Snake Search : méthode de recherche d'empreintes

C. Reynès and **R. Sabatier**

EA 2415 - UFR Pharmacie
15 avenue Charles Flahault, 34093 Montpellier
creynes@univ-montp1.fr - sabatier@univ-montp1.fr

Mots clefs : reconnaissance automatique de signal, alignement de profils, interface graphique.

Dans les démarches *qualité* actuelles, la reconnaissance automatique de substances à partir de leur analyse chimique est un problème important. Pour cela, il est nécessaire d'une part, d'extraire l'information caractéristique de la substance à contrôler, et d'autre part, de pouvoir la reconnaître dans un nouvel échantillon. Ce travail est donc scindé en deux parties : une phase de modélisation du signal et d'apprentissage puis une phase de reconnaissance automatique. La méthode sera appliquée ici à la reconnaissance du venin d'une espèce donnée de serpent.

Pré-traitement

Le jeu de données analysé est constitué de 15 échantillons provenant de 3 lots. Chaque échantillon a été analysé plusieurs fois par chromatographie en phase liquide. On obtient 79 profils. La première étape consiste à pré-traiter les données (cf. Figure). Un algorithme classique de soustraction de la ligne de base a été tout d'abord été utilisé (fonction *baseline* de la librairie *baseline*, méthode *fillPeaks* [1]). Les pics (et épaulements) ont ensuite été identifiés pour chaque profil par la méthode mise au point dans [2]. Elle consiste à effectuer un lissage du signal, puis à calculer la dérivée première de ce signal lissé. Enfin, par un système de fenêtrage, les pics sont successivement identifiés par une annulation de la dérivée associée à un maximum comprise entre deux annulations de la dérivées associées à des minima. On obtient alors une liste de positions des pics (et épaulements) pour chacun des 79 profils.
Une fois les pics identifiés, il est nécessaire de rechercher les correspondances entre profils. Les profils étant assez semblables à l'intérieur d'un lot et plus différents entre lots, l'alignement est réalisé intra-lot dans un premier temps puis inter-lots. Pour cela, trois pics de références sont sélectionnés puis retrouvés dans chacun des profils à l'aide d'une fonction graphique interactive (sous R) nécessitant l'intervention de l'utilisateur. Une fois les pics retrouvés dans tous les profils, un polynôme de degré 3 est ajusté sur les abscisses afin de les aligner.

Sélection de l'empreinte

Après l'alignement, il est possible de savoir quels sont les pics présents dans la majorité des profils. Ces pics seront la base de l'empreinte caractéristique de l'espèce. Afin de choisir uniquement les pics les plus constants donc les plus fiables, nous avons sélectionné ceux qui étaient présents dans au moins 90% des profils étudiés. Un profil moyen est alors calculé pour l'ensemble de ces pics, leur union constituera l'empreinte (cf. Figure).

Test de reconnaissance d'un nouvel échantillon

Dans tous les profils et dans l'empreinte, on repère un pic systématique et isolé dont on identifie la fenêtre dans laquelle se situe l'abscisse originale. Ce pic constituera le point d'ancrage de

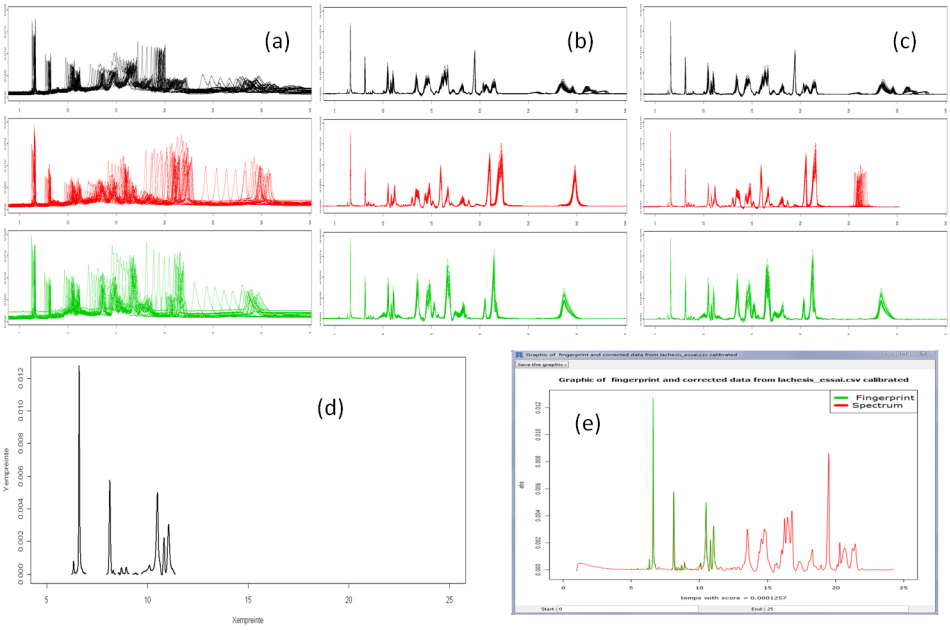

FIGURE 1 – Illustration de la méthode Snake Search. (a) Profils non traités (les 3 graphes correspondent aux 3 lots) (b) Profils après alignement par lot (c) Profils après alignement global (d) Empreinte sélectionnée (e) Sortie graphique pour un nouveau profil.

l'alignement avec l'empreinte. Dans notre exemple, on choisit le pic ayant une abscisse comprise entre 6 et 7.
Lorsque l'on souhaite aligner un nouveau profil, on commence par rechercher la position du maximum dans la fenêtre précédemment identifiée. Ensuite, on cherche à optimiser les paramètres d'une transformation polynomiale permettant d'aligner ce nouveau profil sur l'empreinte. Pour cela, on utilise une grille et on choisit les paramètres minimisant la somme des carrés des écarts mesurés entre les dérivées premières de l'empreinte et du profil à tester. On obtient alors un score que l'on assortit d'une *p-value* calculée grâce à la distribution empirique observée sur l'échantillon d'apprentissage. Cette *p-value* permet de décider de la conformité du nouveau profil à l'empreinte définie.

Conclusion

La partie *reconnaissance* de la méthode proposée a été implémentée dans une interface graphique (cf. Figure) permettant à un utilisateur non spécialiste de R de réaliser automatiquement l'alignement d'un nouveau profil sur l'empreinte précédemment apprise. Il visualise les différentes étapes de l'alignement avec l'empreinte ainsi que le score et la *p-value*. La méthode s'est avérée très performante pour notre jeu de données et devrait très rapidement être généralisées à plus d'une espèce. Nous souhaitons également incorporer le module d'apprentissage d'empreintes dans l'interface graphique.

Références

[1] Liland, K.H. & Mevik, B.-H. (2014). baseline : Baseline Correction of Spectra, *R package version 1.1-3*, http ://CRAN.R-project.org/package=baseline
[2] Reynès, C., Sabatier, R., Molinari, N. & Lehmann, S. (2008). A new genetic algorithm in proteomics : Feature selection for SELDI-TOF data. *Computational Statistics & Data Analysis*, **52**(9), 4380-4394