Algorithmes de classification pour l'identification de bactéries en Spectroscopie Raman

VERONIQUE REBUFFEL[1], ISABELLE ESPAGNON[2], JEAN-CHARLES BARITAUX[1], SOPHIE MORALES[1]

[1] CEA, LETI Laboratoire DTBS/LSIV, MINATEC Campus, F-38054 Grenoble, Cedex 9, France.
[2] CEA, LIST Laboratoire DM2I/LADIS, Saclay, F-91191 Gif-sur-Yvette, France.

[1]veronique.rebuffel @cea.fr, [2]isabelle.espagnon@cea.fr, [1]jean-charles.baritaux@cea.fr,[1]sophie.morales@cea.fr

Résumé - L'identification de bactéries par microscopie Raman est une technique fiable lorsqu'elle est appliquée en conditions de laboratoire, où la culture et l'état des bactéries sont bien maîtrisés. Nous nous intéressons ici à un contexte de menace bactériologique pathogène, où l'identification doit être rapide, robuste et portable sur le terrain. Un système expérimental et les algorithmes de classification associés ont été développés. Dans un premier temps, sur des bactéries en condition de culture relativement standard, une classification par SVM sur les spectres Raman a donné des résultats satisfaisants. Nous avons ensuite considéré des conditions plus difficiles, soit des bactéries dans des états métaboliques variables, ainsi que des espèces bactériennes non apprises dans la base. Une classification de type LDA appliquée de façon hiérarchique associée à une distance de Mahalanobis a donné les résultats les plus robustes, fournissant des informations sur l'espèce bactérienne la plus proche et une indication sur la confiance du résultat.

Abstract - Bacterial identification using Raman spectroscopy is a reliable technique when spectra are acquired in laboratory conditions, when bacteria growth and state are well known. In this study we focus on the context of bacterial pathogen threats. Identification has to be fast, robust, and available for field operations. An experimental system and associated classification algorithms have been developed. For relatively standardly cultured bacteria, an identification using a SVM classifier has shown to provide satisfactory results. When considering more difficult conditions, with variable metabolism bacteria state, and even bacterial species not included in the database, other classification algorithms had to be developed. Based on LDA decomposition applied within a hierarchical frame, followed by a Mahalalobis distance, they allow to provide information on the more probable specie and a confidence level.

1 Introduction

La spectroscopie Raman est une technique bien maîtrisée pour l'identification de micro-organismes dans des cultures de laboratoire. Elle permet, dans ces conditions, une identification au niveau de l'espèce sur une bio-masse allant de quelques cellules à des micro-colonies [1]. Elle est très prometteuse pour une application de recherche d'agents pathogènes dans un contexte de menace biologique. Elle doit alors être applicable sur le terrain, et conduire à une identification rapide et robuste.

Le spectre Raman d'une bactérie dépend de la croissance de la bactérie et de son environnement, conditions que l'on peut maitriser en milieu de laboratoire. Mais qu'en est-il en conditions de terrain ? L'état de chaque bactérie peut varier, et de plus il est impossible de les mettre en culture (ce qui permettrait d'uniformiser cet état) pour des raisons de vitesse d'analyse comme de dangerosité.

Certains travaux sur des bactéries cultivées selon des conditions variables (milieu, T°, âge) ont montré que l'on pouvait réaliser une identification à condition d'apprendre ces variations dans la base de données [2]. Des travaux complémentaires ont analysé l'influence de la matrice environnementale.

Dans cette étude nous considérons des bactéries dans des états de terrain quelconques, ainsi que des bactéries non apprises. Peut-on alors réaliser une identification, même avec des performances dégradées ?

Dans la suite nous présentons rapidement le système instrumental utilisé. Nous introduisons les classes considérées et leur hiérarchie. Un premier algorithme de classification basé sur une méthode SVM (Support Vector Machine) est présenté et évalué sur la base de données. Pour des conditions de terrain plus difficiles, une méthode de type LDA (Linear discrimination analysis) a été également développée, pour une analyse alternative ou complémentaire.

2 Instrumentation et protocole d'acquisition

2.1 Système et spectromètre Raman

Le système d'analyse [3] comporte deux modalités d'imagerie (microscopique et sans lentille) associées à un micro-spectromètre Raman confocal. Une goutte de la solution à analyser est introduite dans l'instrument. Les modalités d'imagerie permettent de localiser puis focaliser sur chaque bactérie présente dans l'échantillon et d'en acquérir un spectre élémentaire d'une durée de 10s environ. Les spectres acquis sont alors confrontés à une base de données pour identification (Figure 1).

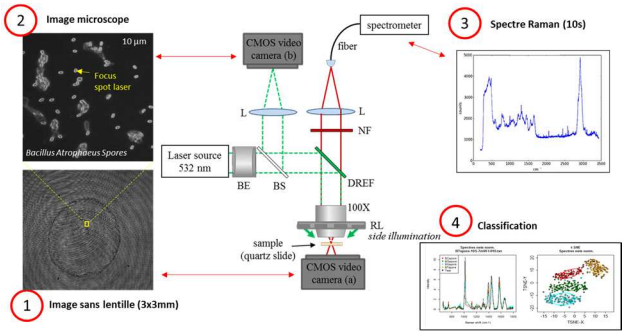

Figure 1 : Dispositif expérimental.

2.2 Les classes considérées

La base retenue est représentative des bactéries pathogènes recherchées dans un contexte de menace biologique, sans être exactement celles-ci pour des raisons de facilité. Elle comporte 4 familles et 12 espèces (Figure 2). La classification visée est au minimum la famille, si possible l'espèce. Certaines espèces sont peuvent être présentes sous deux formes : sporulée ou végétative. Enfin, il est important de noter que l'on peut disposer de plusieurs souches par espèces.

Famille	Espèce	Initiales
Bacillaceae	Bacillus cereus	BC
	Bacillus atrophaeus	BG
	Bacillus subtilis	BS
	Bacillus thuringiensis	BT
Burkholderiaceae	Burkholderia cepacia	BkC
Francisellaceae	Francisella philomiragia	FP
	Francisella tularensis	FT
Enterobacteriaceae	Shigella boydii	SB
	Shigella flexneri	SF
	Shigella sonnei	SS
	Yersinia enterocolitica	YE
	Yersinia pseudotuberculosis	YPs

2 formes possibles : Spore ou Végétative

Figure 2 : Les classes en deux niveaux : familles et espèces.

2.3 Conditions de terrain

Le prélèvement à analyser, typiquement de 1 µl, comporte de quelques unes à quelques dizaines de bactéries. L'utilisateur (le primo-intervenant sur le terrain) souhaite une réponse la plus rapide possible, idéalement une première réponse dès l'analyse de quelques spectres, réponse qui pourra être confortée par l'acquisition et le traitement des spectres suivants.

Dans un prélèvement en milieu de terrain, d'autres objets non biologiques peuvent être présents (poussières, débris,..) qui, même s'ils sont réduits au minimum par un protocole de collecte et préparation [4], devront être gérer par l'algorithme.

Mais surtout, dans un tel contexte, on ne peut maîtriser l'état des bactéries. La spectrosocopie Raman est une méthode phénotypique qui est sensible à l'état métabolique de la bactérie, résulant par exemple de son stress nutritif ou thermique. L'instant où les bactréries supposées pathogènes ont été introduites dans l'nvironnement n'est évidemment pas connu.

Enfin, nous avons également considéré le cas d'espèces bactériennes non apprises dans la base, mais typiques du milieu environnemental.

3 Prétraitement des spectres et classification dans la base de données

3.1 Prétraitement

Les spectres acquis (bruts) sont tout d'abord prétraités. Successivement, on effectue les opérations :

- Lissage par un algorithme de Savitsky-Golay,
- Soustraction de la ligne de base par un algorithme de Clayton,
- Regroupement des canaux, selon traitement
- Sélection de régions d'intérêt (ROI)

Un exemple est donné en Figure 3.

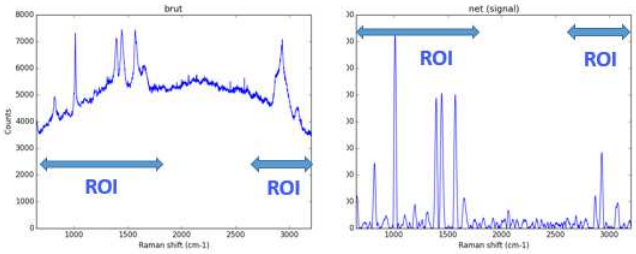

Figure 3 : Spectre initial brut (à gauche) et prétraité (à droite).

3.2 Classification basée sur SVM

Pour constituer la base de données, tous les spectres sont acquis en conditions de laboratoire, mais de façon non excessivement strictes. Un ensemble de spectres des classes présentées en §2.2 a été appris. Afin de classifier une bactérie correspondant à un ensemble de nouveaux spectres, on applique l'algorithme suivant, sur chaque spectre après prétraitement :

- la première étape est basée sur une PLS (« Partial Least Square ») à deux classes *Spore* et *Végétative*, suivie du calcul d'une distance de Mahalanobis, qui est comparée à un seuil. Des tests sur quelques pics caractéristiques sont également effectués. Cette étape permet une première partition en : spore / végétative / bactérie non exploitable (signal trop faible) / non biologique. Dans les 2 derniers cas l'analyse est terminée.

- une seconde étape réalise une SVM (classes 2 à 2) en utilisant soit la base « spore », soit la base « végétative ». Pour les spores la classification est faite au niveau de l'espèce, pour les végétatives au niveau de la famille et de l'espèce, indépendamment.

On dispose alors d'un résultat par spectre, ou par ensemble de spectre (typiquement 10 à 20) que l'on obtient à partir d'un vote majoritaire. L'implémentation est faite en langage R.

3.3 Résultats : matrice de confusion

Les résultats sont présentés sous forme d'une matrice de confusion (Figure 4). La décision est prise ici sur 10 spectres. La base de référence contient 2000 spectres pour les spores et 3000 pour les formes végétatives.

Pour les spores le taux moyen de bonne identification est de 96,5%. Pour les végétatives il est de 94,8% au niveau famille, 78,7% au niveau espèce. Ces performances, si elles étaient obtenues sur le terrain, seraient très satisfaisantes [5].

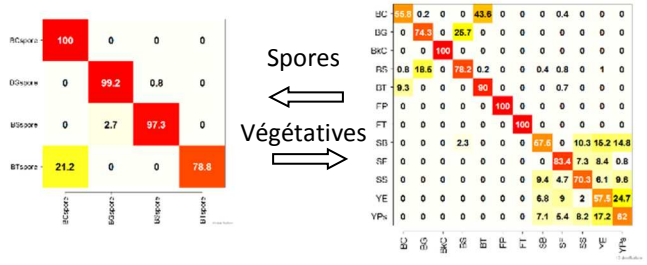

Figure 4 : Matrices de confusion

4 Classification en conditions de terrain

4.1 Limitation de la classification par SVM

Les résultats obtenus se sont révélés nettement moins bons dans les conditions suivantes :

- lorsqu'on ajoute des souches. En effet une souche peut être « plus proche » d'une souche d'une autre espèce que de celles de son espèce. Il peut en résulter des incohérences entre les classifications par famille et par espèce si elles sont conduites indépendamment.
- lors de l'analyse d'une bactérie non apprise dans la base. Dans ce cas l'algorithme SVM n'est pas pertinent, car il conclut nécessairement à la présence d'une espèce de la base.
- enfin, et c'est un cas très fréquent, lorsque les bactéries contenues dans l'échantillon sont dans un état métabolique éloigné de celui de l'apprentissage.

D'autre part, en sortie de SVM, nous ne disposons pas d'une « probabilité » ou d'un score satisfaisant pour l'utilisateur.

4.2 Classification par LDA

Nous avons donc développé un second algorithme de classification. Il est basé sur une LDA appliquée à différents niveaux. Il procède d'une part sur le niveau famille, et le niveau espèce de deux façons, indépendante et hiérarchique. Un niveau intermédiaire regroupant des espèces "confondables" est également effectué. Les différentes conclusions partielles sont confrontées entre elles, et les incohérences trop fortes conduisent à un rejet de décision. Des distances de Mahanalobis, également estimées à différents niveaux, permettent de gérer la confiance selon ces niveaux. Ces distances sont comparées à l'histogramme de leurs valeurs pour les espèces de la base. L'algorithme final fournit diverses sorties possibles, représentées dans la figure 5.

L'analyse par SVM introduite plus haut peut également effectuée en parallèle. Ceci dans un but de comparaison, mais aussi de combinaison éventuelle. Des tests sont en cours pour en démontrer l'intérêt éventuel.

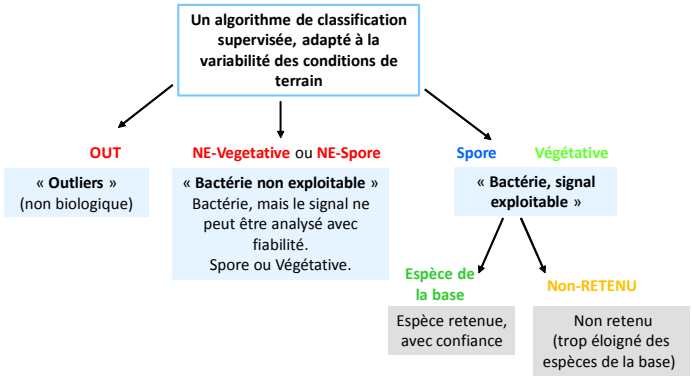

Figure 5 : Sorties de l'algorithme d'identification

La classification présentée ci-dessus s'applique sur un spectre. En fait, on dispose d'un ensemble de spectres. L'analyse est alors effectuée sur chaque spectre, et les conclusions combinées pour l'instant par une règle très simple. Une décision plus élaborée dépend du contexte (par exemple hypothèse d'une seule espèce pathogène présente) et doit être élaborée avec les opérationnels utilisateurs du système.

4.3 Résultats

Présenter un ou quelques cas expérimentaux est nécessairement réducteur. Il suffit d'un écart de quelques degrés de température pour que les conclusions risquent d'être à réviser. De plus, il est difficile, voir impossible, de connaitre la réalité donc d'évaluer, en absolu, les performances de bonnes détections et fausses alarmes. Nous présentons deux cas pour néanmoins illustrer notre méthode et donnons quelques tendances générales dans les conclusions.

Germination de Bacillus atrophaeus (BG)

Des BG après culture sont placées dans un milieu légèrement nutritif (eau peptonée + NaCl) pendant 6h à 30°. Le milieu est donc "propre", contenant peu de bactéries autres que celles introduites, mais celles-ci passent d'un état sporulé à végétatif. Le système (Figure 6) détecte parfaitement le changement d'état. Pendant une période où l'état est intermédiaire, il conclut souvent en "spectres non exploitables".

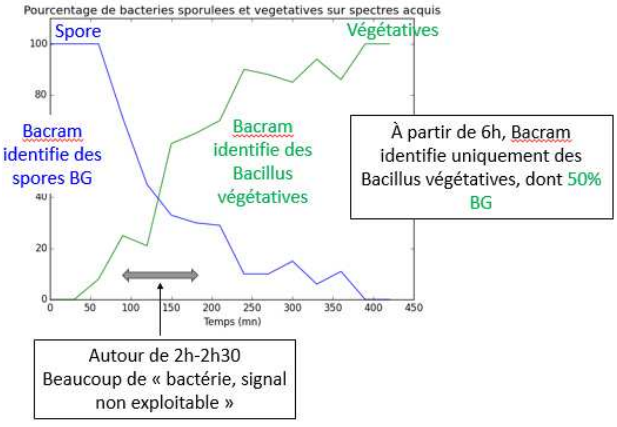

Figure 6 : Germination de Bacillus atrophaeus en milieu propre

Milieu environnemental complexe et BT

Pour ce test on a prélevé l'eau d'un étang péri-urbain donc contenant des débris, algues, cristaux, et des bactéries naturelles inconnues. On y a ajouté des bactéries "environnementales" courantes préalablement cultivées (Pseudomonas fluorescens, Pseudomonas lurida, Pseudomonas rhizosphaerae, Pseudomonas jessenii, Cellulosimicrobium cellulans, Arthrobacter oxydans). Puis enfin des bactéries simulant des pathogènes, à savoir Bacillus thuriengensis ATCC-10792. Le tout a été mis à vieillir et des prélèvements effectués sur plusieurs jours.

Comme le montre la Figure 7, le système trouve au moins 55% de spectres exploitables, qui peuvent donc être confrontés à la seconde étape d'identification.

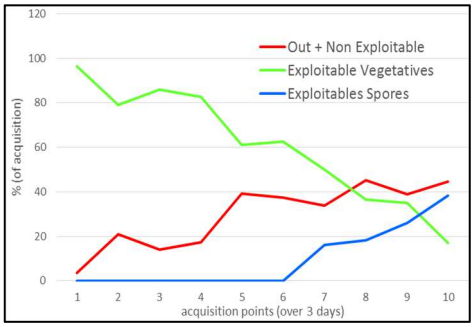

Figure 7 : première étape de l'identification. Milieu complexe et Bacillus thuriengensis.

La seconde étape donne environ 40% de bactéries environnementales (donc non pathogènes), et 20% de "non retenues" (voir les sorties de l'algorithme, Figure 5). Le reste est correctement identifié en BT, et le taux de fausses alarmes quasiment nul.

5 Conclusions

Nous avons donc développé un algorithme permettant d'appréhender des conditions d'analyse différentes des conditions d'apprentissage. Il est clair qu'il est impossible d'apprendre toutes les variations possibles des bactéries en milieu extérieur.

Sur les essais réalisés, l'algorithme s'est montré robuste. Pour un certain nombre de spectres, la décision est "non exploitable" (par exemple bactérie abîmée ?) ou "non retenu" (bactérie trop éloignée de toutes les espèces de la base). Ceci est bien sûr préférable à des fausses alarmes. Le système analysant un ensemble de spectres, une décision peut être fournie à l'utilisateur à partir d'un certain nombre de spectres, nombre dépendant du milieu considéré, plus ou moins "propre".

Enfin, des améliorations pourront être apportées, tout d'abord au niveau de l'identification par utilisation de l'image microscope et donc de l'aspect morphologique de la bactérie. Des réflexions sont également en cours au niveau de l'instrument, aussi bien l'aspect optique que la préparation de l'échantillon – en particulier la concentration permet de définir la sensibilité de l'instrument.

Remerciements

Les auteurs remercient le programme interministériel de R&D NRBC-E pour son soutien au travers du projet BACRAM. Les auteurs tiennent également à remercier M-A. Roncato et L.Bellanger du CEA, DFR/ IBITECS/LI2D pour leur aide à l'acquisition des données expérimentales.
prep

Références

[1] W.E.Huang et al., "Raman microscopic analysis of single microbial cells," Analytical chemistry 76(15), 4452–4458 (2004).

[2] J.C. Baritaux et al., "Fast and robust identification of single bacteria in environmental matrices by Raman spectroscopy", Proc. SPIE BIOS, San Francisco, Feb.2015.

[3] S.A.Strola et al., "Single bacteria identification by Raman spectroscopy" ,Journal of biomedical optics 19(11), (2014).

[4] J.M.Roux et al., « ARCHIBIO: Chaîne analytique de terrain pour la détection de menaces biologiques », Journées scientifiques NRBCE, Saclay, janvier 2017.

[5] V.Rebuffel et al., "A single-cell Raman Spectroscopy device dedicated to fast and robust pathogen threat detection during in the field interventions", Proc. of 2nd CBRNE, Lyon, May 2017

.