ondalys
Donnez du sens à vos données

Etude scientifique

La Chimiométrie et le Machine Learning pour l'analyse des données de chromatographie

Résumé

De façon classique, l'analyse des chromatogrammes consiste à intégrer de façon plus ou moins automatisée les aires des pics chromatographiques identifiés à l'aide du détecteur couplé (UV, spectrométrie de masse, …). Il s'agit donc d'une approche dite ciblée, lorsque l'on sait ce que l'on cherche.

Les méthodes de Chimiométrie et de Machine Learning sont de plus en plus utilisées dans de nombreux domaines d'application en approche non ciblée pour analyser différents types de données sous forme d'empreintes. Elles sont notamment couramment employées dans le cadre de l'analyse multivariée de données spectroscopiques, pour lesquelles chaque variable est associée à une même information dans tous les échantillons.

Cependant, leur application aux données chromatographiques est complexe du fait des décalages de temps de rétention d'un même pic entre les échantillons. Ondalys propose donc dans cette étude un comparatif de plusieurs méthodes d'alignement de chromatogrammes afin d'identifier leurs avantages et inconvénients.

De plus, des pistes d'optimisation de l'alignement pour des cas plus complexes, avec des décalages de pics variables au cours du temps ou sans chromatogramme de référence, sont proposées.

En fin d'article, quelques détails scientifiques sont donnés sur les algorithmes testés DTW (Dynamic Time Warping), Icoshift (Interval Correlation Optimized Shifting) et COW (Correlation Optimized Warping).

ONDALYS S.A.S.
S.A.S. au capital de 7 500 €
RCS Montpellier 447 624 024

Siège Social
4, rue Georges Besse
34830 CLAPIERS - France

Agence Auvergne Rhône Alpes
Rond-point de l'échangeur - Les Levées
69360 SOLAIZE - France

+33 (0)4 67 67 97 87
contact@ondalys.fr
www.ondalys.fr

1 Introduction

L'approche classique pour analyser des chromatogrammes consiste à intégrer de façon plus ou moins automatisée les aires des pics chromatographiques identifiés à l'aide du détecteur couplé (UV, spectrométrie de masse, …). Cependant, cette approche ciblée n'est pas optimale pour comparer des échantillons biologiques complexes contenant de nombreux pics qui ne peuvent pas tous être identifiés. Dans ce cas, une approche multivariée plus avancée de type « fingerprinting », non ciblée, prenant en compte la totalité du signal, est plus adaptée.

La principale difficulté pour l'analyse multivariée de chromatogrammes vient des décalages de temps de rétention qui sont fréquemment observés entre les échantillons. Ces décalages peuvent être causés par de multiples facteurs, par exemple des variations dans les conditions d'analyse, des effets de matrice, une dérive du détecteur ou le vieillissement de la colonne chromatographique. Ces différences entre les signaux ne sont donc pas liées à des variations de composition des échantillons, et peuvent perturber la détection de variations d'intérêt. [1]

En effet, les méthodes d'analyse multivariée supposent que chaque variable est associée à la même information dans tous les échantillons. Ainsi, avant de pouvoir les appliquer à des chromatogrammes, il est nécessaire d'aligner les temps de rétention afin qu'un même pic corresponde au même temps entre les différents échantillons. Dans cette étude, nous nous intéressons au cas particulier de l'alignement de chromatogrammes sans pic de standard interne et sans identification des pics par spectrométrie de masse.

2 Base de données

Les données utilisées pour cette étude sont des signaux de chromatogrammes complexes avec de nombreux pics. Elles ont été obtenues sur des échantillons biologiques par HPLC couplée à un détecteur UV ou MS. Dans le cas du détecteur MS, on utilise comme signal le BPC (Base Peak Chromatogram) ou le TIC (Total Ion Current), afin de proposer une méthode d'alignement générique ne nécessitant pas l'identification des pics.

Avant l'alignement, plusieurs prétraitements des chromatogrammes peuvent être nécessaires :

- Correction de ligne de base, réalisée ici par les algorithmes WLS (Weighted Least Squares) ou ASLS (ASymmetric Least Squares), avec une optimisation des paramètres afin de conserver la forme des pics et ne pas créer d'artefacts.
- Binning pour diminuer le nombre de variables sans perdre d'information, et ainsi réduire le temps de calcul des algorithmes d'alignement.

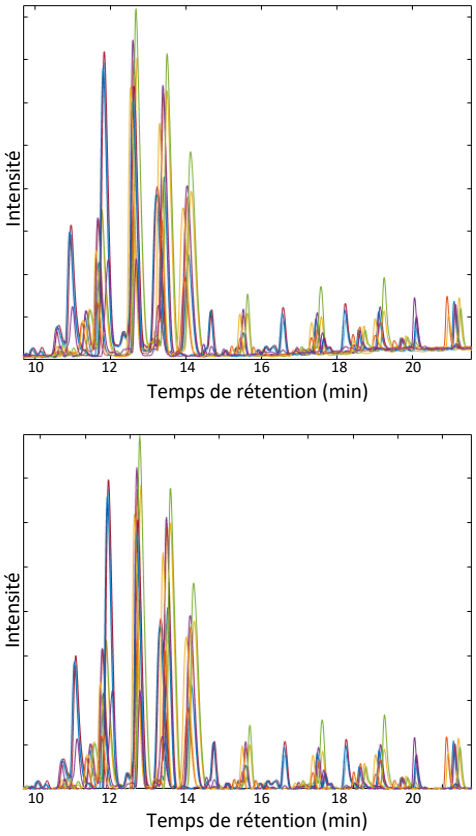

Figure 1 : Exemple de chromatogrammes avant (en haut) et après (en bas) correction de ligne de base et binning

3 Données avec échantillon de référence

3.1 Méthodes testées

Différentes méthodes sont comparées pour aligner les chromatogrammes de plusieurs sessions d'analyse avec le chromatogramme d'un échantillon de référence :

- DTW (Dynamic Time Warping).
- Icoshift (Interval Correlation Optimized Shifting), avec deux étapes d'alignement utilisant des intervalles de tailles différentes.

- COW (Correlation Optimized Warping), avec une optimisation de la taille de fenêtre (segment) et du degré de déplacement maximal (slack) pour l'ensemble du signal.
- COW avec un pré-alignement et une optimisation par zones.

La qualité de l'alignement est évaluée par la valeur de la corrélation moyenne entre les chromatogrammes et la référence. De plus, une inspection visuelle permet de vérifier la bonne conservation de la forme des pics ou au contraire la présence d'artefacts après alignement. La valeur du « peak factor » peut aussi être utilisée comme indicateur de déformation des pics. [2]

3.2 Résultats

3.2.1 Alignement DTW

DTW applique des élongations et compressions à l'axe des temps de rétention de chaque échantillon afin de les aligner avec la référence.

Cette méthode donne de bons résultats en termes de coefficients de corrélation. En effet, les temps de rétention des pics sont bien alignés, comme le montre la Figure 2.

Cependant, DTW ne semble pas adapté dans cette étude pour traiter les chromatogrammes car les pics sont fortement déformés. Des artefacts peuvent être ajoutés, ou au contraire des pics absents de la référence peuvent être supprimés dans les chromatogrammes des échantillons alignés. Ces modifications ne sont pas compatibles avec l'application de méthodes d'analyse multivariée après alignement.

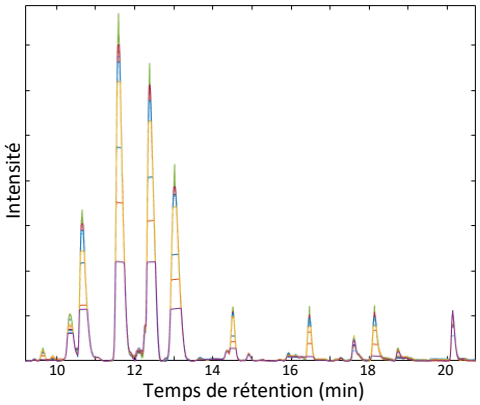

Figure 2 : Exemple de chromatogrammes après alignement par DTW

3.2.2 Alignement Icoshift

Icoshift repose sur le décalage par morceaux de l'axe des temps de rétention et nécessite donc de segmenter les chromatogrammes en intervalles, de façon automatique ou manuelle. La segmentation manuelle est recommandée afin d'optimiser les intervalles pour diminuer le risque de création d'artefacts.

Cette méthode est rapide et donne des coefficients de corrélation globalement satisfaisants. Mais la Figure 3 montre que tous les pics ne sont pas aussi bien alignés, même après deux applications successives avec des tailles d'intervalles de plus en plus fines.

La forme des pics est mieux conservée qu'avec DTW mais quelques artefacts sont tout de même créés, en particulier avec les intervalles plus réduits, lorsqu'un pic se retrouve partagé entre deux intervalles. Icoshift peut donc être utile pour l'alignement de chromatogrammes, mais nécessite une expertise pour la sélection manuelle des limites des intervalles. De plus, cette sélection peut être complexe lorsque de nombreux pics resserrés sont présents.

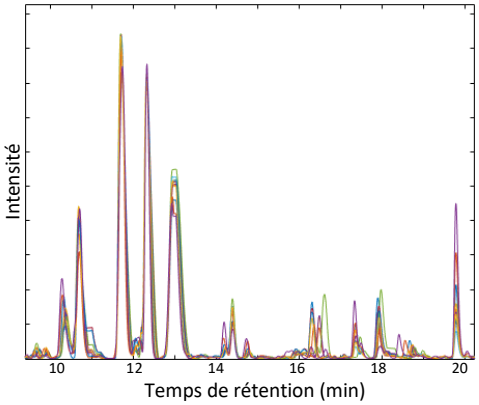

Figure 3 : Exemple de chromatogrammes après alignement par Icoshift

3.2.3 Alignement COW

COW fonctionne par élongations et compressions locales de l'axe des temps de rétention, ce qui nécessite d'optimiser à la fois la taille des segments et le décalage maximal autorisé (« slack »). Cette optimisation peut être automatisée, mais les temps de calculs sont longs. Un binning préalable des chromatogrammes est donc particulièrement recommandé.

Cette méthode est la mieux adaptée aux chromatogrammes dans cette étude, permettant un bon compromis entre l'amélioration des coefficients de corrélation par rapport à la référence et la conservation de la forme des pics. Les temps de rétention sont bien alignés et aucun artefact n'est visible, comme le montre la Figure 4.

En revanche, le temps de calcul est plus long avec cet algorithme, surtout pour l'optimisation, et les pics supplémentaires ou manquants par rapport à la référence peuvent perturber localement l'alignement. Il a aussi été constaté que les valeurs élevées de « slack » augmentent le temps de calcul et le risque de déformation des pics.

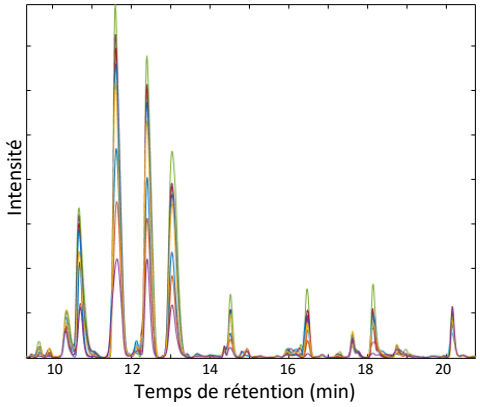

Figure 4 : Exemple de chromatogrammes après alignement par COW

Le Tableau 1 résume les résultats obtenus avec les trois algorithmes d'alignement. Dans ce cas, l'algorithme COW nous semble être le mieux adapté pour l'alignement de chromatogrammes.

Tableau 1 : Comparaison des performances des 3 algorithmes d'alignement

	Corrélation moyenne à la référence	Peak factor
Avant alignement	0.58	1.00
Après DTW	0.98	0.79
Après Icoshift	0.95	0.95
Après COW	0.98	0.96

3.2.4 COW optimisé par zones

Dans des cas plus complexes, avec des amplitudes de décalages de pics variables au cours du temps, l'optimisation des paramètres de segment et slack peut s'avérer difficile. La sélection d'une seule valeur de ces paramètres pour l'ensemble des temps de rétention ne permet alors pas d'aligner correctement tous les pics, et le risque de créer des déformations augmente.

Si les chromatogrammes présentent des régions identifiables de bruit, sans pic significatif, un découpage manuel en plusieurs zones peut être utile. Ceci permet d'optimiser les paramètres de segment et slack indépendamment pour chaque zone et ainsi améliorer les performances de l'alignement par COW.

De plus, en l'absence de pic de standard interne dans les chromatogrammes, l'ajout d'un pré-alignement par simple recalage du dernier pic significatif de chaque zone avant l'alignement par COW réduit les décalages et facilite l'optimisation des paramètres.

Dans l'exemple présenté ici, la présence d'un pic supplémentaire dans l'un des chromatogrammes ne perturbe pas l'alignement des autres pics.

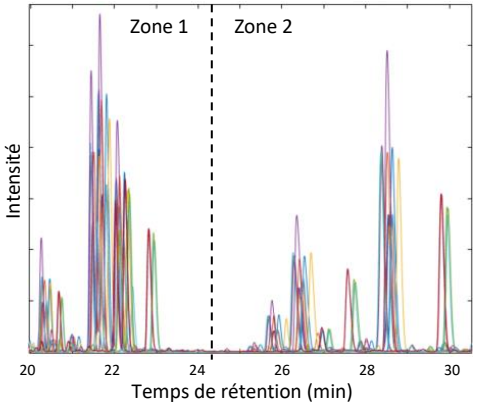

Figure 5 : Exemple de chromatogrammes avec décalage croissant (après correction de ligne de base et binning)

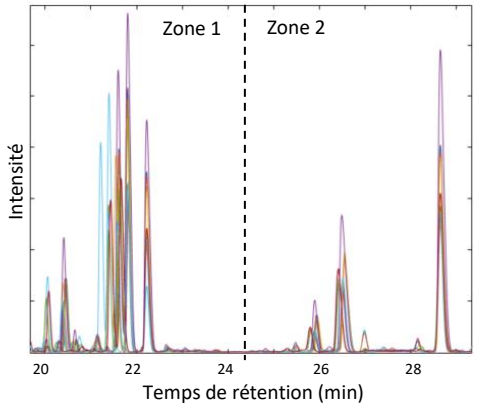

Figure 6 : Exemple de chromatogrammes avec décalage croissant, après pré-alignement par zones

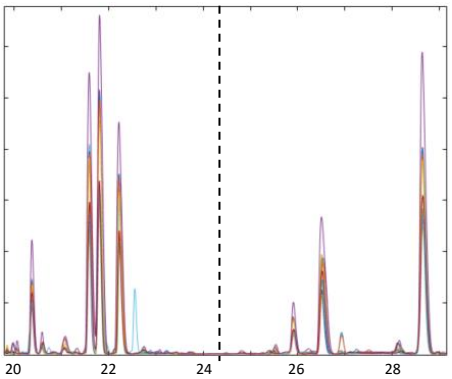

Figure 7 : Exemple de chromatogrammes avec décalage croissant, après pré-alignement et alignement COW par zones

Tableau 2 : Comparaison des performances de COW simple ou par zones

	Corrélation moyenne à la référence	Peak factor
Avant alignement	0.34	1.00
Après COW simple	0.92	0.95
Après pré-alignement et COW par zone	0.95	0.97

Tableau 3 : Comparaison des performances des différentes itérations de COW

		Corrélation moyenne à la référence	Peak factor
Avant alignement		0.82	1.00
Après COW par groupe	It.1	0.89	0.92
	It.2	0.90	0.94
	It.3	0.91	0.95
	It.4	0.91	0.95

4 Données sans échantillon de référence

4.1 Méthodes testées

COW ayant été identifié précédemment comme la méthode la plus efficace, dans cette partie seule cette méthode est utilisée afin de tester son adaptation à une problématique d'alignement sans chromatogramme de référence disponible.

Dans ce cas, un signal moyen peut être calculé à partir des différents chromatogrammes à aligner et cette moyenne est ensuite utilisée comme référence pour l'alignement. Le signal moyen est ensuite recalculé après un premier alignement afin d'affiner ce profil de référence et d'améliorer de façon itérative l'optimisation de COW.

La qualité de l'alignement est ici évaluée par la valeur de la corrélation moyenne entre les chromatogrammes et le signal moyen. Comme précédemment, une inspection visuelle et le calcul du « peak factor » permettent de vérifier la conservation de la forme des pics et l'absence d'artefacts après alignement.

4.2 Résultats

Les chromatogrammes des échantillons sont alignés en prenant le signal moyen comme référence. Cet alignement est affiné en plusieurs itérations : à chaque itération, la moyenne utilisée pour l'itération suivante est recalculée sur les données alignées, mais l'alignement se fait ensuite sur les chromatogrammes de départ pour ne pas cumuler les déformations de pics. L'optimisation s'arrête lorsque l'amélioration entre deux itérations devient minime, dans ce cas après 3 itérations.

5 Conclusion

Cette étude démontre la faisabilité de l'alignement de chromatogrammes, sans pic de standard interne ni identification des pics par spectrométrie de masse. Différentes méthodes d'alignement sont comparées (DTW, Icoshift et COW), avec un résumé donné dans le Tableau 4. L'algorithme COW s'avère être le plus efficace sur les signaux traités dans le cadre de cette étude, et des optimisations sont proposées afin d'améliorer l'alignement dans certains cas spécifiques (décalages non constants, absence de chromatogramme de référence).

Tableau 4 : Comparatif des trois méthodes d'alignement

	DTW	Icoshift	COW
Paramètres à optimiser	Aucun pour DTW basique	Intervalles	Segment et slack
Intervention manuelle	Aucune pour DTW basique	Choix des limites des intervalles	Choix de zones pour optimisation plus avancée
Temps de calcul	Rapide	Rapide	Plus long à optimiser
Intégrité du signal	Déformation de pics Suppression de pics	Déformation de pics (limitée si choix manuel des intervalles)	Bonne conservation des pics
Performance d'alignement	Très bon alignement	Bon alignement	Très bon alignement

6 Les méthodes d'alignement

6.1 DTW

L'algorithme de dynamic time warping repose sur la mesure de similarités de séries temporelles et sur les principes de programmation dynamique afin de minimiser la distance entre deux signaux X (de longueur L_X) et Y (de longueur L_Y) de façon non linéaire. [3] Il a été initialement développé pour l'alignement de spectres sonores (« speech recognition »). [4] L'algorithme DTW calcule la matrice des distances euclidiennes pour chaque paire de valeurs de X et Y, formant une grille de dimensions (L_X, L_Y).

$$d(i,j) = \sqrt{(X_i - Y_j)^2}$$

i parcourant chaque indice de temps de rétention du signal X (i = 1, …, L_X) et j chaque indice de temps de rétention du signal Y (j = 1, …, L_Y)

Ensuite, il cherche le chemin passant par les plus courtes distances, en prenant par exemple le signal X comme référence :

$$\{c(1), \dots, c(i), \dots, c(L_X)\}, avec\ c(i) = [i, j(i)]$$

i étant l'indice de temps de rétention du signal X et j(i) l'indice de temps de rétention aligné du signal Y

Trois principales contraintes encadrent cet alignement :

- Contrainte de limites : le premier et le dernier point du chemin doivent être les premier et dernier points de X et Y.

$$c(1) = [1,1]\ et\ c(L_X) = [L_X, L_Y]$$

- Contrainte de monotonie : l'ordre des points doit être conservé.

$$j(i+1) \geq j(i)$$

- Contrainte de taille de pas : le décalage de position de chaque point est limité.

La construction du chemin d'alignement commence par le dernier point (L_X, L_Y), et à chaque étape le point suivant est choisi parmi les trois plus proches voisins du point actuel jusqu'à atteindre le premier point (1, 1). Ainsi, le chemin optimal est construit itérativement en minimisant à chaque point la distance cumulée :

$$D(i,j) = min \begin{cases} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + d(i,j) \\ D(i,j-1) + d(i,j) \end{cases}$$

6.2 Icoshift

L'algorithme d'interval correlation shifting n'utilise pas la programmation dynamique, mais repose sur le déplacement de segments par insertion ou suppression de points. Il a été d'abord proposé pour l'alignement de spectres RMN. [5] Icoshift cherche à maximiser la corrélation croisée entre les segments de X et Y, en prenant X comme référence.

$$C_Y^X(u) = \int_{-\infty}^{+\infty} Y(t+u)\, X(t)\, dt$$

u étant la valeur du décalage à optimiser pour maximiser C.

Ce calcul fait intervenir la fonction transformée de Fourier ($\mathcal{F}$), en passant d'une représentation en temps (t) à une représentation en fréquence (f).

$$x(f) = \mathcal{F}\big(X(t)\big) = \int_{-\infty}^{+\infty} X(t)\, e^{2\pi i f t}\, dt$$

$$X(t) = \mathcal{F}^{-1}\big(x(f)\big) = \int_{-\infty}^{+\infty} x(f)\, e^{-2\pi i f t}\, df$$

Ainsi, $C_Y^X(u)$ peut être exprimée en fonction de $\mathcal{F}^{-1}(x(f)y(f))$, ce qui permet de calculer les corrélations croisées pour de grandes valeurs de u et d'optimiser les valeurs de décalage de tous les segments simultanément. Ceci réduit fortement les temps de calcul avec cet algorithme.

La définition manuelle des segments est recommandée pour limiter les risques d'artefact et de déformation de pics.

6.3 COW

Comme DTW, l'algorithme de correlation optimized warping repose sur le principe de programmation dynamique mais utilise cette fois l'extension ou la compression linéaire des signaux. [3] Il a été développé spécifiquement afin de limiter les déformations lors de l'alignement de chromatogrammes. [6] Le critère d'optimisation de l'alignement par COW est la corrélation entre les signaux X et Y, qui doit être maximisée. De plus, COW aligne les signaux par segments et nécessite donc le choix de deux paramètres :

- La longueur des segments.
- Le nombre maximal de points (« slack ») sur lequel chaque segment peut être étiré ou compressé.

Les signaux X, utilisés comme référence, et Y sont donc divisés en S segments de longueur L_S, et deux matrices F et U de dimensions (S+1, L_X+1) sont initialisées. Le dernier élément en position (S+1,

L_X+1) est fixé à 0 car le dernier point de X doit être aligné avec le dernier point de Y. Lors de l'optimisation de l'alignement, la matrice F est remplie avec les valeurs de la fonction d'optimisation, c'est-à-dire le coefficient de corrélation entre les segments correspondants de X (s_X) et de Y alignés ($s_{Y'}$) :

$$\frac{(s_X - \bar{s}_X)^T (s_{Y'} - \bar{s}_{Y'})}{std(s_X)\, std(s_{Y'})}$$

avec $\bar{s}$ la moyenne et std(s) l'écart-type de chaque segment.

De plus, les positions possibles des limites de chaque segment sont calculées selon le paramètre de slack t choisi :

$$J_{début} = 1 + \max \begin{cases} (i-1)(L_s - t) \\ L_X - (S - i + 1)(L_s + t) \end{cases}$$

$$J_{fin} = 1 + \min \begin{cases} (i-1)(L_s + t) \\ L_X - (S - i + 1)(L_s - t) \end{cases}$$

i parcourant chaque indice de segment de la matrice F (i = 1, …, S)

Les coefficients de corrélation sont calculés pour chaque position possible et avec tous les nombres de points de déplacement possibles entre -t et +t. En parallèle, la matrice U est remplie avec les valeurs de déplacement des points (« warping ») permettant d'obtenir la meilleure corrélation pour chaque position. Le chemin optimal peut donc être construit en remontant les valeurs de la matrice U, correspondant aux meilleures positions identifiées dans la matrice F.

6.4 Choix de la référence

Quel que soit l'algorithme utilisé, l'alignement cherche à rapprocher chaque chromatogramme d'un signal de référence. Le choix du signal de référence influencera donc l'optimisation et le résultat de l'alignement.

Cette référence peut être :

- Un chromatogramme explicitement identifié comme une référence pour cette analyse.
- Un chromatogramme choisi parmi les différents échantillons analysés, par exemple le premier de la session ou celui qui présente un décalage « moyen » par rapport à tous les autres.
- Une moyenne des différents chromatogrammes à aligner, qui peut être calculée après un premier alignement pour limiter la présence de pics trop larges ou déformés dans ce signal moyen.

7 Implémentation logicielle

Les algorithmes DTW et Icoshift utilisés pour cette étude sont disponibles dans des toolbox publiques développées par l'Université de Copenhague en langages MATLAB® (The Mathworks, USA) et Python® (Python Software Foundation, USA).

L'algorithme COW est disponible en langage MATLAB® dans une toolbox publique développée par l'Université de Copenhague, ainsi que dans la PLS_Toolbox® (EigenVector Research Inc., USA) et dans la version stand-alone SOLO® (EigenVector Research Inc., USA).

Ondalys a développé des scripts d'amélioration de COW dans les langages MATLAB® et Python®.

8 Références

[1] G. Malmquist and R. Danielsson, "Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods," *Journal of Chromatography A,* vol. 687, pp. 71-88, 1994.

[2] T. Skov, F. van den Berg, G. Tomasi et R. Bro, «Automated alignment of chromatographic data.,» *Journal of Chemometrics: A Journal of the Chemometrics Society,* vol. 20, pp. 484-497, 2006.

[3] V. Pravdova, B. Walczak et D. Massart, «A comparison of two algorithms for warping of analytical signals,» *Analytical Chimica Acta,* vol. 456, pp. 77-92, 2002.

[4] H. Sakoe et S. Chiba, «Dynamic Programming Algorithm Optimization for Spoken Word Recognition,» *IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING,* vol. 26, pp. 43-49, 1978.

[5] F. Savorani, G. Tomasi et S. Engelsen, «icoshift: A versatile tool for the rapid alignment of 1D NMR spectra,» *Journal of Magnetic Resonance,* vol. 202, pp. 190-202, 2010.

[6] N.-P. V. Nielsen, J. M. Carstensen et J. Smedsgaard, «Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping,» *Journal of Chromatography A,* vol. 805, pp. 17-35, 1998.