Analyse cartographique et bayésienne de la dégradation d'un antibiotique en traitement de l'Eau

Rachid Ouaret, Ali Badara Minta, Jean-Marc Choubert, Claire Albasi, Antonin Azaïs

Analyse cartographique et bayésienne de la dégradation d'un antibiotique en traitement de l'Eau

Rachid Ouaret[1], **Ali Badara Minta**[1,2], **Jean-Marc Choubert**[2], **Claire Albasi**[1], **Antonin Azaïs**[2]

[1]Laboratoire de Génie Chimique, Université de Toulouse, CNRS, INPT, UPS, Toulouse, France
[2]INRAE, UR REVERSAAL, 5 Rue de La Doua, CS 20244, F-69625, Villeurbanne Cedex, France
rachid.ouaret@toulouse-inp.fr

Abstract

De nombreux projets de recherche menés ces dernières années se sont intéressés au devenir des micropolluants dans les procédés d'épuration. Il a été mis en évidence que la réduction des concentrations de ces contaminants peut s'accompagner de la génération de produits de transformation pouvant conserver ou parfois amplifier des effets toxiques sur la vie aquatique. À ce jour, bien que les moyens analytiques permettent de détecter une large part de ces molécules, leur quantification et leur identification dans des échantillons environnementaux complexes reste un verrou autant technique qu'économique du fait de la disparité des structures, de leur instabilité et de la disponibilité des étalons (analyse ciblée). Différents schémas réactionnels sont proposés dans la littérature, cependant, les données issues de ces schémas sont, à ce jour, inexploitées. Alternative innovante aux limitations expérimentales, la science des données, notamment les techniques de fouille de données sur des graphes, pourrait permettre d'établir les filiations entre les molécules les plus pertinentes. Ce travail, sous un format d'un stage de Master, s'inscrit dans le cadre du projet de recherche ANR TRANSPRO démarré en 2019. Appuyé sur l'analyse d'une bibliographie approfondie regroupant 46 articles pour plus de 140 molécules, ce stage a pour objectif de dresser un profil statistique type des voies réactionnelles observées dans la littérature. Il aura permis la mise en place d'une méthodologie innovante d'analyse cartographique et bayésienne de la dégradation d'un antibiotique. Pour répondre à ces attentes, le formalisme des modèles graphiques probabilistes a été adopté pour "simuler" les chemins réactionnels de la dégradation des molécules en traitement d'eau. Les premiers résultats de cette étude sont très encourageants, surtout ceux obtenus avec les algorithmes à base de score Bayesian Dirichlet Equivalent (BDe).

Introduction

L'eau est une ressource rare, ultime réceptacle des pollutions anthropiques, sa préservation et son traitement font l'objet de nouveaux paradigmes en termes de développement analytique, de procédés, voire de gestion des données générées. Ces dernières décennies la communauté scientifique des sciences analytiques et environnementales a permis la détection de nouvelles molécules de synthèse (produits manufacturés) à l'état de trace, également appelées micropolluants, dans l'ensemble des compartiments aquatiques (eau domestique, de surface et souterraine). Ces observations ont encouragé la mise en place de la Directive Cadre sur l'Eau (DCE), adoptée en Europe en 2000 (2000/60/CE), qui vise à protéger et/ou à restaurer la qualité des écosystèmes aquatiques.

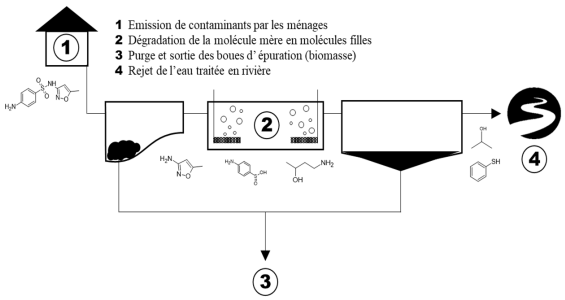


Figure 1: Émission, élimination et rejet de substances émergents en station d'épuration (STEP).

Toutefois, malgré des procédés innovants et une complexification croissante des filières de traitement, la dégradation de ces contaminants émergents ne peut être complète, générant d'autres molécules appelées *"produits de transformation"* (TP) (Figure 1) pouvant conserver ou parfois amplifier des effets toxiques sur l'Environnement (Lai et al. 2018). Très peu d'études se sont intéressées à ces molécules filles, du fait des défis scientifiques que leur détection, quantification et modélisation soulèvent. Pour répondre à ces verrous, le projet ANR TRANSPRO (2019–2022) porté par un consortium de partenaires publics (EPOC, LGC et INRAE) s'est fixé comme objectif d'améliorer les connaissances sur la nature, l'origine et la dynamique des TP.

Les données sur l'occurrence des TP restent parcellaires, mais suffisantes pour empêcher toute tentative de synthèse manuelle. En effet, les schémas réactionnels proposés dans la littérature divergent grandement, autant dans leur structure (pattern) que dans le nombre et la nature des molécules (éléments) représentées. Il est supposé que ces différences sont les résultats d'un concours de causes dont le type de procédé(s) utilisé(s), leurs conditions opératoires, la puissance analytique (résolution) disponible et finalement

l'expertise intrinsèque au consortium de recherche.

Toutefois, même en considérant des études récentes (Chen et al. 2019; Yazdanbakhsh et al. 2020) menées sur le même contaminant (le sulfamethoxazole, un antibiotique) et le même type de procédés, ici d'oxydation (Figure 2), il apparait qu'une seule des molécules filles est commune entre les deux exemples considérés. À ce jour et à la connaissance des auteurs, le bassin de données générés suite à ces études mécanistiques (physico-chimiques) est inexploité. Aucune tentative de consolidation de ces données n'a été menée bien que la science des données apparait comme une alternative innovante permettant d'établir les filiations les plus probables entre les molécules entre elles et d'identifier les mécanismes réactionnels préférentiels. Au vu de l'ensemble des schémas proposés dans la littérature, les techniques de fouille de données de graphes permettraient d'apporter un nouvel éclairage sur les chemins de dégradation d'une substance les plus communs. C'est dans cette perspective que les modèles graphiques probabilistes ont été choisis pour mener cette étude. Outre la création de connaissance et la meilleure compréhension réactionnelle, une bancarisation et un traitement de ces données permettrait de prioriser les produits de transformation (TP) préférentiellement formés et ainsi de les prioriser pour toutes études environnementales et toxicologiques.

En ce sens, un stage de Master a été initié au printemps 2021 ayant pour objectif la bancarisation et le traitement des données issues de la littérature et concernant la dégradation du sulfamethoxazole (SMX) par des traitements biologiques et oxydatifs. Les résultats obtenus au cours de stage, co-encadré par le LGC et l'INRAE, font l'objet de cette soumission.

Méthodologie

Bibliométrie et identification des schémas de dégradation

Cette étude s'articule sur 3 parties complémentaires (Figure 3) : (A) la récolte d'articles de recherche, la collecte des données et la fouille de données bibliométrique à partir des moteurs de recherche dédiés de type ScienceDirect, la partie (B) consiste en l'identification des voies de dégradation d'un polluant cible (ici, SMX) et représentation matricielle des niveaux de dégradation, et enfin la partie (C) représente l'utilisation des informations de différents schémas réactionnels pour l'analyse des Réseaux Bayésiens (RB) ou de connaissance.

- Dans la partie (A), nous avons sélectionné 46 articles sur la base des schémas qu'ils contiennent. Rappelons que les informations des voies réactionnelles sont souvent représentées dans une figure sur laquelle on identifie toutes les molécules ainsi que toutes les voies. À ce jour, aucune méthode automatique d'extraction des informations pour ce type de représentations n'est proposées dans la littérature. Pour cette raison, la partie (B) est conçue pour alimenter une base de données créée à cet effet. Deux différents types de procédés ont été utilisés dans ces articles, avec 30 articles scientifiques basées sur les procédés d'oxydation avancée (AOP) et 16 présentent la dégradation du SMX à travers la mise en œuvre de divers procédés biologiques (BIO). En tout, nous avons identifié 132 chemins réactionnels et 113 molécules pour les procédés oxydatifs et 43 voies pour 57 molécules identifiées dans les procédés biologiques.

- La partie (B) fait appel à l'expert du domaine (physico-chimiste) pour l'extraction des voies de dégradation. Cette partie est fastidieuse, nécessitant une transcription manuelle de toutes les molécules ainsi que de leur niveau, ou position, dans le schéma de dégradation c'est-à-dire transcrire dans la matrice la distance entre les produits de transformation entre eux et avec le SMX. Sur l'ensemble des schémas présentés dans les 46 articles (exemple deux articles de la Figure 2), 141 molécules et 177 voies réactionnelles ont été détectées. Une première représentation matricielle est alors obtenue comme le détaille la Figure 4. Ensuite, une matrice d'adjacence pondérée par le nombre d'arcs observés entres les différents sous-produits est déduite (voir l'exemple dans la Figure 5). En supposant que l'on dispose de deux articles dans lesquels deux schémas de dégradation du SMX menant à la formation des sous-produits suivants : A, B, C, D et E. Dans la matrice, on considère les molécules en colonne comme les molécules d'arrivée issues de la dégradation des molécules de départ qui sont en ligne. Les valeurs dans la matrice correspondent au nombre de fois où la dégradation de la molécule de départ en la molécule d'arrivée est observée dans l'ensemble des articles pour toutes les voies réactionnelles proposées. À ce jour, cette partie d'enrichissement de la base de données est effectuée manuellement. Cette base de données est souvent alimentée par l'actualisation des articles et l'identification des nouvelles molécules et chemins réactionnels.

- La partie (C) consiste à représenter graphiquement les interactions entre la molécule cible (SMX) et un ensemble de produits de transformations. Deux approches ont été utilisées. D'abord par l'intégration systématique des informations sur le nombre de liens directs entre les molécules (poids de l'arc) et le nombre d'arcs sortant des différentes molécules (poids du nœud). Dans ce cas, il s'agit d'un modèle supposé fidèle à la réalité physique. Pour cela, nous avons utilisé le logiciel Gephi (version 0.9.2) qui permet de représenter l'ensemble des voies en mettant en évidence l'influence des molécules intermédiaires (Bastian, Heymann, and Jacomy 2009). Ensuite, à partir de la première matrice (cf. Figure 4), un réseau bayésien est appliqué pour trouver des similitudes avec le modèle de référence. Dans ce dernier cas, très peu d'information sont fournies : un tableau ayant 177 lignes (voies) et 141 variables (molécules). Pour chaque molécule, un codage binaire a été appliqué : molécule observée (VU) pour une voie donnée et molécule non-observé (NP). Pour les procédés BIO, nous disposons de 43 chemins réactionnels et de 57 produits de transformation. Pour les procédés AOP, la matrice est composée de 132 voies observées sur 113 molécules.

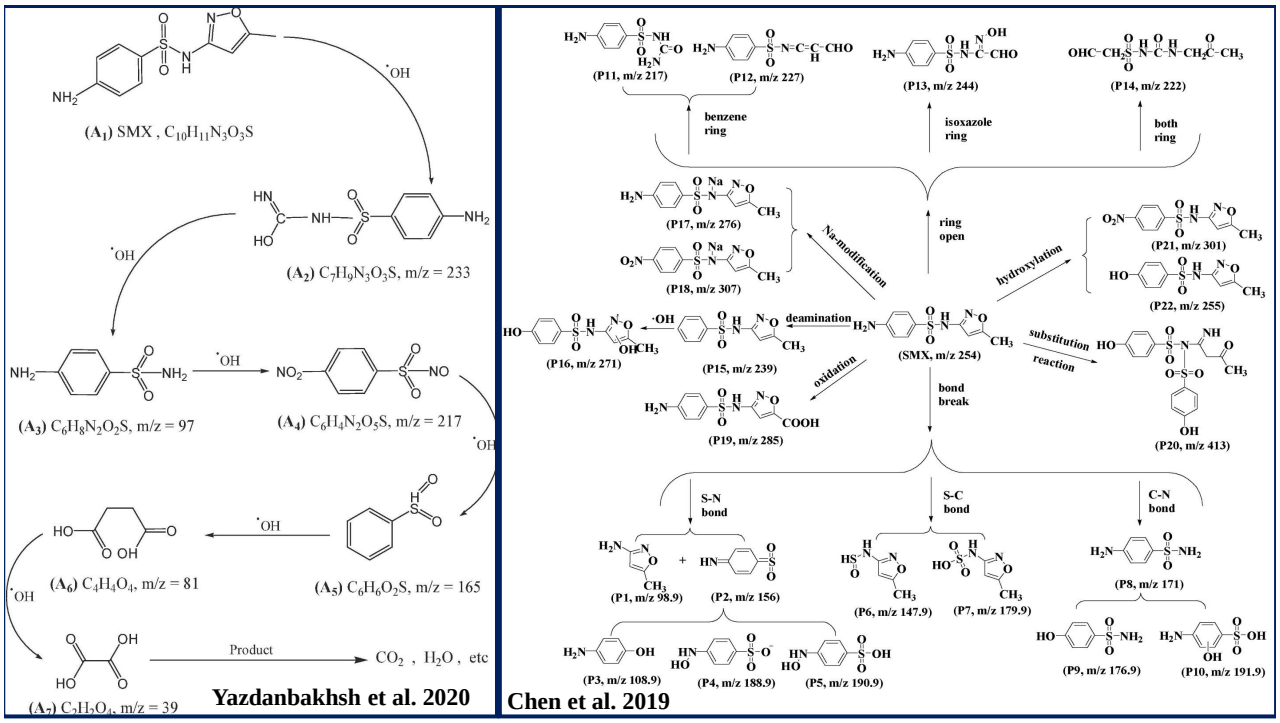


Figure 2: Exemples de schémas de dégradation observés dans la littérature, ici les études de (Chen et al. 2019) et de (Yazdanbakhsh et al. 2020).

Réseaux Bayésiens

le sujet principal de ce travail peut se résumer en une question *"Peut-on imaginer des méthodes de fouille de graphes qui puissent, à partir de schémas réactionnels, extraire les voies les plus probables ?"*. Les modèles graphiques probabilistes (Pearl 1988), et plus précisément les réseaux bayésiens nous paraissent appropriés pour apporter un éclairage nouveau à cette question. De surcroit, un modèle graphique a l'avantage d'être visuel, et selon les propriétés modélisées (indépendance par exemple) peut être plus ou moins puissant pour détecter des relations entre les molécules. À partir du tableau obtenu dans la partie B (*cf.* Section Méthodologie et Figure 4), il est possible d'utiliser un Réseau Bayésien sur des variables nominales. L'utilisation de ce type de modèle nécessite un apprentissage de la structure graphique (nœuds et arcs) et l'association d'une distribution de probabilité conditionnelle à chaque molécule. Formellement, un réseau bayésien $\mathscr{B}(\mathcal{G}, \theta)$ est défini par (*i*) un graphe dirigé $\mathcal{G}(X, E)$ sans circuit dont les nœuds sont associés à un ensemble de variables aléatoires $X = \{X_1, X_2, \ldots, X_n\}$ et E représente l'ensemble des arcs, et (*ii*) un ensemble des probabilités $\theta = \{P(X_i \mid Pa(X_i))\}$ de chaque nœud X_i conditionnellement à l'état de ses parents $Pa(X_i)$ dans $\mathcal{G}$.

Dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée pour l'estimation des probabilités est la méthode de maximum de vraisemblance (MV) qui donne $\widehat{P}(X_i = x_k \mid Pa(X_i) = x_j) = \widehat{\theta^{\text{MV}}_{ijk}} = \frac{N_{ijk}}{\sum_k N_{ijk}}$, où N_{ijk} est le nombre de fois où X_i vaut k et $Pa(X_i)$ prend sa $j-$ième valeur, pour tout k.

En ce qui concerne l'apprentissage du graphe, Il existe deux types de techniques pour construire la structure du réseau bayésien. les *Score based Algorithm* (SBA) et les *Constraint based Algorithm* (CBA). Pour une revue de littérature sur les méthodes d'apprentissage de graphe, nous recommandons la thèse de (François 2006). L'utilisation des algorithmes de type CBA sur les données de réactions physico-chimiques de SMX mènent assez souvent à des constructions très peu fidèles aux structures observées dans la littérature. Pour cette raison, cette étude se concentre sur les méthodes de score et plus particulièrement sur les Algorithmes de types Hill-climbing (HC) (Chickering 2002). Les méthodes à base de score parcourent l'espace de recherche en examinant uniquement les changements locaux possibles dans le voisinage de la solution actuelle et applique celui qui maximise la fonction de score. Très brièvement, ces algorithmes structurent le réseau à partir d'un graphe vide, généré aléatoirement ou prédéfini par l'utilisateur. Ensuite, un score est calculé sur la base de cette initialisation et des modifications dans le graphe (ajout, inversion ou suppression d'une flèche) sont effectuées. Si la modification n'augmente pas le score, on revient à l'état précédent et on effectue une autre modification ; si au contraire le score augmente, on conserve le nouvel état et une nouvelle modification est opérée. L'algorithme s'arrête quand aucune modification ne permet d'augmenter le score. Les méthodes à base de score sont facilement piégées dans les nombreux minima locaux et le graphe final obtenu dépend fortement des conditions initiales. Ainsi, le graphe vide représente le choix le

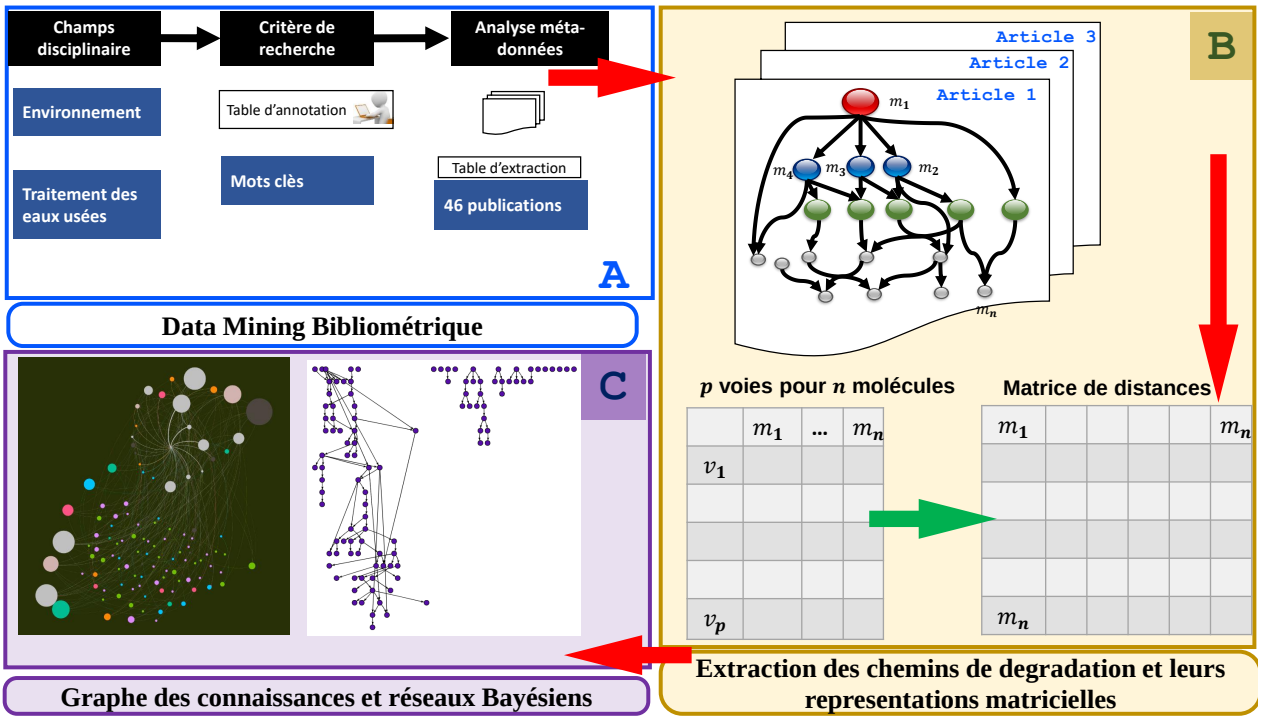
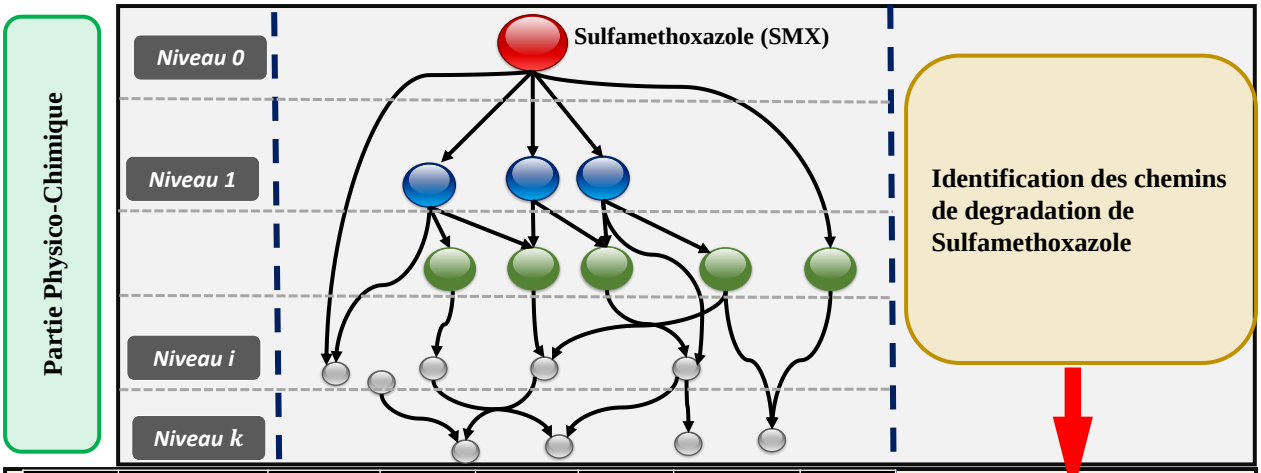
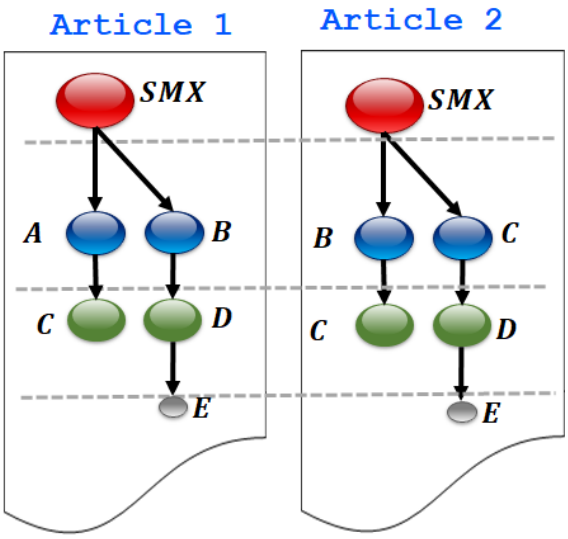


Figure 3: Schéma général de la méthodologie suivie pour l'analyse des voies de dégradation des molécules observées dans la littérature en traitement d'eau. Deux blocs complémentaires (A-B) pour l'enrichissement de la base de données et un bloc (C) d'analyse par les modèles graphiques.



Articles	Voies de dégradation	SMX	Molécule1	Molécule2	....	Molécule $n-1$	Molécule n
Article 1	Voie 1	Procédé 1	Niveau 1	Niveau 5	...	NA	NA
	Voie 2	Procédé 2	Niveau 4	NA	...	Niveau 4	Niveau 2
	Voie 3	Procédé 1	Niveau 3	Niveau 2	...	NA	Niveau 1
		...	...	...	...	...	...
	Voie p	Procédé 2	Niveau 6	Niveau 7	NA	NA	NA

Figure 4: Identification des voies réactionnelles et leur représentation matricielle. Au niveau k, l'ensemble des molécules représente la phase finale de la dégradation.



	SMX	A	B	C	D	E
SMX	0	1	**2**	1	0	0
A	0	0	0	1	0	0
B	0	0	0	1	1	0
C	0	0	0	0	0	1
D	0	0	0	0	0	**2**
E	0	0	0	0	0	0

Figure 5: Exemple simplifié de deux schémas de dégradation avec la matrice d'adjacence associée. Une valeur supérieure ou égale à 1 montre le nombre de fois où on a observé un lien direct entre deux molécules. La valeur 0 code l'absence de lien entre les molécules.

plus fréquent en l'absence de connaissances a priori.

En ce qui concerne le choix du score, le score bayésien (la vraisemblance marginale) de type *"Bayesian Dirichlet Equivalent"* (BDe) proposé dans (Heckerman, Geiger, and Chickering 1995) fournit une structure graphique la plus similaire au graphe obtenu sur Gephi et via la base de données de littérature que nous avons mis en œuvre (*cf. la matrice des réactions illustrée dans la Figure 4*). Soient les valeurs $\{x_{i1}, \ldots, x_{ir_i}\}$, $r_i \geq 1$, $i = 1, \ldots, n$ de l'ensemble que chaque X_i peut prendre, $\boldsymbol{D}$ est la base de données et $\mathscr{G}$ la structure du réseau sur X (ensemble des variables aléatoires). Soit $q_i = \prod_{X \in Pa(X_i)} r_i$ est le nombre de configurations possibles pour les parents de X_i.

Le score BDe et est définie par :

$$P(\mathscr{G}, \boldsymbol{D}) = P(\mathscr{G})\, P(\boldsymbol{D} \mid \mathscr{G})$$

$$P(\boldsymbol{D} \mid \mathscr{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

où $P(\mathscr{G})$) représente la probabilité a priori affectée à la structure $\mathscr{G}$, $\alpha_{ijk} = \eta \times \widehat{P}(X_i = x_k, Pa(X_i) = x_j \mid \mathscr{G}_c)$ avec $\mathscr{G}_c$ est le graphe complètement connecté et η un nombre de pseudo-exemples supplémentaires défini par l'utilisateur.

Résultats et discussion

La représentation graphique de la matrice d'adjacence associée aux procédés biologiques obtenue sur Gephi est présentée sur la Figure 6. les nœuds correspondent aux molécules et les liens représentent les voies réactionnelles de chaque molécule répertoriée. La taille de chaque lien (ou arc) reflète la récurrence des voies considérées donnant une indication sur les voies réactionnelles préférentielles du SMX. Divers algorithmes, Force Atlas, Expansion et Déchevauchement, sont déployés afin d'épurer son maillage et de rendre lisible le réseau associé (Jacomy et al. 2014). Par ailleurs, un filtre sur les poids des liens (de 25 % du poids max.) est appliqué, afin de ne faire apparaitre uniquement les arcs ayant un poids supérieur au quart du poids du lien le plus représenté. Un code couleur est assigné en fonction du degré sortant permettant de rendre compte des molécules centrales ou terminales au sein des diverses filiations. À titre d'exemple (voir le diagramme en bâtons de la Figure 6), les molécules figurées en violet représentent 38 % des nœuds et ne donnent aucune molécule fille. Ces molécules, situées au niveau les plus bas des filiations, sont donc considérées comme terminales et/ou peu réactives chimiquement. En revanche, le SMX représente 1,82% des molécules présentées et donne 43 arcs sortants.

Finalement, sur la base des 16 articles récoltés sur les procédés biologiques, la traduction de ce réseau en schéma de dégradation est représentée (Figure 7). Ce schéma de dégradation unique permet l'identification des chemins réactionnels préférentiels (cf. poids des liens) et de prioriser les études mécanistiques, analytiques et toxicologiques. La même méthodologie est adaptée aux procédés oxydatifs sur lesquels trente articles ont été récoltés et analysés. Toutefois, et à titre d'exemple, seuls les résultats obtenus sur les procédés biologiques sont présentés. En ce sens, la filiation de dégradation la plus probable est la suivante : $SMX \rightarrow TP\_98 \rightarrow TP\_83 \rightarrow TP\_89b \rightarrow TP\_60b$. Le 3-amino-5-methylisoxazole (TP_98) est identifié au premier niveau de dégradation, en accord avec huit articles soit la moitié des articles récoltés pour les procédés biologiques. Cette molécule est produite lors de la mise en œuvre de procédés biologiques comme la pile à combustible microbienne (Wang et al. 2016; Xue, Li, and Zhou 2019). Ce procédé assure la dégradation du SMX jusqu'à la formation de l'isopropanol (TP_60b) voire de méthane (CH4) en molécule terminale (non représentée). La seconde voie d'intérêt est celle affiliée au TP_157c (acide 4-aminobenzenesulfinique) issu de la rupture de la liaison S-N (Zhang et al. 2017). Remarquons que certaines molécules sont produites fréquemment hors du chemin préférentiel de la dégradation du SMX. Par exemple, la molécule TP_93 (aniline) est issue de plusieurs produits de transformation (nœud de grande taille), c'est-à-dire qu'en dehors des chemins préférentiels de SMX, il est fort probable d'observer TP_93 à un niveau de réactionnel particulier. Pour les procédés BIO, le TP_93 (aniline) apparait fréquemment à un niveau allant de 2 à 6. Par contre, il est

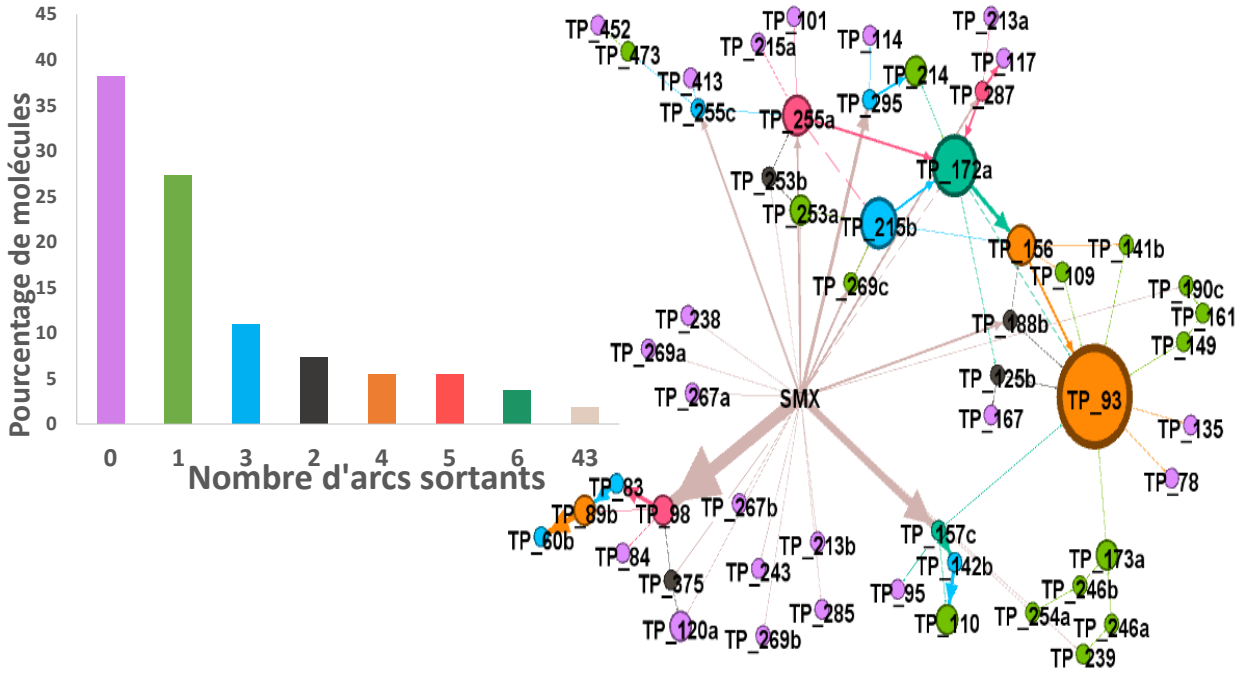


Figure 6: Réseau de dégradation de l'antibiotique sulfamethoxazole (SMX) obtenu pour les procédés biologiques sur la base d'une analyse bibliométrique (Réseau à droite). La taille de chaque lien (ou arc) reflète la récurrence des voies considérées donnant une indication sur les voies réactionnelles préférentielles du SMX. La taille des nœuds représente la probabilité de produire la molécule à partir de plusieurs sous-produits de transformation. Le diagramme en bâton à gauche présente le pourcentage de molécule en fonction d'un nombre maximal d'arcs sortants.

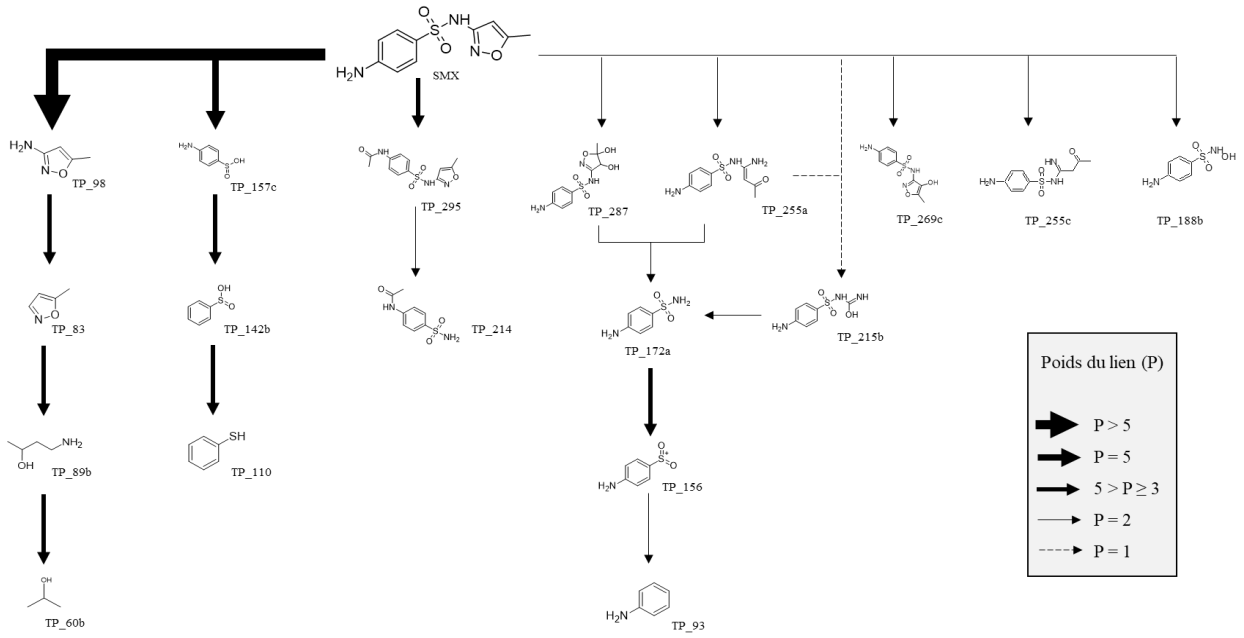


Figure 7: Schémas de dégradation du SMX les plus fréquents basés sur l'interprétation du réseau obtenu à partir de la matrice d'adjacence pour les procédés biologiques.

directement produit par le SMX dans les procédés oxydatifs (niveau 1 de dégradation).

Développer les réseaux bayésiens a également été une perspective de ce stage permettant l'observation des relations de dépendance ou d'indépendance des différentes variables (ici, les molécules). Pour les procédés biologiques, la Figure 8 présente la représentation du réseau bayésien (algorithme HC avec un score BIC) obtenu, ainsi que les filiations communes au graphe de connaissances issues de l'analyse de littérature. Trois codes couleur sont proposés, en vert si la filiation observée est identique et en orange si l'une des deux molécules (mère ou fille) est inversée. Les arcs en noir illustrent les fausses détections ou l'absence de filiation entre les molécules. Il apparait ainsi, que sur le set de données réduit, un réseau bayésien fidèle n'a pu être entièrement obtenu avec le score BIC. En effet, seulement 30 % env. des arcs coïncident entre les liens identifiés dans la littérature (code couleur vert) et moins de la moitié si l'ensemble des filiations mère-fille sont comptabilisées indépendamment de la descendance ou de l'ascendance de l'arc (codes couleur vert et orange). Par ailleurs, notons que les nœuds n'ayant pas de parents directs (voir la Figure 8 et 9) sont considérés comme le premier niveau de dégradation du SMX, donc des arcs (non présentés dans les graphiques) du SMX vers ces molécules sont considérés "valables". Avec le score BDe, le réseau bayésien obtenu améliore sensiblement la détection des filiations. Ainsi, la Figure 9 montre le résultat de l'apprentissage de l'algorithme HC avec le score BDe. Le même code couleur que la Figure 8 a été utilisé pour confronter les liens de parenté avec ceux observés dans la littérature. D'abord, nous remarquons qu'il y a 40% de liens (arcs verts) qui coïncident avec le graphe de connaissance issue de la littérature (*cf.* Réseau de la Figure 6). La molécule TP_93 a été également détectée comme un produit de transformation très fréquent, i.e. plusieurs arcs entrants à partir de différents nœuds. Au vu des résultats obtenus, on peut conforter notre hypothèse sur l'utilité des réseaux bayésiens pour l'identification des schémas réactionnels préférentiels. Sur l'ensemble des scores testés, le BDe donne un bon compromis entre la structure de graphe et son interprétation. Néanmoins, une limitation liée à la construction de la base de données fait qu'on ne peut pas mettre en évidence la molécule cible (SMX) dans les réseaux bayésiens lorsqu'on traite ces données procédé par procédé. Dans ce cas, le SMX ne possède qu'une seule modalité (soit BIO ou AOP), donc elle ne peut pas être considérée comme une variable aléatoire, voir la matrice des réactions illustrée dans la Figure 4. À ce jour, l'amélioration du recouvrement de ces données ainsi que les analyses sur les différents procédés (BIO et AOP) sont en cours.

Conclusion et perspectives

Le suivi et le traitement des pollutions émergentes est le terreau de nombreuses études ces dernières décennies se différenciant par la molécule d'intérêt, le procédé épuratoire, voire le nombre de produits de transformation détectés. Le bassin des données représenté par l'ensemble de ces études paramétriques et physico-chimique est, à ce jour, inexploité. Aucune méthodologie de consolidation et de traitement de cette donnée non numérique n'est proposés dans la littérature. La méthodologie proposée tente de répondre à ce manque, de créer du consensus sur la base d'études où les données sont largement dispersée, de sources hétérogènes, voire contradictoires. Une bancarisation massive de la donnée issue des études menées en sciences environnementales permettrait d'établir des schémas partagés et robustes. Ce stage est une première proposition en ce sens.

Au niveau des réseaux bayésiens, les modèles graphiques apportent un nouvel éclairage dans l'identification des filiations entre les molécules en traitement d'eau. Le choix sur le type d'algorithmes à utiliser pour l'apprentissage des graphes exclu pour l'instant les modèles basés sur des tests sur les données sélectionnées. En effet, la base de données est dans sa phase d'enrichissement, le choix sur les algorithmes pourrait évoluer. Ces premiers résultats jettent également un regard sur le choix de score à utiliser. En effet, le BDe donne un bon compromis entre la structure du graphe et son pouvoir interprétatif. Néanmoins, il reste à tester de nombreuses autres techniques, en particulier parmi les méthodes d'apprentissage de structure sur des données mixtes et étendre ces dernières aux bases d'exemples incomplètes. Il a été également envisagé d'intégrer les propriétés topologiques des molécules dans la phase d'enrichissement de la base de données. Les méthodes d'apprentissage seront ainsi déployées sur une base de données intégrant l'ensemble des procédés de transformations. Les premiers résultats obtenus sur les deux procédés analysés simultanément sont encourageants, mais nécessitent une validation par les experts du domaine. En ce qui concerne l'initialisation, la prochaine étape consiste en l'utilisation d'un graphe d'interaction entre les molécules prédéfinie par l'analyse de la littérature en utilisant le graphe obtenu de la Figure 6. La perspective (la plus ambitieuse) à moyen terme réside dans l'automatisation des tâches d'enrichissement de la base de données. Souvent, les sources d'information fournies dans la littérature sont multiples et hétérogènes (tableaux, Figures, textes). Dans ces conditions, une identification des molécules et de leurs propriétés, serait possible en élargissant notre champ d'investigation en fouille de graphes. La collecte automatique de ces données pourrait harmoniser et structurer les schémas de dégradation d'un produit en traitement d'eau.

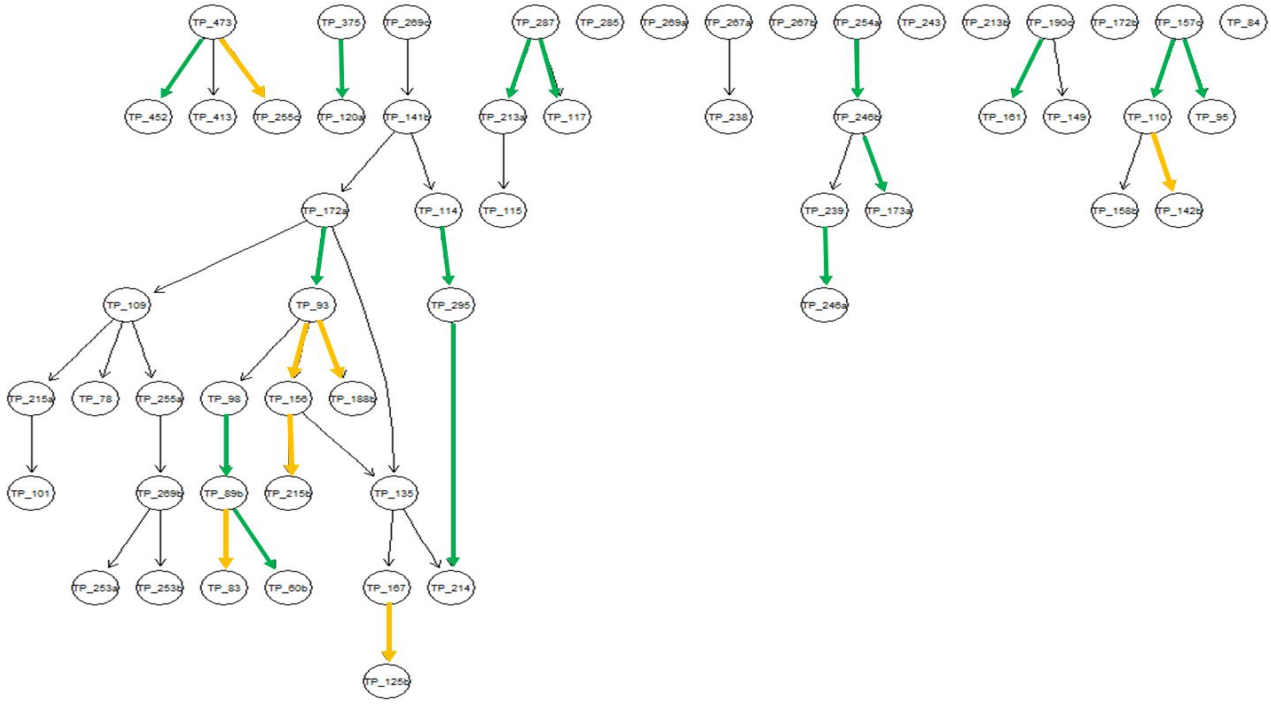


Figure 8: Comparaison du réseau bayésien avec l'algorithme Hill-climbing (score BIC) avec les filiations obtenues par le réseau de connaissance issu de la littérature. Les arcs verts représentent les liens détectés par les RB et qui coïncident avec la littérature des procédés BIO. Les arcs en orange détectent la relation entre les molécules, mais pas le sens de la filiation.

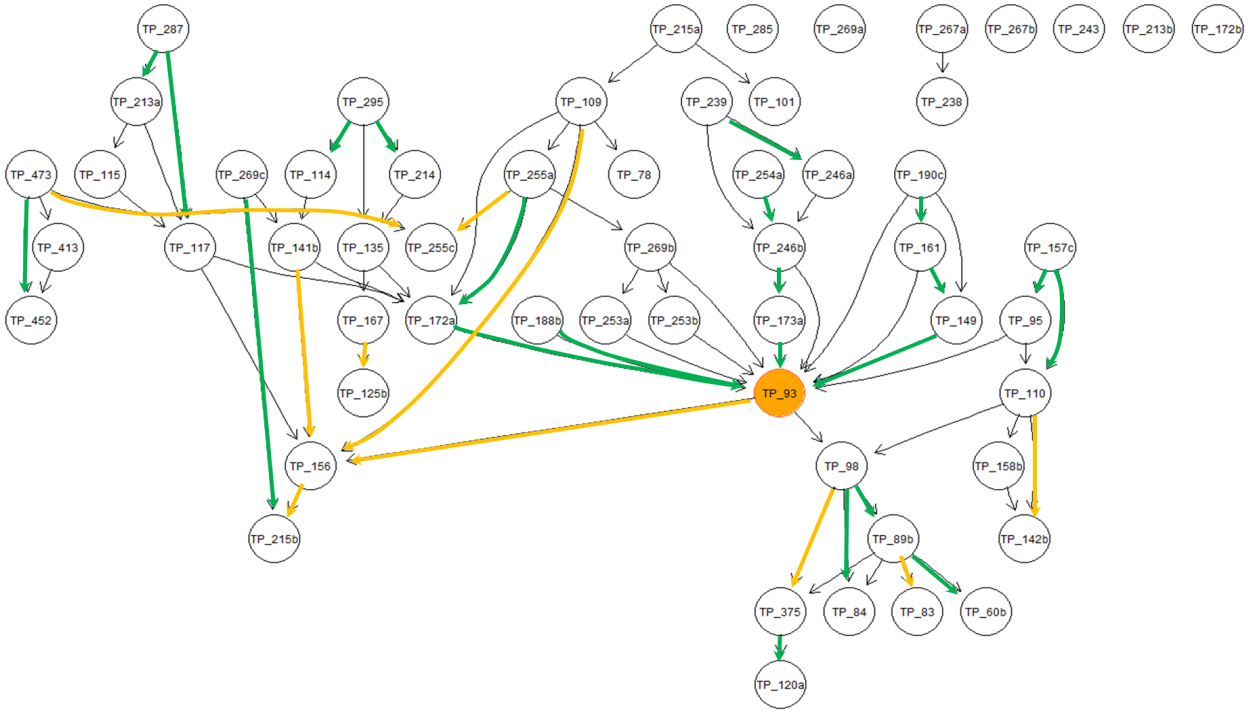


Figure 9: Comparaison du réseau bayésien (algorithme Hill-climbing avec score BDe avec les filiations obtenues par le réseau de connaissance issu de la littérature. Les arcs verts représentent les liens détectés par les RB et qui coïncident avec la littérature des procédés BIO. Les arcs en orange détectent la relation entre les molécules, mais pas le sens de la filiation.

References

Bastian, M.; Heymann, S.; and Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Chen, M.; Guo, C.; Hou, S.; Wu, L.; Lv, J.; Hu, C.; Zhang, Y.; and Xu, J. 2019. In-situ fabrication of Ag/Pg-C3N4 composites with enhanced photocatalytic activity for sulfamethoxazole degradation. *Journal of hazardous materials* 366: 219–228.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of machine learning research* 3(Nov): 507–554.

François, O. 2006. *De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations complètes ou incomplètes.* Ph.D. thesis, INSA de Rouen.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20(3): 197–243.

Jacomy, M.; Venturini, T.; Heymann, S.; and Bastian, M. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9(6): e98679.

Lai, L.; Yan, J.; Li, J.; and Lai, B. 2018. Co/Al2O3-EPM as peroxymonosulfate activator for sulfamethoxazole removal: performance, biotoxicity, degradation pathways and mechanism. *Chemical Engineering Journal* 343: 676–688.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1991 edition.

Wang, L.; Liu, Y.; Ma, J.; and Zhao, F. 2016. Rapid degradation of sulphamethoxazole and the further transformation of 3-amino-5-methylisoxazole in a microbial fuel cell. *Water research* 88: 322–328.

Xue, W.; Li, F.; and Zhou, Q. 2019. Degradation mechanisms of sulfamethoxazole and its induction of bacterial community changes and antibiotic resistance genes in a microbial fuel cell. *Bioresource technology* 289: 121632.

Yazdanbakhsh, A.; Eslami, A.; Massoudinejad, M.; and Avazpour, M. 2020. Enhanced degradation of sulfamethoxazole antibiotic from aqueous solution using Mn-WO3/LED photocatalytic process: Kinetic, mechanism, degradation pathway and toxicity reduction. *Chemical Engineering Journal* 380: 122497.

Zhang, S.; Yang, X.-L.; Li, H.; Song, H.-L.; Wang, R.-C.; and Dai, Z.-Q. 2017. Degradation of sulfamethoxazole in bioelectrochemical system with power supplied by constructed wetland-coupled microbial fuel cells. *Bioresource technology* 244: 345–352.