à jour des composants des matrices de CP. Si l'on s'intéresse à l'unicité de l'estimation des matrices $\tilde{\mathbf{A}}$ et Ψ cela revient à étudier l'identifiabilité de la décomposition bilinéaire $\mathbf{A} = \tilde{\mathbf{A}}\boldsymbol{\Psi}$. D'après l'équation (2.1), nous savons que sans contraintes supplémentaires une telle décomposition n'est pas unique. On propose alors d'appliquer une contrainte de parcimonie sur la matrice $\boldsymbol{\Psi}$. Cependant, la contrainte de parcimonie n'est pas suffisante pour assurer l'unicité de la décomposition bilinéaire. Afin de réduire le nombre de solutions on peut ajouter une nouvelle contrainte de positivité. On sait cette contrainte pertinente car elle a une signification physique dans le cas des signaux de fluorescence.

La méthode proposée est basée sur l'algorithme proposé dans [17] avec la différence que l'étape d'estimation par moindres carrés de $\boldsymbol{\Psi}$ est remplacée par un problème d'optimisation $l_2 - l_1$ [96].

$$\min_{\boldsymbol{\Psi} \geq 0} \left[\|\mathbf{X}_{(1)} - \tilde{\mathbf{A}}\boldsymbol{\Psi}(\mathbf{S} \odot \mathbf{B})^T\|_2^2 + \gamma\|\boldsymbol{\Psi}\|_1\right] \tag{2.8}$$

où l'hyperparamètre γ contrôle le degré de parcimonie de la matrice de contraintes. La minimisation de (2.8) peut être formulée comme un problème LASSO pour laquelle un certain

nombre d'algorithmes efficaces ont été développés récemment (voir [107] et références incluses). Le tableau 2.5 illustre les principales étapes de l'approche proposée, où vec(.), $\otimes$ et $*$ désignent, respectivement, l'opérateur de vectorisation matriciel, le produit de Kronecker.

	`Entrée :` $\boldsymbol{\mathcal{X}}$, γ, F, F_1
1:	`Initialisation` $\tilde{\mathbf{A}}, \mathbf{B}, \mathbf{S}$
2:	$\text{vec}\boldsymbol{\Psi} = \arg\min_{\boldsymbol{\Psi}\geq\mathbf{0}}\{\|\text{vec}\mathbf{X}_{(1)} - [(\mathbf{S}\odot\mathbf{B})\otimes\tilde{\mathbf{A}}]\ \text{vec}\boldsymbol{\Psi}\|_2^2 + \gamma\|\text{vec}\boldsymbol{\Psi}\|_1\}$
3:	$\tilde{\mathbf{A}} = \mathbf{X}_{(1)}(\mathbf{S}\odot\mathbf{B})\boldsymbol{\Psi}^T\left\{\boldsymbol{\Psi}[(\mathbf{B}^T\mathbf{B})(\mathbf{C}^T\mathbf{S})]\boldsymbol{\Psi}^T\right\}^{-1}$
4:	$\mathbf{B} = \mathbf{X}_{(2)}[\mathbf{S}\odot(\tilde{\mathbf{A}}\boldsymbol{\Psi})]\left[(\boldsymbol{\Psi}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\boldsymbol{\Psi})(\mathbf{S}^T\mathbf{S})\right]^{-1}$
5:	$\mathbf{S} = \mathbf{X}_{(3)}[\mathbf{B}\odot(\tilde{\mathbf{A}}\boldsymbol{\Psi})]\left[(\boldsymbol{\Psi}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\boldsymbol{\Psi})(\mathbf{B}^T\mathbf{B})\right]^{-1}$
6:	`Si condition d'arrêt non satisfaite, aller étape 2`
	`Sortie : Estimation de` $\tilde{\mathbf{A}}, \boldsymbol{\Psi}, \mathbf{B}, \mathbf{S}$

TABLE *2.5 – ALS-PARALIND algorithme avec contraintes de parcimonie sur la matrice d'interaction (S-PARALIND)*

Les étapes $3-5$ dans le tableau 2.5 sont facilement obtenues à partir de l'algorithme ALS. Dans la prochaine section, nous comparons S-PARALIND avec l'algorithme ALS-PARALIND présenté dans [17] sur des exemples synthétiques.

Comparaison ALS-PARALIND/S-PARALIND

Les résultats suivants sont obtenus par décomposition d'un lot de données simulées $\boldsymbol{\mathcal{X}}$ de taille $700\times30\times30$. Les données sont générées de manière à représenter les spectres de fluorescence émis par des biosenseurs. Les résultats de la décomposition S-PARALIND sont comparés avec ceux de l'approche ALS-PARALIND proposée dans [17]. Dans cette expérience, le nombre de sources est fixé à 4 ($F=4$) et la taille de la matrice $\tilde{\mathbf{S}}$ est 700×3. La figure 2.16 représente les colonnes des matrices $\tilde{\mathbf{S}}$, $\mathbf{A}$ et $\mathbf{B}$.

Dans un premier exemple, la matrice des interactions est la suivante :

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Dans ce cas, le $rang_k$ de $\mathbf{S}$ est égal à 1 ce qui implique la non unicité de la décomposition CP. Toutefois, les conditions d'unicité décrites dans [39] assurent que la matrice du premier mode est identifiable. Les données sont traitées par les algorithmes ALS-PARALIND et S-PARALIND avec la même initialisation, la même condition de positivité et le même nombre d'itérations. La figure 2.17 représente le résultat de la décomposition par l'algorithme ALS-PARALIND. Sur le

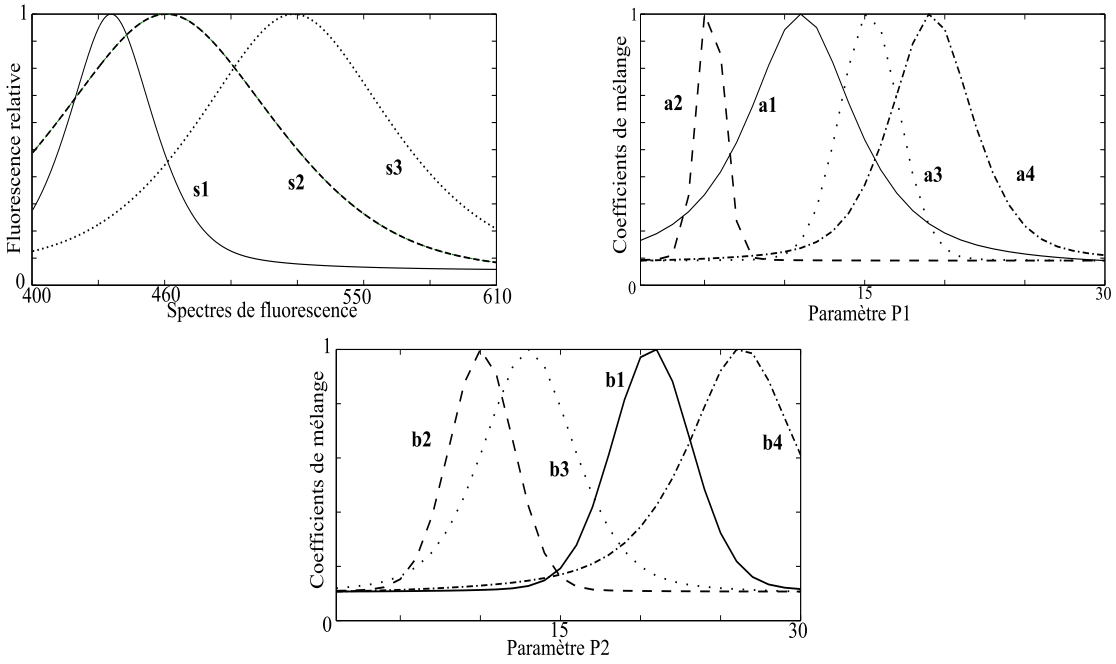


FIGURE *2.16 – Données initiales utilisées pour générer le tenseur* $\boldsymbol{\mathcal{X}}$.

premier graphique, on peut voir l'estimation des sources $\mathbf{s}_1$, $\mathbf{s}_2$ et $\mathbf{s}_3$ regroupées dans la matrice $\tilde{\mathbf{S}}$. Les deux autres graphiques représentent l'estimation des matrices $\mathbf{A}$ et $\mathbf{B}$. De même, la figure 2.18 illustre le résultat obtenu par S-PARALIND. La comparaison de ces deux décompositions indique que la méthode S-PARALIND montre une meilleure estimation de la matrice $\tilde{\mathbf{S}}$ que la méthode ALS-PARALIND. Le spectre $\mathbf{s}_2$ estimé par ALS-PARALIND, représenté figure 2.17, montre qu'il existe encore une indétermination qui se manifeste par une sous estimation de l'intensité de fluorescence sur la plage de longueur d'onde allant de 400 à 460 nm.

Les équations 2.9 et 2.10 correspondent à l'estimation de la matrice $\boldsymbol{\Psi}$ par les deux différentes méthodes. Comme la matrice $\tilde{\mathbf{S}}$ est mieux estimée en utilisant S-PARALIND, l'estimation de la matrice des interactions est également meilleure.

$$\hat{\boldsymbol{\Psi}}_{S-PARALIND} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.008 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{2.9}$$

$$\hat{\boldsymbol{\Psi}}_{ALS-PARALIND} = \begin{bmatrix} 1 & 0.591 & 0.155 & 0.59 \\ 0.003 & 1 & 0.103 & 1 \\ 0.004 & 0.157 & 1 & 0.156 \end{bmatrix}. \tag{2.10}$$

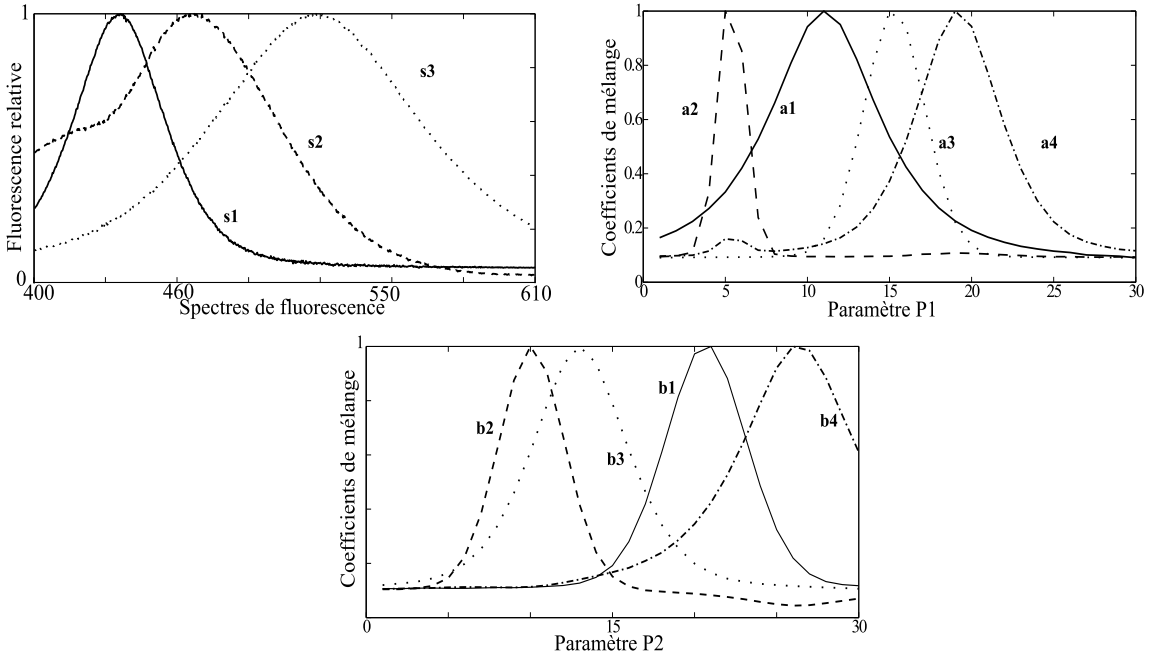


FIGURE *2.17 – Décomposition PARALIND du tenseur $\boldsymbol{\mathcal{X}}$ sans contrainte de parcimonie.*

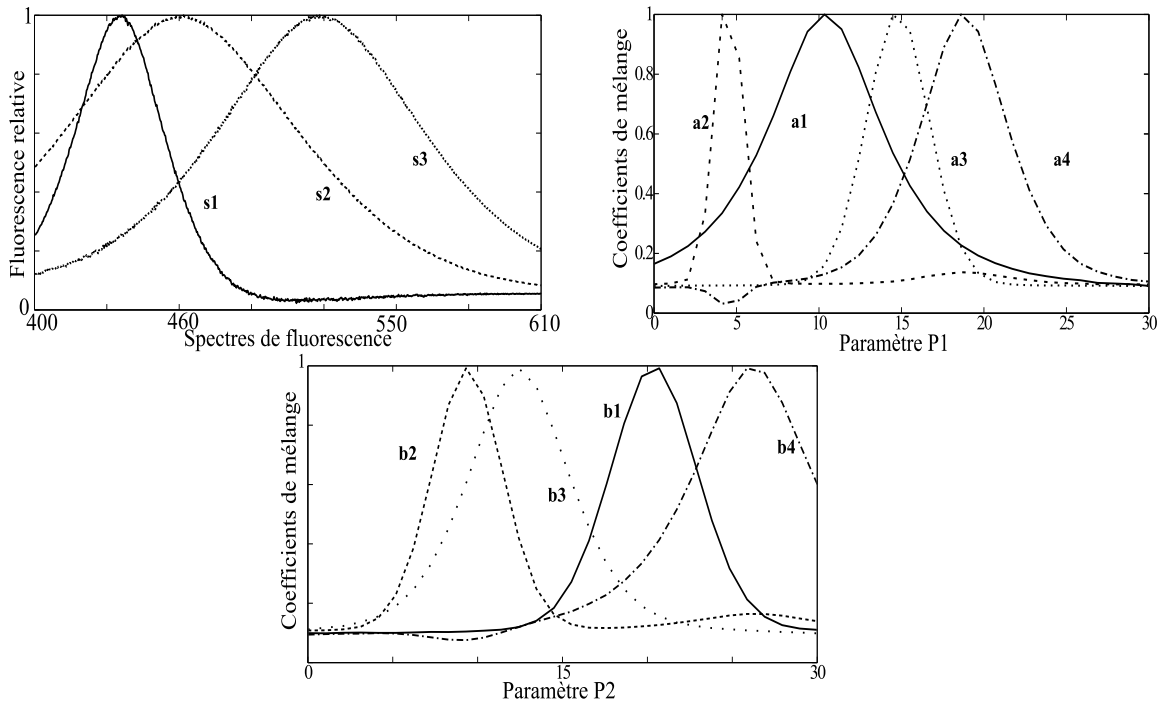


FIGURE *2.18 – Décomposition PARALIND du tenseur* $\mathbf{X}$ *avec contrainte de parcimonie sur la matrice d'interaction (S-PARALIND).*

Considérons le cas où $rang_k(\mathbf{S}) = 2$, la nouvelle matrice d'interaction est alors la suivante :

$$\mathbf{\Psi} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{2.11}$$

D'après cette matrice d'interactions la quatrième colonne de **S** correspond à la somme de la première et de la seconde colonne. Le modèle est identifiable car les deux matrices **A** et **B** sont de rang plein. Les figures 2.19 et 2.20 montrent que les matrices **A** et **B** sont bien estimées par les deux méthodes. Cependant, l'estimation de $\tilde{\mathbf{S}}$ est meilleure par la méthode S-PARALIND.

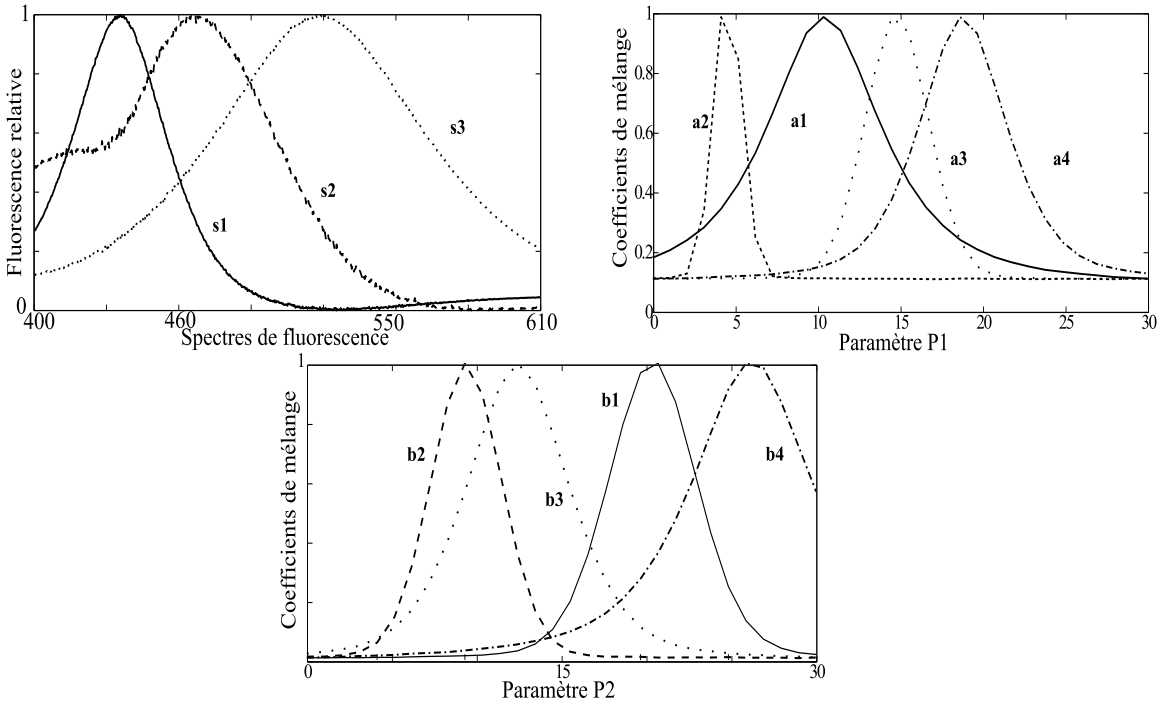


FIGURE *2.19 – Décomposition PARALIND du tenseur $\boldsymbol{\mathcal{X}}$ sans contrainte de parcimonie.*

Les équations 2.12 et 2.13 donnent l'estimation des matrices d'interactions par les deux méthodes.

$$\hat{\mathbf{\Psi}}_{S-PARALIND} = \begin{bmatrix} 1 & 0 & 0 & 0.749 \\ 0.171 & 1 & 0.024 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{2.12}$$

$$\hat{\mathbf{\Psi}}_{ALS-PARALIND} = \begin{bmatrix} 1 & 0.468 & 0.006 & 1 \\ 0.014 & 1 & 0.007 & 0.569 \\ 0.071 & 0.401 & 1 & 0.276 \end{bmatrix}. \tag{2.13}$$

Bien que l'estimation par S-PARALIND soit meilleure, tous les coefficients ne sont pas bien estimés (0.749 au lieu de 1) et il existe des valeurs faibles qui auraient dû être nulles. Le choix d'un coefficient λ trop faible pourrait expliquer cette mauvaise estimation. Dans notre cas, le choix du paramètre γ est empirique car les méthodes d'estimation du paramètre γ optimal étant encore un problème ouvert.

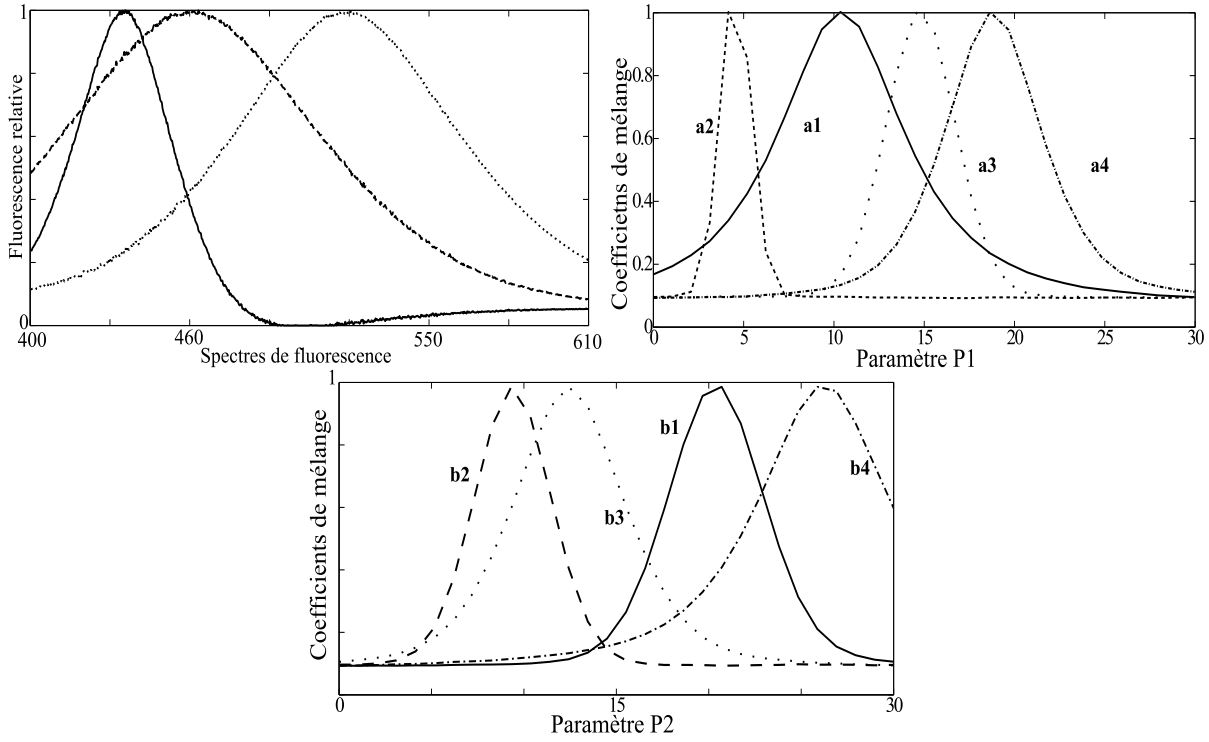


FIGURE *2.20 – Décomposition PARALIND du tenseur $\boldsymbol{\mathcal{X}}$ avec contrainte de parcimonie sur la matrice d'interactions (S-PARALIND).*

2.7 Décomposition CP quadrilinéaire

Précédemment, nous avons montré l'intérêt de l'identification des systèmes bactéries-solutions par des modèles trilinéaires. dans lesquels les différents modes représentent les diversités spectrales des sources de fluorescence, les comportements différentiels des promoteurs réagissant aux variations de concentrations d'un métal et les variations de quantités de biosenseurs soit par l'évolution temporelle soir par des combinaisons variables de biosenseurs. Or, dans certains cas pratique, il est intéressant d'étudier la réponse des promoteurs à plusieurs métaux ou d'associer la réponse cinétique aux variations de ratio de biosenseurs. Dans ces cas, on obtient des données quadrilinéaires que l'on peut représenter sous la forme d'un tenseur de données $\boldsymbol{\mathcal{X}}(I \times J \times K \times L)$ qui contiendra JKL spectres de taille I correspondant à JKL échantillons. Nous disposons alors d'un jeu de données contenant les mesures de fluorescence $x_{i,j,k,l}$ en fonction de quatre paramètres i, j, k, l. Le modèle quadrilinéaire de $\boldsymbol{\mathcal{X}}$ consiste en une combinaison linéaire de F vecteurs $\mathbf{a}_f, \mathbf{b}_f, \mathbf{c}_f, \mathbf{s}_f$ qui s'exprime comme suit :

$$\boldsymbol{\mathcal{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{s}_f \tag{2.14}$$

où $\mathbf{a}_f, \mathbf{b}_f, \mathbf{c}_f$ représente l'évolution de la fluorescence de la source f en fonction de trois paramètres différents et $\mathbf{s}_f$ est le spectre de fluorescence de la source f.

Le principale intérêt de la décomposition quadrilinéaire réside dans les propriétés d'identifiabilité du modèle dans les cas de colinéarité dans un ou plusieurs modes. On peut citer quelques cas pratiques pour lesquels il existe une ou plusieurs colinéarités :

- Lorsque deux gènes promoteurs différents ont la même réponse aux variations d'un élément stressant (*e.g.* métal).
- Lorsque le choix des instants de mesures ne permettent pas de différencier les cinétiques de chaque marqueur (*e.g.* un temps trop long entre deux mesures ou entre le début de l'expérience et la première mesure).
- Lorsque le protocole expérimental génère deux populations de biosenseurs percevant deux environnements différents (*e.g.* un biosenseur mis en présence de particules minérales peut avoir deux comportements. Par exemple, son comportement en suspension dans le milieu sera différent de son comportement en contact direct avec la particule.

Le fait de pouvoir assurer l'unicité du modèle quadrilinéaire dans un cas présentant des colinéarités est un avantage sur les modèles trilinéaires. Dans un cas où il est impossible de résoudre de manière unique un problème trilinéaire le simple ajout d'une diversité peut assouplir les conditions d'unicité et fournir une décomposition unique. Cependant, le choix de nouveaux paramètres doit respecter les propriétés de linéarité avec les précédents paramètres ainsi que les conditions d'identifiabilité des modèles quadrilinéaires.

2.7.1 Application

Pour tester l'obtention d'une unicité partielle à l'ordre 4, une procédure expérimentale simulant le comportement de biosenseurs bactériens par le mélange de trois colorants. Trois colorants fluorescents (Oregon Green 514 (OG514), Rhodamine 6G (R6G), la rhodamine B (RB)) ont été choisis pour le recouvrement de leur spectre d'émission et de leurs réponses différentes à des conditions physico-chimiques comme la température et le pH. L'intensité de fluorescence de OG514 répond au pH tandis que R6G et RB y sont insensibles. L'intensité de fluorescence de RB dépend fortement de la température tandis que l'émission des OG514 et R6G varie peu en fonction de la température.

Trente-six mélanges ont été créés dans les puits d'une microplaque (6 valeurs différentes du pH $\times$ 6 concentrations différentes de R6G). Puis les spectres de fluorescence ont été mesurés à 6 températures différentes. Les données ont été organisées dans un tableau à 4 dimensions (figure 2.21).

Les variations de concentration et pH correspondent, respectivement, aux différents graphiques verticaux et horizontaux. En utilisant la réponse spectrale des colorants et la diversité du plan d'expérience (mélanges des colorants), un modèle CP à 4 modes d'ordre de décomposition 3 peut être proposé :

$$\boldsymbol{\mathcal{X}} = \sum_{f=1}^{3} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{s}_f$$

où $\mathbf{S}$ regroupe les spectres de fluorescence de chacun des colorants. Le choix de protéines

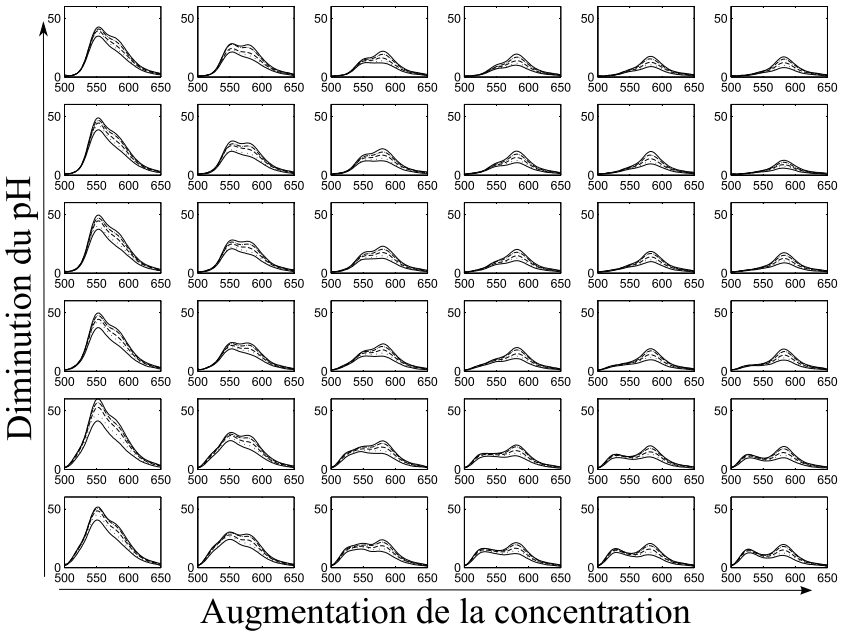


FIGURE *2.21 – Spectres d'émission de fluorescence mesurés (λ de 500 à 650 nm) pour différents mélanges de colorants (Oregon Green 514 (OG514), Rhodamine 6G (R6G), Rhodamine B (RB)) et six différentes valeurs de températures (de 10°C à 45°C) et six valeurs différentes de pH. Chaque graphique représente les spectres obtenus dans un même puits pour les 6 températures différentes.*

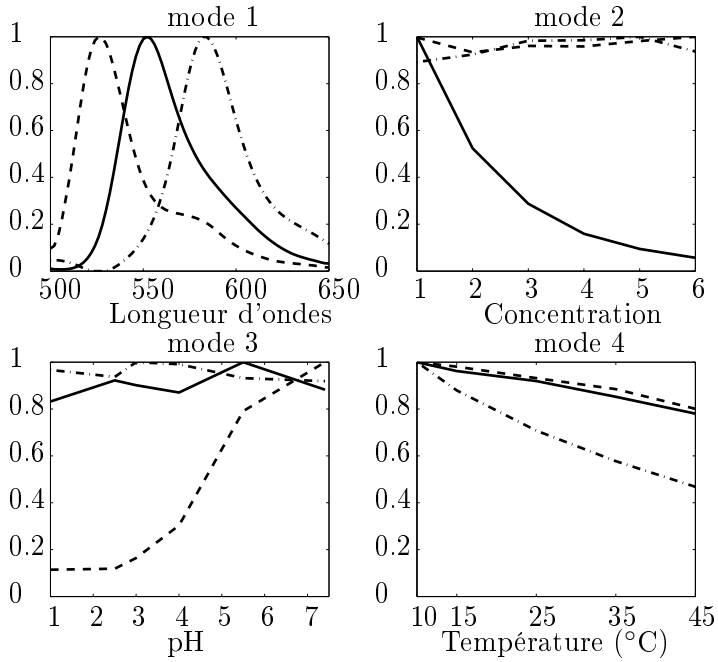


FIGURE *2.22 – Décomposition CP. Le mode 1 représente les trois sources de fluorescence. Les modes 2, 3 et 4 représentent l'évolution de ces trois sources en fonction des paramètres de concentration, de pH et de température.*

fluorescentes différentes assure que la matrice **S** est de rang plein. Chacune des autres matrices représente les réponses des colorants aux autres diversités (concentration de R6G, pH, température) et les colinéarités dans les modes peuvent s'écrire comme suit :

$$\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\ \mathbf{a}_1], \quad r_{\mathbf{A}} = 2, rang_k(\mathbf{A}) = 1; \tag{2.15}$$

$$\mathbf{B} = [\mathbf{b}_1\ \mathbf{b}_1\ \mathbf{b}_2], \quad r_{\mathbf{B}} = 2, rang_k(\mathbf{B}) = 1; \tag{2.16}$$

$$\mathbf{C} = [\mathbf{c}_1\ \mathbf{c}_2\ \mathbf{c}_2], \quad r_{\mathbf{C}} = 2, rang_k(\mathbf{C}) = 1. \tag{2.17}$$

La figure 2.22 montre les résultats obtenus en effectuant la décomposition CP de données organisées sous forme de tableau à 4 modes. Pour chaque mode, les vecteurs sont normalisés à une valeur maximale égale à un. La décomposition a été effectuée en utilisant l'algorithme ALS avec des contraintes de positivité sur tous les modes. La décomposition a été réalisée sans contraintes avec différentes initialisations. Les résultats obtenus sont tous similaires et en accords avec ce qui était attendu. En particulier, le mode 1 représente les spectres des différents colorants. Le mode 2 montre clairement la variation de la concentration de R6G, tandis que les deux autres restent constants. De même, pour le mode 3, la réponse de OG514 varie avec le pH. Et finalement, pour le mode 4, la rhodamine B répond plus fortement que les deux autres colorants. Cette expérience confirme l'unicité du modèle garanti par le théorème énoncé précédemment.

Chapitre 3

Identification et estimation de la réponse de systèmes rapporteurs sensibles aux métaux

Sommaire

3.1 Systèmes rapporteurs utilisés et démarche opératoire

Parmi les éléments métalliques d'importance environnementale, le fer et le cadmium font l'objet d'une attention particulière. Le fer est présent dans de nombreux minéraux des sols et joue un rôle fondamental dans la croissance et l'activité des organismes. Le fer est un oligo-élément essentiel car une carence ou un apport excessif peuvent avoir des conséquences léthales sur l'organisme. En situation de carence, les microorganismes ont développé diverses stratégies d'acquisition du fer. La plus courante est basée sur la production de molécules chélatrices[5] spécifiques du fer : les sidérophores. En situation d'excès, d'autres stratégies sont mises en œuvre comme l'expulsion ou la séquestration des ions ce qui conduit à la réduction de la concentration intracellulaire.

A l'inverse du fer, le cadmium est un élément non-essentiel toxique pour les êtres vivants. En solution, le cadmium se trouve principalement sous forme de cation divalent Cd^{2+}. Sous cette forme, ce métal est très toxique, même à faibles concentrations. En effet, le comportement du cadmium se rapproche de celui d'oligo-élements comme le zinc (Zn^{2+}) et a tendance à former des liaisons covalentes fortes surtout avec le soufre [30].

Ce métal entre dans la cellule bactérienne par diffusion et via des systèmes de transport membranaire spécifique d'ions, tels que les transporteurs du magnésium (Mg^{2+}) [66]. L'efflux du Cd^{2+} est assuré par des pompes membranaires ou des systèmes transporteurs chez les bactéries à Gram négatif, telle que *Pseudomonas putida* (LIM23) utilisée dans ce travail. De façon plus anecdotique, certaines bactéries du genre *Pseudomonas*, résistantes au cadmium, sont capables de le volatiliser sous forme de diméthylcadmium.

Pour les expériences concernant le fer nous avons utilisé la souche *Pseudomonas aeruginosa* PAO1-2Δ comme hôte. Les systèmes rapporteurs sont deux gènes impliqués dans l'homéostasie du fer. Le premier promoteur *pvdA*, régule la synthèse des sidérophores (pyoverdine), et est induit par une carence en fer (voir fig. 3.1). Ce système est destiné à subvenir aux besoins essentiels en fer(III) de la bactérie en condition de carence sévère ([Fe(III)]$< 1\,\mu$M). Le second promoteur, induit par un excès de fer dans le cytoplasme bactérien, *bfrB* régule la synthèse d'une bacteria ferritine, chargé d'immobiliser le fer en excès. Ce système de régulation du fer ferrique intracellulaire fonctionne dans une large gamme de concentrations et stabilise en permanence le statut du fer(III) au sein du cytoplasme.

Pour les expériences concernant le cadmium nous avons utilisé la souche *Pseudomonas putida* comme hôte pour le développement du biosenseur bactérien. La construction du biosenseur consiste à la double fusion des gènes $PPcadA_2$::*gfp* et $P_{A1/04/03}$::*DsRed* au chromosome de la bactérie. Le premier système rapporteur (producteur de la GFP) est construit sur la base d'un promoteur (PP$cadA_2$) d'un gène codant une pompe d'efflux du cadmium. Ce promoteur détecte la présence de cadmium (Voir fig. 3.2). Les mesures effectuées à une seule longueur d'onde (excitation 488 nm, émission 515 nm) montrent que la réponse est dose-dépendante ; elle augmente

5. Complexant avec un ion métallique par au moins deux liaisons.

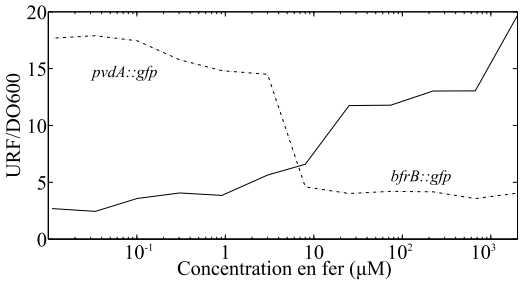


FIGURE *3.1 – Réponse des gènes bfrB et pvdA à différentes concentrations de fer obtenues par dilution de $FeCl_3$. La fluorescence est exprimée en unité de fluorescence relative par rapport à la densité bactérienne estimée à partir d'une mesure d'absorbance à* 600 nm. *La fluorescence est mesurée à* 510 nm *après excitation à* 490 nm.

selon une loi puissance en fonction de la concentration en cadmium. Des expériences ont montrées que l'expression du gène PP$cadA_2$ est également induite par le mercure et plus faiblement par le zinc et le plomb. Le second système rapporteur (producteur de la DsRed) est exprimé constamment et permet d'apprécier l'état physiologique des cellules ou leur nombre. Sa réponse, évaluée par l'émission de fluorescence à 580 nm, ne dépend pas *a priori* de la concentration en cadmium, sauf aux concentrations très élevées (>1mM), lorsque la dose léthale est atteinte.

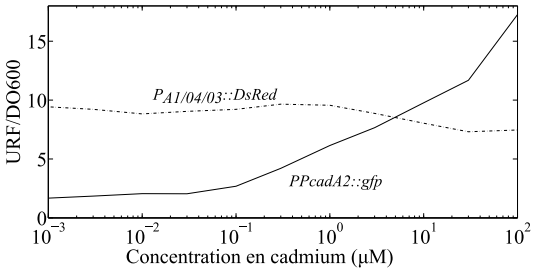


FIGURE *3.2 – Réponse du biosenseur après 4h d'incubation en milieu MOPS en présence de concentrations croissantes de cadmium obtenues par ajout de $CdCl_2$. La fluorescence est exprimée en unité de fluorescence relative par rapport à la densité bactérienne estimée à partir d'une mesure d'absorbance à* 600 nm. *La fluorescence de la GFP est mesurée à* 510 nm *après excitation à* 490 nm. *Pour la DsRed, nous utilisons les longueurs d'onde* 560 nm *(excitation) et* 589 nm *(émission).*

Les biosenseurs utilisés au cours des expériences suivantes sont listés dans le tableau 3.1. La première colonne du tableau regroupe les dénominations des biosenseurs et les colonnes suivantes la construction *promoteur*::*rapporteur* associée au biosenseur.

Habituellement, les réponses des biosenseurs aux stress métalliques (Cd, Fe) sont mesurées individuellement par fluorimétrie, à longueurs d'onde d'excitation et d'émission fixes. Pour évaluer la réponse d'un biosenseur, l'intensité de fluorescence est intégrée dans une plage de longueurs d'onde d'émission donnée pour une longueur d'onde d'excitation donnée. Les bandes passantes choisies encadrent les optimum d'émission et d'excitation. Pour une suspension cellulaire faiblement concentrée ($DO_{600\,\mathrm{nm}} < 1$) ne contenant qu'un seul type de biosenseur, l'intensité de

Biosenseur	Promoteur	Sonde Fluorescente	Construction génétique
$\mathcal{B}_{1G}$	pvdA	GFP	*pvdA*::*gfp*
$\mathcal{B}_{1Y}$	pvdA	EYFP	*pvdA*::*eyfp*
$\mathcal{B}_{2G}$	bfrB	GFP	*bfrB*::*gfp*
$\mathcal{B}_{2Y}$	bfrB	EYFP	*bfrB*::*eyfp*
$\mathcal{B}_{3R/4G}$	$P_{A1/04/03}$ PP*cadA*$_2$	DsRed GFP	$P_{A1/04/03}$::*dsred* *PPcadA*$_2$::*gfp*

TABLE *3.1 – Tableau récapitulatif des biosenseurs et de la nomenclature utilisés dans les expériences suivantes.*

fluorescence mesurée est proportionnelle à la quantité de photons absorbée par les systèmes rapporteurs (Voir eq. (1.2)). C'est une approximation linéaire de la loi de Beer-Lambert dont les limites sont discutables surtout si plusieurs biosenseurs sont employés. La largeur des bandes d'absorption et d'émission ($> 100\,\text{nm}$) ne permet pas de négliger le chevauchement spectral (spectral overlapping en anglais) entre les spectres des différents biosenseurs et autres substances auto-fluorescentes produites par les bactéries ou présentes dans la suspension cellulaire (*e.g.* sidérophore extracellulaire). La prise en compte de cette spécificité contraint à simplifier le milieu de culture (réduction des matières organiques) ou à utiliser des biosenseurs mutants limitant l'auto-fluorescence. L'excitation monochromatique favorise généralement l'émission de fluorescence des biosenseurs dont le maximum d'émission est plus proche de la longueur d'onde excitatrice (déplacement de Stockes faible) provoquant une atténuation de la fluorescence des sources ayant un émission dans les longueurs d'onde les plus grandes. De plus, l'absorption progressive du faisceau excitateur par la suspension cellulaire provoque des phénomènes non linéaires d'écrantage. Le dispositif de mesure ne collecte plus qu'une partie de la lumière émise par l'échantillon. Plus la densité est élevée, plus l'intensité du faisceau excitateur parvenant au centre du puits de mesure est faible.

Pour contourner ces contraintes opératoires, nous avons utilisé un modèle trilinéaire des données spectrales, dont l'inversion (décomposition CP) a été étudiée dans le chapitre 2.

Pour chaque expérience, les résultats des décompositions CP des tenseurs de données sont présentés de la manière suivante. Dans un premier temps, un tableau fournit les indices (CorConDia, énergie résiduelle) permettant, respectivement, de valider la décomposition CP et d'estimer le nombre de sources recherchées. Les calculs de l'indice CorConDia et de l'énergie résiduelle ont été présentés dans le chapitre 2.

La décomposition CP est présentée sous la forme de quatre graphiques présentant l'évolution de chacune des sources suivant les trois modes (longueur d'onde, concentration en élément stressant et ratio de biosenseurs ou temps) et les coefficients de normalisation de chacune des sources. La représentation sous forme de graphique à barres des coefficients de normalisation permettra d'évaluer le poids de chaque composante dans le mélange. Les composantes dans chaque mode sont normalisées par leurs valeurs maximales et le coefficient de normalisation représente

le produit de ces valeurs maximales.

3.2 Estimation des réponses au fer de mélanges de biosenseurs antagonistes

Cette série d'expériences vise à identifier, dans un mélange, les réponses antagonistes de deux biosenseurs répondant au fer. Les deux systèmes rapporteurs sont construits en associant les deux promoteurs (*pvdA* et *bfrB*) avec deux rapporteurs fluorescents (EYFP et GFP). L'approche fournira un modèle d'expression des gènes plus complet qu'une courbe étalon tout en minimisant le nombre d'expériences.

Dans cette première série d'expériences, nous cherchons à séparer les interférences spectrales (autofluorescence, recrouvement spectral) afin d'obtenir une estimation plus réaliste de la réponse du biosenseur. La séparation de la fluorescence du marquage, des autres sources de fluorescence permettra une étude plus poussée de la régulation de l'expression des gènes intervenant dans l'homéostase du fer.

3.2.1 Présentation de l'expérience

Le plan d'expérience a été construit de manière à générer un jeu de données trilinéaire respectant le modèle CP. La fluorescence mesurée est une combinaison linéaire de F sources de fluorescence dont l'intensité varie en fonction de trois paramètres. Dans cette expérience, les trois paramètres sont la concentration en fer, le mélange de biosenseurs et la longueur d'onde. D'après l'équation (2.4), le tenseur de données $\boldsymbol{\mathcal{X}}$ s'exprime par

$$\boldsymbol{\mathcal{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{s}_f \tag{3.1}$$

où $\mathbf{a}_f$ représente l'évolution de la source f en fonction de la concentration en fer, $\mathbf{b}_f$ en fonction du mélange de biosenseurs et $\mathbf{s}_f$ en fonction de la longueur d'onde.

Les variations du premier paramètre (gradient de concentration en fer) sont obtenues par une dilution en cascade d'une solution mère de fer (100 mM) dans un milieu de culture carencé en fer (DCAA). La gamme comporte douze niveaux de concentrations allant de 2 mM à 11 nM obtenue par dilutions successives (raison 1/3) par un robot pipeteur. Le plan de dilution est présenté sur le tableau 3.3.

Les variations du second paramètre sont obtenues par mélange en proportions variables de deux biosenseurs. La souche utilisée est la bactérie *Pseudomonas aeruginosa* dont les gènes promoteurs *pvdA* et *bfrB*, réagissant au fer, ont été marqués par deux systèmes rapporteurs *eyfp* et *gfp*. Le premier biosenseur a été élaboré pour une production de fluorescence (GFP) dose-dépendante, via la fusion d'un gène rapporteur au promoteur *bfrB*. L'expression du gène rapporteur est alors sous contrôle du promoteur *bfrB* dont l'induction est proportionnelle à la

quantité de fer [105]. Le second biosenseur produit également une fluorescence dose-dépendante au fer grâce à la synthèse d'une protèine fluorescente EYFP dont le maximum d'émission est décalé de 18 nm vers le rouge par rapport à la GFP. Le gène rapporteur EYFP est sous contrôle du promoteur *pvdA* dont l'induction est proportionnelle à la carence en fer. Les deux biosenseurs ont donc des réponses antagonistes en fonction de la concentration en fer et spectralement chevauchantes.

Les précultures des deux biosenseurs ($\mathcal{B}_{1Y}$ et $\mathcal{B}_{2G}$) sont réalisées dans un milieu riche (LB) à 37°C sous agitation (240 rpm). Les suspensions de biosenseurs sont ensuite rincées dans le milieu carencé en fer (DCAA) avant d'être distribuées dans une microplaque 96 puits.

Dans chaque puit, une même quantité de biosenseur est ajoutée, en faisant varier le ratio de mélange dans une gamme allant de 0.05 à 0.95 :

$$\alpha_{2G} = \#\mathcal{B}_{2G}/(\#\mathcal{B}_{2G} + \#\mathcal{B}_{1Y})$$

avec $\#\mathcal{B}_{1G}$, $\#\mathcal{B}_{2G}$, $\#\mathcal{B}_{1Y}$ et $\#\mathcal{B}_{2Y}$ correspondant, respectivement, aux quantités de biosenseurs $\mathcal{B}_{1G}$, $\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$ et $\mathcal{B}_{2Y}$. Un plan d'expérience identique est utilisé sur le second mélange de biosenseurs dont le marquage aura été permuté. La synthèse de protéine fluorescente verte (GFP) est sous contrôle du promoteur *pvdA* et alors que l'expression du promoteur *bfrB* est indiquée par la synthèse de la protéine jaune (EYFP). Le ratio de mélange s'exprime comme suit :

$$\alpha_{2Y} = \#\mathcal{B}_{2Y}/(\#\mathcal{B}_{2Y} + \#\mathcal{B}_{1G})$$

Les tableaux 3.2 représentent les deux mélanges de biosenseurs effectués. Chaque tableau représente les biosenseurs utilisés et les comportements théoriques associés des promoteurs et des marqueurs. Les réponses des différents biosenseurs en fonction de la concentration en fer ($\mathbf{a}_i$) sont *a priori* non proportionnelles ($\mathbf{a}_1 \neq \mathbf{a}_2$). De même, les mélanges de biosenseurs et les marqueurs seront choisis de manière à ne pas créer de colinéarités ($\mathbf{b}_1 \neq \mathbf{b}_2$ et $\mathbf{s}_1 \neq \mathbf{s}_2$).

Premier mélange			
	a	b	s
$\mathcal{B}_{1G}$	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{s}_1$
$\mathcal{B}_{2Y}$	$\mathbf{a}_2$	$\mathbf{b}_2$	$\mathbf{s}_2$

Deuxième mélange			
	a	b	s
$\mathcal{B}_{1Y}$	$\mathbf{a}_1$	$\mathbf{b}_1$	$\mathbf{s}_2$
$\mathcal{B}_{2G}$	$\mathbf{a}_2$	$\mathbf{b}_2$	$\mathbf{s}_1$

TABLE *3.2 – Représentation des comportements attendus ($\mathbf{a}_i$, $\mathbf{b}_i$) pour les couples de biosenseurs utilisés dans les deux mélanges réalisés. Dans le second mélange, le marquage des promoteurs a été permuté.*

Le tableau 3.3 représente le plan de plaque de l'expérience sur lequel est indiqué le ratio de biosenseurs et la concentration en fer pour chacun des puits.

Après incubation (24h à 37°C), différents types de spectres de fluorescence ont été mesurés dans chacun des puits de la microplaque :

	1	2	3	4	5	6	7	8	9	10	11	12
A	2 mM 95 :5	667 µM 95 :5	222 µM 95 :5	74 µM 95 :5	25 µM 95 :5	8 µM 95 :5	3 µM 95 :5	0.9 µM 95 :5	0.3 µM 95 :5	0.1 µM 95 :5	34 nM 95 :5	11 nM 95 :5
B	2 mM 80 :20	667 µM 80 :20										11 nM 80 :20
C	2 mM 60 :40		222 µM 60 :40								34 nM 60 :40	
D	2 mM 50 :50			74 µM 50 :50						0.1 µM 50 :50		
E	2 mM 40 :60				25 µM 40 :60				0.3 µM 40 :60			
F	2 mM 20 :80					8 µM 20 :80		0.9 µM 20 :80				
G	2 mM 5 :95						3 µM 5 :95					
H	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc	Blanc

TABLE *3.3 – Plan d'expérience réalisé sur une microplaque 96 puits. En ligne : variation de la concentration en fer sur une gamme de 2mM à 11nM de fer par dilution en cascade d'une solution mère de chlorure de fer(III). En colonne : variation du ratio α_{2G} ou α_{2Y}, allant de 95% à 5%, obtenues par mélange des deux biosenseurs.*

- spectres d'émission de fluorescence dont la mesure a été effectuée entre 506 et 560 nm avec un pas de 2 nm (27 points) pour une longueur d'onde d'excitation de 485 nm,
- spectres synchrones de fluorescence mesurés entre 470 et 530 nm avec un pas de 2 nm (30 points) pour un $\Delta\lambda = 20$ nm.

Ainsi, quatre expériences, en plus de la mesure monolongueur d'onde, ont été réalisées en combinant les mesures spectrales et les constructions de biosenseurs. Les décompositions CP des spectres d'émission et spectres synchrones sont étudiées pour chacun des deux mélanges de biosenseurs ($\mathcal{B}_{1Y}$, $\mathcal{B}_{2G}$) et ($\mathcal{B}_{1G}$, $\mathcal{B}_{2Y}$).

3.2.2 Identification conjointe à partir des données monolongueur d'onde

Pour démontrer l'apport de la méthode CP par rapport à la méthode classique, des mesures de fluorescence à longueur d'onde fixe ont été réalisées. La méthode standard consiste à obtenir les courbes étalons des gènes promoteurs en associant la protéine fluorescente marquant le gène à la longueur d'onde correspondant au maximum de son spectre d'émission. La figure 3.3 représente l'évolution, en fonction de la concentration en fer, de la fluorescence émise aux longueurs d'onde 510 nm (GFP, $\lambda_{exc} = 480$ nm) et 527 nm (EYFP, $\lambda_{exc} = 514$ nm) pour chacun des deux couples de biosenseurs. On retrouve, dans la première expérience avec le couple de biosenseurs ($\mathcal{B}_{1G}$, $\mathcal{B}_{2Y}$) (Fig. 3.3(a)), les comportements antagonistes et dose-dépendantes attendus (Fig.3.1). Cependant, il faut noter l'existence d'une fluorescence basale relativement forte (50% de l'intensité maximale pour $\mathcal{B}_{1G}$). Les courbes de réponses des deux rapporteurs dans la configuration de marquage inverse (Fig.3.3(b)) ne montrent pas de comportement particulier. Les fortes variations entre les réplicats ne permettent pas d'identifier des variations de la fluorescence en fonction de la quantité de fer.

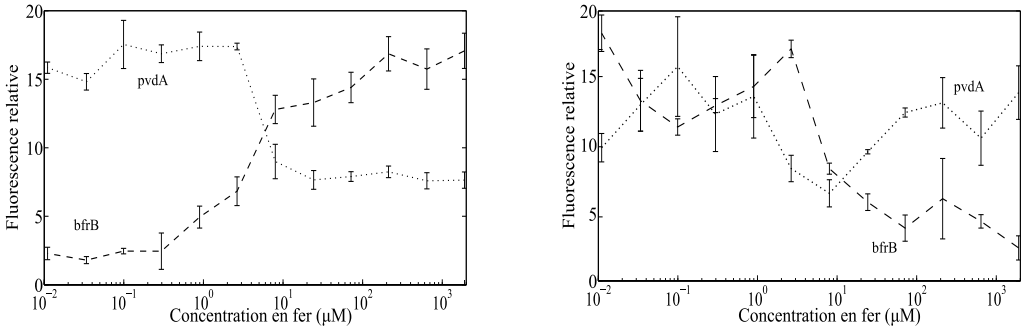


FIGURE *3.3 – Courbes étalons des promoteurs bfrB (tirets gras) et pvdA (pointillés) établis par mesure monolongueur d'onde à* 510 nm *(GFP,* $\lambda_{exc} = 480\,\text{nm}$*) et* 527 nm *(EYFP,* $\lambda_{exc} = 514\,\text{nm}$*) (3 replicats). (a) Couples de biosenseurs* $\mathcal{B}_{2G}$ *et* $\mathcal{B}_{1Y}$*. (b) Couples de biosenseurs* $\mathcal{B}_{2Y}$ *et* $\mathcal{B}_{1G}$*.*

Ces résultats mettent en évidence les biais que peut générer un recouvrement spectral dans l'estimation des réponses. On remarque que la fluorescence basale couplée au recouvrement spectral détériore la limite de détection allant jusqu'à la perte totale de la réponse des promoteurs (Voir fig.3.3 (b)). Les réponses du couple ($\mathcal{B}_{2Y}$, $\mathcal{B}_{1G}$) montrent également que le choix de la construction *promoteur*::*rapporteur* est importante, car la fluorescence est influencée par la force d'induction des différents promoteurs. En aucun cas, nous ne pouvons utiliser ces courbes pour étalonner la réponse conjointe des promoteurs *pvdA* et *bfrB* au fer. Par conséquent, nous proposons d'utiliser la méthode de décomposition tensorielle CP pour accomplir cette tâche.

3.2.3 Analyse de la décomposition CP des spectres d'émission

Sur chaque puits, les spectres d'émission sont mesurés entre 506 et 560 nm avec un pas de 2 nm, pour une excitation à 495 nm. Pour chaque expérience, les mesures sont regroupées dans un tableau tridimensionnel de dimension $27 \times 12 \times 7$. Le premier mode décrit la longueur d'onde du spectre de fluorescence, le second, la concentration en fer et le troisième, le ratio de biosenseurs. Les mesures spectrales obtenues sont présentées sur les figures 3.4 et 3.5 en fonction de la concentration en fer et du mélange de biosenseurs.

Réponse du mélange de biosenseurs ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$)

Le tableau 3.4 regroupe les indices calculés afin d'estimer le nombre de sources dans le mélange. Plusieurs décompositions ont été effectuées, avec différentes initialisations, pour évaluer la qualité de la décomposition et l'évolution de la fraction d'énergie résiduelle en fonction du nombre de sources recherchées. Dans le cas du couple de biosenseurs ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$), nous recherchons *a priori* au moins 2 sources et l'indice CorConDia montre que le modèle CP est valide jusqu'à 7 sources. Néanmoins, l'évolution de l'énergie résiduelle en fonction du nombre de sources

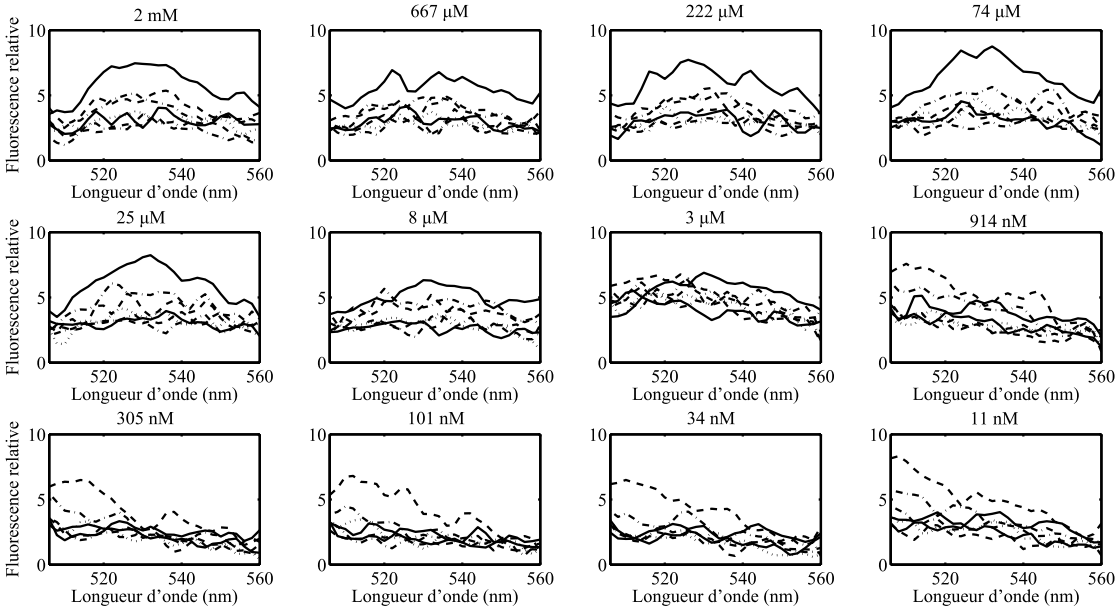


FIGURE *3.4* – *Spectres bruts (non lissés) d'émission de fluorescence mesurés de* 506 *à* 560 nm *avec un pas de* 2 nm *($\lambda_{exc} = 405$ nm). Chaque graphique regroupe les spectres obtenus pour une concentration en fer (colonne de puits). Chaque spectre correspond à l'émission de fluorescence mesurée dans un puits pour un mélange de biosenseurs $\mathcal{B}_{2Y}$ et $\mathcal{B}_{1G}$.*

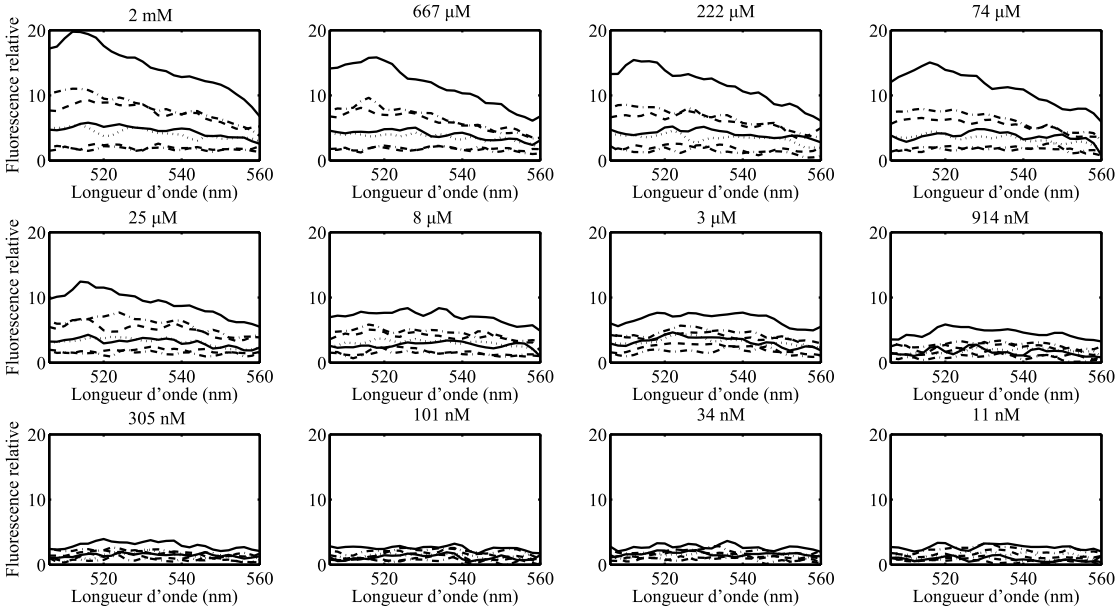


FIGURE *3.5* – *Spectres bruts (non lissés) d'émission de fluorescence mesurés de* 506 *à* 560 nm *avec un pas de* 2 nm *($\lambda_{exc} = 405$ nm). Chaque graphique regroupe les spectres obtenus pour une concentration en fer (colonne de puits). Chaque spectre correspond à l'émission de fluorescence mesurée dans un puits pour un mélange de biosenseurs $\mathcal{B}_{2G}$ et $\mathcal{B}_{1Y}$.*

est faible au-delà de 3 sources, c'est-à-dire que l'ajout d'autres sources ne réduit pas de façon significative l'écart entre le modèle et les mesures. Suite à cette analyse, nous avons fixé le nombre de sources à 3.

Couple de biosenseurs ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$)								
Nb. sources (F)	2	3	4	5	6	7	8	9
CorConDia	100	100	99	99	99	91	71	79
Fraction d'énergie résiduelle	9%	6%	4%	3%	2%	2%	2%	2%

TABLE *3.4 – Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie résiduelle pour la décomposition CP du jeu de données pour des nombres différents de sources recherchées.*

La figure 3.6 montre le résultat de la décomposition CP pour trois sources, obtenu avec le premier mélange de biosenseurs $\mathcal{B}_{2G}$ et $\mathcal{B}_{1Y}$. On peut y voir l'évolution de l'intensité des différentes sources spectrales (Fig. 3.6(a)) en fonction de la concentration en fer (Fig. 3.6(b)) et du ratio α_{2G} du mélange (Fig. 3.6(c)). La figure 3.6(d) représente les coefficients de normalisation des sources. La décomposition montre l'existence d'une source $\#_1$, sans maximum marqué, qui recouvre la quasi totalité du domaine de longueurs d'onde étudié. Cette source croît fortement pour les concentrations supérieures à $1\,\mu$M et n'évolue pas en fonction du ratio α_{2G} de biosenseurs. Les deux autres sources spectrales estimées ($\mathbf{s}_2$, $\mathbf{s}_3$) présentent des maxima à 512 nm et 527 nm respectivement. L'estimation des coefficients de mélange montre que ces sources ont un comportement inverse en fonction de la concentration en fer et du ratio de biosenseur. Les réponses $\mathbf{a}_2$ et $\mathbf{b}_2$ sont croissantes et monotones alors que la réponse $\mathbf{a}_3$ est croissante sur les concentrations $0,1\,\mu$M à $1\,\mu$M puis décroissante pour les concentrations supérieures et que la réponse $\mathbf{b}_3$ est décroissante. Les coefficients de normalisation (Fig. 3.6(d)) montrent une source de fluorescence $\#_2$ plus intense que les deux autres sources dont les niveaux sont équivalents.

En se basant sur les réponses et les spectres estimés, les sources peuvent être identifiées. L'attribution de la source $\#_1$ à un signal autofluorescent intrinsèque des cellules bactériennes semble être judicieuse. Sa diminution pour les faibles concentrations en fer s'explique par un développement bactérien perturbé : peu de fer, peu de croissance. De plus, puisque cette source est constitutive des deux biosenseurs, elle n'évolue pas en fonction du ratio de biosenseurs.

La source spectrale estimée $\mathbf{s}_2$ est conforme au spectre d'émission connue pour la protéine fluorescente GFP dont l'optimum se situe à 510 nm. L'estimation des coefficients de mélange par la méthode CP montre également une augmentation du signal de la GFP avec l'augmentation de la concentration de fer, qui est le comportement attendu pour le promoteur $bfrB$. L'augmentation quasi linéaire de ce signal avec le ratio suit bien l'augmentation du biosenseur $\mathcal{B}_{2G}$ dans le mélange expérimental. Ce résultat atteste de la validité de la décomposition CP.

De même, le spectre estimé $\mathbf{s}_3$ correspond au spectre attendu de la protéine fluorescente EYFP. Sa production dépend de l'induction du gène *pvdA* qui diminue avec l'augmentation de la concentration en fer. Ce phénomène est mis en évidence par la courbe de réponse estimée par la décomposition CP (courbe $\mathbf{a}_3$). On peut également voir une augmentation de l'intensité de

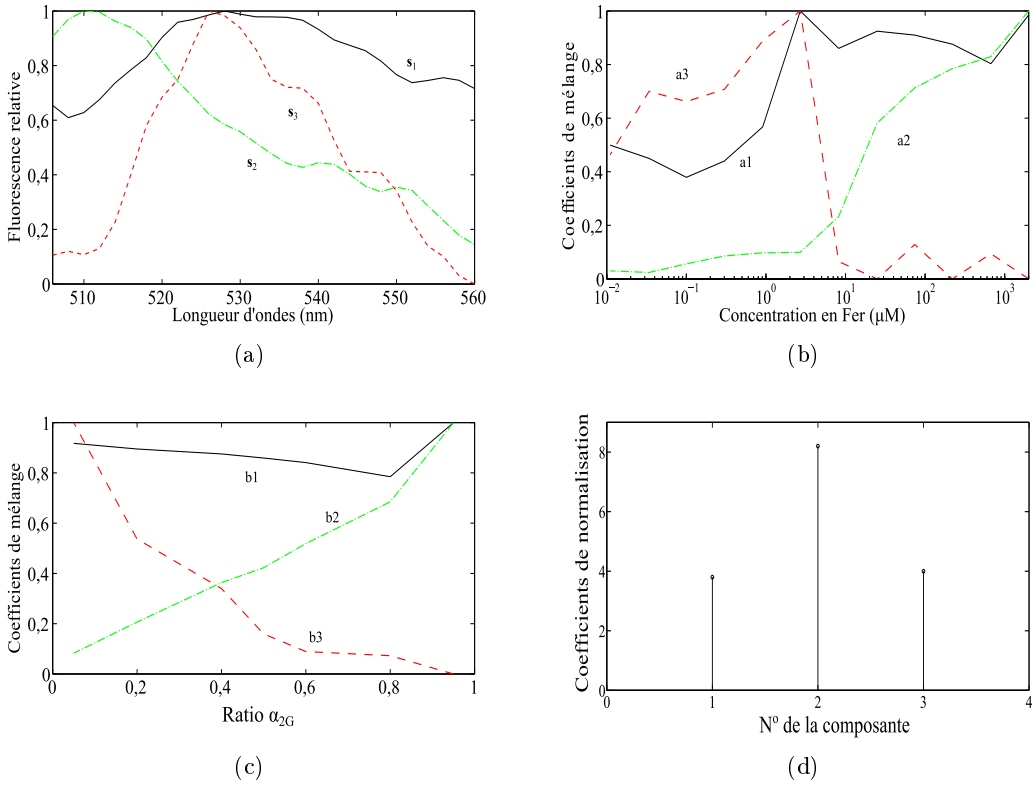


FIGURE *3.6 – Décomposition CP des spectres d'émission de fluorescence mesurés sur le mélange ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$). Sources #1 (autofluorescence), #2 ($\mathcal{B}_{2G}$), #3 ($\mathcal{B}_{1Y}$). (a) Spectres d'émission des sources identifiées. (b) Réponses respectives en fonction de la concentration en fer des sources. (c) Évolution des réponses en fonction du ratio de biosenseurs α_{2G}. (d) Coefficients de normalisation obtenus par mise à 1 de la valeur maximale des réponses modales.*

fluorescence pour les concentrations inférieures à 1 µM. Ceci s'explique par un accroissement de la quantité de bactéries. La diminution des coefficients de mélange ($\mathbf{b}_3$) en fonction du ratio α_{2G} va dans le sens d'une association de la source #3 à la concentration en EYFP.

Réponse du mélange de biosenseurs ($\mathcal{B}_{2Y}$, $\mathcal{B}_{1G}$)

Dans le cas du couple de biosenseurs ($\mathcal{B}_{2Y}$, $\mathcal{B}_{1G}$), nous cherchons également au minimum 2 sources comme pour le couple ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$). L'évolution de l'énergie résiduelle (voir tableau 3.5) est faible à partir de 3 sources et l'étude des décompositions CP avec 2, 3 et 4 sources (non présentés ici) montre que le modèle à 3 sources est le plus adapté. Suite à cette analyse, nous avons fixé le nombre de sources à 3.

La figure 3.7 présente le résultat de la décomposition CP à trois sources des données mesurées sur les mélanges des biosenseurs $\mathcal{B}_{2Y}$ et $\mathcal{B}_{1G}$. On peut y voir l'évolution de l'intensité des trois sources spectrales (Fig. 3.7(a)) en fonction de la concentration en fer (Fig. 3.7(b)) et du ratio α_{2Y} du mélange (Fig. 3.7(c)). La figure 3.7(d) représente les coefficients de normalisation des sources.

Couple de biosenseurs ($\mathcal{B}_{2Y}$, $\mathcal{B}_{1G}$)								
Nb. sources (F)	2	3	4	5	6	7	8	9
CorConDia	100	99	99	99	99	99	96	97
Fraction d'énergie résiduelle	6%	4%	3%	2%	2%	2%	2%	2%

TABLE *3.5 – Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie résiduelle pour la décomposition CP du jeu de données, pour des nombres différents de sources recherchées.*

La décomposition montre l'existence d'une source $\#_1$ dont le spectre $\mathbf{s}_1$ s'étale sur une grande partie de plage de longueurs d'onde. Le signal de cette source est nul pour les concentrations en fer de $0,03\,\mu M$ et $0,1\,\mu M$ et croit fortement pour les concentrations supérieures à $1\,\mu M$. La source spectrale estimée $\mathbf{s}_2$ possède un maximum d'émission à 506 nm. L'estimation des coefficients de mélange par la méthode CP montre une diminution de la fluorescence de cette source avec l'augmentation de la concentration de fer. Le signal de cette source diminue en fonction de l'augmentation du ratio α_{2Y} dans le mélange de biosenseurs. Le spectre estimé $\mathbf{s}_3$ présente un maximum à 532 nm. L'estimation (courbe $\mathbf{a}_3$) de l'évolution de la fluorescence de cette source augmente avec l'augmentation de la concentration en fer. La décomposition CP montre une augmentation des coefficients de mélange ($\mathbf{b}_3$) en fonction du ratio α_{2Y}. La figure 3.7(d) montre une source $\#_2$ dont la fluorescence est plus importante que celle des sources $\#_1$ et $\#_3$.

L'attribution de la source $\#_1$ à un signal autofluorescent intrinsèque des cellules bactériennes semble à nouveau être judicieuse. On retrouve un comportement analogue où l'intensité de fluorescence diminue pour les faibles concentrations de fer. Comme précédemment, le nombre de bactéries est constant dans chaque mélange ce qui explique la faible variation de la source $\#_1$ en fonction du ratio de biosenseurs. Néanmoins, les coefficients nuls pour des concentrations en fer de $0,03\,\mu M$ et $0,1\,\mu M$ et la faible augmentation de la fluorescence pour la valeur 0,6 du ratio peut remettre en cause la qualité de la décomposition pour ce jeu de données.

Malgré une surestimation de la fluorescence à 506 nm qui décale le maximum de fluorescence, le spectre estimé $\mathbf{s}_2$ est conforme à celui attendu pour la GFP. L'estimation des coefficients de mélange par la méthode CP montre une diminution de l'intensité de la source $\#_2$ avec l'augmentation de la concentration en fer dans le milieu. C'est le comportement général attendu pour le biosenseur $\mathcal{B}_{1G}$. L'attribution de cette source à la GFP explique également la diminution de cette source en fonction de l'augmentation du ratio α_{2Y}.

En comparant le spectre estimé $\mathbf{s}_3$ au spectre attendu de la protéine fluorescente EYFP, on remarque une surestimation de l'intensité émise pour les longueurs d'onde inférieures à 520 nm et supérieures à 540 nm ce qui lui donne un aspect plus étalé. La production de EYFP est commandé par le promoteur *bfrB*, dont l'induction augmente avec l'augmentation de la concentration en fer. Ce phénomène est mis en valeur par la courbe de réponse estimée par la décomposition CP (courbe $\mathbf{a}_3$). Cependant, l'évolution de la fluorescence est moins marquée et elle n'évolue pas entre les concentrations 3 nm et 8 nm. Ceci peut s'expliquer par la forte augmentation de l'autofluorescence des bactéries à ces concentrations alors que la séparation des spectres de la

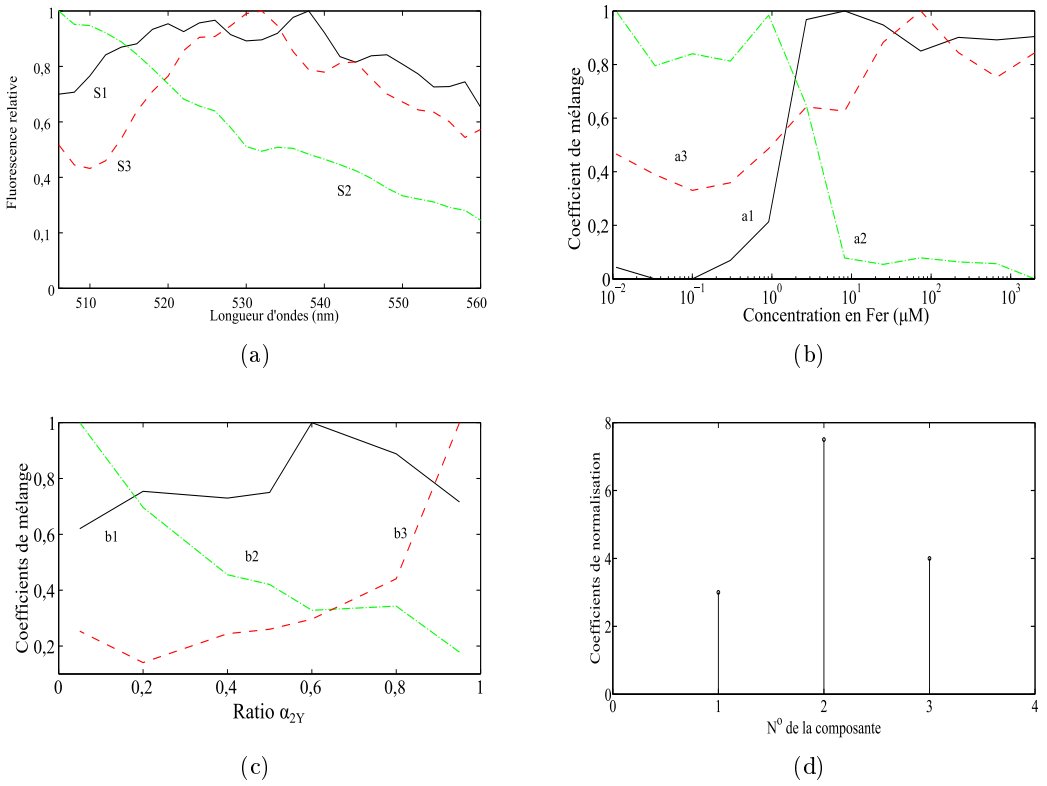


FIGURE *3.7 – Décomposition CP des spectres d'émission de fluorescence mesurés sur le mélange de biosenseurs ($\mathcal{B}_{2Y}$, $\mathcal{B}_{1G}$). Sources #1 (autofluorescence), #2 ($\mathcal{B}_{1G}$), #3 ($\mathcal{B}_{2Y}$). (a) Spectres d'émission des sources identifiées. (b) Réponses respectives en fonction de la concentration en fer des sources. (c) Évolution des réponses en fonction du ratio de biosenseurs α_{2Y}. (d) Coefficients de normalisation obtenus par mise à 1 de la valeur maximale de chaque réponse.*

EYFP et d'autofluorescence est moins bonne et influence l'estimation des coefficients de mélange. L'augmentation des coefficients de mélange ($\mathbf{b}_3$) en fonction du ratio α_{2Y} va dans le sens d'une association de la source #3 à la concentration en EYFP bien que l'augmentation soit sous évaluée pour les ratios allant de 0,5 à 0,8.

En conclusion, dans cette deuxième expérience, les résultats de la décomposition obtenus avec l'interversion des constructions génétiques sont moins satisfaisants que dans l'expérience précédente, notamment en ce qui concerne la séparation de la source EYFP et de l'autofluorescence. L'intensité moyenne des spectres acquis (Figures 3.4 et 3.5) est plus faible pour les faibles concentrations en fer, ce qui rend la séparation du signal de la EYFP plus difficile. Cette contribution de la EYFP au signal s'explique par la longueur d'onde d'excitation qui favorise surtout l'émission de la GFP. Pour enlever cette dépendance à la longueur d'onde d'excitation, une solution consiste à utiliser les spectres synchrones de fluorescence ce qui permettrait d'augmenter l'excitation et le rapport signal à bruit de la EYFP.

3.2.4 Analyse de la décomposition CP des spectres synchrones

De façon similaire aux jeux de données obtenus pour les spectres d'émission, une décomposition CP est réalisée sur les spectres synchrones de fluorescence obtenus sur un plan d'expérience identique au précédent. Les spectres sont mesurés de 470 à 530 nm avec un pas de 2 nm pour un $\Delta\lambda = 20$ nm. Le tableau tridimensionnel de données est de dimension $30 \times 12 \times 7$ pour chacune des deux microplaques. Le premier mode est la fluorescence émise en fonction de la longueur d'onde du spectre synchrone de fluorescence, le second, la concentration en fer et le troisième mode, le ratio de biosenseur. Les données mesurées sont représentées sur les figures 3.8 et 3.9. Nous pouvons remarquer que les sources de fluorescence GFP et EYFP sont plus puissantes et marquées notamment dans les concentrations en fer les plus faibles. Les spectres synchrones obtenus sont également plus piqués et moins chahutés que les spectres d'émission.

Réponse du mélange de biosenseurs ($\mathcal{B}_{1Y}$, $\mathcal{B}_{2G}$)

Dans cette partie, nous présentons les résultats obtenus par décomposition CP des spectres synchrones mesurés sur le mélange de biosenseurs ($\mathcal{B}_{1Y}$, $\mathcal{B}_{2G}$). Le tableau 3.6 regroupe les indices calculés afin de déterminer le nombre de sources recherchées. De la même manière que dans l'expérience précédente, nous évaluons à trois le nombre de sources recherchées. Pour ce nombre de sources, l'énergie résiduelle est faible pour les deux couples de biosenseurs et n'évolue quasiment plus pour un nombre de sources supérieur. De plus, l'indice CorConDia valide la décomposition CP pour ce nombre de sources dans les deux cas étudiés.

Couple de biosenseurs ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$)								
Nb. sources (F)	2	3	4	5	6	7	8	9
CorConDia	100	100	100	100	100	100	100	6
Fraction d'énergie résiduelle	10%	2%	1%	1%	1%	1%	1%	1%

TABLE *3.6 – Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie résiduelle pour la décomposition CP du jeu de données pour des nombres différents de sources recherchées.*

La figure 3.10 montre le résultat de la décomposition CP des spectres synchrones mesurés sur les mélanges des biosenseurs $\mathcal{B}_{2G}$ et $\mathcal{B}_{1Y}$. On peut y voir l'évolution de l'intensité des différentes sources spectrales (Fig. 3.10(a)) en fonction de la concentration en fer (Fig. 3.10(b)) et du ratio α_{2G} (Fig. 3.10(c)). La figure 3.10(d) représente les coefficients de normalisation des sources. Le premier spectre synchrone estimé $\mathbf{s}_1$ a son maximum à 470 nm. La fluorescence de cette source est plus faible pour les faibles concentrations en fer ($\leq 1\,\mu$M). De plus, la fluorescence de cette source croît en fonction de la quantité de $\mathcal{B}_{2G}$ dans le mélange. Le second spectre estimé $\mathbf{s}_2$ possède un maximum à 488 nm. La fluorescence de cette source est croissante dans les deux autres modes et commence à augmenter à partir de 3 µM de fer. La fluorescence du dernier spectre synchrone estimé $\mathbf{s}_3$ est maximum pour la longueur d'onde 512 nm. À l'inverse, sa fluorescence décroit à partir de 3 µM, puis reste constante pour les concentrations supérieures à 8 µM. On

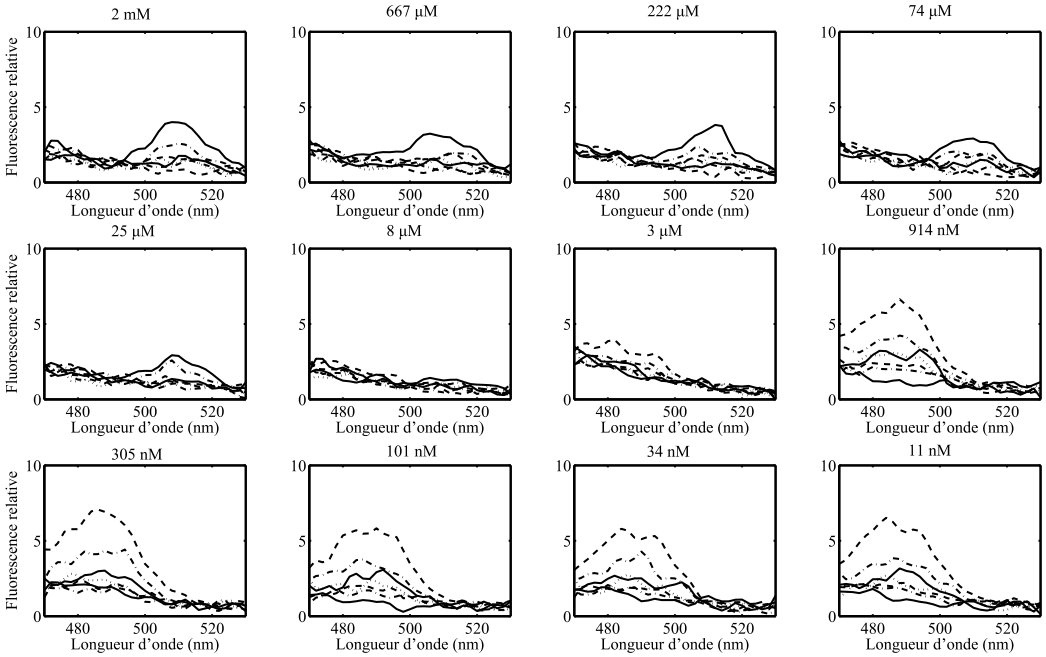


FIGURE *3.8 – Spectres synchrones bruts (non lissés) de fluorescence mesurés de* 470 *à* 530 nm *avec un pas de* 2 nm *et* $\Delta\lambda = 20$ nm. *Chaque graphique regroupe les spectres obtenus pour une concentration en fer (colonne de puits). Chaque spectre correspond à l'émission de fluorescence mesurée dans un puits pour un mélange de biosenseurs* $\mathcal{B}_{2Y}$ *et* $\mathcal{B}_{1G}$.

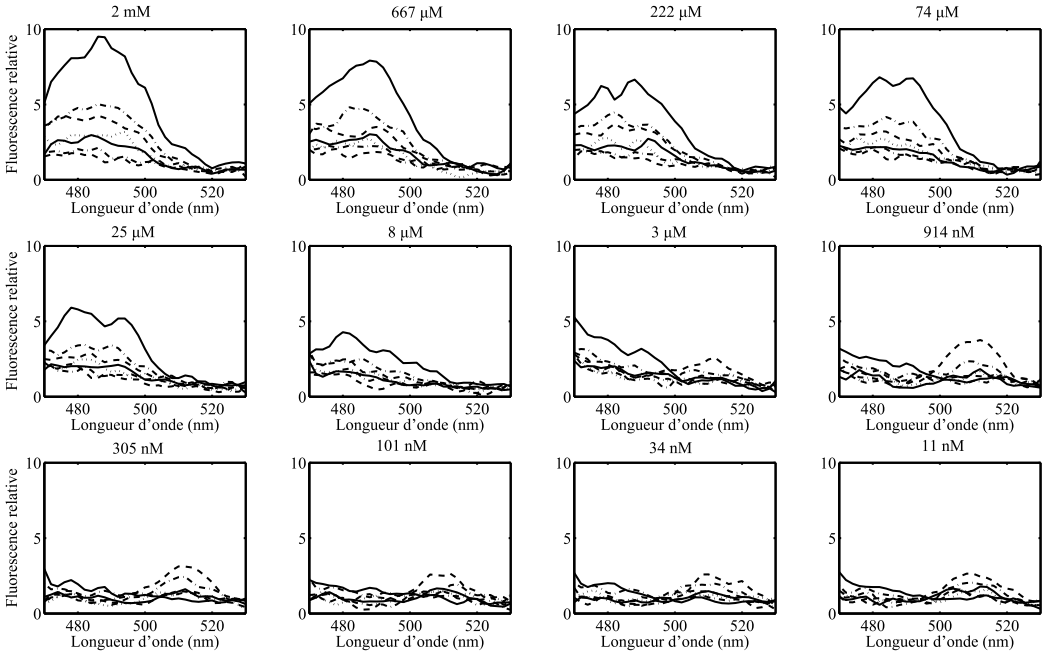


FIGURE *3.9 – Spectres synchrones bruts (non lissés) de fluorescence mesurés de* 470 *à* 530 nm *avec un pas de* 2 nm *et* $\Delta\lambda = 20$ nm. *Chaque graphique regroupe les spectres obtenus pour une concentration en fer (colonne de puits). Chaque spectre correspond à l'émission de fluorescence mesurée dans un puits pour un mélange de biosenseurs* $\mathcal{B}_{2G}$ *et* $\mathcal{B}_{1Y}$.

remarque également (figure 3.10(a)) que la deuxième source présente un niveau de fluorescence plus important que les deux autres.

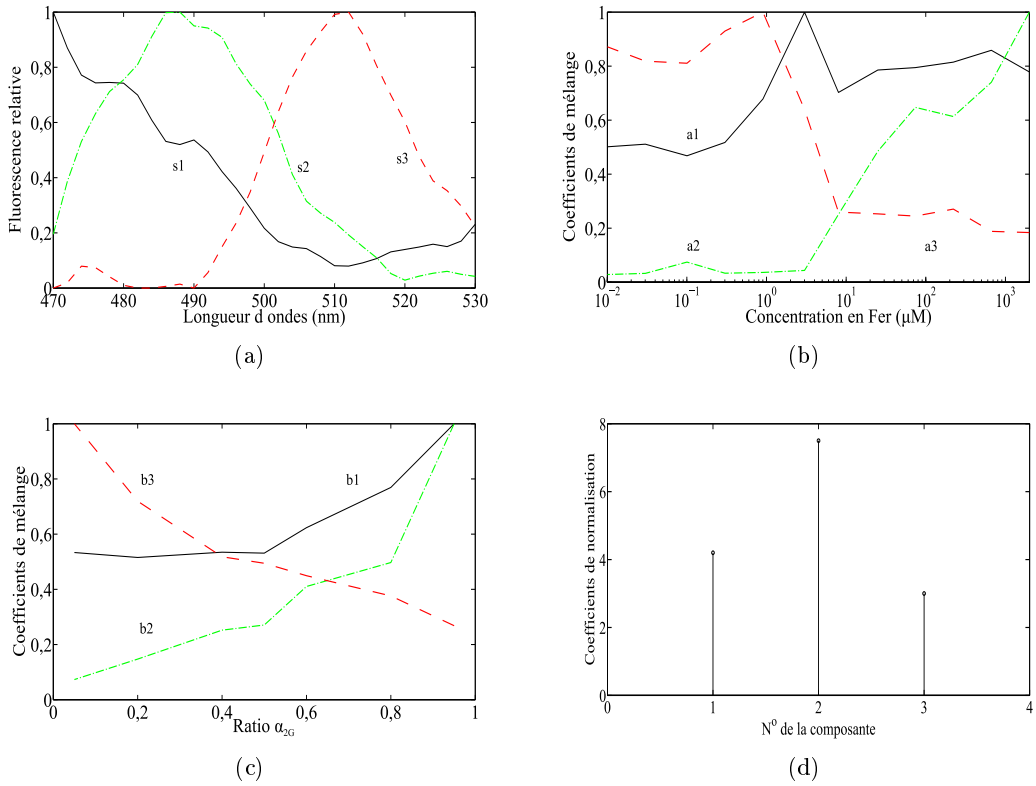


FIGURE *3.10 – Décomposition CP des spectres synchrones du mélange de biosenseurs ($\mathcal{B}_{2G}$, $\mathcal{B}_{1Y}$). Sources $\#_1$ (autofluorescence), $\#_2$ ($\mathcal{B}_{2G}$), $\#_3$ ($\mathcal{B}_{1Y}$). (a) Spectres synchrones de fluorescence des sources identifiées. (b) Réponses respectives en fonction de la concentration en fer. (c) Évolution des réponses en fonction du ratio de biosenseurs α_{2G}. (d) Coefficients de normalisation obtenus par mise à 1 de la valeur maximale.*

Les spectres de fluorescence des protéines fluorescentes GFP ($\mathbf{s}_2$) et EYFP ($\mathbf{s}_3$) sont bien estimés ceci grâce à l'utilisation des spectres synchrones qui permet d'optimiser l'excitation des sondes fluorescentes dans une gamme plus étendue de longueurs d'onde. Néanmoins, la source $\#_1$, associée à l'autofluorescence, montre un comportement anormal en fonction du ratio α_{2G} et de la concentration en fer 3 µM. Ce comportement peut caractériser une séparation incomplète des deux sources $\#_1$ et $\#_2$ ou peut être la conséquence de la normalisation par la valeur maximale qui viendrait amplifier une petite variation du signal.

Réponse du mélange de biosenseurs ($\mathcal{B}_{1G}$, $\mathcal{B}_{2Y}$)

La figure 3.11 montre le résultat de la décomposition CP des spectres synchrones mesurés sur les mélanges des biosenseurs $\mathcal{B}_{2Y}$ et $\mathcal{B}_{1G}$. On peut y voir l'évolution de l'intensité des différentes sources spectrales (Fig. 3.11(a)) en fonction de la concentration en fer (Fig. 3.11(b)) et du ratio

Couple de biosenseurs ($\mathcal{B}_{2Y}$, $\mathcal{B}_{1G}$)								
Nb. sources (F)	2	3	4	5	6	7	8	9
CorConDia	100	100	100	100	100	100	100	3
Fraction d'énergie résiduelle	12%	3%	1,5%	1,4%	1,3%	1,1%	1%	1%

TABLE *3.7 – Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie résiduelle pour la décomposition CP du jeu de données pour des nombres différents de sources recherchées.*

α_{2Y} (Fig. 3.11(c)). La figure 3.11(d) représente les coefficients de normalisation des sources. La décomposition montre l'existence d'une source $\#_1$ dont le spectre $\mathbf{s}_1$ s'étale sur une grande partie de plage de longueur d'onde. Cette source est plus faible pour les concentrations en fer les plus faibles et croît pour les concentrations supérieures à $1\,\mu$M. Le spectre estimé $\mathbf{s}_2$ présente un maximum à 488 nm. L'estimation des coefficients de mélange par la méthode CP montre une diminution de la fluorescence de cette source avec l'augmentation de la concentration de fer. Le signal est constant sur la plage $0,01$ à $0,9\,\mu$M puis diminue pour atteindre un second plateau allant des concentrations 8 à $2000\,\mu$M. De même, il diminue en fonction de l'augmentation de ratio de biosenseur α_{2Y}. Le spectre estimé $\mathbf{s}_3$ présente un maximum à 512 nm. L'estimation (courbe $\mathbf{a}_3$) de l'évolution de la fluorescence de cette source augmente avec l'augmentation de la concentration en fer. La décomposition CP montre une augmentation des coefficients de mélange ($\mathbf{b}_3$) en fonction du ratio α_{2Y}. Comme dans le mélange précédent, la source de fluorescence la plus importante est la source associée à la fluorecence de la GFP.

3.2.5 Conclusion

Dans cette expérience, nous avons abordé le problème de séparation de sources sur un cas concret de fluorescence de biosenseur bactérien. Nous avons mis en évidence les problèmes que peut poser le recouvrement spectral et la présence de sources inconnues, dans l'analyse de la réponse de biosenseur par des méthodes standards. Nous avons ensuite proposé comme solution à ces problèmes d'utiliser une méthode de séparation de sources exploitant la multi-dimensionalité des données (la décomposition CP). Néanmoins, nous avons vu que le choix de la protéine marquant le gène étudié peut influencer la qualité de la décomposition. Le marquage de gène faiblement exprimé par une protéine à faible fluorescence réduit la qualité des séparations de sources et provoque une mauvaise estimation des réponses. L'utilisation des spectres synchrones de fluorescence permet de mieux exciter les sources émettant à des longueurs d'onde plus élevées, ce qui améliore la qualité de l'estimation. Cependant, nous avons vu que le choix du couple *promoteur*::*rapporteur* a une importance dans la qualité de la décomposition et dans la modélisation du biosenseur. Il est alors important d'utiliser les connaissances sur le système afin que le développement des biosenseurs et des protocoles se fasse de manière à améliorer l'estimation des réponses.

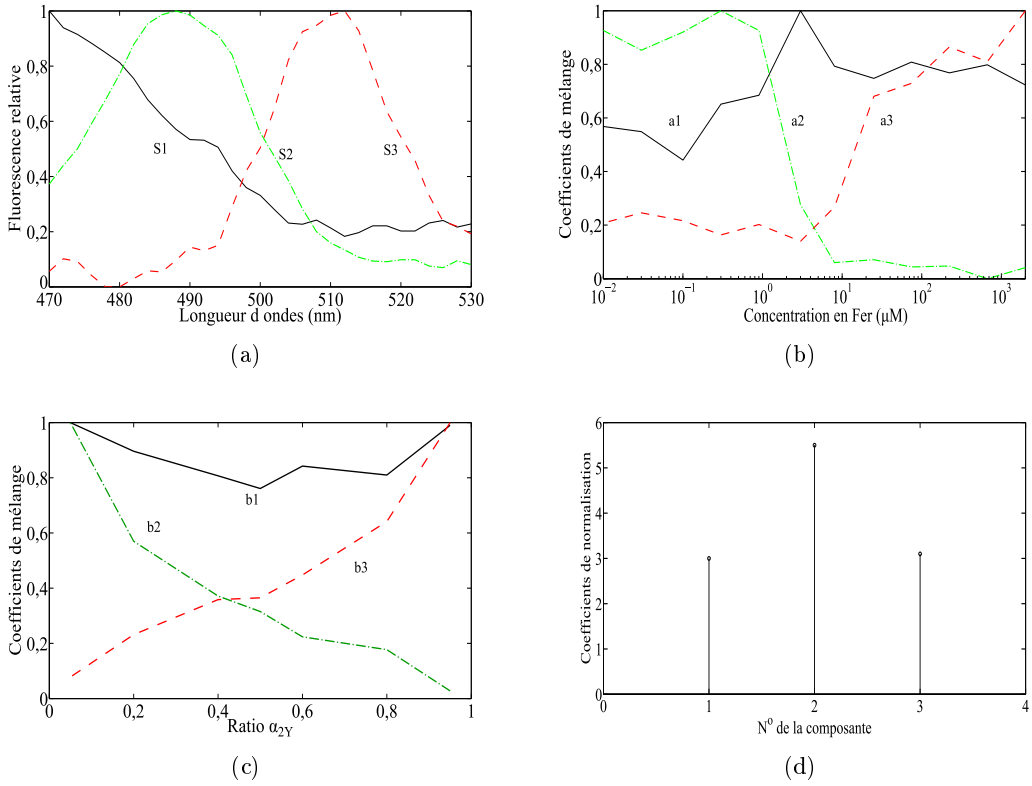


FIGURE 3.11 – *Décomposition CP des données synchrones. Sources* $\#_1$ *(autofluorescence),* $\#_2$ *(*$\mathcal{B}_{1G}$*),* $\#_3$ *(*$\mathcal{B}_{2Y}$*). (a) Spectres synchrones de fluorescence des sources identifiées. (b) Réponses respectives en fonction de la concentration en fer. (c) Évolution des réponses en fonction du ratio de biosenseurs* α_{2Y}*. (d) Coefficients de normalisation obtenus par mise à 1 de la valeur maximale de chaque réponse.*

3.3 Estimation conjointe des réponses temporelles et au cadmium des biosenseurs

Dans cette série d'expériences, nous nous intéressons à la détection de cadmium par le biosenseur PP$cadA_2$ introduit chez la bactérie P. putida (LIM23). Comme nous l'avons vu dans la section 3.1, l'approche courante (à une seule longueur d'onde) fixe un couple de longueurs d'onde d'excitation/émission caractéristique du rapporteur fluorescent marquant le promoteur dont la réponse est dépendante de la concentration en métal.

Or, cette mesure monolongueur d'onde intègre toutes les émissions fluorescentes des cellules et du milieu dans ce domaine de longueur d'onde. Cela a pour conséquence, une augmentation artificielle de la fluorescence de base du senseur et diminue sensiblement la limite de détection de la quantité de métal.

De plus, dans cette méthode la fluorescence est normalisée par l'absorbance mesurée à 600 nm c'est-à-dire à quantité de biosenseur constante. Cependant, cette estimation est peu précise en

milieu complexe car elle est influencée par la présence d'autres particules, substances ou matières absorbantes à cette longueur d'onde.

Les expériences suivantes ont été proposées afin d'étudier la cinétique de réponse d'un gène promoteur réagissant à la présence de cadmium. Les expériences réalisées permettront de générer des tableaux tridimensionnels de données de fluorescence dont les décompositions fourniront des modèles d'expression du gène plus robuste et plus sensible, grâce à l'extraction de la réponse du gène de la fluorescence globale. Les expériences réalisées permettront de tester la robustesse des méthodes multilinéaires de décomposition sur des données réelles.

3.3.1 Présentation de l'expérience

Le plan d'expérience a été construit de manière à générer un jeu de données trilinéaire. Le biosenseur utilisé est le $\mathcal{B}_{3R/4G}$ construit à partir de la bactérie *Pseudomonas putida* dont le gène réagissant au cadmium est marqué par le gène codant la production de la protéine fluorescente verte (GFP : émission max à 505 nm). Le biosenseur a été élaboré pour une production de fluorescence (GFP) dose-dépendante, via la fusion d'un gène rapporteur au promoteur PP$cadA_2$ codant pour une pompe de flux membranaire et inductible par le cadmium. Le biosenseur produit également la protéine rouge (DsRed : émission max à 582 nm) de manière constitutive. La fluorescence mesurée est une combinaison linéaire de F sources de fluorescence dont l'intensité varie en fonction de trois paramètres. Dans ces expériences, les trois paramètres sont la concentration en cadmium, le temps et la longueur d'onde. Ainsi, le tenseur de données $\boldsymbol{\mathcal{X}}$ peut être modélisé par :

$$\boldsymbol{\mathcal{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{s}_f \tag{3.2}$$

où $\mathbf{a}_f$ représente l'évolution de la source f en fonction de la concentration en cadmium, $\mathbf{b}_f$ en fonction du temps et $\mathbf{s}_f$ en fonction de la longueur d'onde.

Les variations du premier paramètre (gradient de concentration en cadmium) sont obtenues par la dilution d'une solution mère de cadmium (100 mM). La solution mère de cadmium est préparée à partir de 0,2 g de chlorure de cadmium ($CdCl_2$) diluée dans 10 mL d'eau ultrapure. Pour la première expérience, la gamme de concentrations comporte huit niveaux allant de 0 à 1 mM obtenus par dilutions successives, à l'aide d'un robot pipeteur, de la solution mère dans un milieu de culture (LB/5). Ce milieu de culture est particulièrement riches en molécules ou substances fluorescentes, qui peuvent perturber les mesures spectrales. Pour la seconde expérience, une gamme de concentrations en cadmium comportant dix niveaux est préparée, allant de 5 nM à 33 µM.

Les variations du second paramètre sont obtenues par la réponse cinétique (en fonction du temps) du biosenseur. Des acquisitions spectrales pour chaque concentration sont effectuées à différents temps d'incubation. Les temps de maturation des marqueurs et la cinétique de réponse des promoteurs génèrent de la diversité dans les mesures.

Le troisième paramètre est la dimension spectrale. Pour la première expérience, les spectres

d'émission de fluorescence sont mesurés entre 440 et 600 nm avec un pas de 4 nm pour une excitation à 390 nm. Dans la seconde expérience, le mode spectral est composé des spectres synchrones de fluorescence relevés de 406 nm à 600 nm avec un $\Delta\lambda = 25$ nm. Les spectres de fluorescence ont été acquis à l'aide d'un spectrofluorimètre FLX-Xenius®SAFAS.

Le tableau 3.9 représente le plan de plaque de l'expérience dans lequel est indiqué le ratio de biosenseurs et la concentration en fer pour chacun des puits.

Expérience Cadmium			
	a	b	s
$\mathcal{B}_{3R/4G}$	$\mathbf{a}_1$ $\mathbf{a}_2$	$\mathbf{b}_1$ $\mathbf{b}_2$	$\mathbf{s}_1$ $\mathbf{s}_2$

TABLE *3.8 – Représentation des comportements attendus ($\mathbf{a}_i$, $\mathbf{b}_i$) pour le biosenseur bicolore $\mathcal{B}_{3R/4G}$ utilisé.*

	1
A	0 M
B	1 nM
C	0,1 µM
D	1 µM
E	10 µM
F	30 µM
G	100 µM
H	1000 µM

N^o tubes	
1	0.005 µM
2	0.015 µM
3	0.046 µM
4	0.14 µM
5	0.41 µM
6	1.24 µM
7	3.71 µM
8	6.67 µM
9	11.11 µM
10	33.33 µM

TABLE *3.9 – Plans des expériences réalisées. Sur une colonne de la microplaque 96 puits, la variation de la concentration en cadmium suit une gamme de 0 à 1 mM de cadmium par dilution en cascade d'une solution mère de chlorure de cadmium ($CdCl_2$). Pour l'expérience réalisée en tubes, la variation de la concentration en cadmium suit une gamme de concentration allant de 5 nM à 33 µM.*

3.3.2 Analyse de la décomposition CP des spectres d'émission

La première 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 puits, on ajoute des quantités variables de cadmium dont les concentrations varient entre 0 et 1mM). Les spectres sont relevés de 440 à 600 nm avec un pas de 4 nm pour une excitation à 390 nm. Le tableau tridimensionnel de données est de dimension $41 \times 8 \times 4$, dont le premier mode est la longueur d'onde du spectre de fluorescence, le second mode la concentration de cadmium et le troisième mode le temps. Dans la représentation des données du tableau tridimensionnel (Fig. 3.12), nous pouvons remarquer que

le signal de fluorescence DsRed de la protéine est très peu visible. On suppose que la longueur d'onde d'excitation unique à 390 nm n'est pas suffisante pour que cette source se dégage des bandes de fluorescence très larges générées par l'autofluorescence.

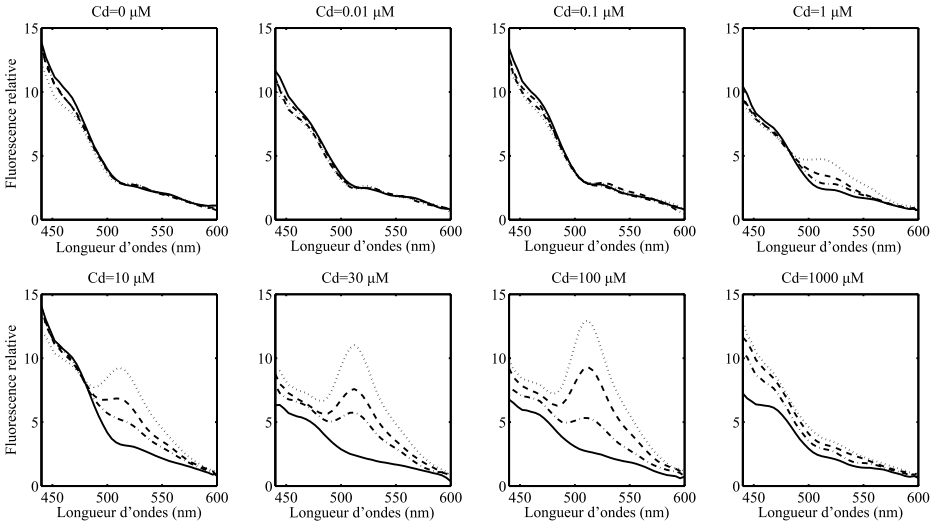


FIGURE *3.12 – Spectres d'émission de fluorescence mesurés de* 440 *à* 600 nm *avec un pas de* 4 nm *pour une excitation à* 390 nm. *Chaque courbe de chacun des graphiques représente le spectre de fluorescence émis dans un puits à un instant t. Chaque puits contient une concentration différente de cadmium.*

Le tableau 3.10 regroupe les indices calculés afin de déterminer le nombre de sources recherchées. Nous recherchons au moins une source (GFP) et l'indice CorConDia montre que le modèle CP est valide jusqu'à 4 sources. De plus, l'évolution de l'énergie résiduelle en fonction du nombre de sources est faible au-delà de 2 sources, ce qui suggére de fixer le nombre de sources à 2 pour les décompositions CP.

Biosenseur $PPcadA_2$::*gfp*								
Nb. sources	1	2	3	4	5	6	7	8
CorConDia	100	100	99	99	16	1	1	0,3
Fraction d'énergie résiduelle	5,5%	0,38%	0,07%	0,05%	0,03%	0,02%	0,01%	0,01%

TABLE *3.10 – Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie résiduelle pour la décomposition CP du jeu de données pour des nombres différents de sources recherchées.*

Le modèle trilinéaire à deux sources explique plus de 99% de l'énergie totale des données initiales. On peut y voir l'évolution de l'intensité d'une source spectrale (Fig. 3.13(a)) en fonction de la concentration en cadmium (Fig. 3.13(b)) et du temps (Fig. 3.13(c)). La décomposition montre l'existence d'un spectre $\mathbf{s}_1$ dont le maximum d'émission se situe à 440 nm et dont l'intensité croit au cours du temps. Le second spectre estimé $\mathbf{s}_2$ montre un maximum d'émission aux alentours de 512 nm. La fluorescence de cette source augmente avec la concentration en cadmium jusqu'à

un seuil de 100 µM et diminue pour les concentrations les plus fortes. La figure 3.13(c), on peut voir l'évolution croissante de la fluorescence des sources $\#_1$ et $\#_2$ en fonction du temps.

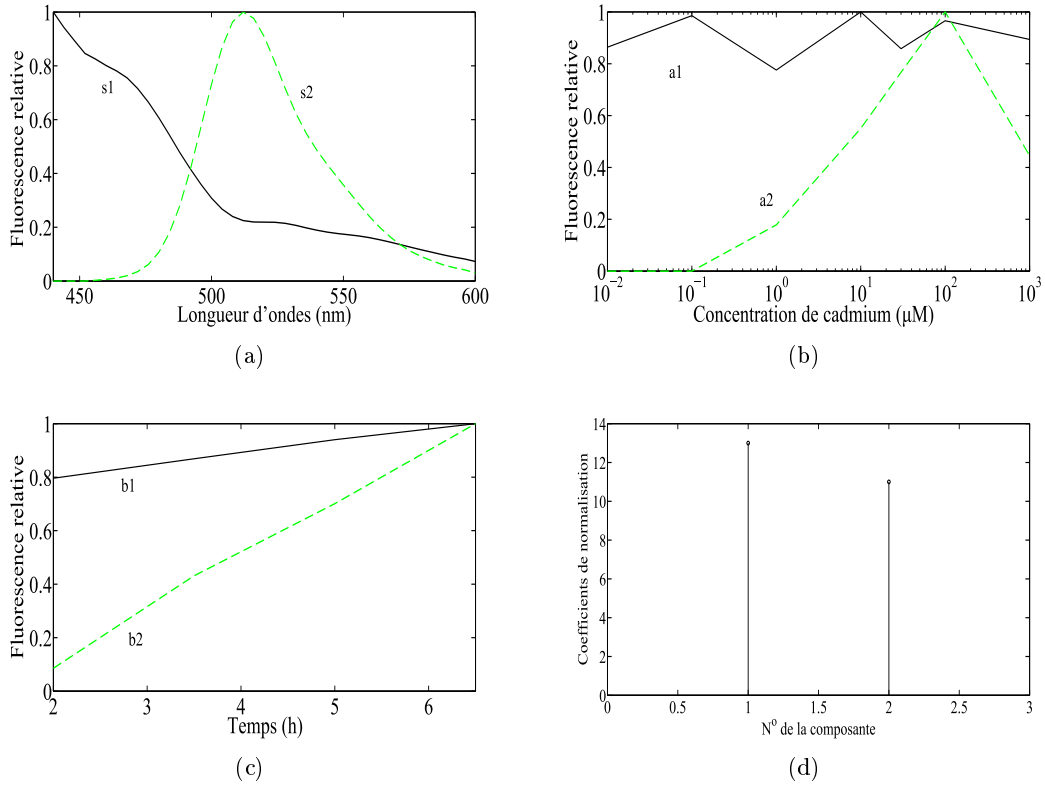


FIGURE *3.13 – Décomposition CP des spectres d'émission de fluorescence mesurés sur le biosenseur $\mathcal{B}_{3R/4G}$. Sources $\#_1$ (autofluorescence), $\#_2$ (GFP). (a) Spectres d'émission des sources identifiées. (b) Réponses respectives des sources $\#_1$ et $\#_2$ en fonction de la concentration en cadmium. (c) Évolution des réponses en fonction du temps. (d) (d) Coefficients de normalisation obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

La source $\#_1$ peut être associée à l'autofluorescence de composants cellulaires bactériens, car son évolution au cours du temps suit celle de la densité bactérienne, que l'on peut estimer par la mesure de la densité optique (à 600 nm fig. 3.17(b)). Le marqueur fluorescent associé au gène promoteur $PPcadA_2$ est représenté par son spectre estimé $\mathbf{s}_2$. Il est conforme à celui attendu pour une GFP avec un maximum d'émission à 510 nm. La concentration en GFP dans la bactérie augmente avec la concentration de cadmium jusqu'au seuil toxique pour la bactérie (entre 0,1 et 1mM). Ces deux phénomènes sont soulignés par la courbe de réponse *en cloche* du gène rapporteur (GFP), estimée par la décomposition CP (ligne discontinu sur la figure 3.13(b)).

L'évolution croissante de la fluorescence des sources $\#_1$ et $\#_2$ en fonction du temps s'explique par la croissance cellulaire. L'accroissement du nombre de cellules 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'aug-

mentation plus rapide du fluorochrome GFP s'explique sans doute par l'accumulation progressive de GFP dans le cytoplasme des cellules entre chaque division cellulaire.

La qualité de ces résultats est validée en comparant les spectres et réponses estimés aux spectres de fluorescence montrés dans le chapitre introductif (1.3) et à la courbe d'expression du gène PP$cadA_2$ [7].

La figure 3.14 montre l'évolution, en fonction de la concentration en cadmium, de la fluorescence associée à la GFP, par une méthode standard [45] (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 $490\,\text{nm} \pm 10\,\text{nm}$ et à un temps donné (15h après mise en contact avec le métal). L'estimation de la fluorescence pour les faibles concentrations de cadmium montre une différence significative entre les deux méthodes. 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 confondue avec l'autofluorescence de la bactérie. La méthode standard, en ignorant l'existence d'autofluorescence, surestime la réponse du rapporteur plus faible. Au contraire, la méthode CP extrait chaque composante et ne fournit que les variations de fluorescence du rapporteur en fonction de la teneur en cadmium quelque soit la temps de mesure. La décomposition CP améliore donc la limite de détection et la sensibilité du biosenseur aux faibles concentrations en cadmium.

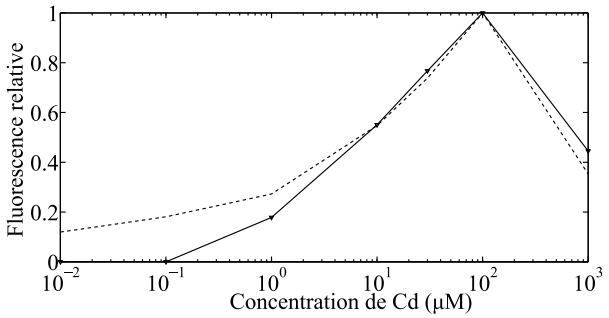


FIGURE *3.14 – Estimation de la réponse du promoteur PP$cadA_2$ en fonction de la concentration en cadmium. En pointillé : estimation classique basée sur une mesure monolongueur d'onde à un temps donné. En trait plein : estimation par la méthode CP.*

L'estimation par décomposition CP donne des résultats qui concordent avec le comportement attendu du biosenseur. Dans le cadre de cette expérience, les paramètres temps et concentration ont été choisis comme diversités car ils sont facilement mesurables. Toutefois, on peut regretter le manque de variabilité de réponse engrendrée par le paramètre temporel. Le lien temporel peut également réduire la robustesse de la décomposition et augmenter le risque d'instabilité en s'éloignant de l'unicité de la solution. Par exemple, la croissance bactérienne induit la même augmentation de fluorescence pour chaque fluorochrome alors que la concentration de cadmium n'agit pas sur l'autofluorescence. La méthode actuelle consiste à choisir des fluorochromes avec des spectres d'émission les plus éloignés possible et d'effectuer des acquisitions à la longueur d'onde maximum de chacun des fluorochromes. Cela permet d'extraire la réponse de chaques

composantes, mais limite le nombre de fluorochromes utilisables dans un biosenseur. Dans le cadre de notre approche, le recouvrement spectral (entre sources ou avec l'autofluorescence) n'est plus 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.

3.3.3 Analyse de la décomposition CP de spectres synchrones

L'expérience consiste à étudier les spectres synchrones de fluorescence d'un biosenseur sensible au cadmium. La culture des biosenseurs est effectuée dans 11 tubes Falcon contenant du milieu LB/5 (18,8 mL), les bactéries (0,9 mL de suspension bactérienne) et pour chaque tube 200 µL d'une des solutions stock de cadmium. L'incubation des biosenseurs est donc réalisée sur une gamme de concentrations en cadmium comportant dix niveaux allant de 5 nM à 33 µM, le onzième tube étant le tube témoin. L'incubation des tubes est faite à 37°C sous agitation (240 RPM) sur portoir incliné. Deux prélèvements sont effectués à intervalles d'environ 1h30 dans chacun des 10 tubes. Le premier prélèvement est de 1 mL de suspension bactérienne. Le second prélèvement est de 1 mL de surnageant, après centrifugation. Les spectres synchrones de fluorescence de chaque prélèvement sont relevés de 406 nm à 600 nm pour un $\Delta\lambda = 25$ nm. Les deux tableaux tridimensionnnels de données obtenus sont de dimension $98 \times 10 \times 6$. Le premier mode est la longueur d'onde du spectre de fluorescence, le second mode la concentration en cadmium et le troisième, le temps (Voir fig.3.15 et fig.3.16.)

La figure 3.17 représente l'évolution de l'absorbance à 600 nm en fonction de la concentration en cadmium (Fig.3.17(a)) et en fonction du temps (Fig.3.17(b)). Ces graphiques montrent une diminution de l'absorbance pour les concentrations allant de 6µM à 33µM mettant en évidence la toxicité de cette élément pour les bactéries. Pour les concentrations inférieures à 6µM le cadmium ne semble pas avoir d'influence sur l'absorbance et l'on retrouve l'évolution temporelle caractéristique du développement cellulaire.

Réponse de la suspension bactérienne

Le tableau 3.11 regroupe les indices calculés afin de déterminer le nombre de sources recherchées. Nous recherchons au moins une source (GFP). L'indice CorConDia montre que le modèle CP est valide jusqu'à 6 sources et l'énergie résiduelle est quasi nulle à partir de 3 sources.

Les modèles trilinéaires à trois sources expliquent plus de 99% de l'énergie totale des données mesurées. La méthode permet conjointement d'indentifier les sources de fluorescence présentes dans la suspension bactérienne et d'estimer leur réponse en fonction de la concentration en métal ou du temps.

L'évolution de l'intensité des sources spectrales (Fig. 3.18(a)) est représentée en fonction de la concentration en cadmium (Fig. 3.18(b)) et du temps (Fig. 3.18(c)).

La décomposition montre l'existence d'une source $\#_1$ dont le maximum d'émission se situe à 408 nm et dont l'intensité évolue peu en fonction de la concentration en cadmium ou du temps. Nous remarquons néanmoins une légère augmentation pour les fortes concentrations en cadmium

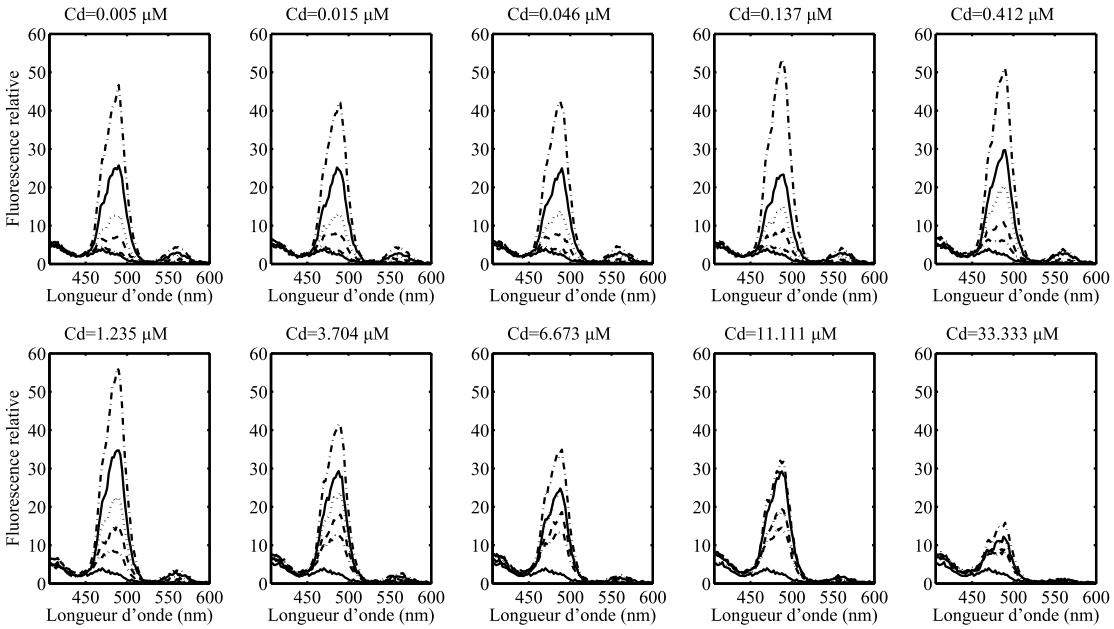


FIGURE *3.15 – Evolution temporelle des spectres synchrones de fluorescence mesurés dans les suspensions bactériennes. Chaque courbe représente le spectre de fluorescence obtenus à un instant t pour une concentration en cadmium.*

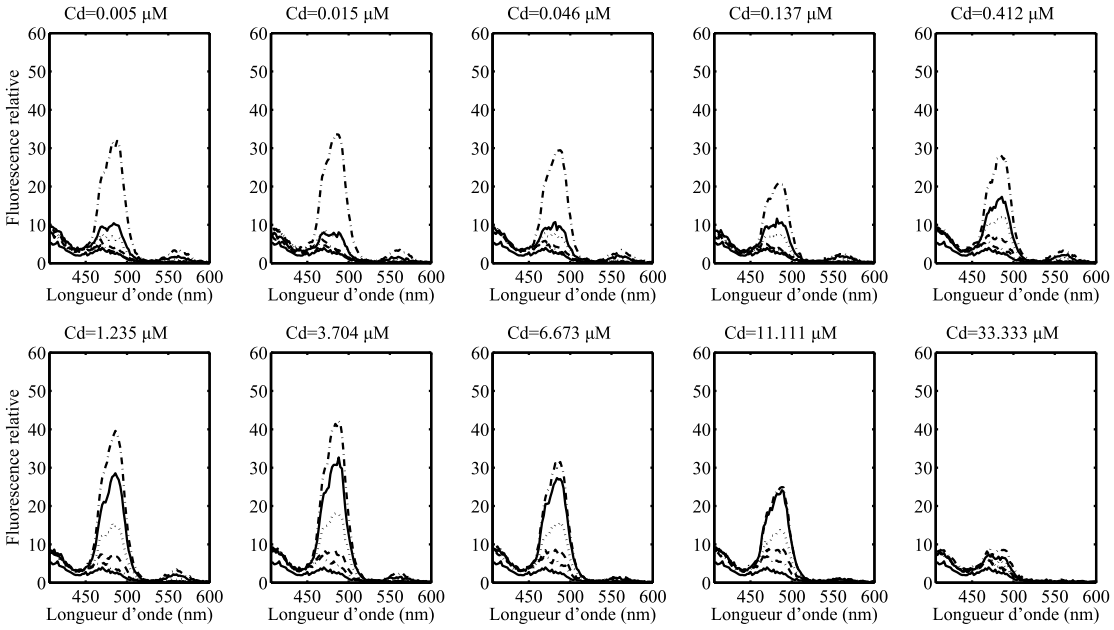
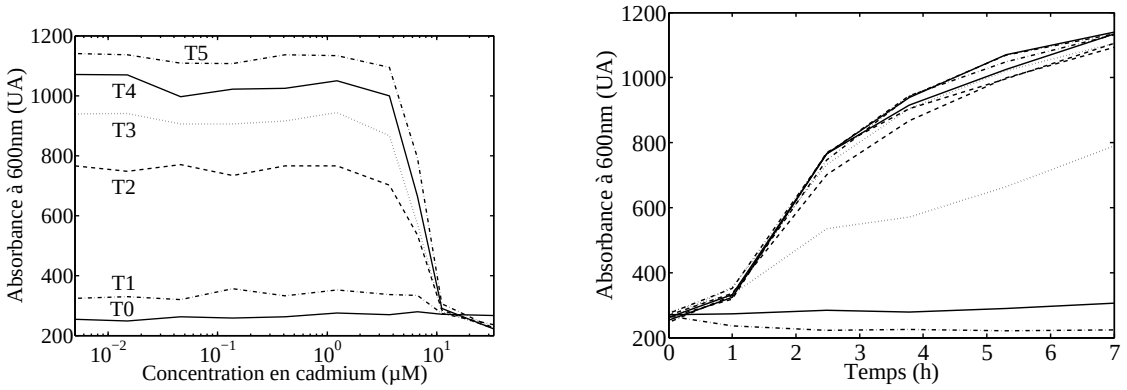


FIGURE *3.16 – Evolution temporelle des spectres synchrones de fluorescence mesurés dans le surnageant filtré après centrifugation (sans bactéries). Chaque courbe représente le spectre de fluorescence à un instant t pour une concentration en cadmium.*



(a) (b)

FIGURE 3.17 – *Évolution de l'absorbance à* 600 nm *mesurée dans les cuves. (a) Évolution de l'absorbance en fonction de la concentration en cadmium. Chaque courbe représente un instant de mesure. (b) Évolution de l'absorbance en fonction du temps. Chaque courbe représente une concentration en cadmium.*

Tenseur de données de la suspension bactérienne								
Nb. sources	1	2	3	4	5	6	7	8
CorConDia	100	100	100	100	100	100	87	72
Fraction d'énergie résiduelle	5%	1,2%	0,3%	0,2%	0,1%	0,1%	0,1%	0,1%
Tenseur de données mesurées après filtration								
Nb. sources	1	2	3	4	5	6	7	8
CorConDia	100	100	100	100	100	93	93	88
Fraction d'énergie résiduelle	8,6%	2,5%	0,3%	0,2%	0,2%	0,2%	0,1%	0,1%

TABLE 3.11 – *Comparaison des valeurs de l'indice de cohérence du cœur, fraction d'énergie résiduelle pour la décomposition CP du jeu de données pour des nombres différents de sources recherchées.*

et une légère diminution au cours du temps. La décomposition montre l'existence d'une source dont le spectre bimodale $\mathbf{s}_3$ possède des maxima locaux à 490 nm et 560 nm. Cette source est quasi constante pour des concentrations de cadmium inférieures à 1 µM puis diminue fortement pour les concentrations supérieures. Le second spectre estimé $\mathbf{s}_2$ montre un maximum d'émission à 490 nm. La fluorescence de cette source augmente avec la concentration en cadmium jusqu'à un seuil de 10 µM et diminue ensuite. Sur la figure 3.18(c), on peut voir l'évolution croissante de la fluorescence des sources $\#_2$ et $\#_3$ en fonction du temps. D'après la figure 3.18(d), la fluorescence de la source $\#_3$ est la plus intense et celle de la source $\#_1$ la plus faible.

De prime abord, l'estimation de l'évolution des sources peut paraître surprenante ; il y aurait deux systèmes sensibles au cadmium et un troisième insensible. En regardant attentivement l'estimation des sources spectrales obtenues, on s'aperçoit que les deux systèmes rapporteurs fluorescents implantés dans la bactérie n'ont pas été séparés. Le protocole expérimental avait

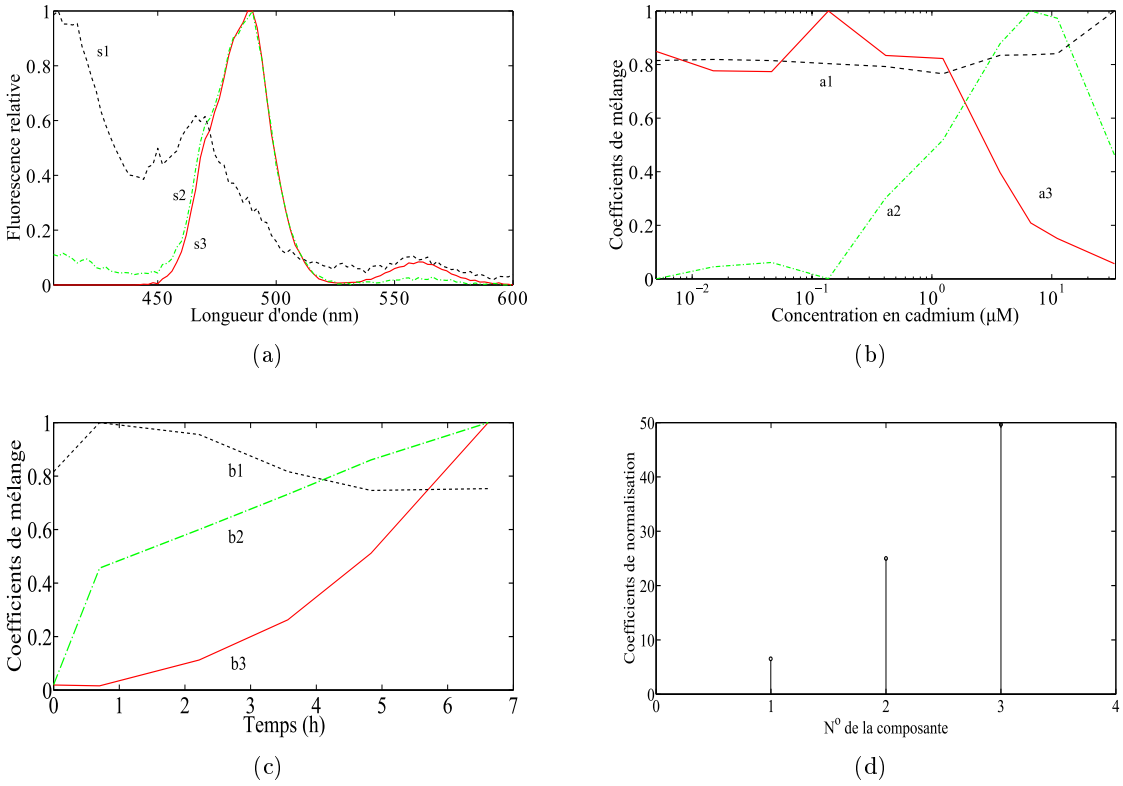


FIGURE *3.18 – Décomposition CP des spectres synchrones mesurés sur les suspensions de biosenseurs bactériens. (a) Spectres synchrones de fluorescence des sources identifiées. (b) Réponses respectives en fonction de la concentration en cadmium. (c) Évolution des réponses en fonction du temps. (d) Coefficients de normalisation obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

été conçu de manière à générer la diversité suffisante pour identifier les systèmes fluorescent de manière exclusive et conduire à la détermination de deux sources spectrales : l'une présentant un maximum de fluorescence à 490 nm et l'autre à 560 nm, conformément aux spectres synchrones connus pour les deux protéines fluorescentes GFP et DsRed.

Comme il n'y a aucune raison légitime de douter de la décomposition obtenue, une autre interprétation est proposée.

La source $\#_1$ est attribuée à l'autofluorescence du milieu dont la contribution est quasi-invariante et indépendante du temps et de la concentration en cadmium. Les faibles variations seraient alors les conséquences d'un changement de densité optique liée à l'augmentation de la quantité de biosenseurs au cours du temps et à une diminution de cette quantité pour les fortes concentrations en cadmium.

La source $\#_2$, dont le spectre est uni-modale, est attribuée au système $PPcadA_2$::*gfp*. Son évolution en fonction de la concentration en cadmium est caractéristique de la réponse du promoteur PP$cadA_2$. De plus, on retrouve dans l'évolution temporelle le temps d'induction du promoteur,

dont l'activité optimale est obtenue au bout d'une heure, et la croissante microbienne qui fait croître la fluorescence au cours du temps.

La source bi-modale $\#_3$ est attribuée à une expression conjointe et corrélée des deux systèmes $PPcadA_2$::*gfp* et $P_{A1/04/03}$::*DsRed*. Ce lien entre les deux systèmes est inattendu et suppose que le système $PPcadA_2$::*gfp* a un fonctionnement basal similaire au système $P_{A1/04/03}$::*DsRed* mais leur dépendance vis-à-vis de la concentration en cadmium et du temps doit être expliquée. La solution retenue est qu'il s'agit de résidus cellulaires post-mortem ou de protéines fluorescentes excrétées dans le milieu. Les réponses observées traduiraient à la fois une dégénérescence progressive des cellules bactériennes au cours du temps et le développement perturbé par les fortes concentrations en cadmium. L'étude de la décomposition du jeu de données obtenus sur les filtrats (Fig. 3.19) permettra de confirmer cette hypothèse.

Réponse du surnageant après centrifugation

La décomposition montre l'existence d'un spectre $\mathbf{s}_1$ dont le maximum d'émission se situe à 406 nm et dont l'intensité n'évolue pas en fonction de la concentration en cadmium. L'évolution de la fluorescence de cette source au cours du temps est très faible hormis entre le début d'expérience et le premier temps de mesure. La seconde source estimée ($\#_2$) posséde un maximum d'émission à 490 nm. L'intensité de fluorescence de cette source augmente progressivement avec la concentration en cadmium jusqu'à un seuil de 3 µM et diminue pour les concentrations supérieures. Sur la figure 3.19(c), on peut voir l'évolution croissante de la fluorescence des sources $\#_2$ et $\#_3$ en fonction du temps. La décomposition montre l'existence d'un spectre $\mathbf{s}_3$ bimodale ayant deux maxima locaux en 490 nm et 560 nm. Cette source décroit pour les concentrations en cadmium comprises entre 0,03 et 0,1 µM et les concentrations supérieures à 3 µM. D'après la figure 3.19(d), les sources $\#_2$ et $\#_3$ ont une fluorescence équivalente et plus intense que celle de la source $\#_1$.

La source $\#_1$ est attribuée à l'autofluorescence du milieu dont nous retrouvons le comportement invariant en fonction du temps et de la concentration en cadmium. Nous remarquons également que l'absence de bactéries supprime les faibles variations de fluorescence pour les fortes concentrations en cadmium et temporelle. C'est un argument pour supposer un effet décrantage (*i.e.* influence de la quantité de bactérie sur l'excitation du milieu).

La source $\#_2$ est attribuée à la fluorescence de la GFP dont l'évolution en fonction de la concentration en cadmium est caractéristique du marquage du gène promoteur PP$cadA_2$. Toutefois, contrairement au résultat précédent, l'évolution temporelle ne permet pas de retrouver l'augmentation de fluorescence au bout d'une heure correspondant à l'induction du promoteur car elle s'effectue à l'intérieur de la bactérie. Néanmoins, nous retrouvons l'évolution correspondant à la dégénéréscence progressive des cellules bactériennes au cours du temps.

La source $\#_3$ est attribuée à l'expression conjointe des deux systèmes $PPcadA_2$::*gfp* et $P_{A1/04/03}$::*DsRed* qui est supposée être le résidu cellulaire post-mortem ou l'excrétion de protéines fluorescentes issues de l'expression basale des promoteurs PP$cadA_2$ et $P_{A1/04/03}$. Selon

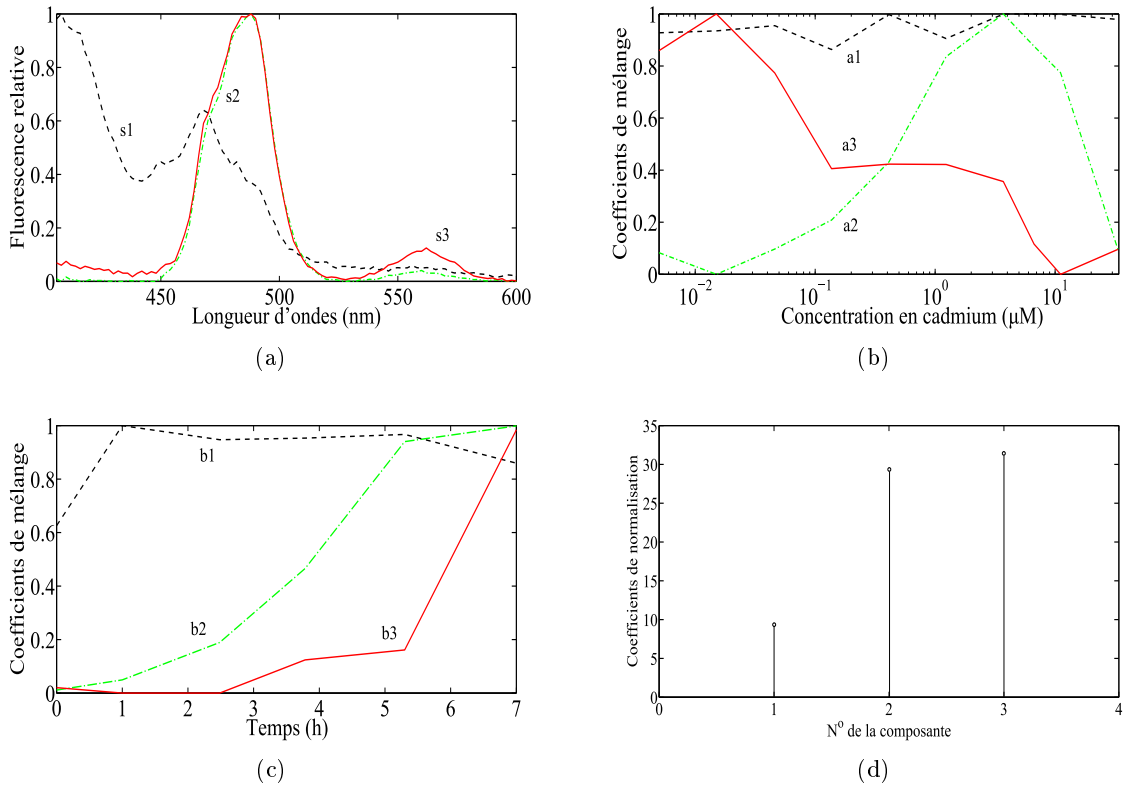


FIGURE *3.19 – Décomposition CP des spectres synchrones mesurés sur les filtrats. (a) Spectres synchrones de fluorescence des sources identifiées. (b) Réponses respectives en fonction de la concentration en cadmium des sources. (c) Évolution des réponses en fonction du temps. (d) Coefficients de normalisation obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

cette hypothèse, la réponse en fonction du cadmium devrait suivre la quantité de biosenseur, donc constante pour les concentrations faibles en cadmium ($\leq 1\,\mu M$) puis diminuer pour les concentrations supérieures. Pourtant la décomposition CP ne permet pas de retrouver ce comportement. De même, l'évolution temporelle de cette source devrait être similaire à l'évolution temporelle de la source $\#_2$, car elles sont toutes deux issues d'un résidu cellulaire post-mortem ou d'une excrétion bactérienne. Cependant, ce lien pourrait expliquer que l'on ne retrouve pas les comportements attendues. S'il existe un lien entre les sources d'un mode l'utilisation d'une contrainte plus forte, telle que l'unimodalité, ou l'utilisation de la décomposition PARALIND s'imposent. Nous proposons alors de tester ces deux approches.

3.3.4 Décomposition PARALIND

Dans cette section, nous proposons de tester la décomposition PARALIND sur les données de fluorescence mesurées dans le milieu après centrifugation et filtration des biosenseurs. La matrice des interactions est placée sur le troisième mode (temporelle). Le nombre de sources de

la décomposition PARALIND est fixé à 3 (R=3) et 2 évolutions temporelles (S=2). La matrice des interactions $\mathbf{\Omega}$ représentant les liens entre les évolutions temporelles des différentes sources et de taille 2×3. La matrice $\tilde{\mathbf{B}}(6 \times 2)$ regroupera les cinétiques des sources. Le tenseur des données s'écrit $\boldsymbol{\mathcal{X}} = [\![\mathbf{S}, \mathbf{A}, \tilde{\mathbf{B}}\mathbf{\Omega}]\!]$.

La décomposition PARALIND des données fournie plusieurs décompositions montrant l'existence de plusieurs minima locaux la solution retenue (fig. 3.20) est la décomposition ayant le minimum d'énergie résiduelle, soit 98% de l'énergie totale expliquée.

Sur la figure 3.20, on peut voir l'évolution des l'intensités des sources spectrales (a) en fonction de la concentration en cadmium (b) et du temps (c). La décomposition montre également l'existence d'un spectre $\mathbf{s}_1$ dont le maximum se situe à 406 nm et dont l'intensité n'évolue pas en fonction du cadmium. Le second spectre estimé $\mathbf{s}_2$ montre un maximum d'émission à 490 nm. La fluorescence de cette source augmente avec la concentration de cadmium jusqu'à un seuil de 10 µM et diminue pour la concentration supérieure. Le troisième spectre $\mathbf{s}_3$ bi-modale montre deux maxima locaux à 490 et 560 nm. Cette troisième source varie fortement avec la concentration en cadmium ; elle diminue sur une première plage allant de 5 nM à $0,1$ µM puis augmente sur la plage de $0,1$ à 1 µM avant de diminuer pour les concentrations supérieures. La figure 3.20(c) représente la matrice $\mathbf{B} = \tilde{\mathbf{B}}\mathbf{\Omega}$ regroupant l'évolution temporelle de la fluorescence des trois sources. On peut y voir l'augmentation au cours du temps des sources $\#_2$ et $\#_3$. Au cours du temps, la source $\#_1$ est quasi-constante. La matrice $\mathbf{\Omega}$ est estimée à

$$\mathbf{\Omega} = \begin{bmatrix} 0,89 & 4,18 & 6,17 \\ 0 & 5,01 & 0,22 \end{bmatrix}. \tag{3.3}$$

La source $\#_1$ est attribuée à l'autofluorescence du milieu dont nous retrouvons le comportement attendu.

La source $\#_2$ est attribuée à la fluorescence de la GFP dont l'estimation du spectre de fluorescence est similiaire au spectre connu de la GFP. Le comportement en fonction de la concentration en cadmium est surestimé, par rapport au décomposition précédente, pour la plage de concentrations allant de $0,046$ à $0,4$ µM.

La source bi-modale $\#_3$ est attribuée à l'émission conjointe des protéines GFP et DsRed. On retrouve le comportement attendu hormis pour la plage de concentration allant de $0,046$ à $0,4$ µM où la fluorescence de la source est sous-estimée.

Les problémes d'estimation des coefficients de mélange $\mathbf{a}_2$ et $\mathbf{a}_3$ sur la plage de concentrations allant de $0,046$ à $0,4$ µM indique qu'il persiste des problèmes liées au lien temporelle entre les sources $\#_2$ et $\#_3$.

3.3.5 Contrainte d'unimodalité sur les spectres

Dans cette section, nous proposons de tester la décomposition CP avec contrainte d'unimodalité sur le mode spectral. La décomposition est appliquée sur les données de fluorescence mesurées sur la suspension bactérienne. La plage de longueurs d'onde allant de 406 à 454 est tron-

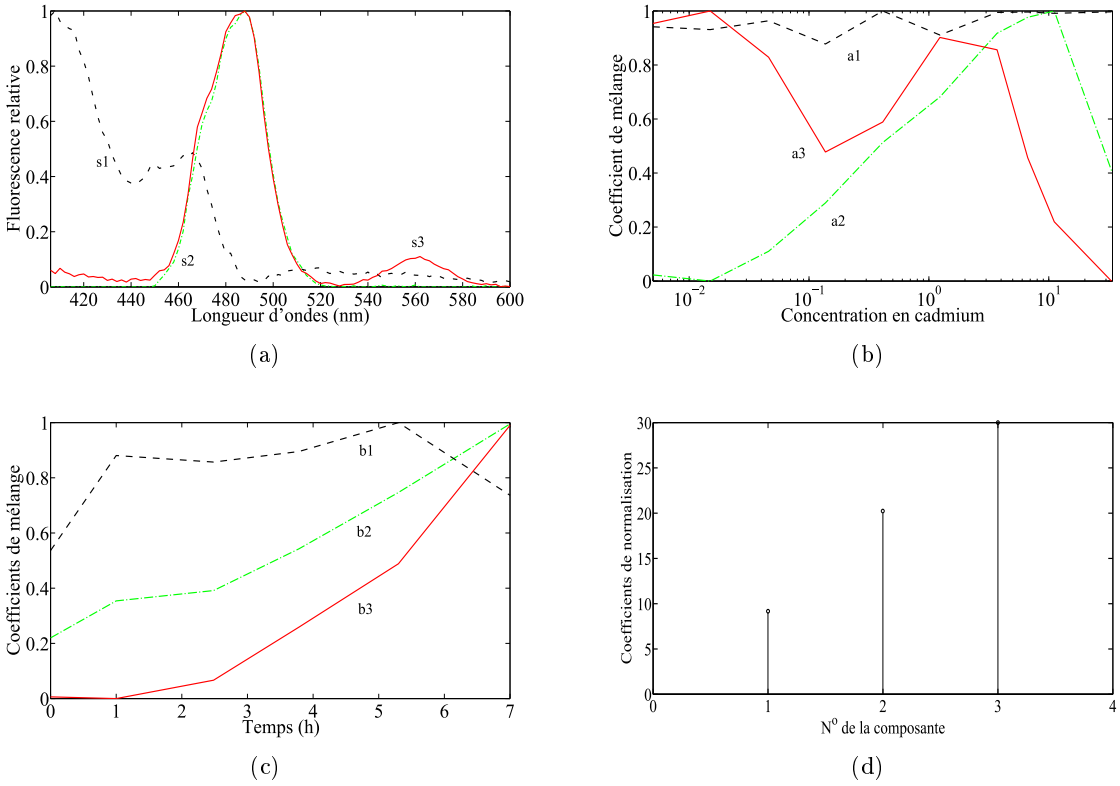


FIGURE *3.20 – Décomposition PARALIND des spectres synchrones mesurés sur les filtrats. (a) Spectres synchrones des sources identifiées. (b) Réponse en fonction de la concentration en cadmium. (c) Évolution en fonction du temps. (d) Coefficients de normalisation obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

quée afin de ne conserver que la partie unimodale de la source correpondant à l'autofluorescence. Le nouveau tableau tridimensionnel regroupant les données est donc de dimension $74 \times 10 \times 6$.

Les résultats de la décomposition CP des données de fluorescence de la suspension bactérienne sont présentées figure 3.21. La décomposition CP a été réalisée sous contrainte de positivité sur les trois modes et avec une contrainte supplémentaire d'unimodalité sur le mode spectral. L'évolution de l'intensité des sources spectrales (Fig.3.21(a)) est représentée en fonction de la concentration en cadmium (Fig.3.21(b)) et du temps ((Fig.3.21(c))). La figure 3.21(d) représente les coefficients de normalisation obtenus par la mise à 1 de la valeur maximale de chaque réponse.

La décomposition CP montre l'existence d'une source dont le spectre $\mathbf{s}_1$ a un maximum en 470 nm. L'intensité de cette source évolue peu en fonction de la concentration en cadmium et en fonction du temps.

Le second spectre estimé $\mathbf{s}_2$ montre un maximum d'émission à 490 nm. L'intensité de fluorescence n'évolue pas dans la plage de concentration en cadmium allant de 5 nM à 1 µM et décroît pour les concentrations supérieures à 1 µM. L'intensité de fluorescence de cette source est croissante au cours du temps.

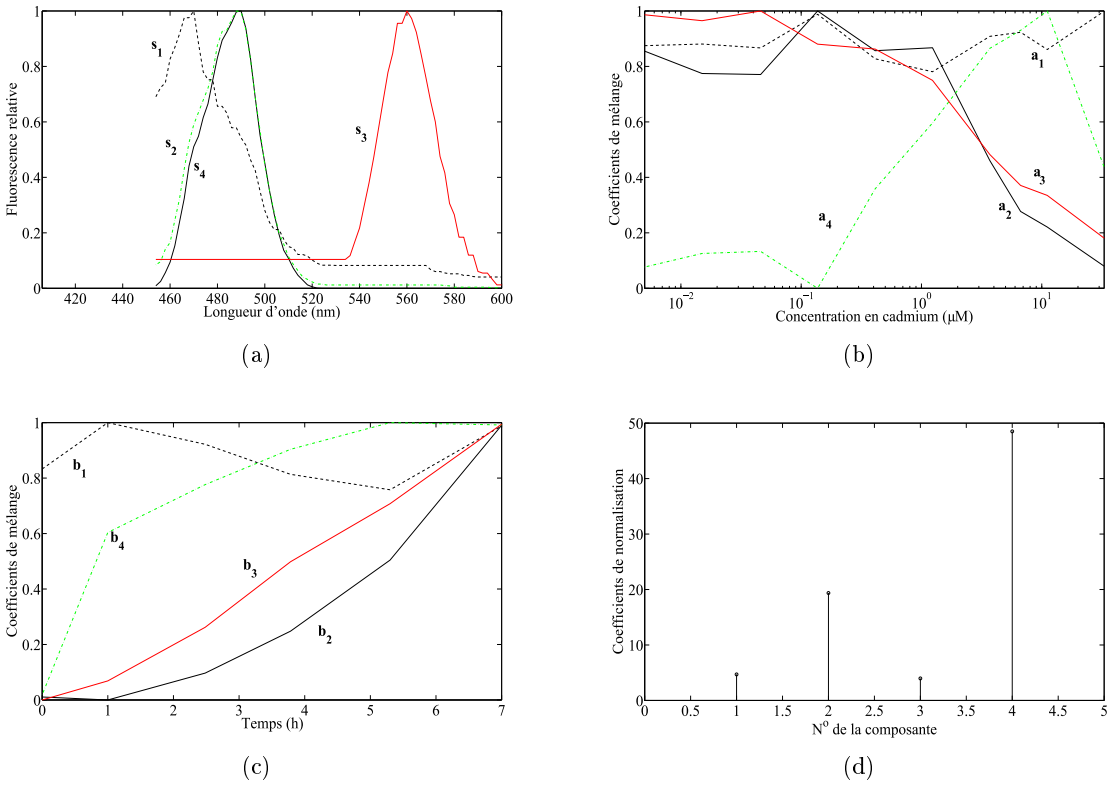


FIGURE *3.21 – Décomposition CP avec contrainte d'unimodalité sur les spectres synchrones mesurés sur les suspensions de biosenseurs bactériens. (a) Spectres de fluorescence de* $\mathbf{s}_1$ *(autofluorescence),* $\mathbf{s}_2$ *(GFP),* $\mathbf{s}_3$ *(DsRed) et* $\mathbf{s}_4$ *(GFP). (b) Réponse en fonction de la concentration en cadmium de* S_1*,* S_2*,* S_3 *et* S_4*. (c) Évolution en fonction du temps. (d) Coefficients de normalisation obtenus par la mise à 1 de la valeur maximale de chaque réponse.*

Le troisième spectre estimé $\mathbf{s}_3$ montre un maximum d'émission à 560 nm. L'intensité de fluorescence de cette source est quasi constante pour les concentrations inférieures à 1 µM et diminue pour les concentrations supérieures. L'intensité de la fluorescence de cette source augmente au cours du temps.

La quatrième source estimée montre une évolution en fonction de la longueur d'onde d'émission similaire au spectre $\mathbf{s}_2$. Cependant, leur évolution en fonction de la concentration en cadmium ($\mathbf{a}_2$ et $\mathbf{a}_4$) et en fonction du temps ($\mathbf{b}_2$ et $\mathbf{b}_4$) sont différentes. Contrairement à la source $\#_2$, la source $\#_4$ est croissante sur la plage de concentration allant de $0,1\,\mu M$ à $10\,\mu M$. La décomposition CP montre également que la fluorescence de la source $\#_4$ augmente plus rapidement que la fluorescence de la source $\#_2$ entre le début de l'expérience et la première heure.

La source $\#_1$ est associée à l'autofluorescence dont nous retrouvons le spectre synchrone de fluorescence tronqué ainsi que les comportements attendus en fonction des paramètres de concentrations en cadmium et de temps. La source $\#_1$ est caractéristique de l'autofluorescence du milieu dont nous supposons que la fluorescence n'évolue pas ou peu en fonction de la concentration

de cadmium ou au cours du temps. Nous pouvons également identifier la source $\#_4$ dont le spectre correspond au spectre de la GFP et dont les réponses estimées $\mathbf{a}_2$ et $\mathbf{b}_2$ correspondent aux réponses attendues du gène promoteur PP$cadA_2$.

L'interprétation des sources $\#_2$ et $\#_3$ est plus subtile. On remarque tout d'abord que les évolutions en fonction des paramètres de concentration en cadmium et du temps de ces deux sources sont très proches. Ce qui explique les difficultés rencontrées pour séparer ces deux sources avec la seule contrainte de positivité. Les spectres estimés $\mathbf{s}_2$ et $\mathbf{s}_3$ sont explicites et ne laissent pas de doute quand à l'association des sources $\#_2$ et $\#_3$ à la fluorescence respective d'une protéine GFP et d'une DsRed.

En étudiant l'évolution de l'absorbance en fonction de la concentration en cadmium (Fig. 3.17), on remarque que les réponses estimées $\mathbf{a}_2$ et $\mathbf{a}_3$ sont similaires, laissant penser à un lien entre ces sources et la quantité de bionsenseur. De plus, on sait que la production de la protéine fluorescente DsRed est liée à l'expression d'un gène promoteur constitutif du biosenseur, donc liée à la quantité de biosenseurs.

Cependant, la cinétique de réponse de ces deux sources ne suit pas l'évolution temporelle de l'absorbance (Fig. 3.17). Pour les concentrations peu toxiques, l'absorbance augmente rapidement dans l'intervalle de temps compris entre 1h et 3h puis, augmente plus lentement. Alors que la réponse $\mathbf{b}_2$ montre une augmentation plus importante au cours du temps et que $\mathbf{b}_3$ croît quasiment de manière constante.

La source $\#_2$ est donc associée aux protéines fluorescentes GFP dans le milieu (extracellulaire). L'augmentation au cours du temps s'explique par une accumulation de la protéine dans le milieu à cause d'une lyse ou d'une excrétion cellulaire. L'expression indépendante de la concentration en cadmium mais dépendante de la quantité de biosenseurs met en évidence que cette source représenterait une expression basale du gène promoteur codant la production de la GFP.

La source $\#_3$ est associée à la protéine DsRed intra- et extra- cellulaire dont les données ne permettent pas de différencier le comportement.

3.3.6 Conclusion

Les expériences réalisées avaient pour but d'étudier la cinétique de réponse d'un gène promoteur réagissant à la présence de cadmium, PP$cadA_2$. Le protocole expérimental avait été construit pour le cas simple d'un biosenseur bi-colore équipé d'un rapporteur fluorescente à maturation rapide (GFP) répondant à la présence de cadmium et d'un rapporteur constitutif à maturation lente traçant la viabilité et la quantité de biosenseur.

Cependant, le comportement du biosenseur s'est avéré plus complexe que prévu rendant l'interprétation de la décomposition CP plus difficile. Néanmoins, l'application de contrainte supplémentaire (unimodalité) et l'utilisation de la décomposition PARALIND a permis de mettre en évidence un relarguage des protéines fluorescente dans le milieu et d'émettre l'hypothèse d'une expression basale du promoteur codant la production de la GFP.

Une centrifugation et resuspension du culot bactérien dans un milieu neuf (ou non autofluorescent) permettrait de mesurer uniquement le signal émis par les cellules bactériennes et valider ces hypothèses.

3.4 Conclusion

Dans ce chapitre, nous avons étudiés les réponses de systèmes rapporteurs sensibles aux métaux. Dans un premier cas, les expériences ont concerné deux gènes impliqués dans l'homéostasie du fer (*pvdA* et *bfrB*). Nous avons montré les problèmes engendrés par le recrouvrement spectral et la présence de sources inconnues, dans l'analyse conjointe des réponses, à différentes concentrations en fer, de ces deux gènes promoteurs. Nous avons alors proposé d'utiliser une méthode de séparation de source exploitant la multi-dimensionnalité des données (décomposition CP). Dans ces expériences, les trois paramètres étaient la longueur d'onde, la concentration en fer et le mélange de biosenseurs, alors que dans les expériences concernant le gène promoteur sensible au cadmium (PP$cadA_2$), les trois paramètres des expériences étaient la longueur d'onde, la concentration en cadmium et le temps. Ce second cas a permis de mettre en évidence les difficultés liées à la diversité temporelle.

Dans ces deux cas étudiés, la décomposition CP a permis d'extraire l'autofluorescence des réponses des gènes promoteurs améliorant ainsi leur estimation. Néanmoins, nous avons vu que le choix du couple *promoteur*::*rapporteur* a une importance dans la qualité de la décomposition CP. Nous avons également vu l'importance du protocole expérimental dans la création de la diversité nécessaire à la séparation des sources. Dans l'expérience sur le cadmium, la suppression de la fluorescence intra-cellulaire des biosenseurs par centrifugation et filtration a eu pour effet de réduire la diversité des données. Cette diminution est importante surtout sur le troisième mode car la diversité du signal repose sur l'induction des promoteurs internes à la cellule. Le lien de linéarité créé entre les sources fluorescentes corrrespondant aux rapporteurs rend impossible la bonne séparation des sources et a pour conséquence la mauvaise estimation des comportements en fonction des concentrations en cadmium. Nous avons également remarqué que l'utilisation de spectres synchrones permettait d'améliorer l'excitation des molécules fluorescentes améliorant le rapport signal à bruit et la qualité des décomposition CP.

Chapitre 4

Discussion générale et perspectives

Sommaire

Ce travail de thèse avait pour objectif d'identifier les réponses fonctionnelles de systèmes bactériens dans des conditions environnementales variées, qu'il s'agisse d'un stress engendré par la présence ou la carence d'un métal. Cette identification est fondée sur l'analyse spectrométrique des signaux de fluorescence émis par des suspensions de biosenseurs bactériens fluorescents, élaborés pour produire des molécules fluorescentes en fonction des métaux disponibles dans l'environnement aqueux des cellules. L'estimation des réponses populationnelles et dose-dépendantes au niveau macroscopique a été réalisée sur différents biosenseurs bactériens ciblant la disponibilité de métaux toxiques ou essentiels, comme le cadmium et le fer. Les senseurs multicolores à cellules entières utilisés ont été développés dans le cadre de l'ANR HAESPRI [6] par insertion chromosomique ou plasmidique de systèmes rapporteurs fluorescents utilisant des promoteurs impliqués dans l'homéostasie cellulaire de ces métaux chez les bactéries :

- promoteur *bfrB* régulant la transcription de gènes codant pour la synthèse d'une bacterioferritine, protéine de stockage du fer intracellulaire en excès,
- promoteur *pvdA* gouvernant la synthèse d'un précurseur de sidérophores, chélateurs organiques puissants du fer produits et excrétés en condition de carence,
- promoteur $cadA_2$ codant une sous-unité protéique d'une pompe d'efflux au cadmium, transporteur protéique membranaire chargé d'expulser les métaux toxiques intracellulaires.

La procédure d'identification des réponses fondamentales exploite la diversité des réponses populationnelles des biosenseurs bactériens dans des conditions expérimentales variées afin d'extraire par des méthodes de séparation de sources, les réponses fonctionnelles des systèmes rapporteurs et par conséquent l'expression de gènes dans des milieux de composition chimique variable. Ces réponses sont estimées et interprétées, avec un minimum d'*a priori*, à partir des signaux spectraux mesurés, grâce à des algorithmes de décomposition multi-linéaire.

Les résultats obtenus lors de l'étude du comportement de ces biosenseurs bactériens fluorescents en présence de métaux, démontrent très clairement l'intérêt des méthodes multi-linéaires et des traitements algorithmiques employés pour :

- l'obtention de modèles multi-paramétriques d'expression des systèmes rapporteurs ou d'activité des promoteurs dans des conditions environnementales variés,
- l'amélioration de la sensibilité et de la robustesse des estimations quelles que soient les constructions génétiques employées, le recouvrement spectral des systèmes rapporteurs employés ou l'autofluorescence des milieux utilisés pour les expériences,
- la possibilité de fournir une information quantitative sur la concentration en métaux par une méthode d'ajouts dosés,
- l'identification des signaux parasites extérieurs ou biologiques inattendus, comme l'autofluorescence du milieu de culture ou la production intra-cytoplasmique ou l'excrétion de substances fluorescentes.

Ils mettent aussi en évidence certaines contraintes et limitations inhérentes à ce type d'approche, en particulier :

6. biosenseurs élaborés par D. Parrello et P. Billard

- le dimensionnement des plans d'expériences et la génération d'une diversité suffisante dans les réponses des biosenseurs,
- la gestion des dépendances entre paramètres expérimentaux et réponses des promoteurs,
- la détermination de l'ordre optimal pour la décomposition (nombre de sources).

4.1 Apport des méthodes multilinéaires à l'étude des réponses fonctionnelles de biosenseurs bactériens fluorescents

Si on dispose de plusieurs expériences utilisant les mêmes biosenseurs mais dans des proportions ou des conditions différentes, il est possible d'estimer leur spectres individuels et leur réponse respective en fonction des paramètres opératoires grâce à des modèles linéaires. Le spectre obtenu au terme de chaque expérience est considéré comme une combinaison linéaire des spectres de chaque biosenseurs ou sources fluorescentes présent dans le système expérimental. Comme nous l'avons vu au chapitre 2, le choix de la méthode de décomposition (itératives ou non itératives, bilinéaires ou multilinéaire) dépend fortement de la nature des données et des contraintes que l'on souhaite imposer dans l'algorithme (positivité, indépendance, orthogonalité, *etc.*).

Dans le cadre de l'étude des signaux émis par des mélanges biosenseurs bactériens fluorescents, nous avons montré dans le chapitre 2, l'intérêt des méthodes tri-linéaires (CP) par rapport aux méthodes bilinéaires (BPSS, SVD, NMF, *etc.*). Nous avons vu que la décomposition bilinéaire nécessite l'application de contraintes afin de réduire le domaine de solutions acceptables. La contrainte d'indépendance statistique imposée par la méthode ICA ne correspond pas à la réalité des réponses attendues des biosenseurs. De plus, il a été montré que la contrainte d'orthogonalité imposée par la décomposition en valeurs singulières (SVD) est inadaptée à des sources spectrales de fluorescence. Il est évident qu'elle ne fournit pas une estimation correcte et interprétable des spectres recherchés, qui sont rarement orthogonaux. Les spectres estimés contiennent des valeurs négatives, ce qui réduit les possibilités de validation des réponses obtenues dans des modèles biologiques. La méthode de factorisation en matrices non-négatives utilise une contrainte de positivité aidant ainsi à conserver la réalité physique des signaux de fluorescence. Cependant, l'unicité de la solution n'est pas garantie. L'approche bayésienne utilisée dans la méthode BPSS n'est pas non plus adaptée. En effet, même si l'utilisation d'une densité gamma comme modèle *a priori* de distribution des sources fluorescentes et des coefficients de mélange permet d'obtenir une solution particulière parmi les solutions admissibles, la représentation des sources spectrales par une distribution gamma n'est pas réaliste. Cette représentation est d'autant plus discutable que les spectres de référence des systèmes rapporteurs utilisés sont *a priori* connus et référencés. A défaut, ils peuvent être déterminés expérimentalement par des mesures spectrométriques sur des systèmes purifiés ou simplifiés comme des suspensions de biosenseurs monofluorescents (unicolores) ou des solutions de protéines purifiées.

En revanche, le traitement des données par des approches tri-linéaires ou quadri-linéaires est plus adaptée. L'unicité de la décomposition peut être garantie sous faibles contraintes liées

aux $rang_k$[7] des matrices sources. De plus, dans le cas où l'unicité de la décomposition n'est pas acquise, on peut facilement imposer des contraintes, physiquement justifiables, de type positivité ou unimodalité permettant de réduire l'espace des solutions admissibles. Si la variabilité des réponses systémique est acquise, l'algorithme de décomposition permet de déterminer les sources élémentaires, identifiant le système biosenseur-milieu ainsi que leur comportement en fonction de paramètres extérieurs.

4.2 Diversité des données spectrales et plans d'expériences

L'utilisation de la décomposition CP nécessite la génération d'un jeu de données multilinéaires. En général, en chimiométrie, le modèle trilinéaire fait intervenir trois modes pour chaque source fluorescente : le spectre d'émission, le spectre d'excitation, et l'abondance ou la concentration des substances d'intérêt. De plus, les matrices d'excitation/émission (MEEF) sont corrigées des interactions entre fluorophores et les échantillons sont très dilués pour satisfaire l'approximation de Beer-Lambert.

Les données obtenues à partir des biosenseurs ne répondent pas exactement au même type d'hypothèses, on dispose rarement pour chaque concentration d'une MEEF. Le plus souvent, on dispose de spectres d'émission, d'excitation ou de fluorescence synchrone. Afin de résoudre le problème d'identification sous faibles contraintes du modèle de fluorescence, il faut ajouter une troisième/quatrième diversité aux jeux de données pour pouvoir utiliser les décompositions multilinéaires.

Nous avons montré que nous pouvions obtenir ce type de données par un plan d'expérience adapté sous condition du respect de la loi de Beer-Lambert, ce qui suppose la conservation d'une concentration faible en protéine fluorescente et un milieu relativement invariant. Cette condition est satisfaite par le principe d'homéostasie à l'intérieur des cellules qui garantit le maintien des conditions physico-chimiques à l'intérieur de chaque cellule bactérienne. De même, en utilisant une densité cellulaire suffisamment faible, on limite les phénomènes non linéaires liés à un écrantage des sources d'excitation ou d'émission par les cellules.

La principale difficulté pour une décomposition CP sur une suspension de biosenseurs est de disposer de données de fluorescence suffisamment diversifiées. La construction du plan d'expérience est donc primordiale. Elle est conditionnée par les possibilités de variations offertes par les paramètres contrôlables (*e.g.* concentration en métaux), fixés (*e.g.* quantités de biosenseurs) ou libres (*e.g.* temps) choisis de manière à générer des signaux fluorescents diversifiés. La création du plan d'expérience doit aussi prendre en compte les conditions d'identifiabilité du modèle CP dans la limite des connaissances *a priori* sur le comportement individuel des biosenseurs et par conséquent des gènes instrumentés.

Par exemple, dans l'expérience concernant le mélange de biosenseurs répondant au fer, les connaissances *a priori* montrent une diminution de l'induction du promoteur *pvdA* pour les

7. Le $rang_k$ d'une matrice équivaut à la plus grande valeur m telle que toutes les sous-matrices de m colonnes soient de rang plein.

concentrations supérieures à 8µM et une augmentation de l'induction du promoteur *bfrB* pour les concentrations supérieurs à 3µM. La plage de concentrations en fer a été élaborée de manière à encadrer ces valeurs.

4.3 Identification des sources

Outre les sources liées aux systèmes rapporteurs implantés dans les bactéries, l'algorithme de décomposition CP permet aussi d'identifier des sources fluorescentes inconnues si leur expression en fonction des paramètres étudiés est différentiable. Nous avons pu aussi tester la capacité des algorithmes à séparer la réponse de rapporteurs fluorescents EYFP/GFP, dont les optima d'émission sont très proches ($\Delta\lambda < 10\,\text{nm}$).

Ces études ont également permis d'identifier certaines limites ou difficultés dans l'utilisation de la décomposition CP pour l'étude des signaux luminescents émis par des biosenseurs bactériens.

Une des limites dans l'utilisation de la décomposition CP réside dans l'estimation du nombre de sources. Aucune méthode ne permet actuellement de définir de manière univoque le nombre de sources appropriées pour la décomposition. Cependant, le calcul de l'énergie résiduelle permet de donner l'ordre de grandeur du nombre de sources et de définir l'ordre des décompositions à analyser. De plus, il n'est pas possible de déterminer de manière iterative les décompositions CP, c'est-à-dire que la décomposition à n sources ne permet pas d'estimer la décomposition à $n+1$ sources. L'approche *ad hoc* retenue consiste à effectuer, des décompositions successives avec un nombre croissant de sources. Le choix final est laissé à l'appréciation de l'expérimentateur et repose sur les critères suivants :

- séparabilité des sources,
- reconnaissance des marqueurs fluorescents utilisés,
- comportements explicables des sources.

L'utilisation de la décomposition CP peut être limitée par l'apparition de colinéarité entre les sources d'un même mode. Bien que le plan d'expérience soit construit de manière à ce que la décomposition soit unique. Les sources et les réponses inattendues peuvent rendre le modèle non identifiable ou partiellement identifiable.

Lors de l'analyse des données obtenues dans les expériences (fer et cadmium), nous avons pu remarquer que le rapport signal à bruit était influencé par de nombreux paramètres, qu'il faut alors prendre en compte lors des expérimentations, comme le couple *promoteur*::*rapporteur* et la position de l'extremum d'émission.

Chaque protéine fluorescente utilisée comme rapporteur a un rendement de fluorescence propre à une longueur d'onde donnée. En conséquence, à quantité de protéines équivalentes produites, le niveau de réponse peut varier d'un rapporteur à l'autre. De plus, les différents promoteurs n'ont pas les mêmes niveaux d'induction ou niveaux basales d'expression, ce qui module les réponses des systèmes rapporteurs qui leur sont associés et par conséquent les quantités de

protéines produites. Il est donc important de ne pas utiliser un couple *promoteur*::*rapporteur* dont le taux d'induction et le rendement de fluorescence sont trop faibles, sous peine de ne pouvoir différencier la source du bruit de mesure.

Il faut ainsi tenir compte de l'optimum d'excitation/émission de chaque rapporteur, surtout dans un mélange de biosenseurs multicolores pour lequel on recherche l'estimation conjointe de chaque réponse. Pour optimiser la contribution de chaque fluorochrome au signal acquis, et limiter l'atténuation progressive des signaux du bleu vers le rouge, il est préférable de travailler avec les spectres synchrones de fluorescence dont les avantages sont les suivants :

- Vitesse d'acquisition plus rapide que celle d'une MEEF,
- Plage spectrale utilisable plus large par une excitation optimale de chaque source de fluorescence,
- Signatures spectrales plus piquées que celle d'un spectre d'émission conventionnel.

Un troisième paramètre pratique, est la quantité de biosenseurs analysées. Cette quantité est dépendante du volume d'excitation, qui peut varier en fonction du support d'expérience (microplaques ou cuves) et de la croissance cellulaire qui dépend du temps d'incubation ou de mesure et de la concentration en métal (effet toxique). En ce qui concerne les expériences menées dans ce travail, l'optimum de réponse se situe dans la plage de 0,1 à 1 unité de densité optique (à 600 nm) pour un volume d'analyte compris entre 200 µL et 1 mL.

4.4 Méthodologie proposée

En nous appuyant sur les simulations et expérimentations réalisées, nous pouvons proposer une méthodologie d'étude des biosenseurs, comprenant trois étapes. Elles sont présentées au niveau de la figure 4.1.

4.4.1 Première étape : Production de données

La première étape de la méthode consiste à définir le plan d'expérience qui permettra de générer un jeu de données trilinéaires respectant le modèle CP. La fluorescence mesurée est une combinaison linéaire de F sources de fluorescence dont l'intensité varie en fonction de trois paramètres modulables.

Le premier paramètre est le domaine de longueur d'onde utilisé pour la mesure spectrale. Lors du réglage du spectrofluorimètre, les choix portent sur la plage de longueur d'ondes, le pas, la longueur d'onde d'excitation dans le cas d'un spectre d'émission ou le $\Delta\lambda$ dans le cas d'un spectre synchrone. Les choix effectués dans la définition de ce paramètre λ conditionneront le temps de mesure et l'intensité des sources de fluorescence, donc le rapport signal à bruit.

Les autres paramètres peuvent être la proportion d'élément stressant (*i.e.* métal), la concentration bactérienne, la proportion de biosenseurs, la température ou la force ionique. Même si le choix de ces paramètres est explicitement guidé par l'objectif de l'étude du biosenseur, il est cependant nécessaire d'en définir la gamme de variation garantissant des réponses diversifiées

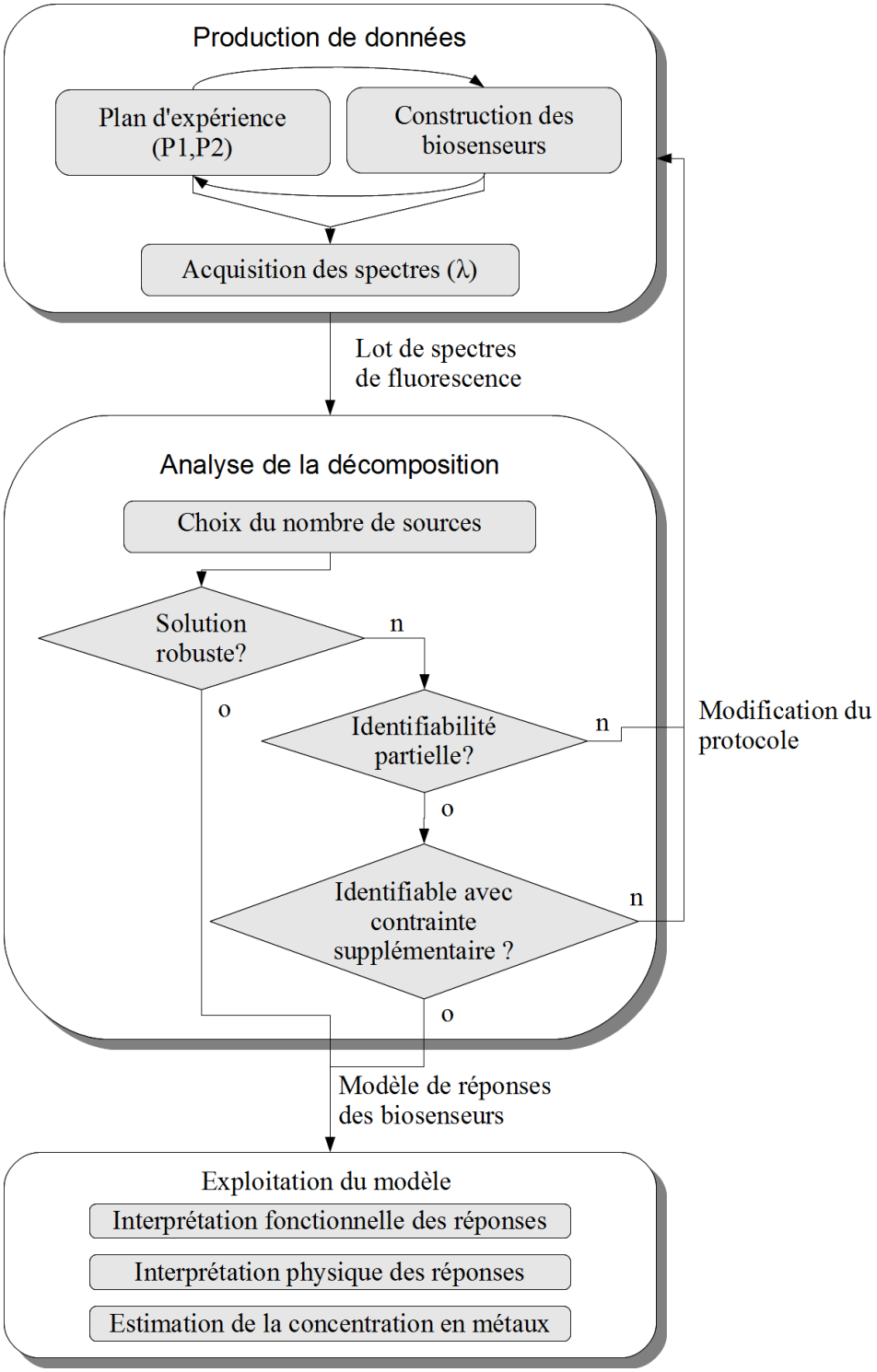


FIGURE *4.1 – Méthode d'analyse des biosenseurs*

et linéairement indépendantes pour les promoteurs étudiés. Comme les quantités de biosenseurs bactériens évoluent dans le temps (croissance, mortalité), le paramètre temporel peut être choisi pour étudier la cinétique des réponses des différents biosenseurs. L'emploi de mélanges controlés

de biosenseurs est aussi une façon d'introduire une diversité dans le plan d'expérience.

Toutefois, le choix du couple *promoteur*::*rapporteur* n'est pas complétement indépendant des paramètres étudiés. Le choix des promoteurs étudiés et des marqueurs associés est intégré au processus afin d'assurer une diversité suffisante et l'identifiabilité du modèle. Il conditionne les réglages optiques du spectrofluorimètre ainsi que la fréquence des mesures, qui dépend de la vitesse de maturation des protéines fluorescentes.

On doit ajouter à cela un aspect pratique : la construction du plan d'expérience. De ce point de vue, les expériences en microplaque sont très faciles à mettre en œuvre. L'avantage des microplaques réside dans la multiplicité des puits et leur ordonnancement matriciel. Il est aisé de produire un jeu de données triparamétriques en mesurant les spectres de fluorescence dans chaque puits et en croisant deux paramètres suivant les colonnes et les lignes de la microplaque. De plus, ces dispositifs sont dimensionnés pour des mesures répétitives et robotisées. Par exemple, l'utilisation d'un robot pipeteur pour remplir les microplaques assure une grande précision et répétabilité dans les volumes délivrés, et permet de produire des combinaisons multiples et variées à façon dans chaque puits. De telles combinaisons sont difficiles à produire par un expérimentateur humain. Cependant, les faibles volumes contenus dans les puits des microplaques posent des problèmes de sensibilité et de détectabilité, surtout lorsque les rendements de fluorescence des rapporteurs sont faibles (protéine fluorescent dans le rouge ou promoteur faiblement exprimé).

4.4.2 Deuxième étape : Analyse des données

La seconde étape de la méthode consiste à analyser les données acquises. Le nombre de sources pour la décomposition CP est choisi par l'étude des variations de l'énergie résiduelle, de l'indice CorConDia et de la décroissance des valeurs singulières. L'estimation du nombre de sources permet de déterminer l'existence de sources supplémentaires inattendues (autres que celles associées aux biosenseurs).

Grâce à la construction du plan d'expérience, nous connaissons le nombre de marqueurs intégrés aux biosenseurs, mais la présence de sources autofluorescentes bactériennes, de la microplaque ou d'éléments fluorescents dans le milieu viennent augmenter le nombre de sources dans le mélange. La valeur seule de l'indice CorConDia ne permet pas de déterminer si le nombre de sources choisi est adéquat, mais une valeur proche de 100 est une condition nécessaire pour valider l'adéquation du modèle CP aux données. Il faut donc étudier la stabilité des solutions proposées pour des décompositions réalisées avec différentes initialisations aléatoires.

Le plan d'expérience est construit de manière à ce que les sources attendues n'aient pas de colinéarités entre elles et que les conditions d'identifiabilité soient respectées. Cependant, des sources inconnues peuvent être identifiées ou une des sources connues peut avoir une réponse inattendue, il est alors nécessaire de valider l'identification et de détecter les colinéarités éventuelles. Dans le cas où l'identifiabilité n'est pas acquise, l'application d'une contrainte supplémentaire sur le mode non identifiable peut résoudre ce problème. Sinon, il est possible de modifier le plan d'expérience afin de générer une diversité supplémentaire respectant les conditions d'identifiabilité

dans le cadre d'un modèle quadrilinéaire.

4.4.3 Troisième étape : Interprétation du modèle obtenu

Le troisième étape est l'interprétation des résultats de la décomposition CP. L'estimation des réponses des biosenseurs améliore la connaissance des interactions métal-bactérie autant du point de vue physique que fonctionnel. De plus, l'utilisation des méthodes de séparation de sources permet de s'affranchir des fluorescences parasites (autofluorescence) autorisant ainsi l'exploitation des données spectrales dans des systèmes naturels. Nous avons également proposé une méthode d'estimation quantitative des métaux bio-disponibles [21]. Dans ce manuscrit, nous avons proposé une méthode permettant l'estimation de la concentration de plusieurs métaux en utilisant des biosenseurs non spécifiques. Pour surmonter le problème d'additivité des réponses des biosenseurs, nous avons proposé d'utiliser des ajouts dosés, couplés à des mélanges de concentrations variables de biosenseurs. Un jeu de données trilinéaires est généré pour chaque polluant. En effectuant la décomposition CP, des données associées à une procédure d'optimisation, il est possible d'estimer les concentrations de métaux, à condition que leur nombre exact dans l'échantillon soit connu.

4.5 Perspectives

Les perspectives des méthodes d'analyse des signaux fluorescents développées dans ce travail laissent entrevoir de nouvelles expériences approfondissant l'analyse des processus géochimiques aux interfaces métal-bactérie-solution ou minéral-bactérie-solution. De plus, les méthodes de spectroscopie employées autorisent des études multi-échelles des réponses de biosenseurs allant de l'étude des comportements de populations bactériennes à l'étude des comportements individuels de chaque cellule. Il est ainsi envisageable d'établir un lien entre la réponse fluorescente globale de populations de biosenseurs et la distribution des réponses cellulaires donnée par l'imagerie hyperspectrale.

Les études menées sur la fluorescence des biosenseurs ont également montré un déficit méthodologique qui nécessite des travaux complémentaires, tant au niveau des algorithmes de décomposition CP, que du point de vue expérimental. Pour la méthode CP, nous avons montré que la prise en compte des interactions pouvait avoir une grande importance dans l'étude des réponses des gènes promoteurs, notamment pour une utilisation en milieu complexe. La nécessité d'un indicateur de l'information/diversité a également été identifié, afin d'établir des plans d'expériences facilitant la décomposition et assurant l'extraction des sources de fluorescence. Du point de vue expérimental, une méthode de génération de tableaux multidimensionnels creux permettrait de réduire le plan d'expérience ou d'augmenter le nombre de paramètres étudiés simultanément. L'utilisation de dictionnaires regroupant les réponses et spectres de fluorescence préétablis réduirait le temps de calcul et améliorerait l'estimation des réponses.

4.5.1 Détection de la présence de métaux par des biosenseurs non spécifiques

Nous pouvons proposer une méthodologie d'étude des biosenseurs autorisant l'utilisation de mélanges de biosenseurs dans des milieux complexes, et ouvrant la perspective du développement de capteur *in situ* de détection de polluants métalliques multiples par des senseurs non spécifiques, comme le montre la simulation suivante.

Expérience multimétaux			
	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$
M_1	$\mathbf{a}_1$	-	$\mathbf{a}_3$
M_2	$\mathbf{b}_1$	$\mathbf{b}_2$	$\mathbf{b}_3$
M_3	$\mathbf{c}_1$	$\mathbf{c}_2$	-

TABLE *4.1 – Représentation des réponses attendus ($\mathbf{a}_i$,$\mathbf{b}_i$,$\mathbf{c}_i$) des biosenseurs non spécifiques $\mathcal{B}_i$ en présence de différents métaux lourds (M_1,M_2,M_3). Le biosenseur $\mathcal{B}_1$ répond aux trois métaux, alors que $\mathcal{B}_1$,$\mathcal{B}_2$ sont insensibles à un d'entre eux.*

En considérant que ces biocapteurs répondent de manière quasi-binaire à la présence d'un métal, on associe à chaque biosenseur une valeur nulle en l'absence de métal ou présence d'un métal auquel il ne réagit pas et une valeur unitaire en présence d'un métal. Pour les différents mélanges de métaux possibles, on obtient le système de codage suivant :

Expérience multimétaux			
Composition du mélange	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$
Aucun des trois métaux	0	0	0
M_1	1	0	1
M_2	1	1	1
M_3	1	1	0
$M_1 + M_2$	2	1	2
$M_1 + M_3$	2	1	1
$M_2 + M_3$	2	2	1
$M_1 + M_2 + M_3$	3	2	2

TABLE *4.2 – Système de codage engendré par les différents mélanges de métaux faisant réagir les biosenseurs $\mathcal{B}_i$.*

Ces combinaisons de biosenseurs non spécifiques permettraient de déterminer la présence d'un ou plusieurs métaux. Par exemple, un biosenseur $\mathcal{B}_1$ non spécifique construit sur la base d'un gène promoteur de type *czc* (gène induit par le cadmium, le zinc et le cobalt) ne permet pas d'assurer l'estimation de la concentration d'un élément dans un milieu complexe (contenant ces trois métaux). En y ajoutant une combinaison de deux autres biosenseurs non spécifiques ($\mathcal{B}_2$ et $\mathcal{B}_3$) bien choisis, c'est-à-dire répondant soit au Cd et Zn, soit au Zn et Co (tableau 4.1), la présence ou l'absence respective de chaque métal est estimable. La figure 4.2 montre les réponses de ces biosenseurs en fonction des métaux présents en solution.

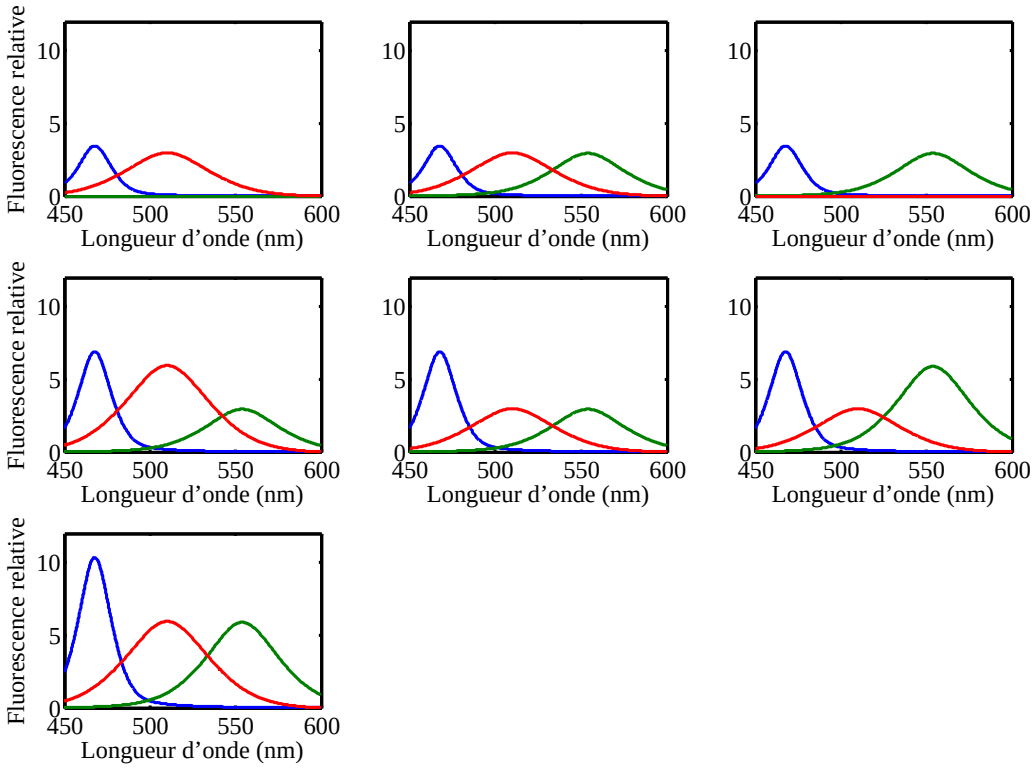


FIGURE *4.2 – Simulation des réponses potentielles de la combinaison de biosenseurs non spécifiques ($\mathcal{B}_1$,$\mathcal{B}_2$,$\mathcal{B}_3$) à des mélanges de métaux.*

4.5.2 Capteurs de polluants *in situ* à base de biosenseurs bactériens

Au-delà de l'approche en laboratoire, on peut envisager l'utilisation des biosenseurs pour la détection *in situ* de polluants. Cette utilisation nécessite encore de nombreux travaux autant dans le domaine de la microbiologie que du point de vue de l'instrumentation. Pour pouvoir utiliser les biosenseurs dans des milieux naturels, il est nécessaire de mieux définir leurs comportements face à des stress environnementaux plus complexes comme la présence de particules minérales et/ou de composés organiques.

Du point de vue des instruments de mesure, l'utilisation de capteurs CCD et de fibres optiques réduirait l'encombrement des spectrofluorimètres. Leur miniaturisation faciliterait la mesure *in situ* voire rendrait possible l'implantation pérenne d'un dispositf complet de mesure, incluant le spectrofluorimètre et un procédé d'immobilisation des cellules bactériennes. Des supports chimiquement inertes, comme des gels de silice, constituerait un point d'ancrage pour les biosenseurs bactériens dont la viabilité serait assuré par la composition du support.

Publications

AN UNIQUENESS CONDITION FOR THE 4-WAY CANDECOMP/PARAFAC MODEL WITH COLLINEAR LOADINGS IN THREE MODES

David Brie, Sebastian Miron

CRAN, Nancy-Université, CNRS
Boulevard des Aiguilletes BP 70239
54506 Vandœuvre-lès-Nancy, France

Fabrice Caland, Christian Mustin

LIMOS, Nancy-Université, CNRS
Boulevard des Aiguilletes BP 70239
54506 Vandœuvre-lès-Nancy, France

ABSTRACT

In this paper we investigate the uniqueness of the 4-way CANDECOMP/PARAFAC (CP) model in the case where the only possible linear dependencies between the columns of the loading matrices take the form of collinear loadings. For this special configuration we state a necessary and sufficient condition for having full column rank of the Khatri-Rao product of two loading matrices. This allows to derive a sufficient condition for uniqueness of the 4-way CP model with collinear loadings in at most three modes. The result is illustrated by analyzing 4-way fluorescence data.

Index Terms— Multilinear algebra, CANDECOMP/PARAFAC, 4-way array, uniqueness, collinear loadings.

1. INTRODUCTION

When dealing with multidimensional signals organized as tensors of order N, a crucial question is to decompose them into a limited number of components from which the main characteristics of the data can be recovered. For order 2 tensors, *i.e.* matrices, this leads to a matrix factorization problem which is known to be ill-posed since an infinite number of possible decompositions yields the same data. In that case, additional constraints such as orthogonality, independence, non-negativity or sparseness have to be imposed to ensure an unique factorization of the data matrix. For $N > 2$, a number of multidimensional extensions of matrix factorizations have been proposed among which we may cite CANDECOMP/PARAFAC (CP) decompositions [1]. A key point in the development of this multidimensional decomposition comes from the fact that adding dimensions, also referred to as diversities, results in multidimensional decompositions admitting an unique solution under mild conditions. This explains the growing interest of these multidimensional decompositions for a wide range of applications including psychometrics, chemometrics and more recently signal and image processing.

Considering the 3-way CP model, the most general uniqueness result is due to Kruskal [2] which provides a sufficient condition. This has been extended to the N-way CP case by Sidiropoulos and Bro in [3]. Ten Berge and Sidiropoulos [4] showed that for the 3-way CP decomposition, Kruskal's sufficient condition is also necessary for tensors of rank 2 and 3. Liu and Sidiropoulos [5] derived general necessary conditions for uniqueness of N-way CP decompositions.

In this paper, we address the CP uniqueness problem when the only possible linear dependencies on the columns of the loading matrices take the form of collinear loadings. This has many practical applications in signal processing [6, 7], chemometrics [8], *etc*.

THIS WORK HAS BEEN SUPPORTED BY THE FRENCH ANR PROGRAM THROUGH GRANT ANR-09-BLAN-0336-04

When a 3-way CP model has two or more collinear factors, uniqueness is no longer achieved [9]. In that case, partial uniqueness, as introduced by Ten Berge [10], can be obtained. It is worth noting that this partial uniqueness is similar to the block Parafac decomposition in $(L, L, 1)$ terms uniqueness introduced by De Lathauwer [11, 12].

Interestingly, the case of 4-way, and more generally N-way CP decomposition with collinear factors is less problematic, since uniqueness can still be achieved. However, to the best of our knowledge, this point has not been studied explicitly, motivating the present work. The paper is organized as follows : in section 2, we introduce the 4-way CP model with *Collinear Loading Only (CLO)*. Sections 3 and 4 present the main results of the paper. More precisely, in section 3 we state the necessary and sufficient condition under which the Khatri-Rao product of two CLO matrices is full column rank. Then in section 4, we give a sufficient condition for having the uniqueness of a 4-way CP model with collinear loadings in three modes. Finally, section 5 gives an illustrative example consisting in analyzing 4-way fluorescence data that mimic the response of bacterial bio-sensors to environmental agents.

2. PROBLEM STATEMENT

2.1. Models and Notations

Consider an $I \times J \times K \times L$ 4-way array $\underline{\mathbf{X}}$ with typical element $x_{i,j,k,l}$ and the quadrilinear CP decomposition of order F

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

for all $i = 1, \cdots, I$, $j = 1, \cdots, J$, $k = 1, \cdots, K$ and $l = 1, \cdots, L$. The equation (1) expresses the 4-way array as the sum of F rank-1 4-way arrays. Similarly to the matrix case, the rank of $\underline{\mathbf{X}}$ is defined as the minimum number of rank-1 4-way arrays needed to decompose $\underline{\mathbf{X}}$.

Defining the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ as :

$$\begin{aligned} \mathbf{A} &= [\mathbf{a}_1 \cdots \mathbf{a}_F] && (I \times F) \\ \mathbf{B} &= [\mathbf{b}_1 \cdots \mathbf{b}_F] && (J \times F) \\ \mathbf{C} &= [\mathbf{c}_1 \cdots \mathbf{c}_F] && (K \times F) \\ \mathbf{D} &= [\mathbf{d}_1 \cdots \mathbf{d}_F] && (L \times F) \end{aligned}$$

where $\mathbf{a}_f, \mathbf{b}_f, \mathbf{c}_f$ and $\mathbf{d}_f$ are column vectors of dimension $(I \times 1), (J \times 1), (K \times 1)$ and $(L \times 1)$ respectively, and denoting by

$\circ$ the outer product, the model (1) can be expressed as :

$$\underline{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{d}_f. \quad (2)$$

For shortening the notations, we will also write model 2 as :

$$\underline{\mathbf{X}} = \mathbf{A}|\mathbf{B}|\mathbf{C}|\mathbf{D}. \quad (3)$$

The 4-way array can be transformed in a matrix $\mathbf{X}$ of dimension $(IJK \times L)$ by the unfolding operation. Depending on the ordering of the unfolding and of the columns, the matrix form of $\underline{\mathbf{X}}$ can yield different matrices $\mathbf{X}$ (see [13] for details). For notational simplicity, in the sequel, the order of the columns in the unfolding operations is lexicographic. Thus, model (1) or (2) yields :

$$\mathbf{X} = (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})\,\mathbf{D}^T \quad (4)$$

where $\odot$ stands for the Khatri-Rao product which is a column-wise Kronecker product denoted by $\otimes$, that is :

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1 \cdots \mathbf{a}_F \otimes \mathbf{b}_F] \quad (5)$$

It can be noticed that $\mathbf{A} \odot \mathbf{B}$ contains the columns $1, F+2, 2F+3, \cdots, (F-1)F+F$ of the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ which is of dimension $(IJ \times F^2)$. An important notion used in the paper is the Kruskal-rank (k-rank) of a matrix $\mathbf{A}$ defined as the largest number $k_{\mathbf{A}}$ such that every subset of $k_{\mathbf{A}}$ columns of the matrix is linearly independent.

2.2. The Collinear Loadings Only (CLO) assumption

We assume that none of the loading matrices has a null column. Let us now introduce the class of 4-way CP model where the possible dependencies causing rank deficiency of the loading matrices can only take the form of collinear loadings. In this context, a loading matrix $\mathbf{A}$ is rank deficient if and only if one (or more columns) of the loading matrix is (or are) proportional to another column. In that case :

$$\exists \quad n \neq m \text{ such as } \mathbf{a}_n = \lambda \mathbf{a}_m \quad (6)$$

which results in $k_{\mathbf{A}} = 1$. As a consequence, under the "Collinear Loadings Only" (CLO) assumption, the k-rank of the loading matrix can only be equal to either F or 1 while its rank may vary between 1 and F.

3. RANK OF THE KHATRI-RAO PRODUCT WITH COLLINEAR LOADINGS ONLY

In this section, a necessary and sufficient condition ensuring the full column rank of the Khatri-Rao product of two CLO matrices is provided. This rank is upper-bounded by $r_{\mathbf{A}\odot\mathbf{B}} \leq r_{\mathbf{A}} \cdot r_{\mathbf{B}}$. Thus, from now on we assume that $F \leq r_{\mathbf{A}} \cdot r_{\mathbf{B}}$. First we prove a necessary and sufficient condition under which two vector Kronecker products are collinear.

Proposition 1 *Let $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}$ be non zero vectors. $\mathbf{x} \otimes \mathbf{u} \neq \lambda \mathbf{y} \otimes \mathbf{v}$ iff $\mathbf{x} \neq \alpha \mathbf{y}$ or $\mathbf{u} \neq \beta \mathbf{v}$.*

This proposition is a direct consequence of rank property of the Kronecker product of two matrices $\mathbf{X}$ and $\mathbf{Y}$: $r_{\mathbf{X}\otimes\mathbf{U}} = r_{\mathbf{X}} \cdot r_{\mathbf{U}}$. Consider $\mathbf{X} = [\mathbf{x} \quad \mathbf{y}]$ and $\mathbf{U} = [\mathbf{u} \quad \mathbf{v}]$, then their Kronecker product is $\mathbf{X} \otimes \mathbf{Y} = [\mathbf{x} \otimes \mathbf{u} \quad \mathbf{x} \otimes \mathbf{v} \quad \mathbf{y} \otimes \mathbf{u} \quad \mathbf{y} \otimes \mathbf{v}]$.

- $r_{\mathbf{X}\otimes\mathbf{U}} = 4 \iff r_{\mathbf{X}} = r_{\mathbf{U}} = 2 \iff \mathbf{x} \neq \alpha \mathbf{y},\ \mathbf{u} \neq \beta \mathbf{v}$. In particular, $\mathbf{x} \otimes \mathbf{u} \neq \lambda \mathbf{y} \otimes \mathbf{v}$

- $r_{\mathbf{X}\otimes\mathbf{U}} = 2 \iff (r_{\mathbf{X}} = 2,\ r_{\mathbf{U}} = 1)$ or $(r_{\mathbf{X}} = 1,\ r_{\mathbf{U}} = 2) \iff (\mathbf{x} \neq \alpha \mathbf{y},\ \mathbf{u} = \beta \mathbf{v})$ or $(\mathbf{x} = \alpha \mathbf{y},\ \mathbf{u} \neq \beta \mathbf{v})$. Then, for the first case : $\mathbf{X} \otimes \mathbf{U} = [\mathbf{x} \otimes \beta \mathbf{v} \quad \mathbf{x} \otimes \mathbf{v} \quad \mathbf{y} \otimes \beta \mathbf{v} \quad \mathbf{y} \otimes \mathbf{v}]$. The vectors $\mathbf{x} \otimes \mathbf{v}$ and $\mathbf{y} \otimes \mathbf{v}$ being respectively collinear to $\mathbf{x} \otimes \beta \mathbf{v}$ and $\mathbf{y} \otimes \beta \mathbf{v}$ and, as $r_{\mathbf{X}\otimes\mathbf{U}} = 2$, we have $\mathbf{x} \otimes \mathbf{u} \neq \lambda(\mathbf{y} \otimes \beta \mathbf{u}) = \lambda \mathbf{y} \otimes \mathbf{v}$. The second case yields the same conclusion □

Now, we give the lemma stating the condition under which the Khatri-Rao product of two matrices satisfying the CLO assumption is full column rank.

Lemma 1 *Consider the two matrices $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_F]$ and $\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_F]$ of size $(I \times F)$ and $(J \times F)$ satisfying the CLO assumption. $\mathbf{A} \odot \mathbf{B}$ is full column rank iff $\forall n \neq m,\ \mathbf{a}_n \neq \alpha \mathbf{a}_m$ or $\mathbf{b}_n \neq \beta \mathbf{b}_m$.*

The proof is done by induction. Suppose that $\mathbf{A}_k$ and $\mathbf{B}_k$ are two matrices of dimension $(I \times k)$ and $(J \times k)$ such as

$$\begin{cases} \mathbf{A}_k \odot \mathbf{B}_k \text{ is full column rank } k, \\ \forall n \neq m,\ \mathbf{a}_n \neq \alpha \mathbf{a}_m, \text{ or } \mathbf{b}_n \neq \beta \mathbf{b}_m \end{cases} \quad (7)$$

Let $\mathbf{A}_{k+1} = [\mathbf{A}_k \quad \mathbf{a}_{k+1}]$ and $\mathbf{B}_{k+1} = [\mathbf{B}_k \quad \mathbf{b}_{k+1}]$ such as $r_{\mathbf{A}_{k+1}\odot\mathbf{B}_{k+1}} = k$ (rank deficient). Combined with the CLO assumption, this implies that $\exists\ n$ such as :

$$\mathbf{a}_{k+1} \otimes \mathbf{b}_{k+1} = \gamma \mathbf{a}_n \otimes \mathbf{b}_n,\ \gamma \neq 0, \quad (8)$$

which, due to proposition 1, is equivalent to $\mathbf{a}_{k+1} = \alpha \mathbf{a}_n$ *and* $\mathbf{b}_{k+1} = \beta \mathbf{b}_n$. In other words we have proven that if $\mathbf{A}_k$ and $\mathbf{B}_k$ are two matrices of dimension $(I \times k)$ and $(J \times k)$ such as $\mathbf{A}_k \odot \mathbf{B}_k$ is full column rank k and satisfies condition (7) then $\mathbf{A}_{k+1} \odot \mathbf{B}_{k+1}$ is full column rank iff $\forall n,\ \mathbf{a}_{k+1} \neq \alpha \mathbf{a}_n$, or $\mathbf{b}_{k+1} \neq \beta \mathbf{b}_n$, from which it turns out that condition (7) is also satisfied by $\mathbf{A}_{k+1}$ and $\mathbf{B}_{k+1}$. The property being true for $k = 2$, it is also true for all $k \leq F$. □

4. UNIQUENESS OF THE 4-WAY CP MODEL WITH COLLINEAR LOADINGS IN THREE MODES

We are now ready to present the main result of this paper which provides a sufficient condition for the uniqueness of the 4-way CP model with collinear loadings in at most three modes. First, we have to introduce a decomposition of the 4-way CP model based on a partition of the loading matrices according the collinear columns arising in one mode.

4.1. Model decomposition

Let us re-write the 4-way Parafac model by separating the set of columns of, say $\mathbf{A}$, in N sub-matrices $\mathbf{A}_i$ containing the collinear columns of $\mathbf{A}$ and a sub-matrix $\breve{\mathbf{A}}$ containing the rest of the non-proportional columns. The matrix $\mathbf{A}$ can be partitioned (after re-ordering) as :

$$\mathbf{A} = [\mathbf{A}_1 \cdots \mathbf{A}_N \ \breve{\mathbf{A}}]$$

with $\mathbf{A_i} = [\mathbf{a}_i \cdots \mathbf{a}_i]$ of dimensions $(K \times F_i)$ and $\breve{\mathbf{A}} = [\mathbf{a}_{\tilde{F}+1} \cdots \mathbf{a}_F]$ of dimensions $(K \times \breve{F})$. We denote $\tilde{F} = \sum_i F_i$ and $\breve{F} = F - \tilde{F}$. The matrices $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ can then be written using the corresponding partition of $\mathbf{A}$ as: $\mathbf{B} = [\mathbf{B}_1 \cdots \mathbf{B}_N \ \breve{\mathbf{B}}]$, $\mathbf{C} = [\mathbf{C}_1 \cdots \mathbf{C}_N \ \breve{\mathbf{C}}]$ and $\mathbf{D} = [\mathbf{D}_1 \cdots \mathbf{D}_N \ \breve{\mathbf{D}}]$. With these notations, the 4-way Parafac model can be written as:

$$\mathbf{A}|\mathbf{B}|\mathbf{C}|\mathbf{D} = \sum_{i=1}^{N} \mathbf{a}_i \circ \mathbf{B}_i|\mathbf{C}_i|\mathbf{D}_i + \breve{\mathbf{A}}|\breve{\mathbf{B}}|\breve{\mathbf{C}}|\breve{\mathbf{D}} \quad (9)$$

4.2. Uniqueness result

The uniqueness result is summarized by the following theorem:

Theorem 1 *Consider a 4-way array* $\underline{\mathbf{X}}$ *given by (9) and admitting an order-*F *CP decomposition. The CP decomposition is* essentially unique *if but not necessarily*

- *one loading matrix (say* $\mathbf{D}$*) is full column rank;*
- *the loading matrices* $\mathbf{A}$*,* $\mathbf{B}$*, and* $\mathbf{C}$ *satisfy the CLO assumption*
- $\forall n \neq m,\ (\mathbf{a}_n \neq \alpha\mathbf{a}_m, \mathbf{b}_n \neq \beta\mathbf{b}_m)$ or $(\mathbf{a}_n \neq \alpha\mathbf{a}_m, \mathbf{c}_n \neq \gamma\mathbf{c}_m)$ or $(\mathbf{b}_n \neq \beta\mathbf{b}_m, \mathbf{c}_n \neq \gamma\mathbf{c}_m)$

Basically, the proof consists in the following two steps :

- proving the partial uniqueness of the corresponding 3-way model obtained by unfolding along one dimension. By doing so, we ensure that the original 4-way problem is transformed into a number of independent 3-way Parafac problems;
- proving the uniqueness of each subproblem.

Partial uniqueness By unfolding the 4-way array, the corresponding 3-way array can be written as $\mathbf{A}|(\mathbf{B} \odot \mathbf{C})|\mathbf{D}$. For the partial uniqueness, according to the partial uniqueness theorem of [8, 10, 12] it is sufficient to have: $r_{\mathbf{D}} = r_{\mathbf{B}\odot\mathbf{C}} = F$ which, thanks to lemma 1 is equivalent to $r_{\mathbf{D}} = F$ and $\forall n \neq m,\ \mathbf{b}_n \neq \beta\mathbf{b}_m$ or $\mathbf{c}_n \neq \gamma\mathbf{c}_m$. In this case, the partial uniqueness theorem guarantees that $\mathbf{A}$ is unique and that $\forall i$, span$(\mathbf{B}_i \odot \mathbf{C}_i)$ and span$(\mathbf{D}_i)$ are unique. In other words, the partial uniqueness ensures the independence of the different subproblems appearing in (9).

Uniqueness of the subproblems To prove the uniqueness of 4-way CP model, we have to show that :

i) $\breve{\mathbf{A}}|\breve{\mathbf{B}}|\breve{\mathbf{C}}|\breve{\mathbf{D}}$ is unique;

ii) $\forall i,\ \mathbf{B}_i|\mathbf{C}_i|\mathbf{D}_i$ is unique;

i) Consider the 4-way array unfolded as $\breve{\mathbf{A}}|(\breve{\mathbf{B}} \odot \breve{\mathbf{C}})|\breve{\mathbf{D}}$, we have: $r_{\breve{\mathbf{A}}} \geq 2$ (no collinear columns in $\breve{\mathbf{A}}$). As $\forall n \neq m,\ \mathbf{b}_n \neq \beta\mathbf{b}_m$ or $\mathbf{c}_n \neq \gamma\mathbf{c}_m$, by lemma 1, $\breve{\mathbf{B}} \odot \breve{\mathbf{C}}$ is full column rank. Similarly, as $\mathbf{D}$ is full column rank, then $\breve{\mathbf{D}}$ is also full column rank. Thus, Kruskal's condition holds for the three way CP model $\breve{\mathbf{A}}|(\breve{\mathbf{B}} \odot \breve{\mathbf{C}})|\breve{\mathbf{D}}$ which ensures the uniqueness of $\breve{\mathbf{A}}|\breve{\mathbf{B}}|\breve{\mathbf{C}}|\breve{\mathbf{D}}$.
ii) As $\mathbf{D}$ is full column rank, $\forall i, \mathbf{D}_i$ is also full column rank. To ensure that each trilinear subproblem is identifiable, it is necessary that both $\mathbf{B}_i$ and $\mathbf{C}_i$ have a k-rank at least equal to 2 (which implies full column rank under the CLO assumption). Thus, on the one hand, $\forall n \neq m$, if $\mathbf{a}_n = \alpha\mathbf{a}_m$, then $\mathbf{b}_n \neq \beta\mathbf{b}_m, \mathbf{c}_n \neq \gamma\mathbf{c}_m$. On the other hand, if $\mathbf{a}_n \neq \alpha\mathbf{a}_m$, only the condition for the full column rank of the Khatri-Rao product is needed, that is $\mathbf{b}_n \neq \beta\mathbf{b}_m$ or $\mathbf{c}_n \neq \gamma\mathbf{c}_m$. Summarizing all this we get : $\forall n \neq m,\ (\mathbf{a}_n \neq \alpha\mathbf{a}_m, \mathbf{b}_n \neq \beta\mathbf{b}_m)$ or $(\mathbf{a}_n \neq \alpha\mathbf{a}_m, \mathbf{c}_n \neq \gamma\mathbf{c}_m)$ or $(\mathbf{b}_n \neq \beta\mathbf{b}_m, \mathbf{c}_n \neq \gamma\mathbf{c}_m)$, which completes the proof. □

5. A PRACTICAL EXAMPLE

This work is part of a project aiming at studying the response of bacterial bio-sensors to different environmental agents [14]. Basically, a bio-sensor is a bacteria genetically modified to produce fluorescent proteins when exposed to some chemical products. When using CP decomposition to analyze the response of different bacterial bio-sensors, it is likely to have bacteria responding similarly to a subset of the studied stimuli. To mimic such a situation, three fluorescent dyes have been selected - Oregon Green 514 (OG514), Rhodamine 6G (R6G), Rhodamine B(RB) - because of their emission spectra and their different responses to temperature and pH. The three fluorescent dyes have different emission spectra. Basically the OG514 responds to the pH while R6G and RB do not. Also RB is responding strongly to temperature while OG514 and R6G only respond moderately and similarly. Thirty-six mixtures have been created in the wells of a microplate (6 differents values of the pH $\times$ 6 different concentration of R6G). Then the fluorescence spectra have been recorded at 6 different temperatures. The data have been organized into a 4-way array (see figure 1). Each subplot shows the spectra obtained in a given well for the 6 different temperatures. The concentration / pH variations correspond to the different vertical / horizontal subplots. By using the available information on the response of the dyes and on the physics of the experiment, the following 4-way CP model of order 3 can be proposed for the data:

$$\underline{\mathbf{X}} = \sum_{f=1}^{3} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{d}_f \tag{10}$$

where :

- $\mathbf{A}$ gathers the emission fluorescence spectra. Thus we can assume that $\mathbf{A}$ is full column rank and $r_{\mathbf{A}} = k_{\mathbf{A}} = 3$;
- each one of the three other matrices is representing the response to the other diversities (concentration, pH, temperature) and is expected to have two collinear loading, that is

$$\begin{aligned}
\mathbf{B} &= [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \mathbf{b}_1], & r_{\mathbf{B}} = 2, k_{\mathbf{B}} = 1;\\
\mathbf{C} &= [\mathbf{c}_1 \quad \mathbf{c}_1 \quad \mathbf{c}_2], & r_{\mathbf{C}} = 2, k_{\mathbf{C}} = 1;\\
\mathbf{D} &= [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \mathbf{d}_2], & r_{\mathbf{D}} = 2, k_{\mathbf{D}} = 1.
\end{aligned}$$

Following theorem 1, the uniqueness of model (10) is guaranteed. This has been confirmed by numerical experiments. The figure 2 shows the results obtained by performing the CP decomposition of data organized as a 4-way array. For each mode, the loadings are normalized to have a maximal value equal to one. The decomposition was performed using the ALS algorithm with non negativity constraints on all modes. We also performed the decomposition using the unconstrained ALS algorithm with different initializations. All our trials yielded similar results.

The obtained results are in good accordance with what was expected. In particular, the mode 1 shows the spectra of the different fluorescent dyes. The mode 2 clearly shows the variation of R6G concentration while the two other remain constant. Similarly for the mode 3, only the response of the OG514 is varying w.r.t. the pH. Finaly, for mode 4, the RB is responding much more strongly than the two other dyes.

6. CONCLUSION

We provide in this paper a sufficient uniqueness condition for a 4-way CP model having collinear loadings in at most 3 modes. The simplifying Collinear Loadings Only (CLO) assumption used in the paper, corresponds to wide range of applications and allows at the same time the derivation of easy to check uniqueness conditions. However, if more complex linear dependencies are assumed, a more difficult combinatorial problem is obtained, that seems hardly tractable. This theoretical result was illustrated on the decomposition of a 4-way fluorescence data aiming at reproducing the expected response of bacterial bio-sensors. Future work will be directed at

studying data corresponding to real bio-sensor responding to environmental solicitations.

7. REFERENCES

[1] R. A. Harshman, "Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis," *UCLA Working Papers Phonetics*, vol. 16, pp. 1–84, Dec. 1970.

[2] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions with application to arithmetic complexity and statistics," *Linear Algebra Applicat.*, vol. 18, no. 2, pp. 95–138, 1977.

[3] N.D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *J. Chemometrics*, vol. 14, no. 3, pp. 229–239, 2000.

[4] J.M.F. ten Berge and N.D. Sidiropoulos, "On uniqueness in CANDECOMP/PARAFAC," *Psychometrica.*, vol. 67, pp. 399–409, 2002.

[5] X. Liu and N.D. Sidiropoulos, "Cramèr-Rao lower bounds for low-rank decomposition of multidimensional arrays," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 2074–2086, 2001.

[6] N.D. Sidiropoulos and G.Z. Dimić, "Blind multiuser detection in w-cdma systems with large delay spread," *IEEE Signal Processing Lett.*, vol. 8, pp. 87?89, 2001.

[7] A.L.F. de Almeida, G. Favier, and J.C.M. Mota, "Constrained tensor modeling approach to blind multiple-antenna CDMA schemes," *IEEE Trans. Signal Processing*, vol. 56, pp. 2417?2428, 2008.

[8] R. Bro, R.A. Harshman, N.D. Sidiropoulos, and M.E. Lundi, "Modeling multi way data with linearly dependent loadings," *J. Chemometrics*, vol. 23, pp. 324–340, 2009.

[9] A. Stegeman and N.D. Sidiropoulos, "On Kruskal's uniqueness condition for the CANDECOMP/PARAFAC decomposition," *Linear Algebra Applicat.*, vol. 420, pp. 540–552, 2007.

[10] J.M.F. ten Berge, "Partial uniqueness in CANDECOMP/PARAFAC," *J. Chemometrics*, vol. 18, pp. 12–16, 2004.

[11] L. De Lathauwer, "Decomposition of a higher order tensor in block terms – part 1: Lemmas for partioned matrices," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 1022–1032, 2008.

[12] L. De Lathauwer, "Decomposition of a higher order tensor in block terms – part 2: Definitions and uniqueness," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 1033–1066, 2008.

[13] T. G. Kolda and B.W. Bader, "Tensor decomposition and applications," *SIAM Rev.*, vol. 8, pp. 455–500, 2009.

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

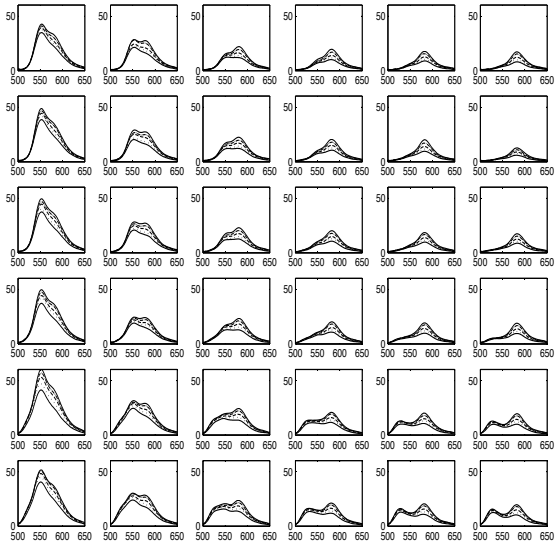


Fig. 1. 4-way fluorescence data

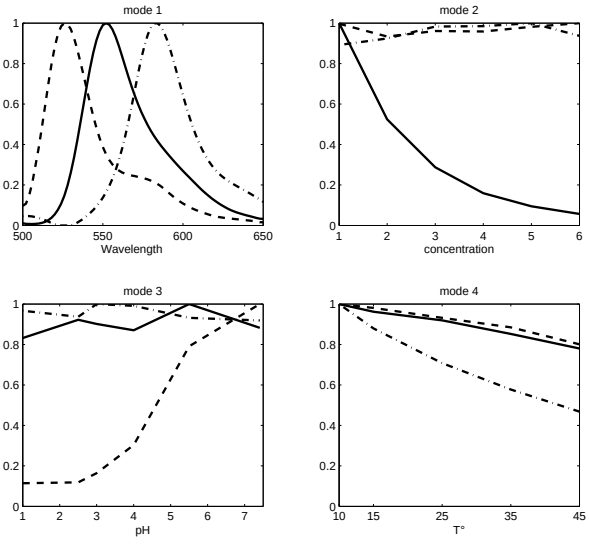


Fig. 2. CP decomposition of the 4-way fluorescence data

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

F. Caland, C. Mustin

LIMOS, Nancy-Université, CNRS
Boulevard des Aiguilettes BP 70239
54506 Vandoeuvre-lès-Nancy, France

S. Miron, D. Brie

CRAN, 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 receptor, 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

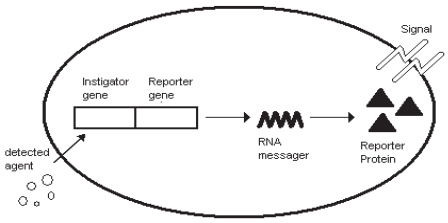


Fig. 1. Schematic diagram of bacterial biosensor

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 pollutants present in a sample. The reminder of the paper is orga-

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.

nized 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 briefly addressed. Performing the CP decomposition yields the three factors representing respectively the fluorescence spectra of each bacterium, the concentration of bacteria and the response curve 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, for a fixed number of polluants, the more types of biosensors are used, the better will be the estimation. In the experiment presented in this article we will use a minimum of types of biosensors F=2.

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. The biosensor concentration is early measurable and controlable, providing a useful diversity. 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. CANDECOMP/PARAFAC 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. In the context of the present application, source separation techniques based on bilinear decompositions, such as ICA [12], cannot succesfully apply. This is because the conditions for decomposition uniqueness are not satisfied by biosensor data. Thus it is the CP uniqueness properties which motivated the design of an experiment yielding three-way data arrays.

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

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

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

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:

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

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 [13]. 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 [14] 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.

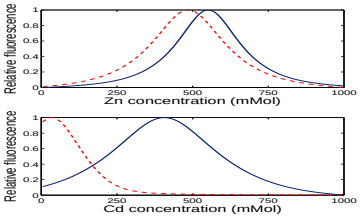


Fig. 2. Calibration curves

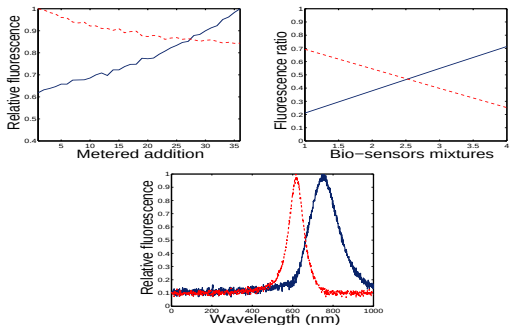


Fig. 3. Results of CP decomposition of the dataset $\mathcal{D}_1$

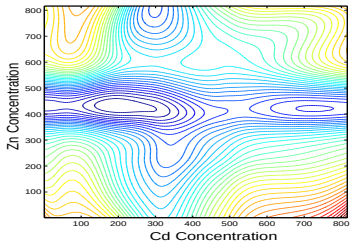


Fig. 4. Criterion $\mathcal{J}(x_0, y_0)$

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.

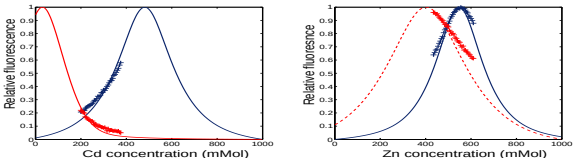


Fig. 5. Result of the concentration estimation

[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.

[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] J. Eriksson and Koivunen V., "Identifiability and separability of linear ica models revisited," 2003, in Fourth International Symposium on Independent Component Analysis and Blind Signal Separation.

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

[14] 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.

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).

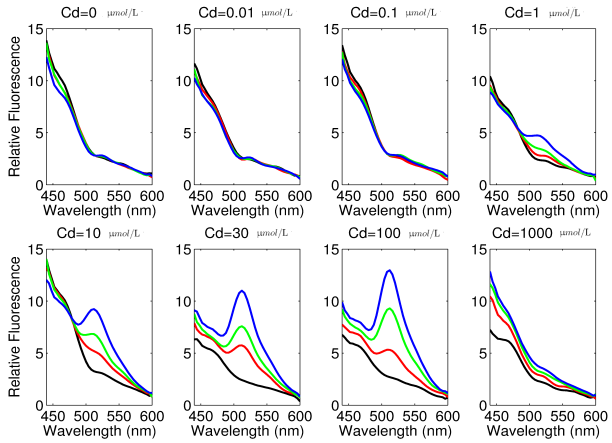


FIGURE 1: Données d'émission de fluorescence. Chaque courbe de chacun des graphiques représente le spectre de fluorescence émis dans un puit à un instant t.

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.

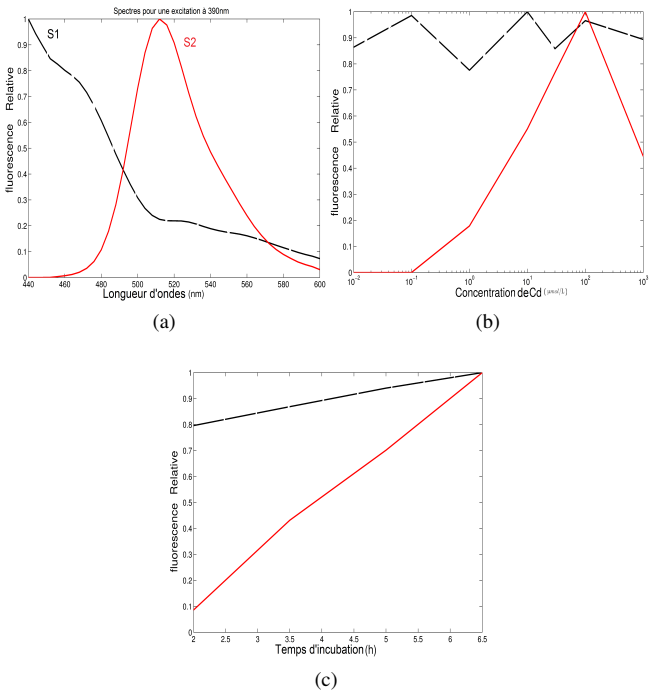


FIGURE 2: Résultats de la décomposition CP des données. Source S_1 (autofluorescence)-Source S_2 (GFP). (a) Spectres de fluorescence de S1 et S2. (b) Réponse en fonction du [Cd] de S_1 et S_2. (c) Évolution en fonction du temps.

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-

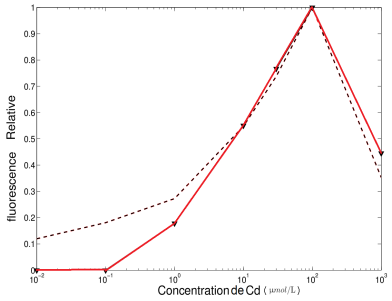


FIGURE 3: Courbe de réponse du biosenseur fluorescent. Mesure à une seule longueur d'onde (en pointillés). Estimation par décomposition trilinéaire (en trait plein).

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.

A BLIND SPARSE APPROACH FOR ESTIMATING CONSTRAINT MATRICES IN PARALIND DATA MODELS

F. Caland[1,2]*, S. Miron*[2]

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

D. Brie[2]*, C. Mustin*[1]

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

ABSTRACT

In this paper we address the problem of estimating the interaction matrices of PARALIND decomposition. In general, this is an ill-posed problem admitting an infinite number of solutions. First we study the gain of imposing sparsity constraints on the interaction matrices, in terms of model identifiability. Then, we propose a new algorithm (S-PARALIND) for fitting the PARALIND model, using a $\ell_2-\ell_1$ optimization step for estimating the interaction matrix. This new approach provides more accurate and robust estimates of the constraint matrices than ALS-PARALIND, thus improving the interpretability of the PARALIND decomposition.

Index Terms— PARAFAC, linear contraints, PARALIND/ CONFAC, sparse, ALS-PARALIND, S-PARALIND

1. INTRODUCTION

PARAFAC-based methods [1, 2] are presently standard tools for factor or component modeling in various domains such as psychometry, spectroscopy, signal processing or telecommunications systems. A general overview of PARAFAC applications can be found in [3, 4]. The PARAFAC decomposition of a $\boldsymbol{\mathcal{X}}$ $(I \times J \times K)$ 3-way array (or tensor) into R rank-1 terms is given by

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r, \tag{1}$$

where $\mathbf{a}_r(I \times 1)$, $\mathbf{b}_r(J \times 1)$ and $\mathbf{c}_r(K \times 1)$ are vectors and "$\circ$" denotes the outer vector product. For simplicity, the noise/error term in (1) is ignored at this point of the presentation. The three dimensions of $\boldsymbol{\mathcal{X}}$ are referred to as "modes". An alternative notation for (1) is

$$\boldsymbol{\mathcal{X}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!], \tag{2}$$

where $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_R]$, $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_R]$ and $\mathbf{C} = [\mathbf{c}_1 \dots \mathbf{c}_R]$ denote the component matrices. The mode-1 matrix unfolding of the PARAFAC model (1) is given by

$$\mathbf{X}_1 = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T, \tag{3}$$

with $\odot$ the Khatri-Rao product of two matrices. The mode-2 and the mode-3 unfolding matrices, $\mathbf{X}_2$ and $\mathbf{X}_3$, can be obtained by switching $\mathbf{A}, \mathbf{B}$ $\mathbf{C}$ in (3).

In some applications, prior knowledge on the existence of linear dependencies between the columns of the component matrices is available. This information can be explicitly taken into account by introducing some constraint (or interaction) matrices $\boldsymbol{\Psi}(R_1 \times R)$, $\boldsymbol{\Phi}(R_2 \times R)$, $\boldsymbol{\Omega}(R_3 \times R)$, containing the linear dependency patterns between the columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, respectively. Thus, instead of $[\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ the decomposition is given by

$$\boldsymbol{\mathcal{X}} = [\![\tilde{\mathbf{A}}\boldsymbol{\Psi}, \tilde{\mathbf{B}}\boldsymbol{\Phi}, \tilde{\mathbf{C}}\boldsymbol{\Omega}]\!], \tag{4}$$

with $\tilde{\mathbf{A}}(I \times R_1)$, $\tilde{\mathbf{B}}(J \times R_2)$ and $\tilde{\mathbf{C}}(K \times R_3)$ full column rank matrices. This type of decomposition was introduced in [5] and previous versions, and named PARALIND[1]. A slightly different version, CONFAC[2], with the constraint matrices having canonical vectors as columns, was proposed in [6, 7]. In order to illustrate PARALIND model we consider a simple example, similar to those given in [5]. Suppose that the columns of the first component matrix $\mathbf{A}$ are $[\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3 \quad \mathbf{a}_2 \quad \mathbf{a}_3 \quad \mathbf{a}_1 + \mathbf{a}_2]$. Then, PARALIND expresses $\mathbf{A}$ as the matrix product of $\tilde{\mathbf{A}} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3]$ and the interaction matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}. \tag{5}$$

In general, the algorithms for fitting the PARALIND model assumes that the constraint matrices are *a priori* known. However, this is not always the case in practice. Moreover, in some real life applications it may be of practical interest to estimate these constraint matrices, as they provide important information on the interactions between the physical mechanisms generating the data. A blind alternating least

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

[1] PARAllel profiles with LINear Dependencies
[2] CONstrained FACtor decomposition

squares (ALS) estimator for the PARALIND model, referred to as ALS-PARALIND, was proposed in [5]. However, for identifiability reasons (as explained in the next section), the interaction matrices estimated by this approach are highly dependent on the algorithm initialization, which limits their practical utility.

In this paper we propose a blind approach for estimating the interaction matrices of the PARALIND model that imposes sparsity of the elements of these matrices. This new approach, called S-PARALIND, when compared to the ALS-PARALIND algorithm, yields more robust estimates of the constraints matrices. Moreover, it also improves the interpretability of the decomposition since the linear dependencies are expressed using a small number of interacting components. The remainder of this paper is organized as follows: in section 2 identifiability issues of PARAFAC and PARALIND are addressed; section 3 introduces the S-PARALIND algorithm and some results on synthetic data are given in section 4. Finally, conclusions are drawn in section 5.

2. IDENTIFIABILITY OF PARAFAC AND PARALIND MODELS

A model is said *identifiable* if all its parameters can be *uniquely* estimated from the data, up to some trivial indeterminacies. Thus, in this paper, identifiability can be understood as a uniqueness problem. For example, the PARAFAC model given by (2) is identifiable if the matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ can be uniquely estimated from $\boldsymbol{\mathcal{X}}$ up to simultaneous column permutation and column-wise rescaling. An attractive feature of this decomposition is its identifiability under mild conditions. The most well-known PARAFAC identifiability condition is due to Kruskal [8] and states that the decomposition in (3) is unique if

$$k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \geq 2R + 2, \tag{6}$$

where $k_{(.)}$ denotes the Kruskal-rank[3] of a matrix.

Following [5], identifiability of the PARALIND model is essentially the same as that of the PARAFAC model. Meanwhile, if the interaction matrices are fixed and known, identifiability conditions specific to PARALIND can be found in [9]. If these interaction matrices are not known, the identifiability problem can be much more complicated. In particular, it may happen that only some components of the three matrices or only one matrix (among the three) are identifiable, resulting in the so-called *partial uniqueness* or *uni-mode uniqueness* results. The interested reader is referred to [10] for details.

Let us now assume that the uniqueness of matrix $\mathbf{A}$ is fulfilled and that we aim at estimating the constraint matrix $\boldsymbol{\Psi}$ together with the full column rank matrix $\tilde{\mathbf{A}}$. The identifiability of $\boldsymbol{\Psi}$ and $\tilde{\mathbf{A}}$ comes down to the uniqueness of the bilinear decomposition $\mathbf{A} = \tilde{\mathbf{A}}\boldsymbol{\Psi}$. Without any further constraints, such a decomposition is not unique since an alternative decomposition can be obtained as

$$\mathbf{A} = \tilde{\mathbf{A}}\boldsymbol{\Psi} = (\tilde{\mathbf{A}}\mathbf{T}^{-1})(\mathbf{T}\boldsymbol{\Psi}) = \tilde{\mathbf{A}}'\boldsymbol{\Psi}',$$

for any non-singular matrix $\mathbf{T}$. In this paper we propose to impose sparsity on the constraint matrix $\boldsymbol{\Psi}$ which should have a minimum number of non-zero entries. This comes down to explaining the rank deficiency of matrix $\mathbf{A}$ by considering the the simplest dependency pattern between its columns. This problem is somewhat connected with the problem of dictionary identification using sparse matrix factorization, which has been intensely studied lately in different papers such as [12]. However, the main difference with these approaches is the fact that the problem addressed in this paper considers only full column rank dictionaries. Therefore, the proposed approach is somewhat closer to sparse singular value decomposition methods [13]. As we do not dispose of any definitive result on the uniqueness of the decomposition using this kind of sparsity constraint, we consider next some examples to illustrate the purpose. Let $\mathbf{A}$ be given by

$$\begin{aligned}
\mathbf{A} &= [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3 \quad \mathbf{a}_1 + \mathbf{a}_2] && (7)\\
&= [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3] \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} && (8)\\
&= [\mathbf{a}_1 + \mathbf{a}_2 \quad \mathbf{a}_1 + \mathbf{a}_3 \quad \mathbf{a}_2 + \mathbf{a}_3] \\
&\quad \begin{bmatrix} 1/2 & 1/2 & -1/2 & 1 \\ 1/2 & -1/2 & 1/2 & 0 \\ -1/2 & 1/2 & 1/2 & 0 \end{bmatrix} && (9)
\end{aligned}$$

As illustrated by (8) and (9), it appears that the sparsest matrix $\boldsymbol{\Psi}$ is obtained by selecting R_1 independent columns of $\mathbf{A}$ to form $\tilde{\mathbf{A}}$. It is worth noting that imposing sparsity of $\boldsymbol{\Psi}$ does not ensure the uniqueness of the bilinear decomposition. For example, another possible decomposition of $\mathbf{A}$ is

$$\mathbf{A} = [\mathbf{a}_1 \quad \mathbf{a}_3 \quad \mathbf{a}_1 + \mathbf{a}_2] \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \tag{10}$$

One can see that $\boldsymbol{\Psi}$ matrix in (10) has the same sparsity degree as the one in (8). From an interpretation point of view, there is no reason to favor either decomposition (10) or (8) since the number of non-zero entries in the constraint matrix is the same in both cases. However, if some additional physical arguments are available, then the number of possible solutions can be reduced. For example it is possible to impose both sparsity and positivity of the entries of the $\boldsymbol{\Psi}$ matrix to ensure uniqueness of the bilinear decomposition. The reader may consult [11] for a discussion on this topic.

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

3. A SPARSE PARALIND ALGORITHM (S-PARALIND)

In this section, we present an algorithm (S-PARALIND) for estimating the PARALIND model with sparse constraints on the interaction matrix. This algorithm is currently implemented for estimating linear dependencies in one mode only, *i.e.* mode-1, but the extension to two or three modes simultaneously is straightforward.

In the presence of noise or model errors, evaluating $\mathbf{\Psi}$ from a PARAFAC estimate of $\mathbf{A}$ (as presented in the previous section) is not judicious because of the error accumulation effect. Therefore, in [5] the constraint matrix is estimated directly within the main ALS loop for updating the PARAFAC matrix components. The approach presented in this section is based on the algorithm proposed in [5] with the difference that the LS estimation step of $\mathbf{\Psi}$ is replaced by the following $\ell_2 - \ell_1$ optimization problem [14]

$$\min_{\mathbf{\Psi}} \left[\|\mathbf{X}_1 - \tilde{\mathbf{A}}\mathbf{\Psi}(\mathbf{C} \odot \mathbf{B})^T\|_2^2 + \lambda \|\mathbf{\Psi}\|_1 \right], \qquad (11)$$

where the hyperparameter λ controls the sparsity degree of the constraint matrix. The minimization in (11) can be formulated as a LASSO problem for which a number of efficient algorithms have been developed lately (see [15] and references therein). Table 1 illustrates the main steps of the proposed approach, where vec(.), $\otimes$ and " $*$ " denotes the matrix vectorization operator, the Kronecker product and the Hadamard (element wise) product, respectively. For simultaneous linear dependencies in all the three modes, each of the steps 4 and 5 in table 1 should be replaced by two other steps (analogous to steps 2 and 3).

	Input : $\boldsymbol{\mathcal{X}}$, λ, R, R_1
1:	Initialize $\tilde{\mathbf{A}}, \mathbf{B}, \mathbf{C}$
2:	$\text{vec}\mathbf{\Psi} = \arg\min_{\mathbf{\Psi}} \{ \|\text{vec}\mathbf{X}_1 - [(\mathbf{C} \odot \mathbf{B}) \otimes \tilde{\mathbf{A}}]\, \text{vec}\mathbf{\Psi}\|_2^2 + \lambda \|\text{vec}\mathbf{\Psi}\|_1 \}$
3:	$\tilde{\mathbf{A}} = \mathbf{X}_1(\mathbf{C} \odot \mathbf{B})\mathbf{\Psi}^T \left\{ \mathbf{\Psi}[(\mathbf{B}^T\mathbf{B}) * (\mathbf{C}^T\mathbf{C})]\mathbf{\Psi}^T \right\}^{-1}$
4:	$\mathbf{B} = \mathbf{X}_2[\mathbf{C} \odot (\tilde{\mathbf{A}}\mathbf{\Psi})] \left[(\mathbf{\Psi}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\mathbf{\Psi}) * (\mathbf{C}^T\mathbf{C}) \right]^{-1}$
5:	$\mathbf{C} = \mathbf{X}_3[\mathbf{B} \odot (\tilde{\mathbf{A}}\mathbf{\Psi})] \left[(\mathbf{\Psi}^T\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}\mathbf{\Psi}) * (\mathbf{B}^T\mathbf{B}) \right]^{-1}$
6:	If stop condition not satisfied, go to Step 2
	Output : Estimated $\tilde{\mathbf{A}}, \mathbf{\Psi}, \mathbf{B}, \mathbf{C}$

Table 1. S-PARALIND algorithm

Steps $3 - 5$ in Table 1 are similar to ALS-PARALIND algorithm presented in [5]. In the next section we compare S-PARALIND with ALS-PARALIND on synthetic examples.

4. RESULTS

In this section, we aim at illustrating the benefit of including a sparsity constraint in the estimation of the $\mathbf{\Psi}$ matrix. To that end, we simulated a positive data array $\boldsymbol{\mathcal{X}}$ of size $700 \times 30 \times 30$ that mimics spectroscopy data. In order to improve the uniqueness properties of the PARALIND decomposition as well as result interpretability, we also imposed non-negativity constraints on the estimated modes. In S-PARALIND, to solve the $\ell_2 - \ell_1$ optimization problem of step 2 in table 1, the LASSO implementation [4] of [14] was used. The results provided by S-PARALIND are compared with those of ALS-PARALIND. In all the experiments the number of sources was set to $R = 4$ and the number of columns of matrix $\tilde{\mathbf{A}}$ to $R_1 = 3$. The same initializations, non-negativity constraints and number of iterations were used for both algorithms. The columns of the simulated matrices $\tilde{\mathbf{A}}$, $\mathbf{B}$ and $\mathbf{C}$ are depicted in Fig. 1.

In a first example the constraint matrix is given by

$$\mathbf{\Psi} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \qquad (12)$$

In this case, the k-rank of $\mathbf{A}$ equals 1, implying the non-uniqueness of the PARAFAC decomposition. However, the uni-mode uniqueness conditions in [10] state that the first mode matrix is still identifiable. Meanwhile $\mathbf{B}$ and $\mathbf{C}$ are subject to rotational ambiguities which, due to the positive offset (background), are not resolved by the non-negativity constraints. Figure 2 and 3 show the three modes estimates for

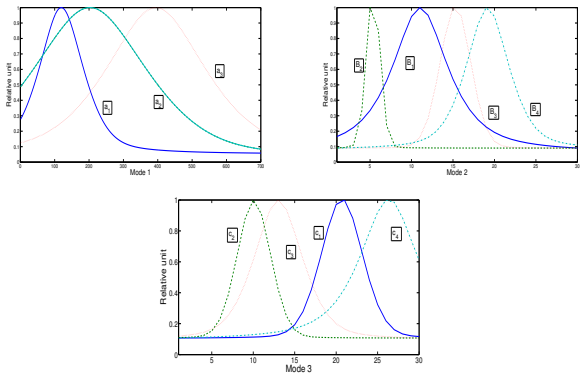


Fig. 1. Simulated data

ALS-PARALIND and S-PARALIND, respectively. One can observe that the S-PARALIND estimates are more accurate than ALS-PARALIND for the first mode. However, for both algorithms, the estimated components of $\mathbf{B}$ and $\mathbf{C}$ present artifacts that are typical for rotational indeterminacies. The effect of the sparse constraints can be clearly seen on the estimated constraint matrices

[4] Available at `http://www.di.ens.fr/ mschmidt/Software/lasso.html`

$$\hat{\mathbf{\Psi}}_{S-\text{PARALIND}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.008 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad (13)$$

$$\hat{\mathbf{\Psi}}_{ALS-\text{PARALIND}} = \begin{bmatrix} 1 & 0.591 & 0.155 & 0.59 \\ 0.003 & 1 & 0.103 & 1 \\ 0.004 & 0.157 & 1 & 0.156 \end{bmatrix}. \quad (14)$$

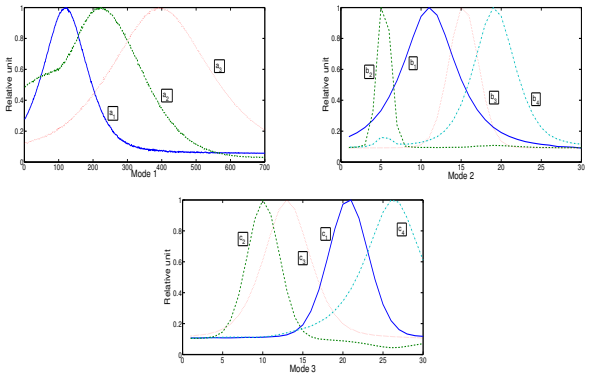


Fig. 2. Results of ALS-PARALIND for the first example

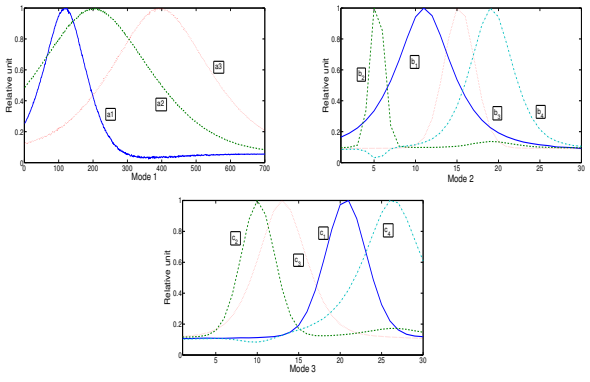


Fig. 3. Results of S-PARALIND for the first example

In a second example matrix $\mathbf{\Psi}$ is fixed to

$$\mathbf{\Psi} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \quad (15)$$

In this case, $k_{\mathbf{A}} = 2$ and the PARAFAC uniqueness is achieved, resulting in a unique estimation of all the three matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. As one can see in fig. 4 and fig. 5, there are no more ambiguities in the estimation of modes 2 and 3. However, in this case also, imposing sparse constraints on $\mathbf{\Psi}$ yields better estimation of the first mode component matrix and constraint matrix

$$\hat{\mathbf{\Psi}}_{S-\text{PARALIND}} = \begin{bmatrix} 1 & 0 & 0 & 0.749 \\ 0.171 & 1 & 0.024 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad (16)$$

$$\hat{\mathbf{\Psi}}_{ALS-\text{PARALIND}} = \begin{bmatrix} 1 & 0.468 & 0.006 & 1 \\ 0.014 & 1 & 0.007 & 0.569 \\ 0.071 & 0.401 & 1 & 0.276 \end{bmatrix}. \quad (17)$$

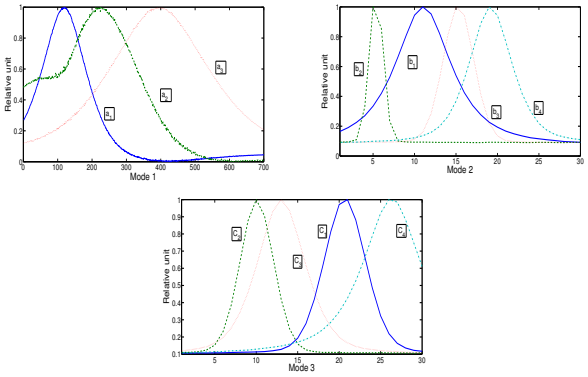


Fig. 4. Results of ALS-PARALIND for the second example

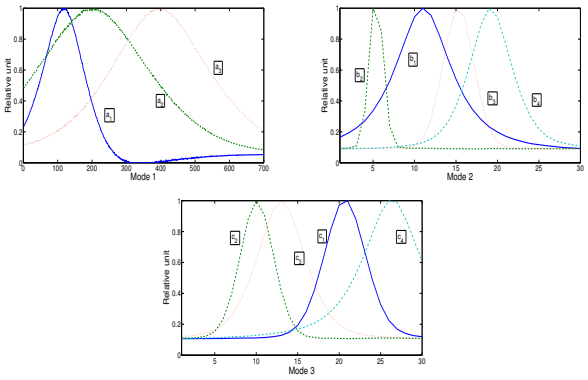


Fig. 5. Results of S-PARALIND for the second example

Next we present the results of the S-PARALIND study of the response of a bacterial biosensor to a varying concentration of IPTG (isopropyl β-D-1-thiogalactopyranoside). The employed biosensor is a bacteria genetically modified to produce two fluorescent proteins when exposed to IPTG. A same gene (lacZ) of the bacteria is instrumented by two other genes encoding the synthesis of different fluorescent proteins. As the production of these fluorescent proteins is controlled by the same gene, they have the same response to IPTG concentration variation, resulting in collinearities in the associated PARAFAC model (mode 1). The second diversity, necessary to source separation, is provided by the spectral mode, each

source having a different fluorescence spectrum. The third diversity in the data is created by considering the time evolution of the fluorescence, as each source has a different maturation time. In other words, the time between the beginning of the synthesis of the protein and the beginning of the fluorescence light emission is different for each protein. The analyzed data contains spectral measurements (between 450 and 600 nm with a step of 3 nm), performed every 45 minutes after an initial time lapse of 1 hour, for 6 different IPTG concentrations. Figure 6 shows the decomposition results of the data using the S-PARALIND algorithm. The red and the green sources correspond to the two fluorescent proteins, while the blue source corresponds to the autofluorescence of the bacteria which is theoretically independent of time and ITPG concentration. The first plot corresponds to the estimated sources as a function of the concentration of IPTG. The other two modes correspond to the estimated biosensor spectra and temporal evolution of fluorescence, respectively. Because the collinearity in the first mode, the red and the green sources are theoretically subject to rotational indeterminacies in modes 2 and 3 (see [10]). This explains the partial overlapping of the source spectra on the second plot, despite the non-negativity constraints imposed in the estimation process. Further constraints, such as unmiodality, can be imposed to tackle this indeterminacy problem. Nevertheless, the S-PARALIND decomposition results comply with theory.

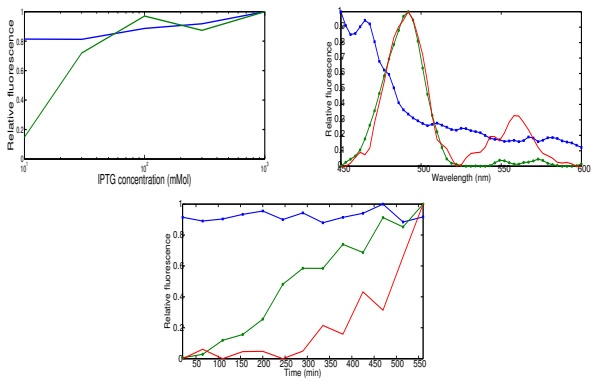


Fig. 6. Results of S-PARALIND decomposition of the dataset.

5. CONCLUSIONS

In this paper, an approach has been proposed for estimating the constraint matrices in PARALIND models. First, the identifiability of the PARALIND model has been investigated and we have provided some evidences showing the interest of imposing sparsity of the constraint matrix. The proposed S-PARALIND approach has been compared to the ALS-PARALIND in two different cases: partially and fully identifiable models. In both cases, the effectiveness of S-PARALIND algorithm has been confirmed. Finally, the proposed algorithm was applied on a real dataset resulted from bacterial biosensors interaction with IPTG.

6. REFERENCES

[1] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sept. 1970.

[2] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, Dec. 1970.

[3] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Sept. 2009.

[4] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE Trans. Knowledge Data Eng.*, vol. 21, no. 1, pp. 6–20, Jan. 2009.

[5] R. Bro, R. A. Harshman, N. D. Sidiropoulos, and M. E. Lundy, "Modeling multi-way data with linearly dependent loadings," *J. Chemometr.*, vol. 23, no. 7-8, pp. 324–340, July-Aug. 2009, Special Issue: In Honor of Professor Richard A. Harshman.

[6] A. L. F. de Almeida, G. Favier, and J. C. M. Mota, "Constrained tensor modeling approach to blind multiple-antenna CDMA schemes," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2417–2428, June 2008.

[7] A. L. F. de Almeida, G. Favier, and J. C. M. Mota, "A constrained factor decomposition with application to MIMO antenna systems," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2429–2442, June 2008.

[8] J. B. Kruskal, "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra Applicat.*, vol. 18, no. 2, pp. 95–138, 1977.

[9] A. Stegeman and A. L. F. de Almeida, "Uniqueness conditions for constrained three-way factor decompositions with linearly dependent loadings," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1469–1499, 2009.

[10] X. Guo, S. Miron, D. Brie, and A. Stegeman, "Uni-mode and partial uniqueness conditions for CANDECOMP/PARAFAC of three-way arrays with linearly dependent loadings," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 1, pp. 111–129, 2012.

[11] S. Miron, M. Dossot, C. Carteret, S. Margueron, and D. Brie, "Joint processing of the parallel and crossed polarized Raman spectra and uniqueness in blind nonnegative source separation," *Chemometr. Intell. Lab.*, vol. 105, no. 1, pp. 7–18, 2011.

[12] R. Gribonval and K. Schnass, "Dictionary identification - sparse matrix-factorisation via ℓ_1-minimisation," *IEEE Trans. Inform. Theory*, vol. 56, no. 7, pp. 3523–3539, July 2010.

[13] M. Li, H. Shen, J.Z. Huang, and J.S. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–1095, Dec. 2010.

[14] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[15] M. Zibulevsky and M. Elad, "L1-L2 optimization in signal and image processing," *IEEE Signal Proc. Mag.*, vol. 27, no. 3, pp. 76–88, May 2010.

Bibliographie

[1] E. Acar and B. Yener. Unsupervised multiway data analysis : A literature survey. *IEEE Trans. Knowledge Data Eng.*, 21(1) :6–20, 2009.

[2] S.K. Alpat, S. Alpar, B. Kutlu, O. Ozbayrak, and H.B. Buyukisik. Development of biosorption-based algal biosensor for cu(ii) using tetraselmis chuii. *Sensors and Actuators B : Chemical*, 128 :273–278, 2008.

[3] A. Amine, H. Mohammadi, I. Bourais, and G. Palleschi. Enzyme inhibition-based biosensors for food safety and environmental monitoring. *Biosensors and Bioelectronics*, 21 :1405–1423, 2006.

[4] C.A. Andersson and R. Bro. The n-way toolbox for matlab. *Chemometrics and Intelligent Laboratory Systems*, 52 :1–4, 2000.

[5] P. Anttila, P. Paatero, U. Tapper, and O. Jarvinen. Source identification of bulk wet deposition in finland by positive matrix factorization. *Atmospheric Environment*, 29(14) :1705 – 1718, 1995.

[6] E. Bigorgne, L. Foucaud, C. Caillet, L. Giamberini, J. Nahmani, F. Thomas, and Rodius F. Cellular and molecular responses of e. fetida cœlomocytes exposed to tio2 nanoparticles. *Journal of Nanoparticle Research*, 14 :959–975, 2012.

[7] P. Billard, C. Mustin, T. Beguiristain, and C. Leyval. Approche innovante pour l'étude de la biodisponibilité des éléments métalliques dans les sols. Technical report, BQR UHP-Région, 2008.

[8] I. Biran, R. Babai, K. Levcov, J. Rishpon, and E.Z. Ron. Online and in situ monitoring of environmental pollutants : electrochemical biosensing of cadmium. *Environmental Microbiology*, 2 :285–290, 2000.

[9] C. Blaise. Microbiotests in aquatic ecotoxicology : Characteristics, utility, and prospects. *Environmental Toxicology and Water Quality*, 6(2) :145–155, 1991.

[10] C. Blaise. Microbiotesting : An expanding field in aquatic toxicology. *Ecotoxicology and Environmental Safety*, 40 :115 – 119, 1998.

[11] D.A. Blake, R.M. Jones, R.C. Blake, A.R. Pavlov, I.A. Darwish, and H. Yu. Antibody-based sensors for heavy metal ions. *Biosensors and Bioelectronics*, 16 :799–809, 2001.

[12] I. Bontidean, J. Ahlqvist, A. Mulchandani, W. Chen, W. Bae, R.K. Mehra, A. Mortari, and E. Csoregi. Novel synthetic phytochelatin based capacitive biosensor for heavy metal detection. *Biosensors and Bioelectronics*, 18 :547–553, 2003.

[13] I. Brian, R. Babai, K. Levcov, J. Rishpon, and E.Z. Ron. Online and in situ monitoring of environmental pollutants : Electrochemical biosensing of cadmium. *Environmental Microbiology*, 2 :285–290, 2000.

[14] R. Bro. Parafac tutorial and applications. *Chemometrics Intell. Lab. Syst.*, 38(2) :149–171, 1997.

[15] R. Bro. *Multi-way Analysis in the Food Industry - Models, Algorithms, and Applications*. PhD thesis, Univ. of Amsterdam, Amsterdam, The Netherlands, 1998.

[16] R. Bro and S. De Jong. A fast non-negativity constrained least squares algorithm. *Journal of Chemometrics*, 11 :393 – 401, 1997.

[17] R. Bro, R. A. Harshman, N. D. Sidiropoulos, and M. E. Lundy. Modeling multi-way data with linearly dependent loadings. *Journal of Chemometrics*, 23(7-8) :324–340, 2009. Special Issue : In Honor of Professor Richard A. Harshman.

[18] R. Bro and H.A. Kiers. A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics*, 17 :274 – 286, 2003.

[19] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1) :34 – 81, 2009.

[20] M.R. Bruins, S. Kapil, and F.W. Oehme. Microbial resistance to metals in the environment. *Ecotoxicology and Environmental Safety*, 45 :198–207, 2000.

[21] F. Caland, S. Miron, D. Brie, and C. Mustin. A candecomp/parafac approach to the estimation of environmental pollutant concentrations using biosensors. *IEEE Statistical Signal Processing Workshop*, 2011.

[22] F. Caland, S. Miron, D. Brie, and C. Mustin. A blind sparse approach for estimating constraint matrices in paralind data models. *20th European Signal Processing Conference, EUSIPCO*, 2012.

[23] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3) :283–319, 1970.

[24] G. Castillo and L. Schafer. Evaluation of a bioassay battery for water toxicity testing : A chilean experience. *Environmental Toxicology*, 15 :331 – 337, 2000.

[25] J.C. Chen. Nonnegative rank factorisation of nonnegative matrices. *Linear Algebra and its Applications*, 62 :207–217, 1984.

[26] P. Comon and C. Jutten. *Séparation de sources*, volume 1 et 2. Hermès - Lavoisier, Paris, 2007.

[27] P. Comon and J. ten Berge. Generic and typical ranks of three-way arrays. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[28] P. Corbisier, D. Lelie, B. Borremans, A. Provoost, V. Lorenzo, N.L. Brown, J.R. Lloyd, J.L. Hobman, E. Csoregi, G. Johansson, and Mattiasson B. Whole cell-and protein-based biosensors for the detection of bioavailable heavy metals in environmental samples. *Analytica chimica acta*, 387 :235–244, 1999.

[29] L. De Lathauwer and A. De Baynast. Blind deconvolution of ds-cdma signals by means of decomposition in rank-(1, l,l) terms. *IEEE Transactions on Signal Processing*, 56 :1562–1571, 2008.

[30] S. Deneux-Mustin, S. Roussel-Debet, C. Mustin, P. Henner, C. Munier-Lamy, C. Colle, J. Berthelin, J.and Garnier-Laplace, and C. Leyval. *Mobilité et transfert racinaire des éléments en traces : influence des micro-organismes du sol*. TEC & DOC, 2003.

[31] O. Domínguez-Renedo, M.A. Alonso-Lomillo, L. Ferreira-Gonçalves, and M.J. Arcos-Martínez. Development of urease based amperometric biosensors for the inhibitive determination of hg (ii). *Talanta*, 79 :1306–1310, 2009.

[32] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts ? In *Advances in neural information processing systems*, volume 16, pages 1141–1148, 2004.

[33] S.F. D'Souza. Microbial biosensors. *Biosensors and Bioelectronics*, 16 :337–353, 2001.

[34] C. Durrieu and C. Tran-Minh. Optical algal biosensor using alkaline phosphatase for determination of heavy metals. *Ecotoxicology and Environmental Safety*, 51 :206–209, 2002.

[35] P. Fochtman, A. Raszka, and E. Nierzedska. The use of conventional bioassays, microbiotests, and some rapid methods in the selection of an optimal test battery for the assessment of pesticides toxicity. *Environmental Toxicology*, 15 :376 – 384, 2000.

[36] J.K. Fredrickson and Y.A. Gorby. Environmental processes mediated by iron-reducing bacteria. *Current Opinion in Biotechnology*, 7(3) :287 – 294, 1996.

[37] J.C. Gayet, A. Haouz, A. Meyer, and C. Burstein. Detection of heavy metal salts with biosensors built with an oxygen electrode coupled to various immobilized oxidases and dehydrogenases. *Biosensors and Bioelectronics*, 8 :177 – 183, 1993.

[38] G.H. Golub and C.F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

[39] X. Guo, S. Miron, and D. Brie. Uni-mode and partial uniqueness conditions for candecomp/parafac of three-way arrays with linearly dependent loadings. *SIAM J. Matrix Anal. Appl*, 33(1) :111–129, 2012.

[40] G. Haferburg and E. Kothe. Microbes and metals : interactions in the environment. *Journal of Basic Microbiology*, 47 :453–467, 2007.

[41] R. A. Harshman. Foundations of the PARAFAC procedure : Models and conditions for an 'explanatory' multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16 :1–84, 1970.

[42] R. A. Harshman and W. S. DeSarbo. An application of parafac to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. *Research methods for multimode data analysis*, pages 602 – 642, 1984.

[43] B. M. Hewitt, N. Singhal, R. G. Elliot, A. Y. H. Chen, J. Y. C. Kuo, F. Vanholsbeeck, and S. Swift. Novel fiber optic detection method for in situ analysis of fluorescently labeled biosensor organisms. *Environmental Science & Technology*, 46(10) :5414–5421, 2012.

[44] S.J. Hug. An adapted water treatment option in bangladesh : solar oxidation and removal of arsenic (soras). *Environmental Sciences*, 8 :467 – 479, 2001.

[45] H. Huot. Utilisation de biosenseurs bactériens fluorescents pour évaluer la biodisponibilité des éléments métalliques dans les sols. Technical report, LIMOS, 2009.

[46] J Jiang, H-l Wu, Y Li, and R-Q Yu. Three-way data resolution by alternating slice-wise diagonalization (asd) method. *Journal of Chemometrics*, 14(1) :15–36, 2000.

[47] N. Jie, Z. Si, J. Yang, Q. Zhang, X. Huang, and D. Yang. Determination of cerium in rare earth ores by fluorescence quenching of rhodamine 6g. *Mikrochimica Acta*, 126 :93–96, 1997.

[48] S. Jouanneau, M.J. Durand, P. Courcoux, T. Blusseau, and Thouand. G. Improvement of the identification of four heavy metals in environmental samples by using predictive decision tree models coupled with a set of five bioluminescent bacteria. *Environmental Science & Technology*, 45 :2925–2931, 2011.

[49] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3) :455–500, September 2009.

[50] D. W. Kolpin, E. T. Furlong, M. T. Meyer, E. M. Thurman, S. D. Zaugg, L. B. Barber, and H. T Buxton. Pharmaceuticals, hormones, and other organic wastewater contaminants in u.s. streams, 1999-2000 : A national reconnaissance. *Environmental Science and Technology*, 36 :1202 – 1211, 2002.

[51] J. B. Kruskal. Three-way arrays : Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2) :95–138, 1977.

[52] J.R. Lakowicz. *Principles of Fluorescence Spectroscopy*. Springer 3rd edition, 2006.

[53] W.H. Lawton and E.A. Sylvestre. Self modeling curve resolution. *Technometrics*, 13 :617 – 633, 1971.

[54] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 :788–791, 1999.

[55] M. Lehmann, K. Riedel, K. Adler, and G. Kunze. Amperometric measurement of copper ions with a deputy substrate using a novel saccharomyces cerevisiae sensor. *Biosensors and Bioelectronics*, 15 :211–219, 2000.

[56] R. Loos, B. M. Gawlik, G. Locoro, E. Rimaviciute, S. Contini, and G. Bidoglio. Eu-wide survey of polar organic persistent pollutants in european river waters. *Environmental Pollution*, 157(2) :561 – 568, 2009.

[57] D.R. Lovley and J.D. Coates. Novel forms of anaerobic respiration of environmental relevance. *Current Opinion in Biotechnology*, 3 :252 – 256, 2000.

[58] X. Luciani. *Analyse numérique des spectres de fluorescence 3D issus de mélanges non linéaires*. PhD thesis, Université Sud Toulon-Var, 2007.

[59] X. Luciani. *Analyse numérique des spectres de fluorescence 3D issus de mélanges non linéaires*. PhD thesis, Université du Sud Toulon Var, 2008.

[60] X. Luciani, S. Mounier, R. Redon, and A. Bois. A simple correction method of inner filter effects affecting feem and its application to the parafac decomposition. *Chemometrics and Intelligent Laboratory Systems*, 96 :227 – 238, 2009.

[61] S. Magrisso, Y. Erel, and S. Belkin. Microbial reporters of metal bioavailability. *Microbial Biotechnology*, 1 :320–330, 2008.

[62] L.S McCarty and C.J. Borgert. Review of the toxicity of chemical mixtures : Theory, policy, and regulatory practice. *Regulatory Toxicology and Pharmacology*, 45 :119 – 143, 2006.

[63] S. Moussaoui, D. Brie, and J. Idier. Non-negative source separation : Range of admissible solutions and conditions for the uniqueness of the solution. In *Proc. ICASSP*, volume 5, pages 289–292, Philadelphia, PA, March 2005.

[64] 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 Trans. on Signal Processing*, 54(11) :4133–4145, November 2006.

[65] A. Mulchandani and A.S. Bassi. Principles and applications of biosensors for bioprocess monitoring and control. *Critical Reviews in Biotechnology*, 15 :105–124, 1995.

[66] D.H. Nies. Microbial heavy-metal resistance. *Applied Microbiology and Biotechnology*, 51 :730–750, 1999.

[67] G. Nucifora, L. Chu, T. K. Misra, and S. Silver. Cadmium resistance from staphylococcus aureus plasmid pi258 cada gene results from a cadmium-efflux atpase. *Proc. Nat. Acad. Sci.*, 86 :3544–3548, 1989.

[68] P. Paatero and U. Tapper. Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2) :111–126, 1994.

[69] S.M. Paixão, L. Silva, A. Fernandes, K. O'Rourke, E. Mendonça, and A. Picado. Performance of a miniaturized algal bioassay in phytotoxicity screening. *Ecotoxicology*, 17 :165 – 171, 2008.

[70] P. Pandard, J. Devillers, A.M. Charissou, V. Poulsen, M.J. Jourdain, J.F. Férard, C. Grand, and A. Bispo. Selecting a battery of bioassays for ecotoxicological characterization of wastes. *Science of the Total Environment*, 363 :114 – 125, 2006.

[71] E.S. Park, C.H. Spiegelman, and R.C. Henry. Bilinear extimation of pollution source profiles and amounts by using multivariate receptor models. *Environmetrics*, 13 :775 – 708, 2002.

[72] M. Rajih, P. Comon, and R. A. Harshman. Enhanced line search : A novel method to accelerate parafac. *SIAM J. Matrix Anal. Appl.*, 30(3) :1128–1147, September 2008.

[73] P. J. Ralph, R.A. Smith, C.M.O. MacInnis-Ng, and C.R. Seery. Use of fluorescence-based ecotoxicological bioassays in monitoring toxicants and pollution in aquatic systems : Review. *Toxicological and Environmental Chemistry*, 89 :589 – 607, 2007.

[74] S. Ramanathan, M. Ensor, and S. Daunert. Bacterial biosensors for monitoring toxic metals. *Trends in Biotechnology*, 15 :500–506, 1997.

[75] I. V. N. Rathnayake, M ; Megharaj, N. Bolan, and Ravi Naidu. Tolerance of heavy metals by gram positive soil bacteria. *World Academy of Science, Engineering and Technology*, 53 :1185–1189, 2009.

[76] C. Rensing, M. Ghosh, and B.P. Rosen. Families of soft-metal-ion-transporting atpases. *J. Bacteriol.*, 181 :5891–5897, 1999.

[77] C. Rensing and R.M. Maier. Issues underlying use of biosensors to measure metal bioavailability. *Ecotoxicol. Environ. Saf.*, 56 :140–147, 2003.

[78] C. Rensing, Y. Sun, B. Mitra, and B.P. Rosen. Pb(ii)-translocating p-type atpases. *J. Biol. Chem.*, 273 :32614–32617, 1998.

[79] T.A. Richmond, T.T. Takahashi, R. Shimkhada, and J. Berndorf. Engineered metal binding sites on green fluorescence protein. *Biochemical And Biophysical Research Communications*, 268 :462–465, 2000.

[80] R. Rojickova-Padrtova, B. Marsalek, and I. Holoubek. Evaluation of alternative and standard toxicity assays for screening of environmental samples : Selection of an optimal test battery. *Chemosphere*, 37 :495 – 507, 1998.

[81] V.J. Ruigrok, M. Levisson, M.H. Eppink, H. Smidt, and van der Oost J. Alternative affinity tools : more attractive than antibodies ? *Biochemical Journal*, 436 :1 – 13, 2011.

[82] I. Satoh. An apoenzyme thermistor microanlaysis for zinc (ii) ions with use of an immobilized alkaline phosphatase reactor in a flow system. *Biosensors and Bioelectronics*, 6 :375 – 379, 1991.

[83] J.A. Scott and S.J. Palmer. Sites of cadmium uptake in bacteria used for biosorption. *Applied Microbiology and Biotechnology*, 33 :221–225, 1990.

[84] T.N. Shekhovtsova, S.V. Muginova, and N.A. Bagirova. Determination of organomercury compounds using immobilized peroxidase. *Analytica chimica acta*, 344 :145–151, 1997.

[85] D.A. Skoog. *Fundamentals of Analytical Chemistry*. Brooks Cole 8 edition, 2003.

[86] A. Smilde, B. Rasmus, and P. Geladi. *Multi-Way Analysis : Applications in the Chemical Sciences*. Wiley, 2004.

[87] A.K. Smilde, H.C.J. Hoefsloot, H.A.L. Kiers, S. Bijlsma, and H.F.M. Boelens. Sufficient condition for unique solutions within a certain class of curve resolution models. *Journal of Chemometrics*, 15 :405–411, 2001.

[88] A. Stegeman and A. L. F. de Almeida. Uniqueness conditions for constrained three-way factor decompositions with linearly dependent loadings. *SIAM J. Matrix Anal. Appl.*, 31(3) :1469–1499, 2009.

[89] J. Stocker, D. Balluch, M. Gsell, H. Harms, J. Feliciano, S. Daunert, K.A. Malik, and J.R. van der Meer. Development of a set of simple bacterial biosensors for quantitative and rapid measurements of arsenite and arsenate in potable water. *Environmental Science & Technology*, 37 :4743 – 4750, 2003.

[90] J.P. Sumner, N.M. Westerberg, A.K. Stoddard, T.K. Hurst, M. Cramer, R.B. Thompson, C.A. Fierke, and R. Kopelman. Dsred as a highly sensitive, selective, and reversible fluorescence-based biosensor for both cu+ and cu2+ ions. *Biosensors and Bioelectronics*, 21(7) :1302 – 1308, 2006.

[91] R. Tauler, B. Kowalski, and S. Fleming. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical chemistry*, 65 :2040 – 2047, 1993.

[92] S. Tauriainen, M. Karp, W. Chang, and M. Virta. Luminiscent bacterial sensor for cadmium and lead. *Biosensors and Bioelectronics*, 13 :931–938, 1998.

[93] R. Tecon and J.R. van der Meer. Bacterial biosensors for measuring availability of environmental pollutants. *Sensors*, 8(7) :4062–4080, 2008.

[94] J. M. F. ten Berge and N. D. Sidiropoulos. On uniqueness in CANDECOMP/PARAFAC. *Psychometrika*, 67(3) :399–409, September 2002.

[95] C. Tibazarwa, P. Corbisier, M. Mench, A. Bossus, P. Solda, M. Mergeay, L. Wyns, and D. van der Lelie. A microbial biosensor to predict bioavailable nickel in soil and its transfer to plants. *Environmental Pollution*, 113 :19–26, 2001.

[96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1) :267–288, 1996.

[97] G. Tomasi and R. Bro. A comparison of algorithms for fitting the parafac model. *Computational Statistics and Data Analysis*, 50(7) :1700 – 1734, 2006.

[98] J.T. Trevors, G.W. Stratton, and G.M. Gadd. Cadmium transport, resistance, and toxicity in bacteria, algae, and fungi. *Canadian Journal of Microbiology*, 32 :447–464, 1986.

[99] J.A. Tropp, A.C. Gilbert, and M.J. Strauss. Algorithms for simultaneous sparse approximation. part i : Greedy pursuit. *Signal Processing, special issue "Sparse approximations in signal and image processing,"*, 86 :572 – 588, 2006.

[100] K.J. Tsai, K.P. Yoon, and A.R. Lynn. Atp-dependent cadmium transport by the cada cadmium resistance determinant in everted membrane vesicles of bacillus subtilis. *J. Bacteriol.*, 174 :116–121, 1992.

[101] M. Valko, H. Morris, and M.T. Cronin. Metal, toxicity and oxidative stress. *Current Medicinal Chemistry*, 12 :1161 –1208, 2005.

[102] E. van der Grinten, M.G. Pikkemaat, E.J. van den Brandhof, G.J. Stroomberg, and M. H.S. Kraak. Comparing the sensitivity of algal, cyanobacterial and bacterial bioassays to different groups of antibiotics. *Chemosphere*, 80 :1 – 6, 2010.

[103] N. Verma and M. Singh. Biosensors for heavy metals. *BioMetals*, 18 :121–129, 2005.

[104] V. Vrabie. *Statistiques d'ordre supérieur : applications en géophysique et électrotechnique*. PhD thesis, INP de Grenoble, 2003.

[105] P.J. Wilderman, N.A. Sowa, D.J. FitzGerald, P.C. FitzGerald, S. Gottesman, U.A. Ochsner, and M.L. Vasil. Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis. *PNAS*, 101(26) :9792–9797, 2004.

[106] H.H. Zeng, R.B. Thompson, B.P. Maliwal, G.R. Fones, J.W. Mosffet, and Fierke C.A. Real-time determination of picomolar free cu(ii) in seawater using a fluorescence based fiber optic biosensor. *Analytical chemistry*, 75 :6807 – 6812, 2003.

[107] M. Zibulevsky and M. Elad. L1-l2 optimization in signal and image processing. *IEEE Signal Proc. Mag.*, 27(3) :76–88, May 2010.

Résumé

La disponibilité et la persistance à l'échelle locale des métaux lourds pourraient être critiques notamment pour l'usage futur des zones agricoles ou urbaines, au droit desquelles de nombreux sites industriels se sont installés dans le passé. La gestion de ces situations environnementales complexes nécessitent le développement de nouvelles méthodes d'analyse peu invasives (capteurs environnementaux), comme celles utilisant des biosenseurs bactériens, afin d'identifier et d'évaluer directement l'effet biologique et la disponibilité chimique des métaux. Ainsi dans ce travail de thèse, nous avons cherché à identifier, à l'aide d'outils mathématiques de l'algèbre multi-linéaire, les réponses de senseurs bactériens fluorescents dans des conditions environnementales variées, qu'il s'agisse d'un stress engendré par la présence à forte dose d'un métal ou d'une carence nutritive engendrée par son absence. Cette identification est fondée sur l'analyse quantitative à l'échelle d'une population bactérienne de signaux multidimensionnels. Elle repose en particulier sur (i) l'acquisition de données spectrales (fluorescence) multivariées sur des suspensions de biosenseurs multicolores interagissant avec des métaux et sur (ii) le développement d'algorithme de décomposition tensoriels. Les méthodes proposées, développées et utilisées dans ce travail s'efforcent d'identifier « sans a *priori* » (a *minima*), la réponse fonctionnelle de biosenseurs sous différentes conditions environnementales, par des méthodes de décomposition de tenseurs sous des signaux spectraux observables. Elles tirent parti de la variabilité des réponses systémiques et permettent de déterminer les « sources » élémentaires identifiant le système et leur comportement en fonction des paramètres extérieurs. Elles sont inspirées des méthodes CP et PARALIND . L'avantage de ce type d'approche, par rapport aux approches classiques, est l'identification unique des réponses des biosenseurs sous de faibles contraintes. Le travail a consisté à développer des algorithmes efficaces de séparations de sources pour les signaux fluorescents émis par des senseurs bactériens, garantissant la séparabilité des sources fluorescentes et l'unicité de la décomposition. Le point original de la thèse est la prise en compte des contraintes liées à la physique des phénomènes analysés telles que (i) la parcimonie des coefficients de mélange ou la positivité des signaux "source", afin de réduire au maximum l'usage d'a priori ou (ii) la détermination non empirique de l'ordre de la décomposition (nombre de sources). Cette posture a permis aussi d'améliorer l'identification en optimisant les mesures physiques par l'utilisation de spectres synchrones ou en apportant une diversité suffisante aux plans d'expériences. L'usage des spectres synchrones s'est avéré déterminant à la fois pour améliorer la séparation des sources de fluorescence, mais aussi pour augmenter le rapport signal sur bruit des biosenseurs les plus faibles. Cette méthode d'analyse spectrale originale permet d'élargir fortement la gamme chromatique des biosenseurs fluorescents multicolores utilisables simultanément. Enfin, une nouvelle méthode d'estimation de la concentration de polluants métalliques présents dans un échantillon à partir de la réponse spectrale d'un mélange de biosenseurs non-spécifiques a été développée.

Mots-clés: biosenseurs, séparation de sources, autofluorescence, données multidimensionnelles, spectrofluorimétrie, pollution environnementales, algèbre multilinéaire, unicité, colinéarité, CANDECOMP/PARAFAC,.

Abstract

Availability and persistence of heavy metals could be critical for future use of agricultural or urban areas, on which many industrial sites have installed in the past. The management of these complex environnemental situations requiring the developpement of new analytical methods minimally invasive, such as bacterial

biosensors, to identify and directly assess the biological effects and the chemical availability of metals.

The aims of this thesis was to identify the responses of fluorescent bacterial sensors invarious environmental conditions, using mathematical tools of algebra multilinear, whether stress caused by the presence of high dose of a metal or a nutrient deficiency caused by his absence. This identification is based on quantitative analysis of multidimensional signals at the bacterial population-scale.

It is based in particular on (i) the acquisition of multivariate spectral data on suspensions of multicolored biosensors interacting with metals and (ii) the development of algortihms for tensor decomposition.

The proposed methods, developed and used in this study attempt to identify functional response of biosensors without *a priori* by decomposition of tensor containing the spectral signals. These methods take advantage of the variability of systemic responses and allow to determine the basic sources identifying the system and their behavior to external factors. They are inspired by the CP and PARALIND methods. The advantage of this approach, compared to conventional approaches, is the unique identification of the responses of biosensors at low constraints.

The work was to develop efficient algorithms for the source separation of fluorescent signals emitted by bacterial sensors, ensuring the sources separability and the uniqueness of the decomposition. The original point of this thesis is the consideration of the physical constraints of analyzed phenomena such as (i) the sparsity of mixing coefficients or positivity of sources signals in order to minimize the used of a priori or (ii) the non-empirical determination of the order of decomposition (number of sources).This posture has also improved the identification optimizing physical measurements by the use of synchronous spectra or providing sufficient diversity in design of experiments. The use of synchronous spectra proved crucial both to improve the separation of fluorescent sources, but also to increase the signal to noise ratio of the lowest biosensors. This original method of spectral analysis can greatly expand the color range of multicolored fluorescent biosensors used simultaneously. Finally, a new method of estimating the concentration of metal pollutants present in a sample from the spectral response of a mixture of non-specific biosensor was developed.

Keywords: Biosensors, sources separation, autofluorescence, multiway data, spectrofluorimetry, environmental pollution, multilinear algebra, 4-way array, uniqueness, collinear loadings, CANDECOMP/PARAFAC.