Université des Sciences et Technologies de Lille

THESE

pour obtenir le grade de
Docteur de l'Université des Sciences et Technologies de Lille
Mention : Instrumentation et analyses avancées

Présentée par
Cyril Ruckebusch

Spectroscopie infrarouge et chimiométrie pour l'instrumentation en chimie analytique des procédés
Application au suivi de l'hydrolyse d'hémoglobine bovine

Soutenance publique le 19 décembre 2000 devant le jury :
Mr Pierre Legrand, Professeur, Directeur de l'EUDIL, Villeneuve d'Ascq

Mr Désiré-Luc Massart, Professeur, Vrije Universiteit Brussel, Bruxelles
Mr Nguyen Quy Dao, Directeur de Recherche, Ecole Centrale de Paris, Châtenay-Malabry

Mme Mireille Bayart, Professeur, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq
Mr Daniel Bougeard, Directeur de Recherche, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq
Mr Didier Guillochon, Professeur, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq
Mr Jean-Pierre Huvenne, Professeur, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq

A Séverine, à mes parents et amis.

A la mémoire de ma mère.

Remerciements

Ce travail, financé par le Ministère de l'éducation, de l'enseignement supérieur et de la recherche, a été effectué au Laboratoire de Spectrochimie Infrarouge et Raman.

Mes plus vifs remerciements vont bien entendu à Monsieur le Professeur Huvenne qui m'a accueilli dans l'équipe caractérisation moléculaire et chimiométrie. Je lui suis reconnaissant de la confiance qu'il me porte.

J'insiste sur l'honneur que me font Messieurs Massart, Professeur à l'Université Libre de Bruxelles et Nguyen Qui Dao, directeur de recherche à l'Ecole Centrale de Paris, d'accepter de juger ce travail.

Je remercie également, Monsieur le Professeur Legrand, Directeur de l'EUDIL, qui préside ce jury.

Je n'oublie pas les rapporteurs, Madame le Professeur Bayart, du LAIL, Monsieur Bougeard, directeur du LASIR, ainsi que Monsieur le Professeur Guillochon responsable du LTSN.

Enfin, la sympathie et les compétences de tous les membres du laboratoire, et notamment de mes plus proches collaborateurs, ont fait de ce travail une occupation quotidienne agréable et motivante.

Sommaire

Introduction

Dans le domaine de l'analyse des procédés, les modèles d'étalonnage multivarié, et plus globalement les méthodes de chimiométrie, offrent des alternatives intéressantes pour la détermination indirecte de propriétés à partir d'un certain nombre d'observations. A condition d'être généralisables à des données nouvelles, ces modèles permettent le contrôle quantitatif des systèmes.

L'objectif global de ce travail est la mise au point, à l'échelle du laboratoire, d'un outil permettant d'estimer, si possible en ligne, la progression d'une hydrolyse d'hémoglobine bovine. Il s'agit d'une application complexe pour laquelle la modélisation empirique peut apporter une information concernant l'avancement de la réaction, et donc le contrôle des produits. L'analyste s'affranchit alors des mesures physico-chimiques classiques, coûteuses et longues à mettre en œuvre, qui condamnent tout espoir de contrôle réactif du réacteur biologique.

La spectrométrie infrarouge à transformée de Fourier convient pour l'analyse in situ de molécules aussi compliquées que les protéines. En effet, les informations spectrales concernent les vibrations des liaisons inter-atomiques et sont, à ce titre, aussi précieuses que complètes. Néanmoins, leur complexité exclut toute interprétation directe. Ainsi, au carrefour de plusieurs disciplines, le travail notamment consiste à :

- mettre au point l'observation spectrale qui doit être de la meilleure qualité possible.
- choisir un type de modèle, par l'analyse de la structure des données. La présence de relations non-linéaires, de classes ou encore l'absence de relations mathématiquement établies prônent l'utilisation de méthodes neuronales.
- estimer les paramètres du modèle de telle sorte que les causes et les effets soient pris en compte, à partir d'exemples imposés ou choisis. L'optimisation du modèle doit permettre d'identifier l'information pertinente. D'un point de vue mathématique, il s'agit d'un problème à la fois multivarié, continu, non-linéaire et non contraint.
- établir la précision de la prédiction, notamment lors du contrôle de répétitions inconnues du procédé, pour juger de l'adéquation du modèle aux spécifications.
- améliorer la compréhension du procédé par l'étude, la manipulation ou l'observation des modèles.

La présentation du manuscrit propose, dans une première partie, une synthèse des concepts de la chimie analytique des procédés, entre autres : la spectrométrie infrarouge et les réseaux de neurones artificiels. Nos intérêts sont néanmoins du point de vue de l'application. Les résultats

expérimentaux concernent donc spécifiquement les développements pour le contrôle d'un procédé d'hydrolyse d'hémoglobine bovine. Néanmoins, puisque le succès d'une modélisation par réseaux de neurones dépend de facteurs expérimentaux, des données, de leur pré-traitement et de la topologie, nous détaillerons les heuristiques nécessaires pour l'amélioration de l'apprentissage. Par ailleurs, nous insisterons notamment sur l'emploi d'outils de visualisation qui permettent une analyse exploratoire des données.

Concepts &

étude bibliographique

Introduction

L'émergence de la chimie analytique des procédés répond à de nouvelles exigences de compétitivité, de productivité mais aussi à la volonté de contrôle et d'analyse des systèmes. Elle se distingue principalement de la chimie analytique traditionnelle par deux aspects. Le premier concerne la localisation des analyseurs qui, pour être opérationnels, sont placés directement sur les systèmes. L'autre point de vue est lié à l'instrumentation qui joue un rôle supplémentaire en chimie analytique des procédés. En effet, elle doit fournir un résultat qui permet d'ajuster le procédé ou tout au moins d'agir sur celui-ci.
Bien que ce soit le procédé qui détermine l'analyse, de nombreuses avancées technologiques ont facilité la création d'outils pour le contrôle de systèmes. La première approche consiste à étudier les possibilités de diverses méthodes analytiques susceptibles de fournir des résultats avant d'envisager des techniques plus créatives. D'un point de vue instrumental, la chimie analytique constitue une approche globale, intégrée. Sur le plan humain, elle requiert un travail en équipe et des connaissances pluridisciplinaires. Les efforts du chimiste, de l'analyste, de l'ingénieur sont associés pour produire une solution. Aspect contradictoire, un tel investissement n'est la plupart du temps récompensé que par une solution dédiée ; un capteur est un outil spécialisé.

Les philosophies des domaines de la chimie analytique des procédés, de l'instrumentation ou de l'automatisme se partagent le concept de mesure. Pour l'analyste, mesurer consiste à déterminer directement ou non une valeur analytique expérimentale alors que pour l'automaticien, il s'agit de la fonction d'un instrument, d'un capteur. La métrologie réunit un ensemble de moyens logiciels et matériels qui concourent à la production d'informations fiables. Une image la plus fidèle possible de la grandeur observée doit être obtenue et des traitements numériques des données brutes doivent fournir des informations validées voire évoluées. Ces différentes notions font partie intégrante du concept de modélisation, elles seront développées dans le premier chapitre.
Les évolutions en chimie analytique des procédés suivent naturellement celles des domaines qui la composent. On cite facilement l'informatique ou la microélectronique mais ce ne sont pas les seules. C'est ainsi que pour les techniques spectroscopiques, la tendance est au remplacement des instruments de laboratoire par des outils plus spécifiques. Les données sont

la matière première de la modélisation. Nous les étudierons au chapitre 2. En outre, la spectroscopie de vibration comme outil analytique est possible puisqu'un groupement chimique produit des caractéristiques spectrales propres.

Enfin, l'échantillonnage des procédés est incontestablement entré dans l'ère de l'information. Les flots d'informations brutes constituent un formidable réservoir de connaissances à exploiter, à valoriser. Ainsi, les méthodes de chimiométrie permettent la détermination empirique des corrélations entre les paramètres du procédé et la performance au sens général du terme. Entre autres qualités, les systèmes d'analyse multidimensionnelle assurent le passage d'une information classique à un résultat permettant des prises de décisions. Lorsque les mécanismes réactionnels sont inconnus, complexes, lorsqu'ils se déroulent dans des environnements difficiles et/ou changeants, la modélisation du comportement d'un procédé nécessite d'utiliser des techniques robustes. Celles basées sur les réseaux de neurones artificiels, dont les fondements seront détaillés au chapitre 3, permettent la construction de systèmes élaborés de calcul.

Chapitre 1

Contrôle des procédés

Pour répondre à des besoins divers, les systèmes de production sont de plus en plus automatisés et la convergence des évolutions dans plusieurs domaines technologiques a conduit à l'apparition d'instruments intelligents. Leurs rôles et leurs responsabilités sont accrus par rapport aux capteurs classiques. Les services rendus par les capteurs intelligents concernent principalement les activités d'exploitation des procédés.

Le chimiométricien tire parti des informations provenant aussi bien de la communauté des statisticiens que de celle des chercheurs de la chimie analytique. Il profite donc à la fois de l'étude de l'aspect dynamique des variables et de l'analyse multivariée.
Les observations génèrent des mesures qui peuvent être utilisées pour contrôler le comportement d'un procédé. Un premier aspect consiste à utiliser des cartes de contrôle pour apprécier la stabilité des procédés. Cela permet de suivre les paramètres statistiques de la caractéristique mesurée. Néanmoins, la tendance est à l'utilisation de techniques de modélisation qui facilitent l'étude de tous les systèmes, y compris des systèmes dynamiques et discontinus.
La représentation d'un procédé prend en général la forme d'un modèle empirique basé sur un historique des données. Le comportement d'un procédé est modélisé en utilisant des données obtenues lorsque celui-ci est sous contrôle. Les évènements futurs sont estimés par comparaison avec le modèle.

1.1 Instrumentation

La première fonction remplie par l'instrumentation est la mesure. Celle-ci fournit des informations sur le système physico-chimique. A partir de ces informations, des stratégies de surveillance ou de gestion sont éventuellement élaborées. En conséquence, l'analyste considère le capteur comme un objet finalisé dont la fonction est de délivrer ou plutôt d'élaborer une mesure. Un autre point de vue, celui de l'automaticien, valorise plutôt le capteur en tant que constituant d'un système automatisé de production, ce qui lui confère des propriétés de décision vis-à-vis de la manipulation d'actionneurs.[1]

1.1.1 La mesure

La connaissance analytique d'un échantillon procure des informations sur un objet ou un procédé. Il est donc primordial de pouvoir accéder à la valeur d'une grandeur. L'élaboration d'une grandeur significative met en jeu un ensemble d'opérations que regroupe la fonction "mesurer".[2] Les grandeurs observées permettent ainsi d'estimer un phénomène. La grandeur physique, objet de l'observation, est appelée mesurande (notée m) et la fonction du capteur est donc de traduire un ensemble de mesurandes en une information (notée s) utilisable.

La fonction "mesurer" consiste à utiliser des méthodes, spectroscopiques, physiques ou chimiques, combinées à des traitements (la chimiométrie par exemple) pour l'analyse d'un système. On distingue deux cas de figure, la mesure pouvant être directe ou indirecte. Le premier cas concerne les analyseurs fournissant une image de la grandeur physique ou chimique d'intérêt. Lorsqu'il n'existe pas d'instrument permettant d'accéder directement à cette grandeur, il convient d'engendrer celle-ci à partir de la grandeur observée, on parle alors de mesure indirecte ou élaborée.

En outre, il faut disposer d'une certaine connaissance du procédé ou du problème analytique pour juger de la pertinence d'effectuer une mesure. Notons qu'un contrôle efficace nécessite une bonne qualité des mesures et que la simplicité de la méthode employée est souvent un critère important.

---

[1] M. Staroswiecki, M. Bayart, *Actionneurs intelligents*, Hermès (1994)
[2] M. Bayart, *Journées d'études SAPID*, SAPID (1995)

1.1.2 Capteurs

Les capteurs sont les instruments dont la fonction est de délivrer une mesure. Ils élaborent des images du procédé de production qui doivent représenter au mieux celui-ci. Ils sont principalement composés de deux constituants.

- Le transducteur est la partie la plus sensible au phénomène physique ; il module une grandeur relative à celui-ci.
- Le conditionneur est l'élément de mise en forme de la grandeur produite par le transducteur.

Le modèle le plus simple décrit les mécanismes de transformation du mesurande en une image directe de celui-ci, il explicite donc une relation du type f (s, m) = 0 en reprenant les notations définies au paragraphe précédent.

Comme tout système instrumental, un capteur est caractérisé par ses performances et ses limites de fonctionnement[3] : son utilisation est soumise à des conditions particulières, les grandeurs obtenues sont liées aux conditions environnementales, elles-mêmes caractérisées par des grandeurs d'influence. Finalement, l'efficacité d'un capteur n'est garantie que dans un certain domaine de validité.

D'autre part, une amélioration possible pour les capteurs est la découverte de nouveaux transducteurs, permettant d'optimiser les mesures pour une application donnée. Néanmoins, c'est surtout la réalisation d'instruments capables de fournir à l'opérateur, de manière indirecte, des grandeurs qui lui étaient auparavant inaccessibles, qui focalise les activités de recherche.

1.1.3 Capteurs virtuels

Certains capteurs clés sont trop complexes ou trop coûteux pour être utilisés en ligne. Il est alors intéressant, à partir d'une campagne de mesures, de générer des capteurs virtuels.

Les capteurs virtuels[4] constituent une opportunité pour deux raisons :

- ils permettent de pallier le manque de transducteur, soit parce que l'effet physico-chimique

[3] A. L. Gehin, *Thèse de doctorat*, Université des Sciences et Technologies de Lille (1994)
[4] CIAME, *Les capteurs intelligents – Réflexions des utilisateurs*, AFCET (1987)

n'est pas techniquement mesurable, soit parce qu'il s'agit d'un état difficile à exploiter : le goût, l'odeur ou l'avancement d'un procédé.

- ils rendent possible la production de grandeurs élaborées, telles que des historiques de preuves procurant des aides à la détermination ou à la connaissance de l'état d'un système d'exploitation.

Les capteurs virtuels sont des outils fournissant à l'opérateur des fonctions et des paramètres cruciaux non accessibles auparavant à partir des grandeurs mesurées. En ce sens, ils participent à l'amélioration de la fonction "mesurer". Ce sont ces aspects "intelligents"[4] que développent analystes et statisticiens. Ils modélisent, par l'intermédiaire de méthodes mathématiques de corrélation, les propriétés d'un procédé à partir de grandeurs observées.

Ces mécanismes d'élaboration de grandeurs évoluées dépendent bien entendu du type de l'application à laquelle on destine l'instrument de mesure, en particulier du procédé que l'on veut contrôler et de son niveau du contrôle.

1.2 Procédés et contrôles

Les observations des procédés analytiques peuvent être utilisées pour différentes raisons : la vérification de la qualité d'un produit, le diagnostic d'un système ou le contrôle statistique d'un procédé. Une utilisation de la mesure qui améliore la connaissance fondamentale des mécanismes d'un système est également envisageable. Néanmoins, la fonction la plus utile est sans doute le retour sur le procédé. On entend par-là l'ajustement, en temps réel, des paramètres en entrée du procédé pour obtenir des spécifications particulières en sortie.

1.2.1 Procédés

L'estimation de l'état d'un système est basée sur les résultats fournis par les actions que nous avons établies au paragraphe précédent, et qui définissent la mesure.

Types de procédés

On distingue les procédés discontinus (*batch* en anglais) des procédés continus. Un procédé discontinu est défini comme un système de durée finie dans lequel un réacteur est chargé de « matière » afin de réaliser une recette spécifiée. La réaction se déroule pendant un certain temps au cours duquel les variables sont mesurées. Finalement, les produits résultants seront déchargés. Un procédé continu est au contraire approvisionné en matière première, de telle sorte qu'une production ininterrompue dans le temps soit assurée. Dans la pratique, la distinction est rarement aussi tranchée et de nombreux procédés sont dits *semi-batch*.

Etat d'un procédé

La fonction "mesurer" peut prendre un certain temps. Le résultat analytique ne représente alors l'état actuel du système qu'avec une certaine latence. Autrement dit, le système peut avoir évolué entre l'enregistrement de l'observation et la production d'un résultat utilisable. On parle parfois de temps mort pour désigner la durée séparant l'échantillonnage de l'action de contrôle.

La probabilité de voir l'état du système changer dépend de la cinétique du procédé. Il est donc nécessaire de connaître l'évolution des systèmes et la rapidité d'acquisition des dispositifs de mesure. D'une manière générale, l'incertitude entourant l'état d'un procédé est corrélée à la différence entre l'état réel, inconnu au sens analytique, et l'état estimé.

Qualité de l'estimation

La pensée statistique est très différente de l'esprit phénoménologique et causal de la chimie analytique. Le concept statistique de population est un idéal représentant la collection totale des objets d'une classe particulière. En réalité, seul un sous-ensemble de la population est observé et les propriétés statistiques telles que la moyenne ou la dispersion sont estimées à partir de celui-ci.

La qualité des mesures est en étroite relation avec la qualité des procédés. Pour quantifier la dispersion des mesures, les statisticiens utilisent la variance qui vérifie la propriété d'additivité.[5] Néanmoins, l'écart-type, racine carrée de la variance, possède l'avantage d'être plus facilement interprétable puisqu'il est exprimé dans la même unité que la grandeur mesurée. Idéalement, la population des objets produits doit être centrée autour d'une valeur-cible, assimilable à la valeur vraie, qui caractérise une quantité parfaitement définie dans les conditions de l'observation. Néanmoins, un résultat analytique s'écarte inévitablement de la valeur vraie. On distingue le critère justesse des mesures, représentatif des erreurs systématiques, du critère précision des mesures qui estime les contributions des erreurs aléatoires.

Parmi les autres sources d'erreurs possibles, on peut citer :

- les valeurs aberrantes qu'il est nécessaire, d'un point de vue statistique, de détecter et d'éliminer. Il est d'ailleurs préférable de travailler avec des méthodes robustes c'est-à-dire qui résistent à ces valeurs atypiques. De plus, les méthodes robustes ne nécessitent aucune supposition concernant la distribution de la population.
- les dérives qui indiquent que le procédé n'est pas sous contrôle statistique, moyenne et/ou dispersion des résultats n'étant pas constantes.
- les variations du signal de blanc, sources d'erreurs qui n'affectent pas le procédé, et sont appelées variations de ligne de base.

[5] H. Martens, T.Naes, *Multivariate Calibration*, John Wiley & Sons (1989)

1.2.2 Suivis de procédés

Le type de contrôle recherché, la nature du procédé et la technique d'échantillonnage sont liés. En conséquence, on peut envisager l'échantillonnage d'un processus pour différents objectifs.

• A un instant donné, décrire la composition ou l'état d'un procédé.

L'analyse permet de juger du bon fonctionnement du procédé : une répétition à un instant donné répond-elle aux spécifications ? Si c'est le cas, le procédé est dit *capable*.

• Au cours du temps, contrôler un état.

Le système s'écarte-t-il d'une valeur-cible et risque-t-on de dépasser une limite donnée? Pour obtenir un produit de qualité attendue, les variables observées doivent décrire certaines trajectoires qui représentent leur mode de fonctionnement normal. Des déviations de ces trajectoires produisent de mauvaises prédictions. L'analyseur est donc un indicateur de tendance qui estime la stabilité du procédé.

• Anticiper en dirigeant un état.

Cela consiste à réaliser une série d'actions consécutives dans le but de manufacturer un produit répondant à certaines spécifications. Cette fonction est la motivation du contrôle de procédés.[6] En effet, l'efficacité du procédé n'est conservée que lorsqu'une action en temps réel sur les entrées est possible.

L'aspect discontinu (ou continu) du système est à prendre en considération. Les procédés discontinus nécessitent souvent une instrumentation pouvant supporter des variations importantes des paramètres du procédé au cours d'un essai. Les systèmes continus exigent un intervalle dynamique moins important mais l'étalonnage doit être plus précis.

Procédés stables

L'utilisation de cartes de contrôle, proposée par Shewart[7], est un moyen graphique permettant de déceler les variations systématiques d'une caractéristique que l'on veut contrôler. Ces cartes amènent à constater si le procédé est *capable* ou non.

---

[6] B. M. Wise, N. B. Gallagher, *Journal of Process Control*, 6 (1996) p. 329
[7] W. A. Shewhart, *Economic Control of Quality of Manufactured products*, van Nostrand (1931)

Pour ce faire, des caractéristiques statistiques sont observées. Elles ne doivent pas évoluer de manière perceptible. Ces conditions peuvent parfois être assouplies, il faut alors s'assurer que les limites de tolérance définies sont respectées

Procédés évolutifs

Actuellement, la tendance est à la représentation de la connaissance d'un système par un modèle à partir duquel des répétitions quasi-identiques, mais inconnues du point de vue analytique, seront estimées. Des procédés quasi-identiques sont des systèmes qualitativement semblables mais dont la caractérisation analytique varie légèrement à cause de phénomènes de fluctuations, d'interférences ou d'infimes variations des conditions. C'est souvent le cas pour les procédés discontinus qui ne sont pas, la plupart du temps, exactement répétables. On réalise alors une modélisation de la dynamique du procédé qui fait explicitement partie du système de contrôle.[8]

S'il est représentatif du procédé, ce modèle est ajouté au flux. On peut considérer que les contrôles basés sur des modèles sont alimentés de l'avant. Cela signifie que la méthode identifie la dynamique du procédé et y participe effectivement. C'est principalement pour cette raison que les contrôles basés sur des modèles sont plus efficaces pour les procédés complexes.

[8] D. Seborg, T. Edgar, D. Mellichamp, *Process Dynamic and Control*, Wiley & Sons (1989)

1.3 Modèles de procédés

Un modèle constitue une description mathématique d'un processus réel. La construction d'un modèle est une démarche interdisciplinaire et les objectifs suivis doivent être clairs pour chacun. Il peut s'agir, entre autres, de l'analyse de phénomènes afin d'approfondir leur compréhension, de la détermination des grandeurs pour lesquelles aucun capteur n'est disponible ou de la prédiction du comportement évolutif d'un système. Dans tous les cas, le modèle sera jugé sur la façon dont il permet d'atteindre un but fixé.

1.3.1 Modélisation

Définitions

Un système est constitué de la partie de l'univers que l'expérimentateur a choisi, plus ou moins arbitrairement, d'appréhender comme un tout avec laquelle il interagit. L'action sur le système se traduit par le biais de grandeurs connues qui forment les entrées. De même, certaines grandeurs du système sont observées et les résultats de ces observations forment les sorties.
Un modèle du système est une règle qui permet de calculer des grandeurs d'intérêt. L'analyste effectue ce calcul à partir de grandeurs connues ou mesurées sur le système.

Principe

Construire un modèle revient à estimer un ensemble de paramètres inconnus à partir des données recueillies sur le système et des connaissances disponibles a priori. La qualité d'un modèle est en général évaluée en observant l'erreur de sortie, différence entre la valeur produite par le système et celle fournie par le modèle. Cette erreur est estimée par le biais d'un critère qu'il faut optimiser et dont le choix traduit le but de la modélisation.[9]
Le rôle de l'algorithme d'optimisation est alors de minimiser ce critère de façon à calculer les

[9] E. Walter, L. Pronzato, *Identification de modèles paramétriques*, Masson (1994)

paramètres du modèle. Cette étape délicate nécessite de prendre en considération le niveau d'incertitude sur les mesures ainsi que certaines précautions concernant, entres autres, l'initialisation des paramètres ou l'arrêt des itérations.
Il faut noter qu'il n'existe pas réellement de modèle idéal mais plutôt un ensemble de modèles acceptables. Une analyse critique des résultats est donc toujours nécessaire. En effet, les paramètres du modèle ne sont pas, à eux seuls, suffisants pour le valider. Ainsi, on s'intéresse plutôt aux capacités de prédiction du modèle. Celles-ci sont analysées par la présentation d'entrées nouvelles. De la même manière, il est alors important de s'assurer que le modèle est suffisamment robuste. De petites perturbations aléatoires des entrées ne doivent pas nuire au comportement global.

1.3.2 Caractérisation des modèles

Le choix de la structure d'un modèle définit un ensemble de comportements possibles et un ensemble de valeurs admissibles pour les paramètres. Ce choix ne résulte jamais d'une solution univoque, il implique au contraire des décisions plus ou moins arbitraires aux conséquences importantes.

Modèles phénoménologiques et comportementaux

Les modèles phénoménologiques sont construits sur des lois bien connues de la physique ou de la chimie[10]. Ils se prêtent bien à la prise en compte de l'information connue a priori.
Les modèles comportementaux ne nécessitent, au contraire, aucune connaissance particulière sur le procédé ayant généré des données. Ils reproduisent un comportement observé. Parmi les structures possibles pour ce type de modèle, certains réseaux de neurones présentent la possibilité de modéliser n'importe quel comportement continu, avec une précision arbitraire.[11, 12]

[10] G. E. P. Box, W. G. Hunter, *Technometrics*, 7 (1965) p. 23
[11] A. N. Kolmogorov, *Mathematical Social Transactions*, 28 (1963) p. 55
[12] V. Kurkova, *Neural Networks*, 5 (1992) p. 501

Linéarité

Une structure de modèle est linéaire si ses sorties vérifient le principe de superposition.[9] On distingue deux cas :
- les structures linéaires par rapport aux entrées du système. C'est le point de vue des automaticiens.
- les structures linéaires par rapport aux paramètres du modèle. C'est le point de vue des statisticiens.[13]

Déterminisme

Un modèle est déterministe si les sorties ne sont décrites que comme des fonctions des entrées, et uniquement de la valeur vraie de celles-ci. Cette vision idéale est, dans la plupart des cas, irréaliste. Il convient de décrire l'influence du bruit associé à la mesure ou des perturbations agissant sur le système.

1.3.3 Applications

Le contrôle peut être effectué par le biais de cartes, les scores et les résidus d'une analyse en composantes principales[5] (PCA dans la littérature anglo-saxonne) permettent de prédire la qualité finale des variables, ainsi que les intervalles de contrôle pour les répétitions futures.[14] Néanmoins, ces méthodes simples ne s'intéressent qu'à la recherche de déviations du comportement par rapport à un état défini.[15] A ce titre, l'analyse en composantes principales multivoies (MPCA) est une des méthodes permettant de prendre en compte les aspects dynamiques.[16] Cela revient à considérer que les données de référence de procédés discontinus évoluent dans le temps. Cette méthode est statistiquement conforme à l'analyse en composantes principales[17], mais elle doit disposer des données complètes de la dynamique

[13] R. I. Jenrich, M. L. Raltson, *Annual Review of Biophysics & Bioengineering*, 8 (1979) p. 195
[14] R. Boqué, A. K. Smilde, *AIChE Journal*, 45 (1999) p. 1504
[15] M. J. Piovoso, K. A. Kosanovich, P. K. Pearson, *IEEE Transactions on Instrumentation & Measurement*, 41 (1992) p. 1361
[16] R. Nomikos, J. Macgregor, *AIChE Journal*, 40 (1994) p. 1361
[17] S. Wold, P. Geladi, K. Esbensen, J. Ohman, *Journal of Chemometrics*, 1 (1987) p. 41

que l'on veut caractériser.[18]

En corrélant les données d'entrée aux valeurs attendues pour un certain nombre d'exemples, un modèle qui ne prend en compte que les variations significatives, c'est à dire intéressantes pour la mesure, des variables du procédé est construit. Une fois validé, ce modèle peut être utilisé pour estimer le comportement de nouvelles répétitions. Ainsi, de nombreuses approches basées sur des modèles expérimentaux ont donc été étudiées ces dernières années pour permettre l'apprentissage des procédés industriels.[19]

Les modèles expérimentaux procurent un regard initial sur la chimie et le mécanisme du processus. En particulier, ils permettent l'identification du nombre de composantes significatives et de leurs contributions respectives au sein du lot de données. Certaines solutions récentes font appel à des techniques dites d'intelligence artificielle.

- Les approches basées sur des connaissances « expert »[20] et les techniques de logique floue qui définissent des fonctions d'apprentissage.

Celles-ci convertissent les variables du procédé en variables symboliques.[21] Ces techniques nécessitent de trouver et d'organiser la connaissance d'un système.

- Les méthodes basées sur des exemples (*self-modeling* en anglais).

Elles ignorent les limitations énoncées ci-dessus. En effet, elles se procurent leur connaissance du système de manière induite.

Les procédés biotechnologiques dépendent fortement des matières premières, des enzymes, des variables environnementales, et nécessitent un contrôle ayant des capacités d'adaptation.[22] En outre, les mécanismes réactionnels sont complexes, souvent inconnus, et ne peuvent être mis en équation. Enfin, la plupart de ces procédés sont discontinus, caractérisés par des variables assez sévèrement corrélées et certaines données peuvent être manquantes, bruitées ou entachées d'erreurs. Cela justifie l'utilisation des méthodes neuronales, sous réserve de conditions appropriées notamment concernant les données d'apprentissage.

Reste que la modélisation ne peut être mise en œuvre qu'une fois les données représentatives du procédé acquises. La technique utilisée pour cette étape est en grande partie déterminée par

[18] S. Chen, T. J. Mc Avoy, *Journal of Process Control*, 8 (1998) p. 409
[19] K. Gokarajv, K. Raju, C. L. Cooney, *AIChE Journal*, 44 (1998) p. 2199
[20] P. N. Penchev, G. N. Anorev, K. Warmuza, *Analytica Chimica Acta*, 388 (1999) p. 145
[21] K. B. Konstantinov, T. Yoshida, *Biotechnology & Bioengineering*, 39 (1992) p. 479
[22] J. Sternby, *IEEE Transactions on Control Systems Technology*, 4 (1996) p. 11

les facilités d'interfaçage. La tendance est à l'utilisation de systèmes ne nécessitant pas d'échantillonnage élaboré, une sonde distante insérée directement dans le flux du procédé à analyser est souvent considérée comme l'interface idéale. Cela donne a priori un avantage de poids aux méthodes spectroscopiques.

Chapitre 2

Les données pour le contrôle

Les analyseurs pour le contrôle nécessitent principalement des qualités de robustesse, de fiabilité, de facilité et de sécurité. Dans ce cadre, les techniques spectroscopiques présentent un potentiel d'application important pour la chimie analytique des procédés. En effet, il s'agit généralement de systèmes de contrôle non destructif qui, de plus, ne requièrent aucun contact avec l'échantillon. Ces techniques procurent des données consistantes, significatives et répétables avec des critères de rapidité et de sûreté satisfaisants.

La spectrométrie infrarouge, notamment, est une méthode analytique très utile et universelle. L'infrarouge analytique regroupe des méthodes d'identifications et de dosages non destructifs d'échantillons liquides, solides ou gazeux, organiques ou inorganiques. L'énergie impliquée provient des radiations électromagnétiques dont les fréquences sont voisines, mais inférieures, aux fréquences du visible. Les données correspondantes, les spectres, fournissent une très grande quantité d'informations[23] exploitables par le physicien, le chimiste, l'organicien ou le biologiste. Dans ce contexte, le rôle de la chimiométrie sera d'exalter les informations de faible niveau, d'élucider les signaux complexes et de déduire des renseignements latents.

L'utilisation de la transformée de Fourier en spectrométrie a constitué la révolution instrumentale la plus importante mais des progrès ont aussi été accomplis dans d'autres domaines. Notre intérêt concerne l'intégration des observations dans un dispositif de capteur intelligent capable de fournir des informations fiables et de tirer parti de la spécificité de la spectrométrie moyen infrarouge.

[23] A. Lee Smith, *Applied Infrared Spectroscopy*, Wiley-interscience (1979)

2.1 Spectroscopie et chimie analytique des procédés

Les premiers analyseurs de procédés étaient en réalité des instruments de laboratoire modifiés pour permettre l'introduction automatique d'échantillons. Aujourd'hui, ils nécessitent parfois des investissements lourds et des modifications instrumentales importantes.[24] Néanmoins, les techniques spectrométriques combinent des propriétés de robustesse et d'adaptation qui permettent souvent d'utiliser, sans engager de modifications importantes, des appareils commerciaux pour le suivi de procédé.

2.1.1 Echantillonnage

Le Tableau 1 présente les différents types d'analyse en fonction des caractéristiques de leur échantillonnage et de la localisation de l'analyseur.[25]

Type d'analyse	Caractéristique(s) de l'échantillonnage	Caractéristique(s) de l'analyse
Off-line	Manuel	En laboratoire
At-line	Manuel	Sur place
On-line	Automatique	Le long de la ligne, automatique
In-line	Automatique	Sur la ligne, automatique

Tableau 1 : Types d'analyse.

2.1.2 Instrumentation

Les techniques spectroscopiques, généralement faciles à interfacer, sont de fait des candidates idéales aux analyses *in-line* ou *on-line*. La Figure 1 représente un tel système d'analyse de

[24] V. Lopez-Avila, H. H. Hill, *Analytical chemistry*, 69 (1997) p. 289R
[25] D. C. Hassell, E. M. Bowman, *Applied Spectroscopy*, 52 (1998) p. 18A

procédé en état de fonctionnement. Une étape préalable consiste à utiliser les éléments **1** et **2** pour construire un modèle. L'absence d'éléments de transport et de conditionnement des échantillons entre les éléments notés **1** et **2** est caractéristique des analyses spectroscopiques.

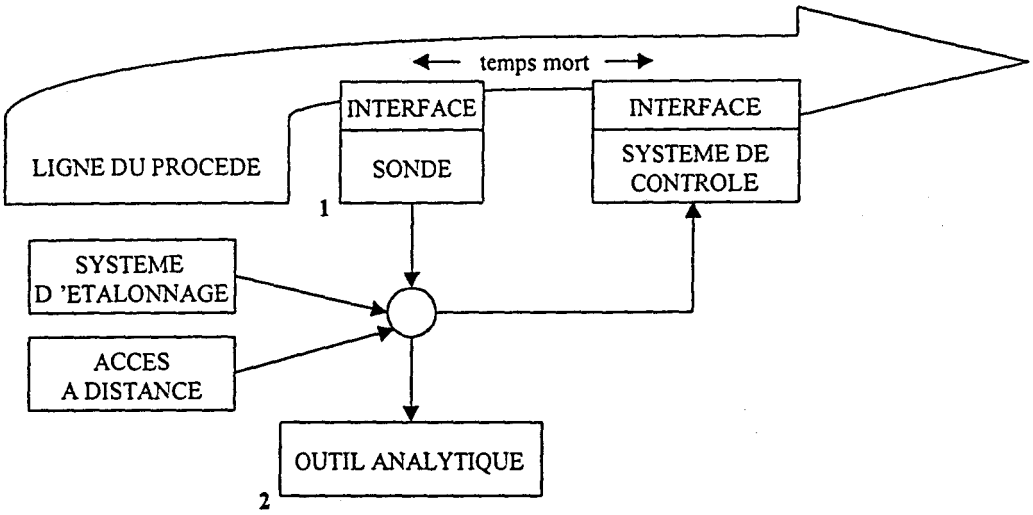


Figure 1 : Système d'analyse de procédé.

Les instruments optiques utilisés sur les procédés sont, soit des photomètres basés sur des filtres interférentiels ou des réseaux, soit des spectromètres à transformée de Fourier.

2.1.3 Spectrométries pour le contrôle

Les applications de la spectroscopie à la chimie analytique des procédés concernent la plupart des régions du spectre électromagnétique.

• L'emploi de la spectrométrie infrarouge à transformée de Fourier (IRTF) permet d'utiliser les spectres de vibration pour détecter des modifications au cours de procédés. Citons à titre d'exemple le suivi de la dégradation de la matière organique d'un compost[26] ou le contrôle de la fermentation de l'éthanol.[27] La haute sensibilité de l'infrarouge moyen est très utile pour les analyses en phase gaz mais a longtemps été un handicap en ce qui concerne les analyses en phase liquide ou solide.

• Les spectres de la région du proche infrarouge sont composés d'harmoniques et de bandes

[26] D. Y. Tseng, R. Vir, S. J. Traina, J. J. Chalmers, *Biotechnology & Bioenginnering*, 52 (1996) p. 661
[27] P. Fayolle, D.Picque, G. Corrieu, *Vibrational Spectroscopy*, 14 (1997) p. 247

de combinaison. Cette région du spectre n'est pas aussi sensible que celle de l'infrarouge moyen. L'un des domaines d'utilisation les plus importants est l'agroalimentaire.[28,29] Des fibres optiques performantes facilitent la réalisation d'analyseurs dédiés au contrôle laitier en cours de traite[30] ou au suivi de polymérisation.[31]

• La spectroscopie Raman utilise des excitations électromagnétiques aux longueurs d'onde plus courtes. Cela induit des situations plus favorables du point de vue instrumental permettant notamment des mesures *in situ* à distance.[32] Parmi les principaux avantages, on peut insister sur le fait que l'eau soit transparente en Raman ce qui s'avère une propriété intéressante pour les analyses biologiques par exemple. Cela fait de la spectroscopie Raman une technique complémentaire de celles présentées précédemment. De nombreuses introductions à l'utilisation en ligne de la technologie Raman existent.[33,34] Cette technologie a été récemment appliquée au contrôle de l'hydrolyse d'un composé organique[35] ou au suivi en ligne d'une biotransformation.[36]

• Les avancées dans le domaine des spectroscopies UV-visible et de la fluorescence[37] reposent sur les développements de sondes et d'optiques. *Eichhorn et al.* ont proposé le suivi d'une fermentation par le biais de la fluorescence d'une protéine.[38]

---

[28] W. W. Blaser, R. A. Bredeweg, R. S. Harner, M. A. Leugers, D. P. Martin, R. J. Pell ; J. Workman, L. G. Wright, *Analytical Chemistry,* 67 (1995) p. 47
[29] I. K. Brookes, B. N. Gedge, S. V. Hammond, *Proceedings of the 7th International Conference on Near Infrared Spectrosopy*, A. M. C. Davies & P. C. Williams (1996)
[30] L. Duponchel, *Thèse de doctorat*, Université des Sciences et Technologies de Lille (1997)
[31] R. Reshadat, S. Desa, S. Joseph, M. Mehra, N. Stoev, S. T. Balke, *Applied Spectroscopy*, 53 (1999) p. 1412
[32] Nguyen Quang Huy, M. Jouan, Nguyen Quy Dao, *Applied Spectroscopy*, 47 (1993) p. 2013
[33] F. Adar, R. Geiger, J. Nooman, *Applied Spectrosopy*, 32 (1997) p. 45
[34] I. R. Lewis, P. R. Griffiths, *Applied Spectroscopy*, 50 (1996) p. 12A
[35] O. Svensson, M. Josefson, F. W. Langkilde, *Chemometrics & Intelligent Laboratory Systems*, 49 (1999) p. 49
[36] A. D. Shaw, N. Kaderbhai, A. Jones, A. M. Woodward, R. Goodacre, J. J. Rowland, D. B. Kell, *Applied Spectroscopy*, 53 (1999) p. 1419
[37] J. M. Andrews, S. H. Lieberman, *Analytica Chimica Acta*, 285 (1994) p. 237
[38] L. R. Eichdorn, C. R. Albano, J. Sipior, W. E. Bentley, G. Rao, *Biotechnology & Bioengineering*, 55 (1997) p. 921

2.2 Spectroscopie de vibration

Toute méthode spectroscopique est basée sur l'interaction de rayonnements électromagnétiques avec la matière, les molécules ou les atomes. Elle doit satisfaire une condition de résonance. La différence d'énergie entre deux états stationnaires doit égaler l'énergie du photon mis en jeu :

$$\text{Équation 1} \qquad \Delta E = h\nu .$$

La spectroscopie infrarouge, tout comme la diffusion Raman, permet principalement l'observation des états énergétiques correspondant aux vibrations des atomes d'une molécule. Les paragraphes suivants résument donc les notions permettant une approche rigoureuse de la spectroscopie de vibration. Nous décrirons succinctement d'une part l'état d'une molécule isolée et, d'autre part, le champ électrique en l'absence de toute molécule. Finalement, nous aborderons les interactions champ-matière.

2.2.1 Vibrations moléculaires

Introduction

L'état d'une molécule peut se décrire à partir des fonctions d'ondes rotationnelles, vibrationnelles et électroniques à condition de négliger les faibles contributions des mouvements de translation. L'approximation dite de Born-Oppenheimer[39] rend possible la séparation des termes de l'Hamiltonien. Compte-tenu de l'ordre de grandeur du rapport des masses de l'électron et du proton et de la séparation des mouvements de vibration et de rotation des noyaux,[40] la fonction d'onde Φ peut raisonnablement s'écrire sous la forme de l'Équation 2.

$$\text{Équation 2} \qquad \Phi = \Phi_{rot}\Phi_{vib}\Phi_{él} \; ; \; E = E_{rot} + E_{vib} + E_{él} .$$

Trouver les fonctions d'onde décrivant les énergies des niveaux de vibration de la molécule

[39] M. Born, R. Oppenheimer, *Annalen der Physik*, 84 (1927) p. 457
[40] H. Eyring, J. Walter, G. E. Kimball, *Quantum Chemistry*, Wiley (1944)

revient à résoudre l'équation de Schrödinger vibrationnelle :

Équation 3 $$H\Phi_{vib}(x,y,z) = E\Phi_{vib}(x,y,z).$$

Aspect vibrationnel de l'équation de Schrödinger

Afin de résoudre intégralement le terme vibrationnel de l'équation de Schrödinger, il faut développer le potentiel au voisinage du minimum d'énergie.

En première approximation, il est courant d'envisager une description harmonique du potentiel. Ce modèle très simple ne permet pas d'envisager qualitativement la fonction d'énergie potentielle pour le problème diatomique. En effet, il est juste que le potentiel doit présenter un minimum à la distance d'équilibre entre les atomes et croître autour de ce point. Néanmoins, comme le montre la Figure 2, la liaison doit être rompue pour les élongations importantes alors qu'aux courtes distances inter-atomiques dominent les forces de Van der Waals. Une étape supplémentaire est donc nécessaire pour décrire cette fonction et le potentiel de Morse peut conduire à une solution approchée. Il est préférable pour mieux considérer l'anharmonicité d'inclure les termes cubiques du développement de Taylor dans la forme du potentiel. Ils sont traités par la méthode des perturbations.[41] Cela revient à considérer que les solutions sont voisines de celles obtenues lors de la résolution exacte, mais harmonique, du problème. L'énergie de vibration tenant compte des contributions anharmoniques est ainsi :

Équation 4 $$E = \left(v + \frac{1}{2}\right)hc\bar{\nu} - \left(v + \frac{1}{2}\right)^2 h\chi\bar{\nu}$$; $\Delta v = \pm 1, \pm 2, \ldots$; χ est appelée constante d'anharmonicité.

Les niveaux d'énergie représentés Figure 3 ne sont plus équidistants, la transition $\Delta v = \pm 1$ caractérisée dans l'approximation harmonique reste la plus observée mais d'autres combinaisons sont possibles et obtenues par intégration des fonctions d'onde vibrationnelles orthogonales.

[41] C. Cohen-Tannoudji, B. Diu, F. Laloë, *Mécanique Quantique*, Hermann (1977)

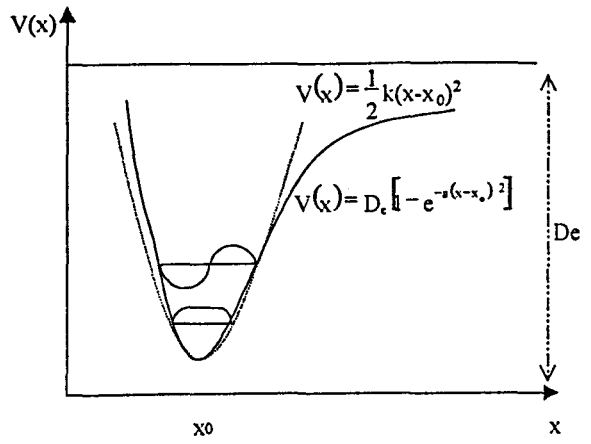


Figure 2 : Potentiel harmonique et potentiel de morse *(D_e est la profondeur du puits de potentiel, x la distance internucléaire et a représente la mesure de la courbure du potentiel au voisinage du minimum, k étant la constante de force de la liaison).*

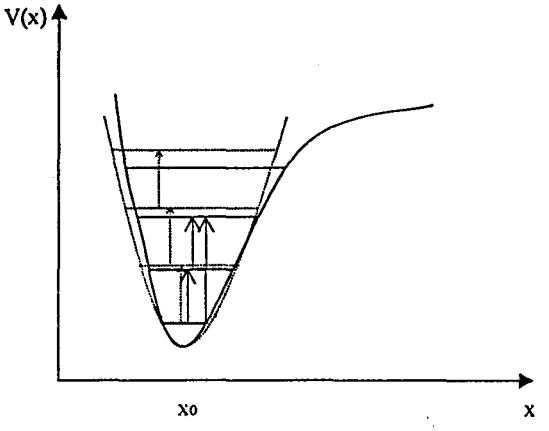


Figure 3 : Niveaux d'énergie et transitions permises selon l'harmonicité du potentiel.

Il faut noter qu'une forme encore plus réaliste prendrait en compte les termes d'ordres plus élevés du développement et en premier lieu la présence d'un couplage vibration-rotation.

Généralisation aux molécules polyatomiques

Limitons-nous au point de vue de la mécanique classique, la modélisation d'un système polyatomique par un ensemble de points massiques permet d'introduire le concept de modes normaux de vibration. Un mode normal est une des 3N solutions de l'Équation 5 où tous les atomes oscillent avec la même fréquence, en phase mais avec des amplitudes différentes.

Équation 5 $$\sum_{i=1}^{3N}\left(f_{ij}-\delta_{ij}\lambda\right)A_{ik}=0.$$

Notons que le traitement complet permettant d'écrire l'Équation 5 est obtenue à partir de l'équation de Lagrange.[42] Parmi les 3N solutions obtenues, 6 correspondent à des valeurs propres de fréquence nulle représentant les mouvements de translation et de rotation d'ensemble de la molécule. Le concept de coordonnées normales est alors introduit : une coordonnée normale Q est associée à chacun des 3N-6 modes de vibration.

[42] J. D. Graybeal, *Molecular Spectroscopy*, McGraw-Hill (1988)

2.2.2 Champ électromagnétique

Introduction

Une des particularités de la théorie des rayonnements est de reconnaître la dualité de la nature de l'énergie électromagnétique. Dans certains cas, il est préférable de considérer le rayonnement comme un flux de photons. Pour d'autres aspects, la description se fait par l'intermédiaire de paquets d'ondes.

Equations de Maxwell

La théorie générale concernant les radiations électromagnétiques a été développée par Maxwell au milieu du XIX^{ème} siècle et résumée par les équations suivantes :

Équation 6 $\quad \vec{\nabla}\cdot\vec{E}+\frac{1}{c}\frac{\partial\vec{H}}{\partial t}=\vec{0} \qquad \nabla\cdot\vec{E}=4\pi\rho$, ρ représente la densité de charge.

Équation 7 $\quad \vec{\nabla}\cdot\vec{H}-\frac{1}{c}\frac{\partial\vec{E}}{\partial t}=\frac{4\pi}{c}i \quad \nabla\cdot\vec{H}=0$, i représente la densité de courant.

Les notations ∇ et $\vec{\nabla}$ désignent respectivement les opérateurs divergence et rotationnel. Ces équations sont simplifiables si l'on introduit un potentiel scalaire φ et un potentiel vecteur $\vec{A}$, reliés à $\vec{E}$ et $\vec{H}$ par les équations suivantes.

Équation 8 $\quad \vec{H}=\vec{\nabla}\cdot\vec{A} \qquad \vec{E}=-\nabla\varphi-\frac{1}{c}\frac{\partial\vec{A}}{\partial t}.$

Les potentiels $\vec{A}$ et φ n'étant déterminés de manière unique, la résolution des équations de Maxwell nécessite l'introduction d'équations simplificatrices. Ainsi, les *jauges*[43,44] sont des choix de $\vec{A}$ et φ qui laissent $\vec{E}$ et $\vec{H}$ invariants :

Équation 9 $\quad \varphi=0 \qquad \nabla\cdot\vec{A}=0 \qquad \nabla^2\vec{A}=\frac{1}{c^2}\frac{\partial^2\vec{A}}{\partial t^2}.$

[43] W. Heitler, *The Quantum theory of Radiation*, Oxford Clarendon Press (1954)
[44] H. F. Hameka, *Advanced Quantum Chemistry*, Addison Wesley (1965)

Hamiltonien du champ électromagnétique

On retiendra que l'hamiltonien H peut s'écrire comme la somme de λ hamiltoniens $H_\lambda(Q_\lambda, P_\lambda)$ où Q_λ représente les coordonnées et P_λ les moments conjugués :

$$\text{Équation 10} \quad \phi = \prod_\lambda \psi_{n\lambda} \text{ et } E = \sum_\lambda E_{n\lambda}$$

où $\psi_{n\lambda}$ et $E_{n\lambda}$ sont fonctions propres et valeurs propres de H_λ et n est un nombre quantique. On obtient, après calculs, l'équation suivante :

$$\text{Équation 11} \quad H_\lambda = \frac{1}{2}\left(P_\lambda^2 + \omega_\lambda^2 Q_\lambda^2 - \hbar\omega_\lambda\right) \text{ et } E_\lambda = \sum_\lambda n_\lambda \omega_\lambda \hbar .$$

La fonction d'onde ϕ décrit le champ sur un ensemble de nombres quantiques n_λ, chacun d'entre eux étant relié à un oscillateur de fréquence $\frac{\omega_\lambda}{2\pi}$.

En d'autres termes, la fonction d'onde $\phi(n_{\lambda 1}, n_{\lambda 2}, \ldots)$ décrit un champ contenant $n_{\lambda 1}$ photons d'espèces λ_1, $n_{\lambda 2}$ photons d'espèces λ_2... De plus, il existe des opérateurs qui transforment la fonction d'onde $\phi(n_{\lambda 1}, n_{\lambda 2}, \ldots, n_{\lambda i}, \ldots)$ en $\phi(n_{\lambda 1}, n_{\lambda 2}, \ldots, n_{\lambda i \pm 1}, \ldots)$ et décrivent la création ou l'annihilation des photons d'énergie $\hbar\omega_{\lambda i}$.

2.2.3 Interaction rayonnement-matière

Introduction

Sous l'influence d'une perturbation électromagnétique, une transition atomique ou moléculaire a lieu si :

- l'énergie du champ perturbateur correspond à la différence d'énergie entre les états du système. Celle-ci est déterminée par la dépendance temporelle de l'équation de Schrödinger.
- le moment de transition de vibration, fonction de la géométrie de la molécule et de la polarité des atomes, est non nul. L'absence de symétrie centrale de charge procure à une molécule un moment dipolaire électrique permanent, affecté par les vibrations des atomes.

C'est le carré du moment de transition qui constitue l'observable. Il détermine l'intensité de la

bande du spectre d'absorption. L'autre information, la position en terme de nombres d'onde, correspond aux valeurs propres de l'hamiltonien obtenues par la résolution de l'équation vibrationnelle de Schrödinger stationnaire.

Equation de Schrödinger dépendant du temps

Pour décrire une transition, procédé par lequel un système évolue, sous l'effet d'une perturbation électromagnétique, d'un état stationnaire ψ_m à un autre ψ_n d'énergie supérieure, il est nécessaire de prendre en considération l'aspect temporel de l'équation de Schrödinger (Équation 3). La stratégie traditionnellement employée pour le calcul consiste à développer les fonctions d'ondes dépendant du temps sous la forme suivante, qui permet d'appliquer la méthode des perturbations :[41]

$$\text{Équation 12} \quad \Phi(x,y,z,t) = \sum_k c_k(t) e^{iwt} \varphi_k(x,y,z).$$

Les c_k sont des coefficients dépendant du temps qui détermine l'évolution de la population de l'état excité. Ils s'expriment finalement sous la forme du produit du moment de transition par un terme dépendant du temps, quantitativement très important à la résonance.

Coefficients d'Einstein pour l'absorption et l'émission

La mécanique quantique assimile le carré d'une fonction d'onde à une mesure de la probabilité de présence d'un système dans un état excité. Par analogie, le carré des coefficients c(t) exprime donc la probabilité de transition entre états. Or, à la résonance, c(t) étant proportionnel à t, il vient que l'intensité d'une transition dépend du temps d'exposition au rayonnement. Ce n'est évidemment pas en accord avec l'observation et le modèle précédent ne décrit visiblement pas tous les phénomènes.
L'explication de cette inadéquation est liée au fait que, sous l'action d'un champ électromagnétique, absorption et émission sont des phénomènes simultanés qui nivellent l'intensité de la transition. Ce sont les coefficients d'Einstein qui décrivent cette théorie en terme de cinétique.[41,45]

[45] M. Diem, *Introduction to Modern Vibrational Spectroscopy*, John Wiley & Sons (1993)

2.3 Données spectrales

Toutes les mesures spectrométriques conduisent au même document de base, ensemble de valeurs numériques, qui définissent le spectre, dont l'interprétation doit permettre la caractérisation des échantillons. Pour chaque longueur d'onde ou chaque fréquence, l'observation d'intérêt (en présence de l'échantillon) rapportée à une observation de référence produit une valeur numérique.
Un traitement statistique de ces valeurs nécessite la connaissance des propriétés mathématiques du vecteur associé à un spectre.

2.3.1 Spectres de l'infrarouge moyen

Les mesures de transmission T, rapport des intensités transmises avec et sans échantillon, sont transformées en unités d'absorbance $A = -\log T$. Afin de faire correspondre les abscisses à une échelle linéaire en énergie, les longueurs d'ondes sont converties en nombres d'onde $\bar{\nu}(cm^{-1}) = 1/\lambda$. Ceux-ci sont compris dans l'intervalle 4000-400 cm^{-1}.

Description

Il y a en première approximation une relation entre les positions des maxima d'absorption de certaines bandes et la présence de fonctions particulières. Néanmoins, l'interprétation rigoureuse des spectres de produits évolués est délicate voire impossible à cause de leur complexité, mais également pour des raisons plus précises :
- la loi de Beer-Lambert est un bon modèle de comportement mais les intensités ne sont pas systématiquement exactement proportionnelles à la concentration du groupe en question.
- les fréquences de certains groupes présentent des déplacements dus à leur environnement ainsi qu'à des influences qui ne sont pas toujours facilement identifiables.

Les composés ont donc tous des spectres différents, même si les différences sont parfois très subtiles. Ce qui apparaît de prime abord comme un désavantage est, en fait, gage de l'unicité du spectre et permet d'inférer des informations sur la structure moléculaire. Notons à ce

propos que les seules molécules transparentes en infrarouge sont les molécules constituées d'un seul type d'atome et les molécules homopolaires.

Les représentations mathématiques des spectres de transmission ou d'absorbance sont des vecteurs dont nous précisons les caractéristiques.

- La dimension des vecteurs correspond au nombre de pas utilisés pour l'échantillonnage du spectre.
- Les valeurs numériques des coordonnées, les intensités, sont généralement converties dans l'intervalle [0,1].

Après traitement des spectres, elles peuvent être translatées vers d'autres intervalles, comme [-1,1] si l'on considère les formes dérivées des spectres.

- Les valeurs des intensités pour des nombres d'onde voisins sont très corrélées.

En effet, pour des raisons fondamentales, le principe d'incertitude d'Heisenberg, et instrumentales, les bandes de vibration ont une largeur intrinsèque.

Facteurs d'influence internes

Si les liaisons chimiques étaient toujours les mêmes, quel que soit le composé dans lequel on les trouve, les problèmes de dynamique moléculaire pourraient être résolus facilement. De nombreux facteurs subtils influencent la longueur, la direction, la force ou la polarité des liaisons. Du point de vue de la spectroscopie infrarouge, plusieurs facteurs affectent les vibrations.

- Les changements dans les masses atomiques.

Fréquence et masse réduite sont en effet reliées simplement dans l'approximation harmonique.

- Les phénomènes de couplage des vibrations.

Des groupements voisins, présentant des similitudes en termes de symétrie et de fréquence, peuvent interagir pour produire d'autres vibrations.

- Les effets d'association.

En phase condensée, les interactions soluté/solvant et soluté/soluté induisent des changements visibles sur les spectres. Les liaisons hydrogène représentent l'effet associatif le plus commun.

Facteurs d'influence externes

Les influences externes sont, jusqu'à un certain point, contrôlables par l'analyste. Chacune des variables suivantes peut avoir une influence importante sur les positions des fréquences d'un groupe.

- Les différents états physiques d'un même échantillon donnent lieu à des spectres spectaculairement différents.
- Les fréquences et les intensités des vibrations de certains groupes sont affectées lorsque le solvant change à cause d'interactions spécifiques ou des modifications de propriétés macroscopiques.
- Les changements de température modifient l'intensité, la position et la largeur des bandes d'absorption.

2.3.2 Traitement des spectres

Le but du traitement du signal est d'éliminer le bruit ainsi que les sources de variations qui ne sont pas directement associés à la variable mesurée. Il permet éventuellement de rendre aux données leurs caractéristiques linéaires mais dépend énormément de nature du signal. N'importe quel type de pré-traitement a pour objectif la simplification et/ou l'amélioration du modèle mathématique ultérieurement construit. Bien que les méthodes de compression ou de réduction soient parfois considérées comme des pré-traitements,[46] nous ne les abordons pas en détail ici.

Correction de ligne de base, dérivation

Les effets de ligne de base sont généralement dus, soit à l'influence de la matrice contenant l'échantillon, soit à la validité du spectre de référence. Ce dernier doit être enregistré à chaque fois que les conditions environnementales de la mesure changent.
Les corrections les plus simples de ligne de base sont effectuées en dérivant les spectres, à l'ordre 1 pour les décalages, à l'ordre 2 pour les tendances. Cette méthode est générale et peut

[46] J. M Andrade, M. V. Garcia, P. Lopez-Mahia, D. Prada, *Talanta*, 44 (1997) p. 2167

être appliquée aussi bien en MIR[47] qu'en UV[48] ou pour d'autres régions spectrales. Cependant lorsque l'ordre augmente, la procédure de dérivation multiplie le niveau de bruit par plusieurs ordres de grandeur. Le calcul est en conséquence souvent associé à une méthode de lissage qui doit limiter cette dégradation du rapport signal sur bruit. La technique la plus simple consiste à remplacer chaque point du spectre par la moyenne des points d'un intervalle qui se déplace (*moving average* en anglais) sur celui-ci.

Une autre méthode, très utilisée, a été développée par Savitsky et Golay[49]. Brièvement, elle consiste à effectuer une régression par les moindres carrés entre un polynôme de degré k et au moins k+1 points du spectre autour de chaque nombre d'onde où la dérivée est calculée. Celle-ci est alors évaluée par la dérivée du polynôme *fitté* en ce point. Remarquons que cette méthode tronque le spectre à ses extrémités.

Outre ces effets sur la ligne de base, la dérivée permet de mettre en valeur les différences spectrales en exaltant les points d'inflexion des spectres,[50] ce qui permet éventuellement de décomposer des massifs larges d'absorption.[51]

Normalisation

L'utilisation la plus courante de la normalisation vise à s'affranchir des problèmes d'interférences ainsi qu'à mettre en valeur certaines contributions du signal par rapport au signal total. Une méthode très répandue en chimie analytique est la normalisation par référence interne.[52,36] On peut aussi mentionner la division par la norme euclidienne, fréquemment utilisée pour normaliser les données au sein d'un même lot. Il existe d'autres possibilités moins intuitives comme la MSC (*multiplicative scatter correction*) développée par Martens *et al.*[53,54] Cela consiste en la régression de chacun des spectres d'un lot de données sur le spectre moyen dans le but de corriger les données spectrales des effets multiplicatifs et additifs liés à la diffusion.[55]

---

[47] N. Dupuy, L. Duponchel, J. P. Huvenne, B. Sombret, P. Legrand, *Food Chemistry*, 57 (1996) p. 245
[48] R. A. Dalteur, J. R. Hurtubise, *Analytical Chemistry*, 56 (1984) p. 819
[49] A. Savitsky, M. J. E. Golay, *Analytical Chemistry*, 36 (1964) p. 1627
[50] Y. Ozaki, T. Miura, K. Sakurai, T. Matzunaga, *Applied Spectroscopy*, 46 (1992) p. 875
[51] A. Dong, P. Hiung, W. S. Caughey, *Biochemistry*, 29 (1990) p. 3303
[52] H. Martens, T. Naes, *Multivariate Calibration*, Wiley & Sons (1989)
[53] H. Martens, S. A. Jensen, P. Geladi, *Proceedings of the Nordic Symposium of Appied Statistics*, (1983) p. 205
[54] J. L. Ilari, H. Martens, T. Isakson, *Applied Spectroscopy*, 42 (1988) p. 722
[55] T. Isakson, T. Naes, *Applied spectroscopy*, 39 (1988) p. 491

La méthode de pré-traitement la plus efficace est celle qui produit le meilleur résultat. Le choix peut être orienté par la nature physique des échantillons, la répétabilité des mesures, la correction recherchée ou d'autres facteurs.

2.4 Instrumentation

Quasiment tous les laboratoires analytiques utilisent pour les analyses de routine et de recherche des spectromètres d'absorption infrarouge à transformée de Fourier (IRTF). Pour ce qui est de l'aspect pratique, notre intérêt est porté principalement aux techniques d'échantillonnage.

2.4.1 Introduction

Historique

L'existence d'un rayonnement thermique en dehors du spectre de la lumière a été découvert par Herschel.[56] Les progrès ont ensuite été corrélés au développement de matériaux permettant d'améliorer la détection.

- 1830 : Apparition des détecteurs basés sur le principe des thermocouples. Le principe de base est la production d'une force électromotrice par effet thermoélectrique.
- 1880 : Découverte des matériaux photorésistants.
- 1870-1920 : Développement des premiers détecteurs quantiques.

Les détecteurs photoconducteurs et photovoltaïques utilisent des matériaux semi-conducteurs, leurs sensibilités sont élevées et leurs temps de réponse brefs.

- 1880-1910 : Premier enregistrement systématique de spectres, observation de la corrélation entre certains groupements et certaines fréquences.[57, 58]
- 1947 : Apparition du premier spectrographe double faisceau permettant la représentation directe de $T=f(\lambda)$.[59]
- 1960 : Apparition des détecteurs quantiques Mercure-Cadmium-Tellure (MCT).
- 1970 : L'utilisation du phénomène de réflexion totale atténuée (ATR) est envisagée.[60]

---

[56] W. Herschel, *Philosophical Transactions of the Royal Society*, 90 (1800) p. 284
[57] W. W. Abney, E. R. Festing, *Philosophical Transactions of the Royal Society*, 172 (1882) p. 887
[58] W. W. Coblenz, *Investigation of Infrared Spectra*, Carnegie Institute (1905)
[59] N. Wright, L. W. Herscher, *Journal of Optical Society of America*, 37 (1947) p. 211
[60] N. J. Harrick., *Internal Reflection Spectroscopy*, Interscience (1967)

Chaîne de mesure

Différents types de spectromètres ont été mis au point au cours de ces 50 dernières années.

• Les outils séquentiels pour lesquels l'information est collectée au cours du temps.

Ils utilisent un détecteur unique et les éléments spectraux sont balayés au fur et à mesure.

• Les outils spatiaux qui utilisent plusieurs détecteurs afin d'enregistrer plusieurs longueurs d'onde en même temps.

• Les outils basés sur le principe du multiplexage en longueur d'onde.

Un détecteur reçoit simultanément plusieurs informations spectrales. Il convient alors de décoder le signal.

Bien que la domination des spectromètres à transformée de Fourier soit incontestable, la tendance des développements actuels (spectroscopie par diode laser, imagerie CCD) concerne les deux premiers types d'analyseurs. Néanmoins, tous les analyseurs infrarouge ont certains éléments en commun.

• La source qui est généralement équivalente à un corps noir du point de vue de son rayonnement.

Les sources les plus populaires sont les Globar et les filaments de Nernst.[61]

• Le système optique dont le rôle est de guider la radiation.

Dans la région spectrale considérée, les lentilles ont historiquement d'abord été remplacées par des miroirs qui permettaient d'éviter les aberrations chromatiques ainsi que les problèmes liés à la dispersion. La tendance est actuellement à l'utilisation de lentilles de séléniure de zinc.

• Le détecteur qui est certainement l'élément le plus critique.

Les caractéristiques spécifiques des différents systèmes et notamment leur détectivité spécifique sont récapitulées dans de nombreuses références.[62] On signalera simplement que pour être efficaces dans la gamme complète de l'infrarouge moyen, les détecteurs quantiques doivent, le plus souvent être, maintenus à la température de l'azote liquide. Néanmoins, le choix d'un détecteur dépend aussi des conditions de fonctionnement du système : domaine spectral, niveau d'énergie, fréquence de modulation du signal...

---

[61] M. W. P. Cann, *Applied Optic*, 8 (1969) p. 1645

[62] G. Gaussorgues, *la thermographie infrarouge : principe, technologies, applications, technique et documentation*, Lavoisier (1984)

Principe de l'interférométrie

Le spectromètre IRTF intègre simultanément toutes les longueurs d'onde. Afin d'obtenir la distribution spectrale $I(\bar{\nu})$, une lumière modulée est présentée au détecteur. Cette figure d'interférences est ensuite décodée par transformation de Fourier inverse.
La plupart des spectromètres utilisent des interféromètres de Michelson. Les variations de chemin optique sont échantillonnées par la figure d'interférences sinusoïdale du laser He-Ne. En conséquence, la précision sur la détermination du nombre d'onde est donnée par la précision de la radiation laser (~0.01 cm^{-1}). Cette propriété est connue sous le nom d'avantage de Connes.
La spectroscopie IRTF se voit attribuée deux autres propriétés. L'avantage de Fellgett concerne le multiplexage alors que l'avantage de Jacquinot caractérise l'étendue du faisceau donc la quantité d'énergie disponible.

2.4.2 Réflexion totale atténuée

Généralités

L'analyste dispose d'une grande variété de possibilités d'échantillonnage. Bien que la méthode soit principalement dictée par la nature de l'échantillon, des choix restent possibles. Le mode de présentation des échantillons conditionne la faisabilité d'une analyse quantitative par la répétabilité, la sensibilité et la spécificité des informations spectrales.
La transmission nécessite bien entendu que les échantillons soient relativement transparents dans la région spectrale de l'analyse. L'atténuation du faisceau suit plus ou moins la loi de Beer-Lambert. Il est cependant possible d'améliorer la précision des informations obtenues par ajustement de la concentration et/ou du chemin optique. Des ordres de grandeur typiques pour ce dernier sont la dizaine de micromètres dans le cas des phases condensées, le mètre voire plus dans le cas des gaz. Pour les molécules biologiques qui sont souvent solubles dans l'eau, le chemin optique permettant de compenser les absorptions du solvant est typiquement inférieur à 10 µm.[63] Néanmoins, la réflexion totale atténuée (ATR) permet de résoudre

[63] A. Dong, P. Huang, W. S. Caughey, *Biochemistry*, 29 (1990) p. 3303

certains problèmes d'échantillonnage.

Principe

Une meilleure compréhension des composants optiques passe par la connaissance de leurs champs proches.[64] Celui-ci se situe à une distance de l'ordre de grandeur de la longueur d'onde de la radiation utilisée. Le champ proche comporte à la fois une partie évanescente et une partie propagative qui ne peuvent être physiquement dissociées. Le terme évanescent signifie que les ondes restent confinées au voisinage de l'objet qui leur a donné naissance. Toute perturbation de l'objet qui engendre les ondes évanescentes modifie fortement celles-ci. C'est cette caractéristique que l'on utilise pour réaliser des capteurs. Parmi les systèmes optiques permettant la création d'un champ évanescent, citons à titre d'exemple la microscopie en champ proche[65] et intéressons-nous à la réflexion totale représentée Figure 4.

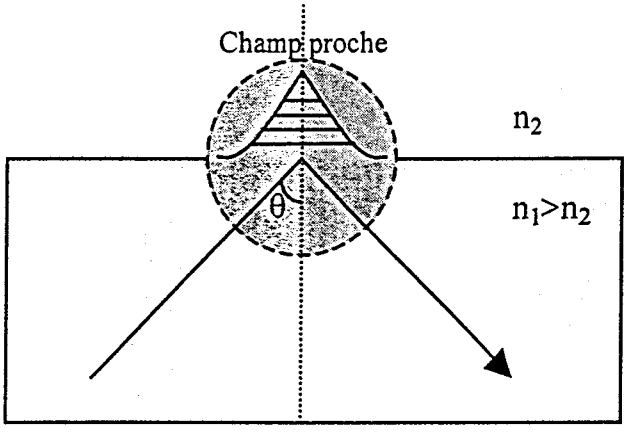


Figure 4 : Réflexion totale à l'interface de deux milieux.

Les lois de Snell-Descartes établissent l'existence de la réflexion totale lors du passage d'une onde plane d'un milieu d'indice n_1 à un milieu d'indice n_2 moins réfringent pour des incidences excédant un angle θ_c limite. L'écriture des champs $\vec{E}$ et $\vec{H}$, ainsi que de leurs équations de continuité au voisinage des interfaces, permettent les descriptions du champ en fonction de l'angle d'incidence et de la profondeur.

[64] F. de Fornel, *Les ondes évanescentes en optique et en optoélectronique*, Eyrolles (1998)
[65] F. E. Lytle, *Applied Spectroscopy*, 53 (1999) p. 212A

La profondeur de pénétration d_p dans l'Équation 13 traduit la rapidité de la décroissance du champ évanescent.

$$\text{Équation 13} \quad d_p = \frac{\lambda}{2\pi\sqrt{n_1^2 \sin^2\theta - n_2^2}}.$$

Une réflexion à l'interface entre le cristal de séléniure de zinc (ZnSe) et l'eau à la longueur d'onde de 10 µm sous un angle θ de 60° correspond à une profondeur de pénétration de l'ordre de 1 µm. Cet exemple permet d'apprécier le confinement du champ évanescent.

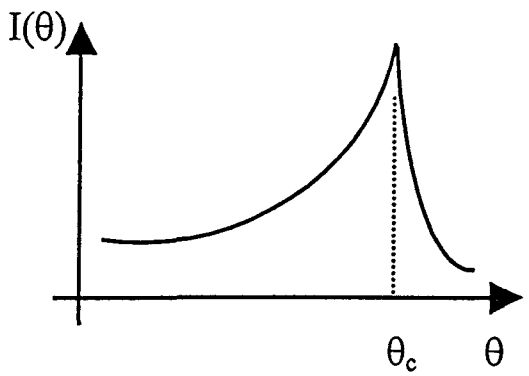


Figure 5 : Intensité sur l'interface selon l'incidence.

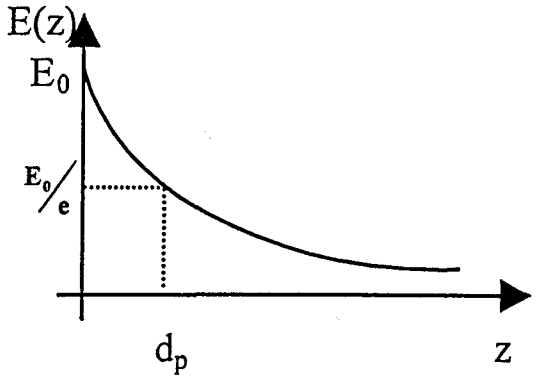


Figure 6 : Décroissance exponentielle du champ évanescent.

Pratique

Une application très importante consiste à réaliser des mesures spectroscopiques par réflexion totale. Le milieu d'indice élevé est en général un cristal tel que le ZnSe qui présente l'intérêt d'être inaltérable au contact de l'eau. La quantité de lumière réfléchie dépend de l'absorption (due à des transitions vibrationnelles) du matériau en contact optique à l'intérieur duquel les ondes évanescentes ont pénétré. Cette radiation réfléchie transporte un spectre analogue au spectre de transmission de l'échantillon.

Pour les études concernant les protéines, les systèmes d'ATR permettent de compenser les interférences dues aux absorptions des solvants[66]. Les résultats obtenus dépendent néanmoins de nombreux paramètres. La longueur d'onde, par exemple, possède une influence sur la

[66] F. N. Fu, D. B. Deoliveira, W. Trumble, H. K. Sarkar, B. R. Singh, *Applied Spectroscopy*, 48 (1994) p. 1432

profondeur de pénétration dans le milieu. Des profils en profondeur d'un matériau sont ainsi réalisables.[67] Il existe donc des systèmes aux géométries variées : ATR plat[68], cylindrique...
Par ailleurs, l'utilisation de fibres optiques permet de délocaliser le compartiment échantillon du spectromètre. Une sonde présente un avantage indéniable sur les méthodes d'échantillonnage traditionnelles. Des mesures *in situ* peuvent ainsi être effectuées au sein même de systèmes très variés.

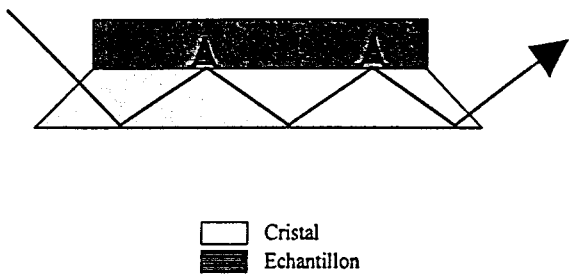


Figure 7 : ATR plat.

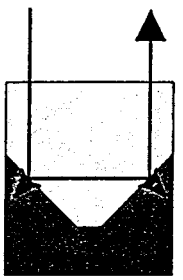


Figure 8 : ATR prisme.

Parmi les matériaux utilisés dans l'infrarouge et permettant la réalisation de fibres, seuls les verres de chacogénures[69] transmettent quasiment toute la gamme du moyen infrarouge. Malheureusement, ils n'autorisent qu'une faible longueur de câble, de l'ordre du mètre, en raison de pertes élevées par absorption en particulier autour de 2250 cm^{-1}. Ces fibres peuvent être couplées à un cristal usiné en forme de prisme permettant des mesures par ATR.[70,71] Aux pertes intrinsèques du matériau s'ajoutent celles dues au couplage avec le spectromètre et le rendement total d'une sonde est toujours inférieur à 10%. Un tel système n'améliore donc en aucun cas la sensibilité d'une mesure difficile et les circonstances d'application s'en trouvent limitées.
Un autre type d'application connue sous le nom de FEWS (pour *fiber optic evanescent wave spectroscopy*, en anglais) consiste à utiliser les fibres optiques elles-mêmes comme élément de réflexion interne. Cela n'est possible que lorsque l'indice de réfraction de la fibre est plus élevé que celui du milieu analysé, comme dans le cas d'analyses sanguines.[72]

[67] S. Ekgasit, H. Ishida, *Applied Spectroscopy*, 51 10 (1997) p. 1448
[68] D. Dufour, D. Bertrand, T. Haertlé, *Journal of Protein Chemistry*, 13 (1994) p. 143
[69] T. Miyashita, T. Manabe, *IEEE Journal of Quantum Electronic*, 18 (1982) p. 1342
[70] N. Dupuy, L. Duponchel, J. P. Huvenne, B. Sombret, P. Legrand, *Food Chemistry*, 57 (1996) p. 245
[71] R. Gotz, B. Mizaikoff, R. Kellner, *Applied Spectroscopy*, 52 (1998) p. 9
[72] Y. Gotshal, R. Simhi, B. A. Seca, A. Katzir, *Sensors & Actuators*, B42 (1997) p. 157

Chapitre 3

Chimiométrie des procédés, réseaux de neurones artificiels

Les technologies du domaine de l'intelligence artificielle utilisent les données disponibles pour atteindre la mesure d'un procédé, à un niveau d'abstraction significatif. La qualité, la consistance des observations améliorent les possibilités de mesure des systèmes traditionnels. Des instruments basés sur des fonctions avancées peuvent ainsi être élaborés. Les réseaux de neurones artificiels (RNA) sont caractéristiques de l'évolution des traitements de données dans cette direction. Il s'agit de réseaux d'entités de calcul locales, connectées par des liaisons paramétrées qui véhiculent des valeurs numériques.[73]

Nous situerons, dans un premier temps, l'utilité des RNA pour la modélisation d'étalonnages multivariés. En effet, les méthodes traditionnelles traitent les signaux d'entrée suivant des algorithmes strictement déterministes. Au contraire, les RNA apprennent une solution à partir d'exemples. Ce traitement des données permet d'étalonner le comportement de systèmes complexes en ne faisant appel à aucune connaissance a priori. L'accent est porté sur l'intérêt des méthodes neuronales dans le cadre de la chimie analytique des procédés et particulièrement pour l'exploitation des mesures spectroscopiques. La connaissance empirique des utilisateurs concernant la technique, le procédé, la physico-chimie des variables mesurées est nécessaire à l'utilisation des RNA. C'est ce qu'englobe la chimiométrie, définie comme la méthodologie utilisant les mathématiques, les statistiques et/ou toute méthode de logique formelle pour construire des procédures de mesure optimales et trouver un maximum d'informations chimiques pertinentes pour l'analyse des données.[74] Nous expliciterons enfin les formalismes mathématiques associés à ces outils de modélisation, en particulier ceux des réseaux de Kohonen et des réseaux *multicouches*.

---

[73] S. Haykin, *Neural Networks : a Comprehensive Foundation*, Macmillan (1994)
[74] D. L. Massart, B. G. M. Vandenginste, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Seyers-Verbeke, *Handbook of Chemometrics & Qualimetrics : Part A*, Elsevier (1997)

3.1 Méthodologie pour la chimiométrie

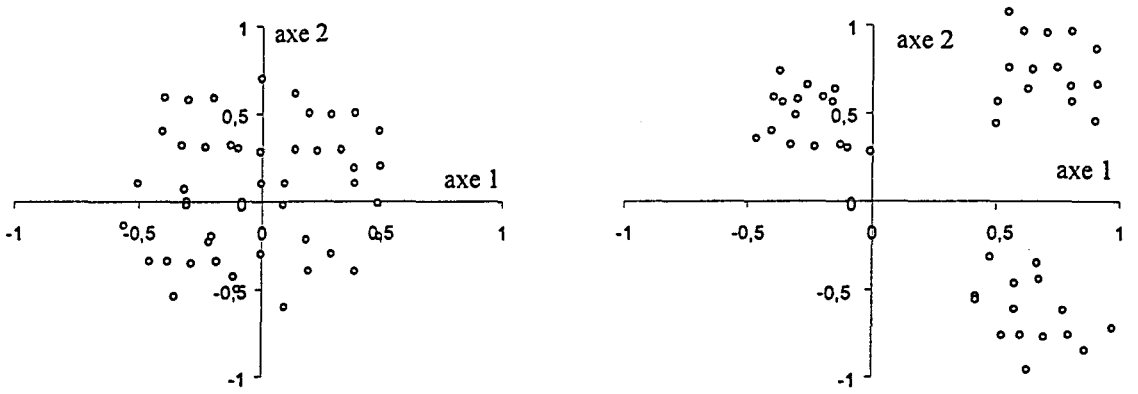
3.1.1 Structure des données

De manière générale, on enregistre un ensemble de p variables pour chacun des n échantillons dont on dispose. Ces données sont ensuite regroupées dans une matrice **X** de dimensions (n×p). De plus, chaque échantillon se trouve caractérisé par la valeur d'une mesure de référence, une ou plusieurs concentrations par exemple. Ces mesures sont regroupées sous forme d'un vecteur colonne **y** de dimension n ou d'une matrice **Y**.
Le principe de toute régression multivariée est le développement d'un modèle **Y**=f(**X**) qui pourra être utilisé pour estimer les valeurs y d'échantillons inconnus.
Dans la suite du manuscrit, les concepts sont illustrés sur des axes arbitraires, notés axe 1 et axe 2 dans l'espace des données d'entrée, y lorsqu'il s'agit des données de sortie.

Homogénéité des données

Un groupe de données (*cluster* en anglais) correspond à une représentation homogène des échantillons dans l'espace des **x** ou des **y**. La présence de groupes distincts au sein d'un même lot indique que les échantillons appartiennent à des populations différentes. C'est souvent le cas lorsque les mesures ne sont pas suffisamment reproductibles ou lorsque les conditions initiales et environnementales sont changeantes. Le modèle de régression est alors plus complexe et moins robuste que lorsque les données sont homogènes. Techniquement, il est possible de construire autant de modèles de régression qu'il existe de classes d'échantillons. Néanmoins, cela multiplie le travail à fournir et le nombre d'échantillons nécessaires pour l'étalonnage.
L'existence de classes est la plupart du temps détectée par simple observation visuelle du lot de données dans un espace de représentation adéquat, comme le montre la Figure 9. Pour ce faire, on utilise souvent des méthodes de condensation du lot de données. Dans le cas de l'analyse en composantes principales (PCA),[5] des combinaisons linéaires des variables initiales sont construites de telle sorte que la variance du lot soit optimisée. La représentation obtenue, plus facilement interprétable du fait de la réduction du nombre de facteurs, est souvent une mesure abstraite de la structure des données.[52] Une autre possibilité est fournie

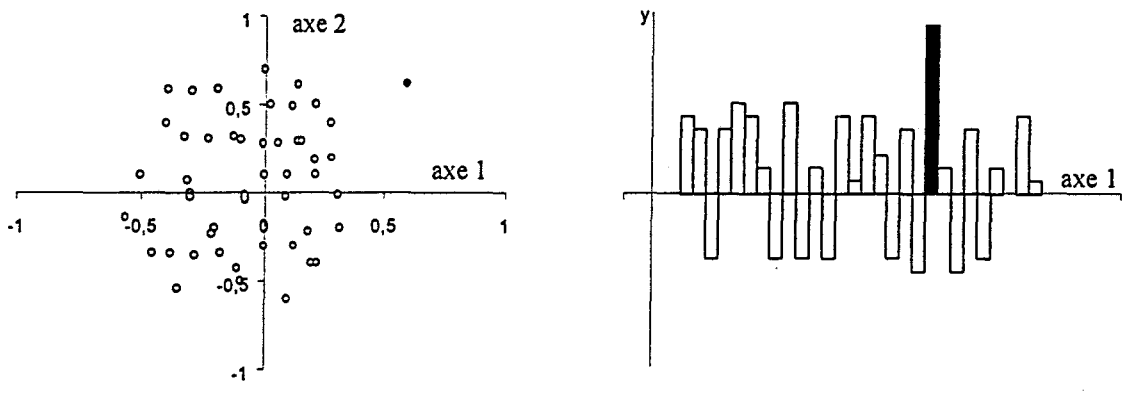
par les cartes de Kohonen, outils de projection permettant une visualisation de la totalité de l'information, que nous étudierons en détail ultérieurement.



A/ B/

Figure 9 : Représentation des échantillons : A/ Lot homogène B/ Groupes d'échantillons.

L'observation de la structure des données présente l'avantage supplémentaire de permettre la détection des points aberrants (*outliers* en anglais) dans l'espace des x ou des y. Il s'agit de points aux caractéristiques extrêmes, visibles par observations cumulées des données dans un espace représentatif des variables x (Figure 10A/) et y (Figure 10B/).



A/ B/

Figure 10 : Points aberrants : A/ Sur les données X B/ Sur les données Y.

Il est préconisé de supprimer ces points aberrants qui peuvent constituer une entrave au développement d'un modèle d'étalonnage multivarié.

Linéarité des données

De manière générale, les non-linéarités concernent tous les lots de données où la représentation entre les données de sortie **Y** et certaines valeurs de la matrice **X** n'est pas linéaire. On parle de non-linéarités apparentes.[75] Cela comprend les situations où les variables ne sont pas analytiquement séparables, les données étant, soit corrélées entre elles, soit corrélées aux perturbations du système. D'autres sources de non-linéarité peuvent être liées aux propriétés des échantillons ou à l'instrumentation. Des méthodes robustes de modélisation sont alors à encourager.

Les non-linéarités apparentes sont à distinguer des non-linéarités vraies qui ne caractérisent pas les données mais désignent plutôt des modèles. Les RNA, dont les propres paramètres sont déterminés non linéairement, en représentent un exemple typique.

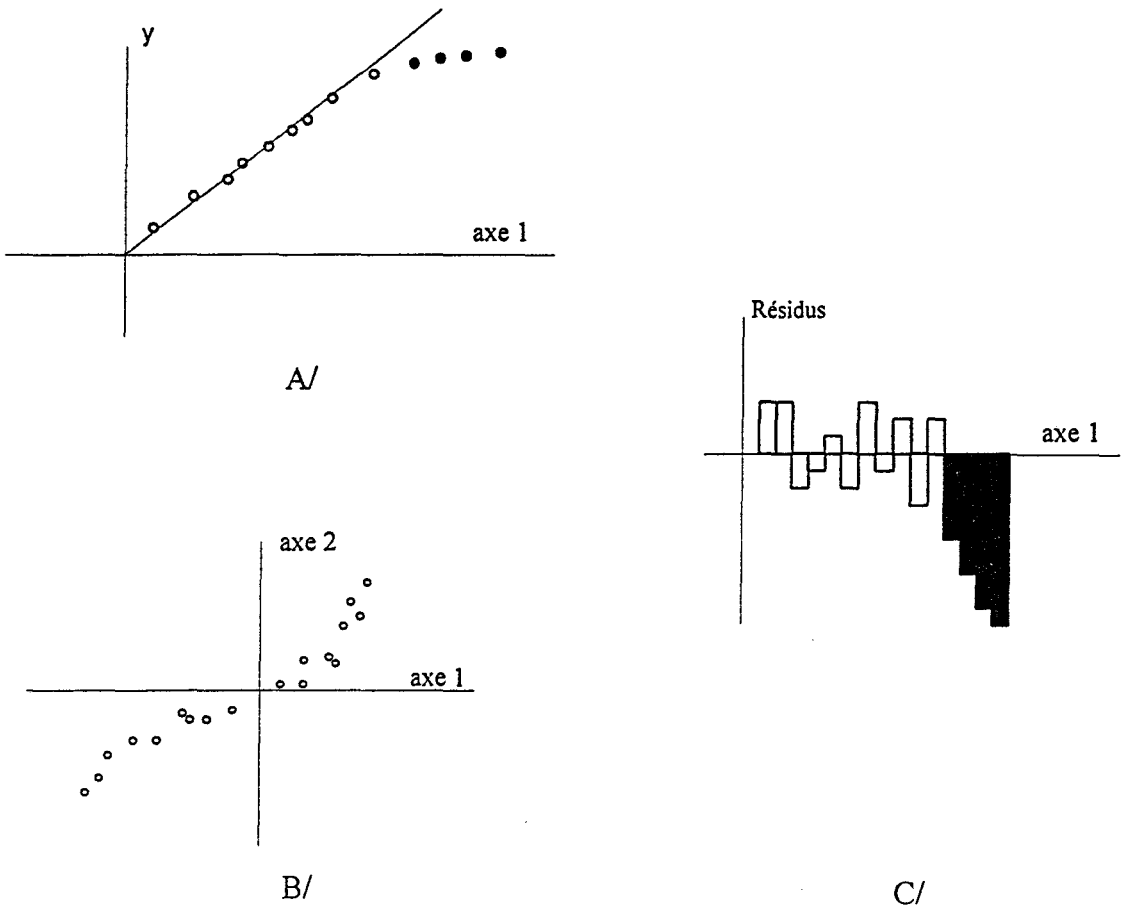


Figure 11 : Détection de relations non-linéarités : A/ *A priori* concernant les variables d'entrée et de sortie B/ *A priori* concernant les variables d'entrée C/ *A posteriori*.

[75] F. Despagne, D. L. Massart, *Analyst*, 123 (1998) p157R

Dans la majorité des applications, la forme de la fonction analytique est inconnue et toute linéarisation des données est impossible. Bien que des tests statistiques de détection des non-linéarités existent, c'est essentiellement l'analyse visuelle qui est utilisée.

La Figure 11A/ représente une propriété d'intérêt y en fonction d'une des variables explicatives du phénomène étudié. Elle montre qu'il est possible de détecter a priori certains effets non-linéaires entre les variables d'entrée et de sortie. La Figure 11B/ permet le même constat à propos des corrélations entre les variables d'entrée elles-mêmes. D'autre part, une fois un modèle d'étalonnage ébauché, l'analyse des résidus peut fournir de nombreuses informations sur la linéarité (Figure 11C/).

Représentativité des lots de données

La finalité d'un modèle d'étalonnage est la prédiction de variables y d'échantillons inconnus. Il convient donc de construire ce modèle à partir d'un lot de données le plus représentatif possible de la population. Dans la mesure du possible, le lot de test doit être inclus dans le lot utilisé pour l'étalonnage, ce dernier prenant en compte toutes les sources existantes de variation.

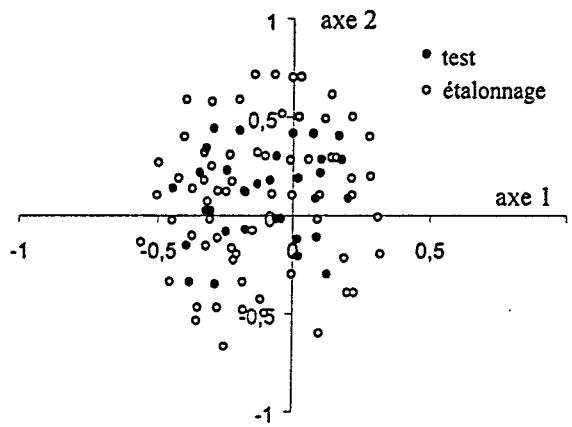


Figure 12 : Répartition des échantillons pour l'étalonnage et le test.

Si une distribution idéale d'échantillons telle que celle présentée Figure 12 ne peut pas être obtenue, une certaine forme d'extrapolation des données est alors inévitablement requise lors de la phase de prédiction.

Les modèles basés sur les RNA sont plutôt robustes et conservent en général de bonnes

capacités prédictives en cas d'extrapolation des données **X**. Par contre, toute extrapolation des données de **Y** conduit à des résultats bien plus désastreux pour les modèles non-linéaires (Figure 13A/) que pour les régressions linéaires (Figure 13B/).

Enfin, le nombre d'échantillons utilisés pour l'étalonnage doit être le plus élevé possible afin de minimiser l'erreur de prédiction.[76] Malheureusement, ce desideratum est rarement compatible avec le coût des analyses de référence.

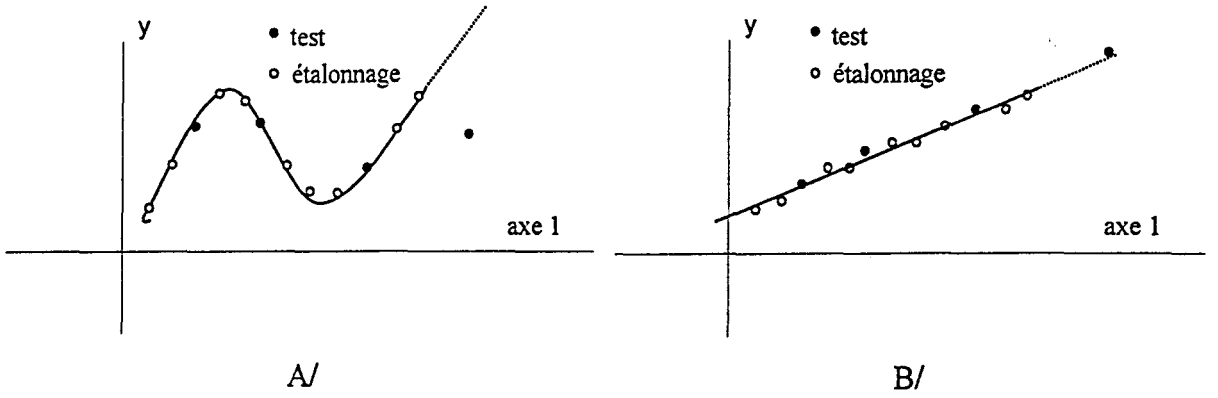


Figure 13 : Extrapolation des données y : A/ Modèle non linéaire B/ Modèle linéaire.

3.1.2 Etalonnages multivariés

Les techniques statistiques conventionnelles et les méthodes neuronales ont des fonctions similaires et partagent les mêmes fondements. Néanmoins, bien que les modèles basés sur les RNA ne constituent en aucun cas une solution "universelle", ils peuvent apporter éventuellement une amélioration des résultats pour certains lots de données.

Aparté

Leur intérêt commun étant l'analyse des données, la distinction entre les techniques statistiques et les réseaux neuronaux n'est que théorique, voire culturelle. En effet, la méthodologie appliquée à ces derniers est très largement inspirée des techniques statistiques. Les réseaux multicouches à apprentissage supervisé ne sont finalement que des modèles

[76] A. Lorber, B. R. Kowalski, *Journal of Chemometrics*, 2 (1988) p. 67

d'étalonnage non-linéaires. Ce sont des estimateurs non paramétriques performants même lorsque les problèmes sont non-linéaires ou que les conditions d'obtention des données sont statistiquement pauvres.

Reste qu'on peut considérer les méthodes neuronales comme des techniques exploratoires alors que les statistiques sont plutôt utilisées pour confirmer ou réfuter une hypothèse, l'univers du problème étant connu. Si le lot de données est représentatif du problème, les capacités d'approximation universelle des RNA doivent permettre de découvrir la structure du modèle.

Choix d'une méthode

Il existe de nombreuses méthodes de régressions multivariées dont certaines sont explicitement non-linéaires.[77] Néanmoins, les méthodes les plus connues, par leur fréquence d'utilisation ou pour des raisons pédagogiques et historiques, sont les méthodes globales de régression.

• La régression multilinéaire MLR (*Multilinear Regression*).

Basée sur la sélection de variables, elle doit être préférée dans les cas les plus simples.[52]

• Les régressions qui utilisent les projections des échantillons sur des variables latentes telles que PCR (*Principal Component Regression*) et PLS (*Partial Least Square*).[52]

De manière générale, des étalonnages performants peuvent être obtenus en projetant les nombreuses variables $\mathbf{X} = \{x_k,\ k=1...K\}$ de l'espace de départ sur les variables $\mathbf{T}=\{t_a,\ a=1...A<K\}$ et en effectuant la régression des variables y de sortie sur les variables latentes T. Il faut définir les facteurs T les plus significatifs à la fois du point de vue de l'interprétation et de la prédiction. Les variables latentes sont des combinaisons linéaires des variables originales. Dans le cadre de la régression PCR, il s'agit des vecteurs propres issus de la PCA sur X. Pour la régression PLS, c'est la covariance entre les données X et Y qui représente le critère à optimiser.

La méthode PLS est une régression bilinéaire (la matrice X est décomposée sous la forme d'un produit de deux matrices linéaires) qui, à la différence des méthodes plus simples, utilise les variables à prédire y dans la décomposition de X. En utilisant les informations en X et en Y, cette méthode réduit l'impact de variations importantes en X, mais non significatives du

[77] B. D. Ripley, *Pattern recognition & Neural networks*, Cambrige University Press (1996)

point de vue de l'étalonnage. Une régression de type MLR est ensuite effectuée sur les projections des données sur ces nouvelles variables, appelées facteurs *score*. Ces méthodes sont linéaires même si des capacités à prendre en compte certains comportements modérément non-linéaires ont été constatées et des variantes des algorithmes proposées.[78,79] Existent également des méthodes de régressions locales,[80] qui appliquent les régressions PLS ou PCR en favorisant les données les plus proches des points à prédire. Des résultats probants ont été obtenus notamment en spectroscopie proche infrarouge.[81]

La comparaison entre ces méthodes n'est pas simplifiée par leur diversité mais certaines études ont cependant été effectuées.[82,83,84] Récemment, *Centner et al.*[85] ont proposé une revue exhaustive des techniques multivariées pour le traitement de données proche infrarouge. En outre, le choix de la méthode est toujours guidé par le problème à résoudre ainsi que par la forme et par la qualité des lots de données. Les données bruitées, corrélées en **X** et en **Y**, pour lesquelles la forme mathématique de la fonction de corrélation n'est pas connue et les non-linéarités sont avérées, plébiscitent l'utilisation des RNA. Ceux-ci peuvent néanmoins être mis en concurrence avec les modèles locaux lorsqu'on est en présence de données inhomogènes. Ces derniers sont moins fréquemment utilisés et n'ont été étudiés que dans des proportions limitées. De plus, ils ne présentent pas les avantages de flexibilité des RNA.[75] Cependant le nombre d'échantillons à disposition est un facteur limitant pour les méthodes neuronales étant donné la quantité de paramètres à ajuster lors de l'apprentissage. Enfin, aucune technique n'étant universelle, chaque application requiert connaissance, pratique et expertise, à défaut d'une méthodologie bien établie.

3.1.3 Applications en chimie analytique des procédés

Les paradigmes basés sur les RNA, caractérisés par leurs facultés d'accommodation sont

[78] S. D Oman, T. Naes, A. Zube, *Journal of Chemometrics*, 7 (1993) p. 195
[79] S. Wold, N. Kettaneh-Wold, B. Skaberger, *Chemometrics & Intelligent Laboratory systems*, 7 (1989) p. 53
[80] B. G. M. Vandenginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Seyers-Verbeke, *Handbook of Chemometrics & Qualimetrics : Part B*, Elsevier science (1998)
[81] T. Naes, T. Isakson, B. Kowalsky, *Analytical Chemistry*, 2 (1990) p. 664
[82] S. Sekulics, B. R. Kowalski, Z. U. Wang, *Analytical Chemistry*, 66 (1994) p. 249
[83] B. R. Bakshi, U. Utojo, *Analytica Chimica Acta*, 384 (1999) p. 227
[84] S. Q. Liu, W. W. Wang, *Chemometrics & Intelligent Laboratory Systems*, 45 (1999) p. 131
[85] V. Centner, J. Verdu-Andres, B. Walczak, D. Jouan-Rimbaud, F. Despagne, L. Pasti, R. Poppi, D. L. Massart, O. De Noord, *Applied Spectroscopy*, 54 (2000) p. 608

capables de trouver des solutions correctes à des problèmes difficiles.[86] Les stratégies que ces systèmes mettent en place reposent sur un apprentissage à partir d'exemples. Le comportement d'un procédé est ainsi mis à jour par le biais de déductions, d'analogies ou de généralisations qui sont des capacités généralement attribuées aux raisonnements humains et qui justifient la dénomination de traitement intelligent de l'information tel qu'il a été défini au Chapitre 1.

Mesure

L'utilisation des RNA dans le domaine de l'instrumentation génère de nombreuses applications liées à l'analyse et à la manipulation des données.

• L'implémentation ou l'amélioration d'un capteur.

Des systèmes d'identification de composés[87] ou de visualisation des propriétés d'échantillons industriels[88] ont été construits à partir des spectres IRTF par le biais de RNA différents.

• La fusion de plusieurs outils pour la construction d'un capteur de niveau d'abstraction plus élevé.

Des études de faisabilité concernant un système de surveillance à partir des informations fournies par plusieurs diodes électroluminescentes ont été présentées.[89] Un système de RNA a été utilisé pour la détermination simultanée de concentrations en alcool et sucres et la mise en valeur d'une réponse élaborée.[90]

• L'identification de systèmes pour la prédiction.

Gemperline[91] propose un ensemble de stratégies pour le développement de méthodes robustes de prédiction de la composition d'échantillons à partir de méthodes spectroscopiques. Ce type d'applications joue un rôle important dans les développements instrumentaux actuels. Le comportement d'un système peut être capturé à partir d'observations, même bruitées, de celui-ci.

• Le remplacement de mesures complexes et coûteuses par une instrumentation intelligente.

---

[86] C. Alippi, A. Ferrero, V. Piuri, *IEEE Instrumentation & Measurement magazine*, (1998) p. 9
[87] H. Yang, P. R. Griffiths, *Analytical Chemistry*, 7 (1999) p. 3356
[88] L. Dolmatova, C. Ruckebusch, N. Dupuy, J. P. Huvenne, P. Legrand, *Chemometrics & Intelligent Laboratory Systems*, 36 (1997) p. 125
[89] M. N Taib, R. Narayanaswary, *Sensors & Actuators*, B 38-39 (1997) p. 29
[90] C. Bessant, S. Saini, *Analytical Chemistry*, 71 (1999) p. 2806
[91] P. J. Gemperline, *Chemometrics & Intelligent Laboratory Systems*, 39 (1997) p. 29

Cela représente l'enjeu des techniques dont nous discutons. Les capacités des RNA ont été démontrées sur un processus industriel de fermentation et un système global d'optimisation de la productivité a été développé.[92] Très récemment, *Chen et Peng*[93] ont élaboré un système intelligent de contrôle de procédés. Il s'agit là de l'étape ultérieure du développement instrumental, on recherche le maintien de conditions de bon fonctionnement d'un procédé à partir d'une fonction complexe (transparente pour l'analyste) des observations.

Ce spectre large d'applications est lié aux propriétés intrinsèques des RNA en termes d'adaptabilité à un problème donné, conséquence de la diversité et de la souplesse de leurs architectures. De plus, les capacités de calcul nécessaires à l'optimisation de leurs paramètres lors de la phase d'entraînement sont désormais largement satisfaites par les ordinateurs personnels commerciaux. Enfin, la chimie analytique des procédés tirera sans doute profit dans les années à venir de la relative compacité du lot de paramètres représentant un RNA entraîné. Cet atout devrait faciliter les transferts d'étalonnage vers des circuits intégrés pour les applications en temps réel.

Spectroscopie moyen infrarouge

La construction d'un modèle de RNA ne doit se faire que si des non-linéarités sont suspectées. En ce qui concerne la spectroscopie infrarouge, de nombreuses sources de déviations influencent les observations.[94] Les non-linéarités du détecteur et des éléments optiques du spectromètre sont à prendre en compte sur le plan instrumental. D'un point de vue analytique, des distorsions peuvent être dues à des échantillons très absorbants ou à des recouvrements de signaux. Enfin, dans certaines applications, des variations de température ou de composition du solvant affectent le spectre infrarouge.
Concernant l'utilisation des RNA pour l'interprétation de spectres infrarouges, *Robb et Munk*[95] font office de précurseurs. Dans un premier temps, ils montrèrent l'applicabilité d'un simple modèle neuronal linéaire. Les résultats furent ensuite améliorés par l'incorporation d'une couche cachée de neurones[96] prouvant l'utilité de prendre en considération certains

[92] P. Lednicky, A. Meszaros, *Bioprocess Engineering*, 18 (1998) p. 427
[93] C. T. Chen, S. T Peng, *Journal of Process Control*, 9 (1999) p. 493.
[94] P. J. Gemperline, J. R. Long, V. G. Gregoriou, *Analytical Chemistry*, 63 (1991) p. 2313
[95] E. W. Robb, M. E. Munk, *Mikrochimica Acta*, 1 (1990) p. 131
[96] M. E. Munk, M. S. Madison, E. W. Robb, *Mikrochimica Acta*, 2 (1991) p. 505

aspects non-linéaires. A la même époque, les premiers travaux significatifs concernant les reconnaissances de propriétés par classification des données spectrales furent publiés par *Zupan et Gasteiger.*[97] Ces auteurs ont généralisé la méthode au monde de la chimie analytique.[98]

Les capacités des réseaux à représenter des propriétés complexes entre des données d'entrée et des données de sortie ont par la suite été largement exploitées. Parmi les études les plus récentes, on peut citer la classification ou l'identification de lots de données de grandes dimensions[99,87] ainsi que la détermination de structures à partir de données spectrales.[100] Il faut noter la variété des outils neuronaux utilisés pour ces applications. Au laboratoire, nous nous sommes intéressés de plus près aux analyses quantitatives et qualitatives d'échantillons industriels caractérisés par leurs spectres infrarouge.[88,101]

Néanmoins, l'utilisation des spectres complets pour des applications basées sur des méthodes de régression utilisant les RNA pose un problème en terme de dimension. Les spectres sont généralement constitués de mesures d'absorbance à plusieurs centaines de longueurs d'onde, le réseau est en conséquence composé d'un nombre de paramètres adaptables du même ordre de grandeur. Le problème n'est alors déterminé que si le nombre d'échantillons dont on dispose est considérable, assurant un rapport nombre d'échantillons sur nombre de paramètres élevé. La plupart des analyses quantitatives utilisent donc des données spectrales transformées et réduites[102] par une méthode de sélection de variables. Cette étape est une phase d'optimisation critique qui doit être discutée lors du développement d'un réseau.

[97] J. Zupan, J. Gasteiger, *Analytica Chimica Acta*, 248 (1991) p. 1
[98] J. Zupan, J. Gasteiger, *Neural Networks for chemists : an introduction*, VCH Publishers (1993)
[99] C. Cleva, D. Cachet, D. Cabrol-Bass, *Analusis*, 27 (1999) p. 81
[100] M. Novic, J. Zupan, *Journal of Chemical Information & Computer Science*, 35 (1995) p. 454
[101] L. Dolmatova, C. Ruckebusch, N. Dupuy, J. P. Huvenne, P. Legrand, *Applied Spectroscopy*, 52 (1998) p. 3
[102] W. Wu, D. L. Massart, *Chemometrics & Intelligent Laboratory Systems*, 35 (1996) p127

3.2 Réseaux de neurones artificiels

Il existe une variété importante de types de RNA et de nouveaux systèmes sont proposés constamment. Ces réseaux sont classés en fonction du principe qu'ils utilisent pour l'optimisation de leurs paramètres et on oppose en général deux types de méthodes d'apprentissage.

D'une part, les méthodes supervisées qui exploitent des couples de données (**X** ; **Y**) mesurés ou connus. L'apprentissage du réseau consiste à minimiser la différence entre les sorties attendues et les sorties calculées pour les exemples qui lui sont présentés.

D'autre part, l'apprentissage non supervisé, qui ne nécessite aucune connaissance a priori des données **Y**. Le réseau, autonome, découvre des propriétés au sein du lot de données **X** et organise ses sorties dans un espace de dimension réduite.

3.2.1 Historique

L'histoire des systèmes neuronaux est concentrée sur les cinquante dernières années et peut être résumée par quelques évènements importants.[80,98]

• En 1943, *McCulloch et Pitts*[103] initient les recherches. Le fonctionnement du système nerveux est décrit par l'intermédiaire d'unités élémentaires interconnectées.

• Les prémices des principes de l'apprentissage sont découverts en 1949.[104] Néanmoins, le *perceptron* programmé par *Rosenblatt*[105] en 1958 ne réalise que certaines tâches linéaires.

• L'année 1969 marque un ralentissement brutal du nombre de travaux dans le domaine. Les limitations des structures linéaires de type *perceptron* sont en effet clairement établies.[106]

• Plus tard, une nouvelle règle d'apprentissage, la *rétro-propagation* (*back-propagation* en anglais), dont les fondements sont attribués à *Werbos*[107], est publiée en 1986 par *Rumelhart.*[108] Elle permet d'envisager le traitement de systèmes non-linéaires.

---

[103] W. S. McCulloch, W. Pitts, *Bulletin of Mathematical Biophysics*, 5 (1943) p. 115
[104] D. O. Hebb, *The Organisation of Behaviour*, Wiley (1949)
[105] F. Rosenblatt, *Psychological Review*, 63 (1958) p. 386
[106] M. L. Minsky, S. A. Papert, *Perceptrons*, MIT Press (1969)
[107] P. Werbos, *PhD Thesis*, Univresity of Harvard (1974)
[108] D. Rumelhart, G. E. Hinton, R. J. Williams, *Parallel Distributing Processing : exploration in the nomenclature of cognition*, MIT Press (1986)

Depuis cette découverte, la croissance des activités de recherche ne s'est pas ralentie.

3.2.2 Le neurone artificiel

Le neurone artificiel est l'unité de base des RNA, le comportement global du système étant déterminé par l'ensemble des neurones interconnectés. Les connexions sont paramétrées par des poids. Ce sont des nombres réels qui représentent la force de la liaison entre deux neurones artificiels, modulant la quantité de signal pouvant atteindre une unité. De leurs valeurs dépend en grande partie le bon fonctionnement du réseau. Ils représentent, d'une certaine manière, la connaissance distribuée de celui-ci.

La représentation usuelle d'un neurone est proposée Figure 14. Les notations x_1, ..., x_p représentent les coordonnées du vecteur correspondant au signal d'entrée **x** et les paramètres w_1, ..., w_p du vecteur **w** désignent les poids correspondants. Chaque neurone artificiel peut recevoir simultanément un certain nombre de signaux qu'il pondère et somme, formant ainsi son propre signal d'entrée, $E = \sum_i x_i w_i$. Le rôle de la fonction F, appelée fonction de transfert, est alors de convertir ce stimulus en un signal de sortie S.

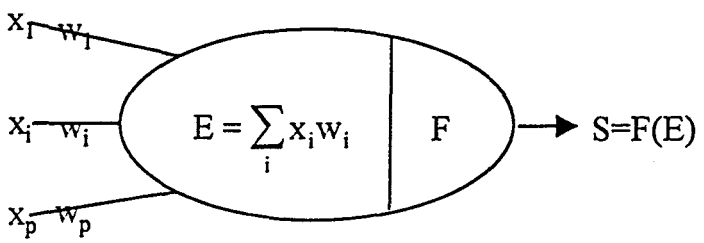


Figure 14 : Représentation formelle d'un neurone artificiel.

3.2.3 Organisation des neurones

La propagation du signal est déterminée par les connexions entre les neurones et par les poids qui leurs sont associés. Les différents types de RNA sont ainsi identifiés par le schéma d'interconnexion des neurones artificiels qui les composent.

Une couche (*layer* en anglais) de neurones est constituée d'unités produisant des sorties S simultanées. Tous les neurones d'une couche possèdent le même nombre de poids, déterminé

par le nombre d'unités des autres couches.

Réseaux

La Figure 15 présente l'arrangement de deux couches adjacentes au sein d'un RNA dit *multicouche*. Ces réseaux opèrent séquentiellement, les neurones d'une couche j ne reçoivent un signal que lorsque les neurones de la couche antérieure j-1 ont terminé leurs traitements. Un réseau est constitué d'au moins deux couches de neurones, à savoir la couche d'entrée et la couche de sortie.

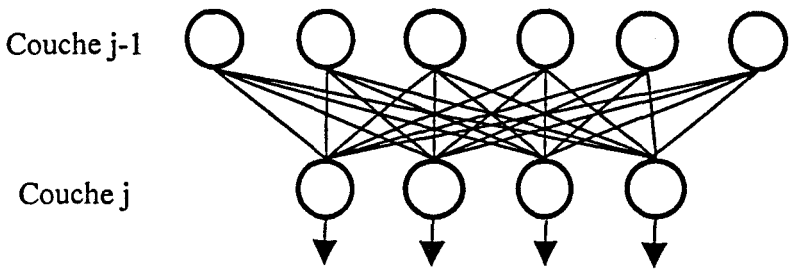


Figure 15 : Couches adjacentes de neurones artificiels.

Les unités de la couche d'entrée se contentent de collecter les signaux provenant des observations du système à analyser (Figure 16). Ces unités ne sont donc pas actives au sens du traitement de l'information.

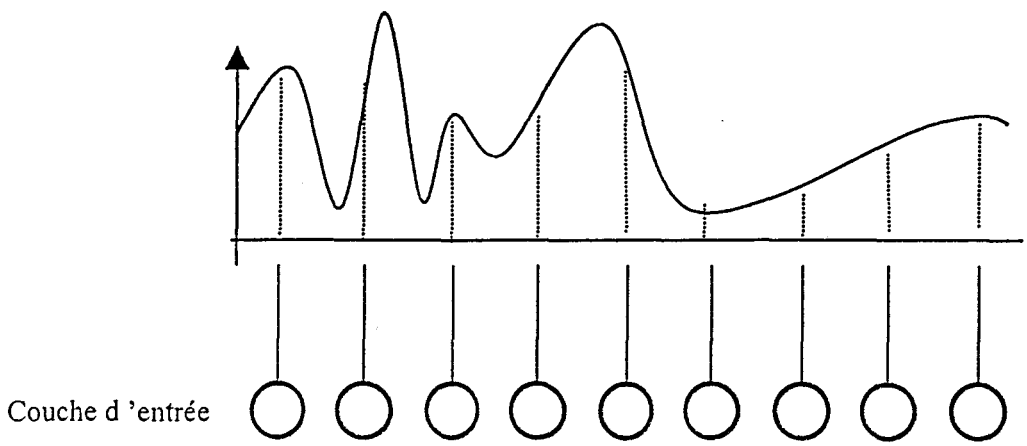


Figure 16 : Couche d'entrée : échantillonnage du signal.

Les neurones des autres couches, au contraire, doivent produire un signal élaboré. Celles-ci

sont donc composées d'unités de traitement qui transforment, par le biais des fonctions de transfert F, les signaux E qu'elles reçoivent.
Les couches de neurones supplémentaires intercalées entre les niveaux d'entrée et de sortie, appelées couches cachées, sont ainsi composées exclusivement d'unités de traitement.

Architecture

L'architecture d'un réseau désigne l'organisation globale des neurones pour le transfert, le stockage et le traitement de l'information.[109] Ce terme englobe les notions de structure et de topologie. L'architecture est généralement résumée par une représentation schématique. Le réseau de la Figure 17 est composé de trois couches contenant respectivement p, q et s unités et sa structure est en conséquence désignée par la notation (p×q×s).

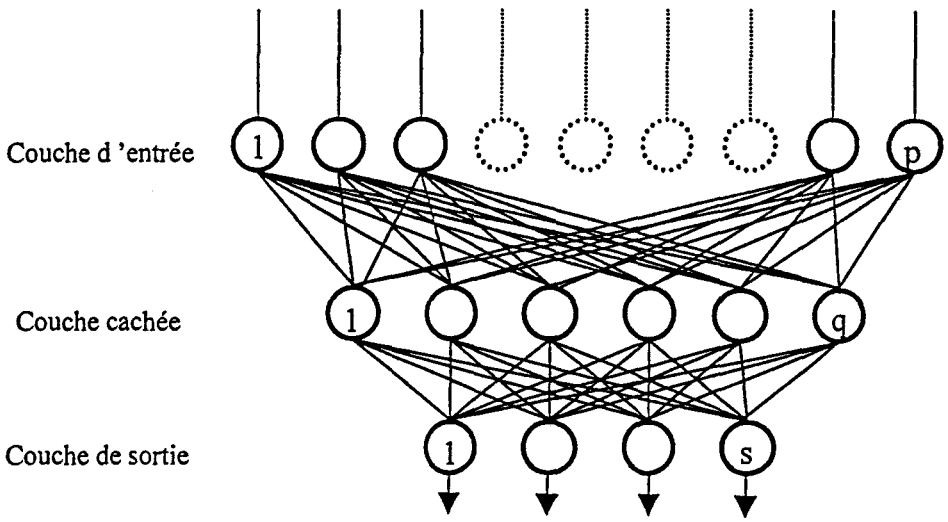


Figure 17 : Architecture d'un réseau multicouche.

3.2.3 Vers des machines plus performantes

Les machines linéaires de type *perceptron* permettent de déterminer une droite frontière D séparant les 2 classes d'objets représentées Figure 18 dans un espace bidimensionnel.
Une fois les poids w_1 et w_2, ainsi que le terme de biais θ déterminés par le processus

[109] T. Kohonen, *Self-Organizing Maps*, S[nd] Ed., Springer-Verlag (1995)

d'entraînement du réseau,[80,98] la droite D est déterminée. La Figure 19 représente le fonctionnement schématique d'un tel outil de classification. La fonction de transfert utilisée est une fonction seuil, bien adaptée au codage binaire.

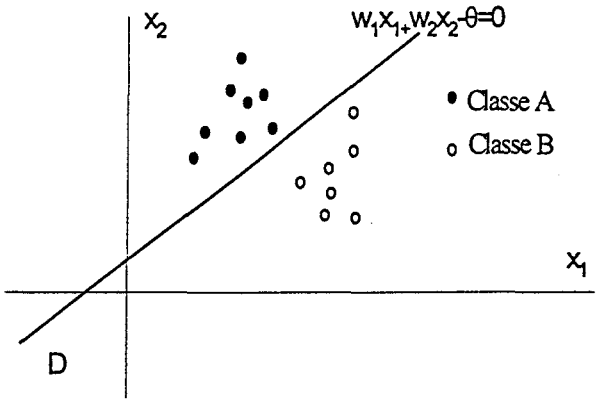


Figure 18 : Frontière séparant deux classes dans un espace à deux dimensions.[80]

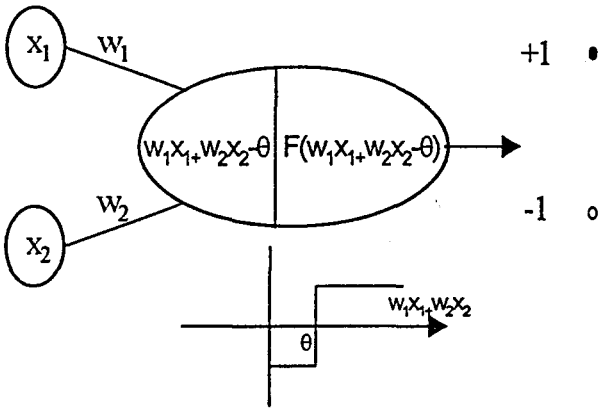


Figure 19 : Machine permettant les classifications linéaires. [80]

Apports d'une couche cachée

Les limitations des RNA ne possédant pas de couche cachée sont dévoilées par l'exemple de classification proposé Figure 20. Ce problème de logique est connu dans la littérature sous le nom de problème XOR (*exclusive or* en anglais). Il est impossible de définir une et une seule droite frontière permettant la discrimination des deux couples d'échantillons en utilisant un outil tel que celui décrit Figure 19.

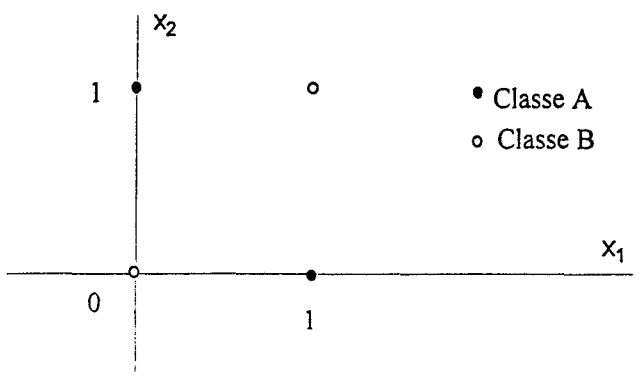


Figure 20 : Le problème de classification xor.

Par contre, l'ajout d'une couche de neurones cachés autorise la résolution. La Figure 21 et la Figure 22 montrent le rôle des deux unités cachées dans la détermination des frontières.

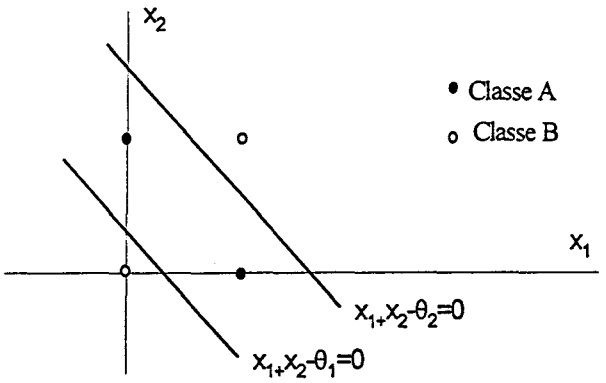


Figure 21 : Une solution au problème xor. [80]

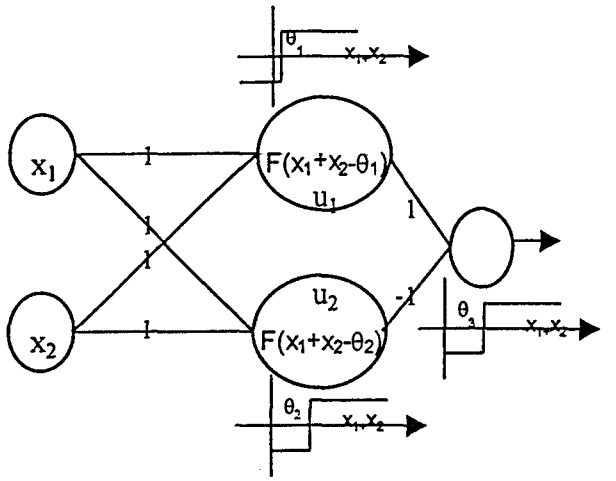


Figure 22 : RNA (2×2×1) à une couche cachée. [80]

De plus, si on représente les 4 objets dans l'espace défini par les valeurs des sorties des deux neurones cachés u_1 et u_2, on obtient la Figure 23 sur laquelle la droite frontière est définie par le neurone de sortie. Le rôle des neurones cachés est donc aussi de définir un sous-espace des représentations du problème.

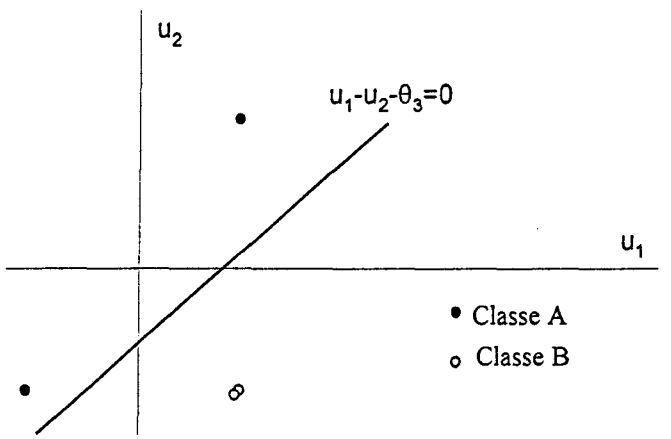
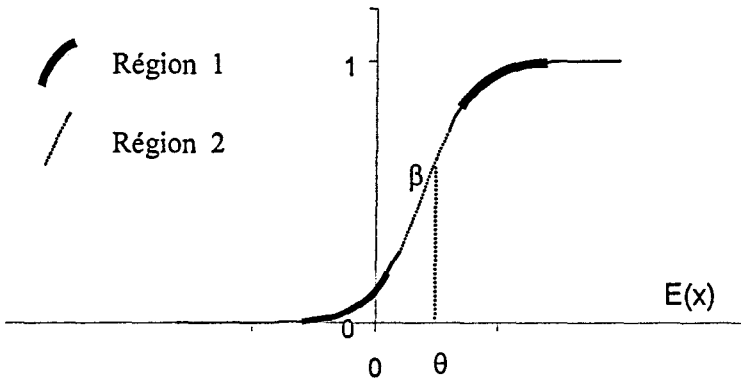


Figure 23 : Représentation dans l'espace des valeurs de sortie des neurones cachés (u_1 ; u_2).

Apports de la fonction de transfert sigmoïde

Certes, l'ajout d'une couche de neurones intermédiaires permet d'augmenter la complexité de la tâche de classification. Cependant, cela ne confère au réseau aucune capacité concernant d'éventuels traitements non-linéaires des données. Pour ce faire, il faut utiliser des fonctions de transfert F différentes des fonctions à seuil que nous avons considérées jusqu'ici.
La fonction sigmoïde permet la modélisation de relations non-linéaires entre l'espace de départ et l'espace d'intérêt. La Figure 24 montre clairement comment ce rôle est rempli par les portions de la courbe correspondant à la région 1.



$$F = \frac{1}{1 + e^{-\beta(E(x)-\theta)}}$$

$$E(x) = \sum_i w_i x_i$$

Figure 24 : Fonction sigmoïde.

Cette fonction sigmoïde est la forme mathématique la plus couramment utilisée. En effet, elle présente l'avantage de produire des valeurs de sortie dans l'intervalle]0,1[et d'être dérivable dans tout le domaine. De plus, en dépit des propriétés décrites ci-dessus, l'approximation quasi-rectiligne de la courbe dans la région 2 permet de conserver de bonnes capacités linéaires de modélisation. Il est ainsi possible de décrire correctement des systèmes linéaires à partir de bases de fonctions sigmoïdes.[110]

[110] W. J. Melssen, L. M. C. Buydens, *Analytical Proceedings*, 32 (1995) p. 53

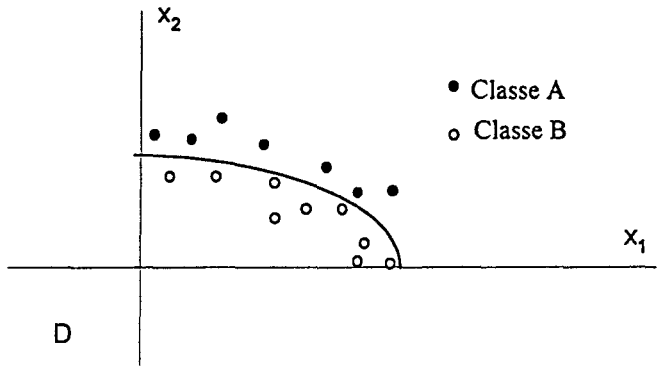


Figure 25 : Problème de classification non-linéaire.

La combinaison adaptée, par l'intermédiaire des poids, de plusieurs fonctions sigmoïdes permet de modéliser des relations non triviales. Le nombre d'unités cachées est ainsi corrélé à la complexité de la relation à modéliser. Dans le cadre de notre exemple de classification à deux dimensions, une frontière du type de celle dessinée Figure 25 peut ainsi être établie à partir d'un réseau contenant deux neurones cachés.

3.3 Apprentissage non supervisé : réseaux de Kohonen

Les réseaux de Kohonen appartiennent à la classe des cartes auto-organisatrices de données. Celles-ci sont définies comme les résultats de régressions non paramétriques.[111] Ces cartes sont principalement utilisées pour représenter, sur un espace bidimensionnel, des vecteurs de grandes dimensions reliés entre eux par des propriétés non-linéaires. Le résultat produit est une classification non supervisée des échantillons. Cela signifie que la distribution est construite sur la seule connaissance des valeurs analytiques x des entrées **x** du système ainsi que sur les relations existant entre les objets eux-mêmes.
L'algorithme basé sur un apprentissage compétitif des données, conduit à la formation de représentations des échantillons qui conservent la topologie de l'espace d'entrée.

3.3.1 Concepts généraux

Représenter un lot d'entrées multidimensionnelles sur un espace de sortie de dimension significativement plus faible est un problème de compression des données. Le dilemme consiste évidemment à obtenir la compression maximale pour une perte d'information minimale. Le problème de conservation de la topologie lors de cette étape de projection est la base conceptuelle des RNA de Kohonen.

Conservation de la topologie

Les propriétés topologiques de l'information concernent les relations entre les données, c'est-à-dire relatives à leur structure, plutôt qu'aux valeurs algébriques qui leurs sont attribuées. Un exemple didactique du concept de conservation de la topologie est retranscrit Figure 26. Il s'agit de la représentation de la surface d'un objet tridimensionnel structuré, en l'occurrence une main, sur une surface bidimensionnelle de taille arbitraire.

[111] T. Kohonen, *Neural Networks*, 1 (1988) p. 3

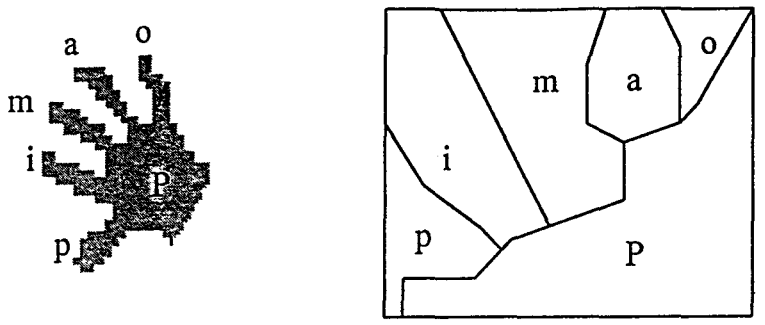


Figure 26 : Préservation de la topologie lors de la projection sur une carte de neurones artificiels.[98]

Du point de vue des RNA, ce sont les connexions entre unités voisines qui définissent la topologie. Pour en faciliter la visualisation, on peut considérer que chaque neurone est connecté à tous ses voisins. C'est alors la disposition des neurones les uns par rapport aux autres dans le plan de la carte qui révèle la topologie du voisinage.

Figure 27 : Exemples de voisinages topologiques d'un neurone.[109]

Architecture

Les réseaux de Kohonen sont composés d'une couche de neurones d'entrée ainsi que de neurones de sortie ordonnés sur une carte de faible dimension. Il peut s'agir d'une carte à une dimension, à deux dimensions (ce qui est la représentation la plus courante) ou à plus de deux dimensions (l'intérêt pour l'analyste est moins immédiat !). Chaque neurone de ce niveau, dit actif, est localement connecté à un certain nombre d'unités proches, constituant son voisinage.

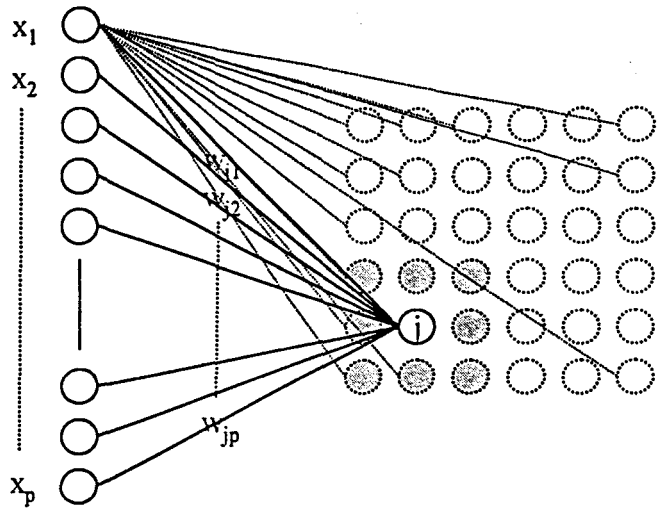


Figure 28 : RNA de Kohonen – Carte auto-organisatrice à deux dimensions.

La Figure 28 montre comment les données sont distribuées par la couche d'entrée sur tous les neurones de la couche de traitement qui reçoivent donc les mêmes entrées multidimensionnelles. Chaque neurone j de cette couche est associé à un vecteur paramétrique de référence $\mathbf{w}_j = [w_{j1}, w_{j2}, \ldots, w_{jp}]^T$ de $\Re^p$. Ce vecteur contient les poids des liaisons avec les neurones de la couche d'entrée.

Le concept élaboré par Kohonen établit que des signaux d'entrée similaires, quelle que soit la nature de cette ressemblance, activent des neurones voisins. Pour ce faire, chaque donnée d'entrée $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$ est comparée à tous les vecteurs **w**. La position du neurone pour laquelle ces deux valeurs sont les plus ressemblantes, dans une certaine métrique, est calculée. C'est ce neurone, que l'on appelle vainqueur, qui se voit attribué la localisation de la réponse au vecteur **x** présenté.

3.3.2 Apprentissage des réseaux de Kohonen

Le procédé d'apprentissage met en place la projection non-linéaire des données. Au cours de celui-ci, tous les neurones d'un voisinage défini sont activés chaque fois qu'un vecteur x est présenté à un neurone de la couche active. Cela se traduit par une adaptation des vecteurs poids associés à ces unités. En répétant itérativement ce procédé pour l'ensemble X des

données, on obtient un ordonnancement global des échantillons sur les neurones de la carte.

Apprentissage compétitif

L'apprentissage compétitif consiste à sélectionner un seul neurone parmi ceux de la couche active après présentation des données. Cette méthode est connue sous le nom de " *winner takes it all* " en anglais.

Les étapes suivantes décrivent l'algorithme d'apprentissage compétititif.

• Initialiser tous les poids à l'aide de valeurs aléatoires. On impose en général une "normalisation" du type de l'Équation 14.

$$\text{Équation 14} \quad \sqrt{\sum_{i=1}^{p} w_{ji}^2} = 1 \text{ pour chaque j.}$$

Pour chaque itération, pour chaque vecteur d'entrée :

• Calculer une mesure de similarité entre le vecteur poids de chaque neurone j et une donnée d'entrée particulière $\mathbf{x_e}$:

$$\text{Équation 15} \quad \sum_{i=1}^{p} (x_{ei} - w_{ji})^2 .$$

• Activer l'unité la plus similaire au signal d'entrée. Il s'agit du neurone vainqueur, c, dont les paramètres vérifient l'Équation 16.

$$\text{Équation 16} \quad \sum_{i=1}^{p} (x_{ei} - w_{ci})^2 = \min\left(\sum_{i=1}^{p} (x_{ei} - w_{ji})^2\right).$$

• Adapter les poids du neurone vainqueur selon la loi décrite par l'Équation 17. Chaque itération, rapproche la réponse du neurone c de la valeur $\mathbf{x_e}$ de l'objet.

$$\text{Équation 17} \quad w_{ci}(t+1) = w_{ci}(t) + \eta(x_{ei} - w_{ci}(t)) \text{ pour tout i.}$$

Le paramètre η représente un taux d'apprentissage.

• Corriger, en utilisant l'Équation 18, les poids des unités se trouvant dans un proche voisinage de c.

$$\text{Équation 18} \quad w_{ij}(t+1) = w_{ij}(t) + \eta A(t,r)(x_{ej} - w_{ij}(t)) .$$

A(t,r) est une fonction de voisinage prédéfinie, r représentant la distance au centre de cette fonction.

Une procédure complète d'apprentissage consiste en un certain nombre de ces cycles (un cycle correspondant au passage de toutes les données d'entrée) au cours desquels la fonction de voisinage est réduite afin d'assurer la convergence du réseau.

Adaptation des poids

L'Équation 17 décrit l'évolution du vecteur poids du neurone vainqueur vers le vecteur d'entrée et la Figure 29 en procure une représentation schématique. A chaque itération, le vecteur poids se rapproche du vecteur d'entrée.

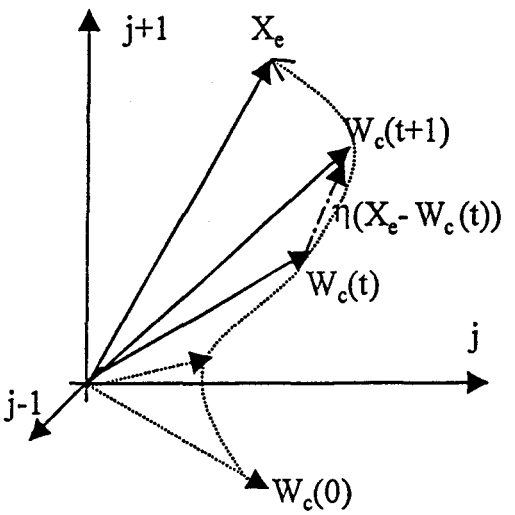


Figure 29 : Evolution du vecteur poids au cours de l'apprentissage.

L'apprentissage est un procédé stochastique, pas tout à fait déterministe. La précision de la représentation des données obtenue dépend du nombre d'itérations permettant d'atteindre la convergence. Parmi les propriétés intéressantes, on notera que la dimension de l'espace d'entrée n'a aucune influence sur le nombre d'itérations[109] (mais pas sur le temps de calcul d'une itération !). Des vecteurs d'entrée de taille importante peuvent en conséquence être utilisés.

Fonction de voisinage

L'adaptation des paramètres du neurone vainqueur se fait en entraînant les vecteurs poids des

neurones voisins. Cette correction locale représentée Figure 30 par la fonction de voisinage A(r,t) doit respecter deux conditions :

- elle n'est pas la même pour tous les neurones du voisinage en question ; A=A(r).
- elle est de plus en plus faible au fur et à mesure de l'avancement de l'apprentissage ; A=A(t) et une mesure d de l'influence de A doit vérifier d(t+1)<d(t).

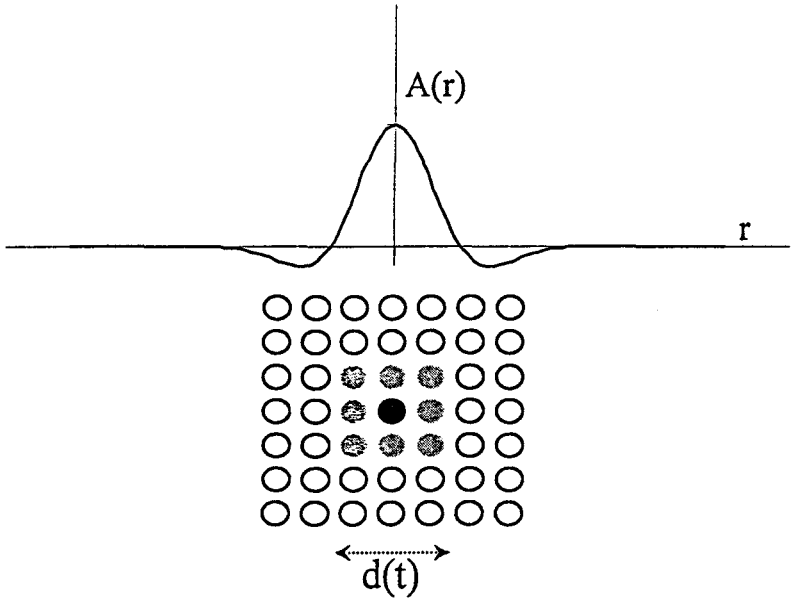


Figure 30 : Fonction de voisinage centrée sur le neurone vainqueur pour la correction locale des poids.

3.3.3 Interprétation des cartes de Kohonen

Les cartes de Kohonen sont des techniques de projection conservant la topologie de l'information. Elles sont principalement utilisées pour l'examen de données à propos desquelles on dispose de peu de connaissance a priori. Les choix concernant la taille de la carte ou la dimension des données d'entrée sont entre autres gouvernés par l'application étudiée,[112] le nombre de classes attendues[99] ou le nombre d'échantillons à représenter.[113]
La Figure 31 représente un RNA de Kohonen sous la forme d'un arrangement de vecteurs colonnes dans une matrice carrée. Chaque colonne contient les p poids associés à un neurone.

[112] W. J. Melssen, J. R. M. Smits, L. M. C. Buydens, G. Kateman, *Chemometrics & Intelligent Laboratory Systems*, 23 (1994) p. 267
[113] R. Goodacre, J. Pygale, D. B. Kell, *Chemometrics & Intelligent Laboratory Systems*, 33 (1996) p. 69

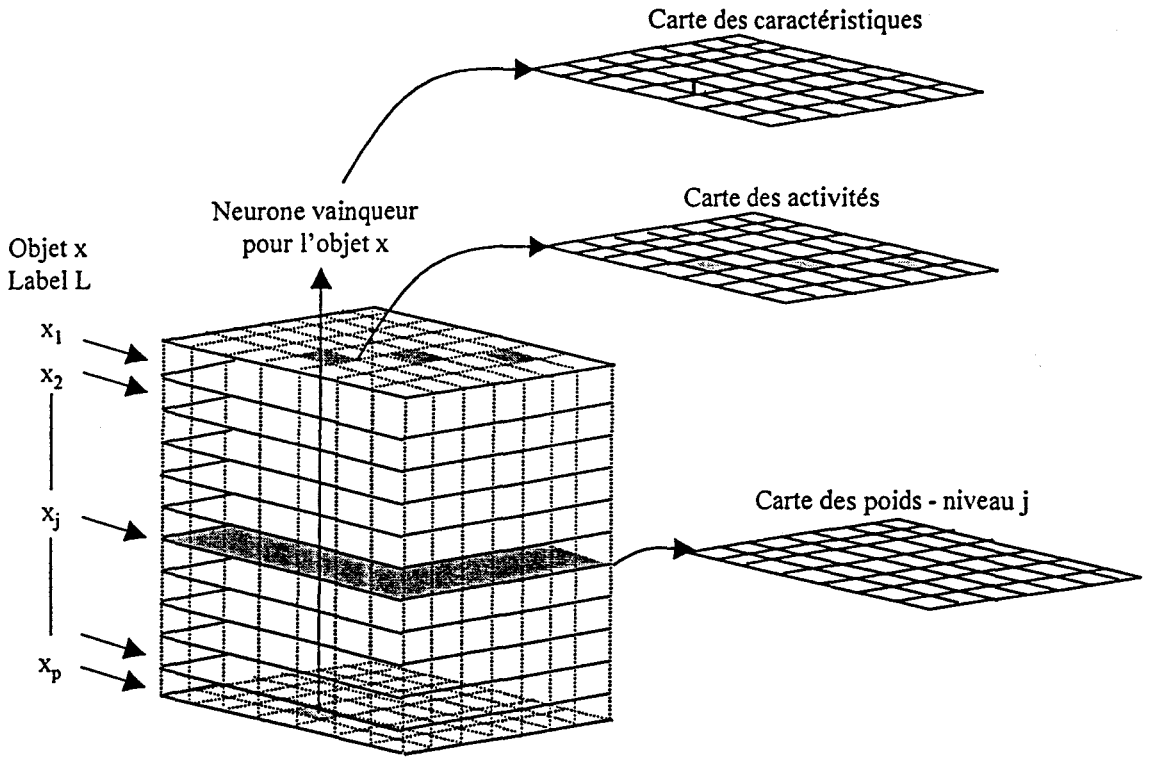


Figure 31 : Représentation d'un RNA de Kohonen[114] - Interprétation des cartes.

Plusieurs interprétations complémentaires des réseaux de Kohonen, à des niveaux d'abstraction différents, sont envisageables.

Cartes des caractéristiques

Une fois l'entraînement terminé, on présente une nouvelle fois au réseau l'ensemble des vecteurs du lot d'entraînement **X**. Les labels L des neurones vainqueurs sont alors résumés sur la carte. Ces labels correspondent généralement aux numéros d'identification des échantillons ou à certaines de leurs propriétés. On associe donc à chaque neurone vainqueur un ou plusieurs labels qui permettent éventuellement la détermination de propriétés communes pour les échantillons. La carte des caractéristiques ainsi formée est une aide à la visualisation de la réponse du réseau de Kohonen aux stimuli d'entrée. On peut même envisager la prédiction des propriétés de nouveaux échantillons par projection de ceux-ci sur la carte. Une autre application peut être la sélection d'échantillons représentatifs au sein d'un lot très fourni en

utilisant la fréquence de distribution des objets sur la carte. Cela peut faciliter la construction d'un modèle supervisé.[114]

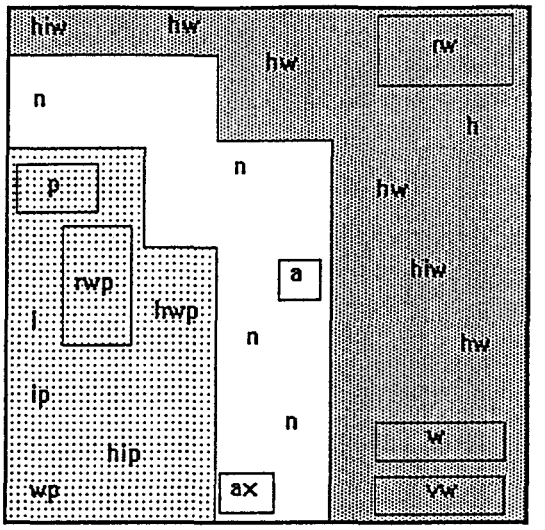


Figure 32 : Carte 15*15 de caractéristiques. [101]

Nous présentons Figure 32 un exemple d'utilisation de ce type de carte pour la discrimination de propriétés physico-chimiques d'amidons modifiés à partir des spectres infrarouge.[101] Chaque lettre correspond à une modification particulière et certaines combinaisons sont possibles. Un échantillon labellisé "hwp" est à la fois "h", "w" et "p". Cette carte permet de distinguer sans ambiguïté 3 régions, correspondant aux niveaux de gris représentés.

Cartes des poids

Chaque neurone possède le même nombre de poids. Ainsi, à une altitude fixée dans la matrice des poids (Figure 31), ce sont toujours les données provenant d'une même variable qui sont traitées. En conséquence, une fois l'apprentissage accompli, une carte montrant la distribution des valeurs pour une variable particulière peut être obtenue. Néanmoins, en pratique, ce sont plus souvent les combinaisons de ces cartes qui permettent de retrouver une distribution attendue des qualités des échantillons.[114]

[114] J. Zupan, M. Novic, I. Ruisanchez, *Chemometrics & Intelligent Laboratory Systems*, 38 (1997) p. 1

Cartes des activités de sortie

L'observation de l'adéquation d'un vecteur x avec plusieurs vecteurs poids conduit à définir l'activité des neurones que l'on représente sur une carte. L'activité d'un neurone est la mesure d'une distance séparant son vecteur poids w du vecteur x. Il s'agit donc d'une mesure de similarité et la carte des activités de sortie permet l'inspection des régions qui ont des poids similaires pour une entrée donnée. Les activités sont la plupart du temps visualisées par le biais d'un code couleur ou de niveaux de gris.

La Figure 33 représente la carte des activités correspondant à une étude préliminaire effectuée en vue d'une analyse quantitative des constituants du couchage de papiers.[88] Les zones noircies de la carte correspondent aux activités les plus importantes. Elles permettront de visualiser les échantillons ayant des concentrations extrêmes (0% et 100%) en carbonate.

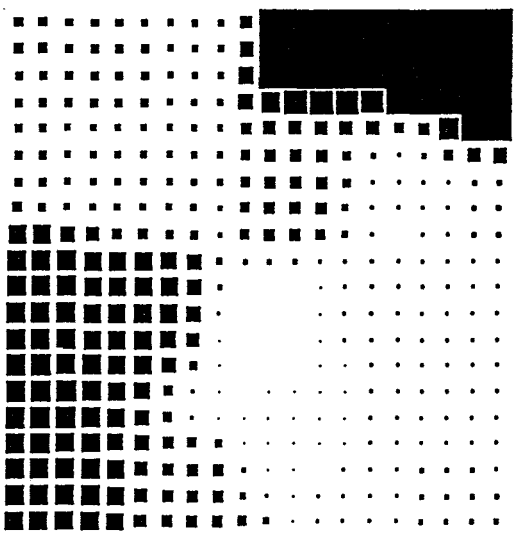


Figure 33 : Carte 20*20 des activités d'un RNA de Kohonen.[88]

Mais, comme le montre la Figure 34, l'observation simultanée de plusieurs types de cartes de résultats améliore souvent la connaissance du problème et des données.

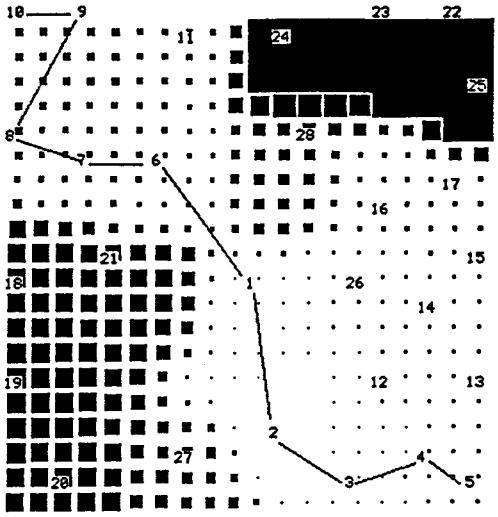


Figure 34 : Superposition des cartes 20*20 de caractéristiques et des activités.[88]

3.4 Apprentissage supervisé : réseaux multicouches

Notre intérêt réside dans les réseaux *multicouches* pour lesquels la propagation des données est réalisée en direction de la couche de sortie (RNA-FF), c'est-à-dire vers l'avant (FF pour *feed-forward* en anglais). Au sein de ces réseaux, la distribution des poids est optimisée par un apprentissage supervisé.
Les caractéristiques des RNA-FF autorisent la résolution de problèmes très variés. Il faut néanmoins respecter certaines conditions lors de la phase d'entraînement ainsi que certaines restrictions lors de l'utilisation en prédiction. La construction d'un outil efficace, délicate, dépend du problème et des données et nécessite de nombreux choix uniquement guidés par des heuristiques.

3.4.1 Description des réseaux feed-forward

Les RNA-FF sont aussi appelés *perceptrons multicouches* ou RNA à rétro-propagation de l'erreur. Ce sont les réseaux les plus populaires et les plus utilisés dans une large variété de problèmes de classification ou de modélisation,[115] y compris en chimie.[98] Leur force principale est leur capacité d'approximation universelle.[116] Ils peuvent, si l'on ne restreint pas leur taille, modéliser n'importe quel type de relation continue entre les données d'un espace d'entrée et les données d'un espace de sortie. Ils utilisent à cet effet les connaissances **X** et **Y**, dont on dispose a priori sur les échantillons exemples, pour capturer l'information.
L'apprentissage est donc supervisé, ce qui suppose que l'information significative est contenue implicitement dans les données.

Structure

Les unités des RNA-FF sont arrangées en couches. La Figure 35 représente l'architecture la plus courante, l'utilisation d'une couche cachée unique faisant la quasi-unanimité au sein des

[115] J. R. Smits, W. J. Melssen, L. M. C. Buydens, G. Kateman, *Chemometrics & Intelligent Laboratory Systems*, 22 (1994) p. 165
[116] G. M. Maggiora, D. W. Elrod, R. G. Trenary, *Journal of Chemical Information & Computer Science*, 32 (1992) p. 732

travaux relatés.

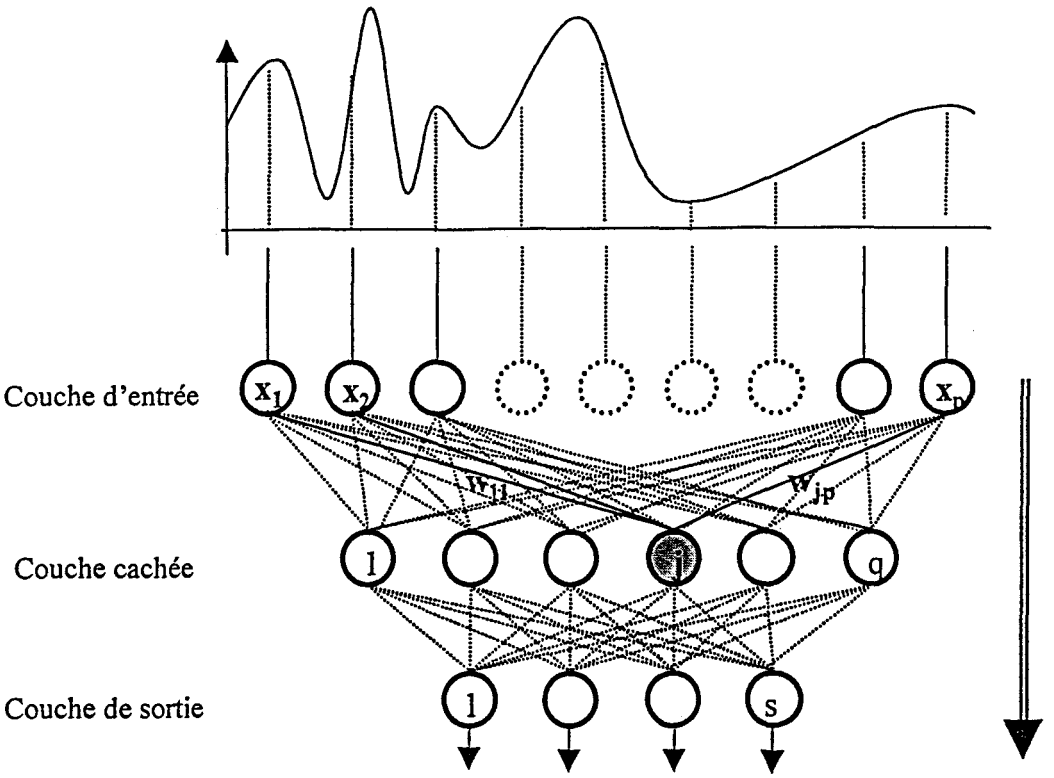


Figure 35 : Réseau multicouche *feed-forward.*

Toutes les unités d'un niveau sont connectées à l'ensemble des unités de la couche suivante. Le réseau est donc composé d'un nombre de poids égal à q×(p+s). L'architecture peut éventuellement autoriser des liaisons directes de la couche d'entrée vers la couche de sortie.

La propagation du signal

De manière générale, chacun des p neurones de la couche d'entrée reçoit une variable x_i échantillonnée à partir du signal **x**, obtenu pour un objet du lot **X**. Cette information transite ensuite par la couche cachée où elle est traitée par q unités dont les fonctions de transfert sont généralement des formes sigmoïdes. Les résultats sont ensuite distribués sur le niveau de sortie. Celui-ci, qui produit les réponses du réseau, contient un nombre s de neurones correspondant au nombre de variables y de Y. Les fonctions de transfert des neurones de la couche de sortie peuvent être, selon la nature du modèle, soit des formes sigmoïdes soit des fonctions linéaires.

Lors de ce processus, les signaux sont uniquement propagés vers l'avant, c'est le sens de la flèche tracée sur la Figure 35. Les retours vers des couches supérieures et les transferts au sein d'une même couche de l'information sont interdits par l'algorithme.

3.4.2 Apprentissage

Au cours de l'entraînement d'un réseau FF, les poids sont adaptés en respectant une règle d'apprentissage choisie. La plus courante et la plus compréhensible est basée sur la propagation de l'erreur en sens inverse du signal (*back-propagation* en anglais), de la couche de sortie vers la couche d'entrée.[108] L'apprentissage du réseau consiste à présenter itérativement les couples d'exemples au réseau et à modifier les poids en fonction de la différence entre les vecteurs de sortie produits **o** et ceux attendus **t**. Il s'agit donc d'un problème d'optimisation : trouver le minimum d'une surface multidimensionnelle par une approche basée sur des calculs de gradient. Cette surface représente, dans l'espace des poids, l'erreur de modélisation du réseau.

Algorithme de back-propagation

Les principales étapes de l'algorithme peuvent être résumées comme suit.[115] Hormis l'étape d'initialisation, elles sont répétées pour toutes les données d'entraînement et pour le nombre d'itérations nécessaires à la convergence.

- Initialiser les poids à des valeurs faibles autour de zéro.
- Calculer pour les s neurones j la différence entre la valeur attendue et la valeur obtenue. Cette erreur ε est une fonction des poids.

$$\text{ÉQUATION 19} \qquad \varepsilon = \frac{1}{2}\sum_{j=1}^{s}\left(t_j - o_j\right)^2 .$$

t_j et o_j respectivement les valeurs de sortie attendue et obtenue pour le neurone j.

- Effectuer les corrections Δw_{ji} des poids du neurone j, la Figure 36 rappelle les notations utilisées :

Équation 20 $\Delta w_{ji} = \eta \delta_j o_i$

η est une constante de [0,1] appelée taux d'apprentissage.

δ_j est un terme basé sur le gradient $\frac{\partial \varepsilon}{\partial \omega_{ji}}$ de la fonction d'erreur qui dépend de l'index l du niveau.

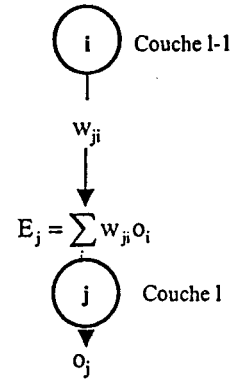


Figure 36 : Notations utilisées.

• Le gradient de la fonction d'erreur, représenté schématiquement Figure 37, est calculé comme suit :

1- Pour la couche de sortie :

Équation 21 $\delta_j = (t_j - o_j) \cdot f_j'(E_j)$.

$E_j = \sum_i w_{ji} x_{ji}$ pour le neurone j.

f_j' représente la dérivée de la fonction d'activation en E_j.

2- Pour une couche cachée :

Équation 22 $\delta_j = f_j'(E_j) \cdot \sum_k \delta_k w_{jk}$

k se réfère à un neurone de la couche suivante.

La présentation de toutes les données au réseau constitue une itération.

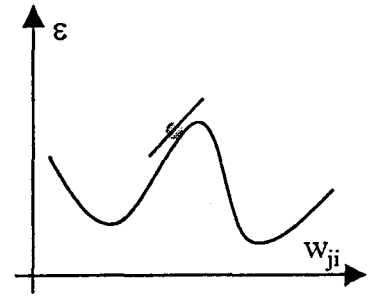


Figure 37 : Profil de la surface d'erreur sur w_{ji}.

Le taux d'apprentissage η, qui intervient dans l'Équation 20, détermine la performance de l'entraînement dont deux comportements typiques sont représentés Figure 38. Si la valeur de η n'est pas assez élevée, la convergence vers un optimum est lente et risque d'être interrompue par un minimum local de la fonction d'erreur. Au contraire, si ce paramètre reçoit une valeur excessive, le système a tendance à osciller.

Le choix d'une valeur appropriée du taux d'apprentissage est déterminé par l'application.

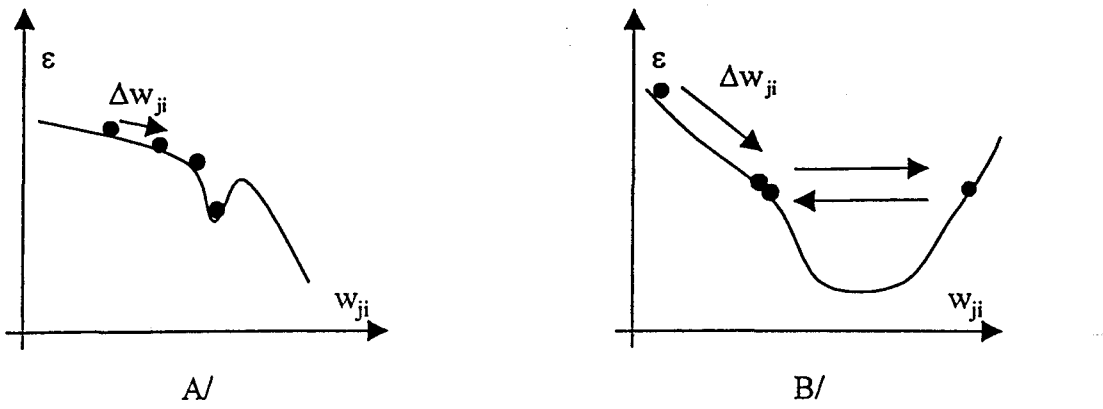


Figure 38 : Choix du taux d'apprentissage : A/ η trop faible. B/ η trop élevé.

Vers des algorithmes plus performants

La plupart des algorithmes existants pour l'apprentissage des RNA-FF sont des variantes de l'algorithme de *back-propagation* standard.
La forme généralisée de ce dernier contrôle les oscillations de la Figure 38B/ par le biais d'un terme de moment μ. Il permet d'adapter la taille du pas de correction Δw_{ji} en fonction de l'erreur obtenue au cycle précédent :

$$\text{Équation 23} \quad \Delta w_{ji}(t+1) = \eta \delta_j o_i + \mu \Delta w_{ji}(t).$$

Ce terme de moment accélère donc la convergence et rend éventuellement la procédure d'apprentissage plus stable. Néanmoins, tout comme pour η, les valeurs optimales de μ dépendent des données.

Algorithme de resilient propagation

Un des algorithmes récents très efficaces est connu sous le nom de RPROP[117] (propagation élastique ou *resilient propagation* en anglais). Il a été inventé pour corriger un inconvénient notoire des algorithmes présentés précédemment.
L'amplitude de la correction Δw_{ji}, quelle soit décrite par l'Équation 20 ou par l'Équation 23, reste fonction de la valeur de la dérivée partielle. Cela signifie qu'une irrégularité de la surface

[117] M. Riedmiller, H. Braun, *Proceedings of the IEEE International Conference on Neural Networks*, (1993) p. 56

d'erreur ε peut donner lieu à des comportements inattendus. Pour s'en affranchir, l'optimisation des poids effectuée par l'algorithme RPROP ne se base pas sur les valeurs des dérivées partielles mais plutôt sur leurs signes.

Pour chaque itération t, on répète les procédures suivantes :

• Calculer les amplitudes des modifications au rang t, notées $\Delta_{ji}(t)$:

Équation 24

$$\Delta_{ji}(t) = \begin{cases} \eta^{+}\Delta_{ji}(t-1) & \text{si } \dfrac{\partial\varepsilon(t-1)}{\partial w_{ji}} \cdot \dfrac{\partial\varepsilon(t)}{\partial w_{ji}} > 0; \\ \eta^{-}\Delta_{ji}(t-1) & \text{si } \dfrac{\partial\varepsilon(t-1)}{\partial w_{ji}} \cdot \dfrac{\partial\varepsilon(t)}{\partial w_{ji}} < 0; \\ \Delta_{ji}(t-1) & \text{si } \dfrac{\partial\varepsilon(t-1)}{\partial w_{ji}} \cdot \dfrac{\partial\varepsilon(t)}{\partial w_{ji}} = 0. \end{cases}$$

$$0<\eta^{-}<1<\eta^{+}$$

Chaque fois que la dernière modification des poids a été trop importante, la dérivée partielle de la fonction d'erreur change de signe. Cela signifie qu'un minimum local a été survolé et l'amplitude de la modification est alors réduite par le facteur η^{-} inférieur à 1. Dans le cas contraire, la convergence est accélérée par η^{+}, supérieur à l'unité.

• Corriger les poids :

Équation 25

$$\Delta w_{ji}(t) = \begin{cases} -\Delta_{ji}(t) & \text{si } \dfrac{\partial\varepsilon(t)}{\partial w_{ji}} > 0; \\ -\Delta_{ji}(t) & \text{si } \dfrac{\partial\varepsilon(t)}{\partial w_{ji}} < 0; \\ \Delta_{ji}(t-1) & \text{si } \dfrac{\partial\varepsilon(t)}{\partial w_{ji}} = 0. \end{cases}$$

Un avantage supplémentaire de l'algorithme RPROP provient de la minimisation d'une fonction d'erreur plus générale que la somme des moindres carrés. Elle est décrite par l'Équation 26.

Équation 26 $$\varepsilon = \sum_{j}\left(t_j - o_j\right)^2 + 10^{-\alpha}\sum_{j} w_{ji}^2 .$$

α est un coefficient positif. Il assure l'affaiblissement systématique des poids les moins utiles. Cela permet d'obtenir un réseau plus structuré,[101] un apprentissage plus uniformément réparti sur le réseau, [117,118] ainsi que des performances de calcul supérieures.

3.4.3 Aspects méthodologiques

Bien que les RNA ne nécessitent aucune supposition sur les données, ils ne peuvent être appliqués qu'à des lots de données suffisamment importants. Or, dans la plupart des applications réelles, on doit se contenter d'un nombre fixé d'exemples et on ne peut en conséquence pas utiliser un réseau délibérément large.
Pour améliorer les capacités de généralisation[119] d'un réseau, il faut optimiser le rapport entre le nombre de poids et le nombre d'échantillons. Cela revient à trouver un compromis au dilemme biais-variance[120] et à appliquer le principe de parcimonie.[121,122] Ce dernier privilégie le modèle de plus basse complexité permettant de modéliser suffisamment correctement les données.[123]

Assurer la généralisation

On parle de capacité de généralisation lorsque le modèle d'étalonnage construit reste approximativement correct pour des données qui n'ont pas participé à l'entraînement. Cela nécessite trois conditions.

- Les données utilisées pour l'apprentissage doivent évidemment contenir suffisamment d'informations pertinentes pour que la relation existant entre les entrées et les sorties puisse être correctement établie.
- La fonction inconnue que l'on cherche à modéliser doit être continue et relativement lisse.

Une légère variation des données d'entrée ne doit avoir que de faibles conséquences sur les données de sortie.

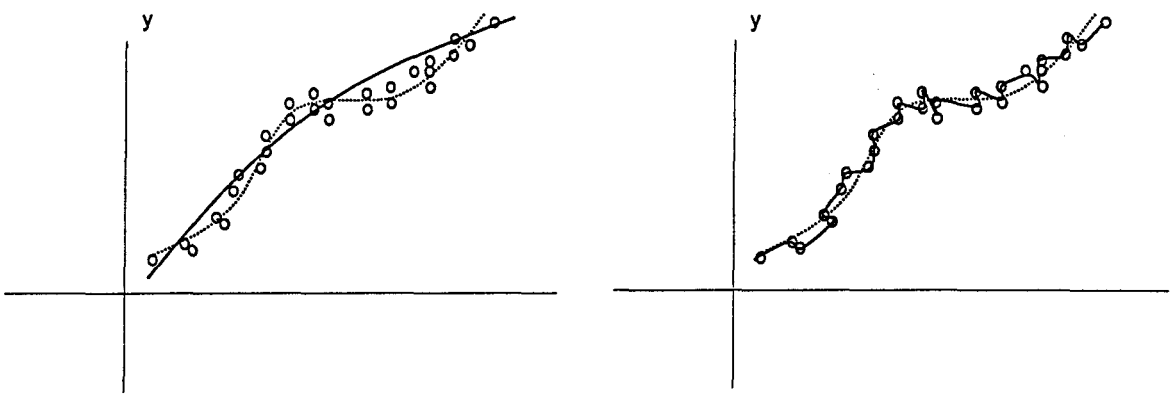
---

[118] M. Riedmiller, *Computer Standards & Interfaces*, 16 (1994) p. 265
[119] A. Wieland, R. Leighton, *1st IEEE International Conference on Neural Networks*, 3 (1987) p. 387
[120] S. Geman, E. Bienenstock, R. Doursat, *Neural Computation*, 4 (1992) p. 1
[121] P. Stoïca, T. Söderström, *International Journal of Control*, 36 (1982) p. 404
[122] M. B. Seasholz, B. R. Kowalski, *Analytica Chimica Acta*, 277 (1993) p. 165
[123] R. Q. Yu, J. H. Jiang, *Chemometrics & Intelligent Laboratory Systems*, 45 (1999) p. 191

• Le lot de données utilisé pour l'apprentissage doit être suffisamment fourni et le plus représentatif possible.

Cela permet d'éviter les extrapolations lors de la phase de prédiction. Il existe des critères qui estiment la borne inférieure du nombre d'échantillons suffisant pour obtenir des capacités de généralisation,[124] même lorsque la dimension spatiale des données d'entrée est importante.[125]

Les capacités de généralisation sont intimement liées aux problèmes de sous- et sur-adaptation (*underfitting* et *overfitting* en anglais) représentés Figure 39, la fonction en pointillés simulant une modélisation correcte du lot de données.



A/ B/

Figure 39 : Modèle A/ sous-adapté. B/ sur-adapté.

Un réseau sous-adapté n'est pas suffisamment complexe pour détecter la structure de la relation entrée-sortie propre au lot de données particulier. Cela donne lieu à une représentation du type de la Figure 39A/. Au contraire, le phénomène de sur-adaptation concerne les RNA dont la structure est trop complexe par rapport à l'information intrinsèque. Ceux-ci disposent de suffisamment de degrés de liberté pour modéliser le bruit en plus du signal (Figure 39B/). Il s'agit probablement de l'aspect le plus sournois de l'entraînement des RNA. En effet, l'erreur d'entraînement associée à ces modèles peut prendre des valeurs idéalement faibles même si les capacités de prédiction sont mauvaises.

Une approche rigoureuse de ce problème est formulée par le compromis biais-variance[120] dont l'étude théorique est encore d'actualité.[126] Le problème se résume par l'écriture de l'erreur

[124] D. R. Hush, B. G. Horne, *IEEE Signal Processing Magazine*, 1 (1993) p. 8
[125] J. L. Yuan, T. L. Fine, *IEEE Transactions on Neural Networks*, 9 (1998) p. 266
[126] M. Faber, *Journal of Chemometrics*, 13 (1999) p. 185

quadratique moyenne de prédiction sous la forme d'une somme de deux termes :[52]

$$\text{Équation 27} \quad E(y-\hat{y})^2 = E(\hat{y}-E(\hat{y}))^2 + (E(\hat{y})-y)^2 .$$

Le premier terme correspond à la variance du modèle de régression, c'est-à-dire au degré d'incertitude qui lui est associé, alors que le second représente la valeur quadratique du biais. Si l'on trace les contributions respectives de ces deux termes à l'erreur de prédiction en fonction de la complexité du modèle, c'est à dire du nombre de neurones cachés d'un réseau FF, on obtient la Figure 40.

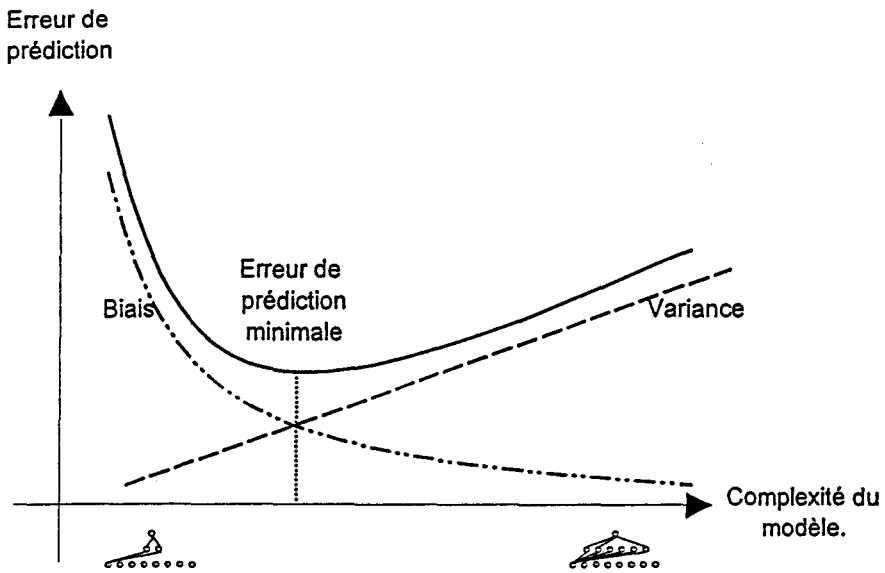


Figure 40 : Evolution de l'erreur de prédiction en fonction de la complexité du modèle.

Lorsque la complexité du modèle augmente, de plus en plus d'informations sont prises en compte par le modèle et la contribution du biais diminue. Néanmoins, il est nécessaire d'estimer un nombre croissant de paramètres ce qui, pour un nombre d'exemples constant, induit inévitablement une augmentation de la variance. Par répétitions successives de l'entraînement pour des modèles plus ou moins complexes, on parvient à définir un optimum qui minimise l'erreur quadratique moyenne de prédiction.

Principe de parcimonie

Afin de satisfaire aux conditions de généralisation, il faut rendre le quotient du nombre d'échantillons sur le nombre de paramètres le plus élevé possible. La dimension du vecteur

représentant les données d'entrée est donc un facteur limitant, connu sous le nom de *curse of dimensionality*, en anglais. Pour représenter un espace d'entrée sur un espace de sortie, il faut donc disposer d'un nombre de neurones proportionnel à la dimension de l'espace d'entrée. Cela impose, pour la majorité des problèmes, un nombre d'échantillons élevé.

Lorsque le nombre d'échantillons est fixé par l'application, la seule parade consiste à réduire au maximum le nombre de neurones d'entrée du réseau. Cela se fait, soit en projetant préalablement les données d'entrée dans un sous-espace représentatif de l'espace de départ, soit en sélectionnant les variables les plus représentatives ou les plus pertinentes de celles-ci. La transformation des données d'entrée est quasiment incontournable. Ce sont souvent des combinaisons de ces deux solutions qui sont utilisées dans la pratique et qui permettent d'éliminer une proportion non négligeable d'information insignifiante. On procède généralement à partir d'une couche d'entrée délibérément large que l'on réduit, essai après essai.

Chaque algorithme de sélection présente ses propres limitations.[127] Par ailleurs, toute méthode de transformation consiste en un changement de base qui permet de modifier la représentation des données, c'est à dire de calculer un nouveau jeu de coefficients, dans le but de la rendre plus adaptée aux traitements ultérieurs. Pour assurer la compression des données, la représentation du signal est généralement tronquée de telle sorte que le maximum d'information significative soit conservé. La méthode de transformation la plus utilisée, depuis de nombreuses années, semble être l'analyse en composantes principales.[5] Les composantes construites, combinaisons linéaires de variables initiales, ont un caractère global. Des analyses de Fourier ou des transformations d'Hadamard sont également envisageables.[128] Récemment, un grand nombre d'études dans le domaine de la chimie ont prouvé l'utilité de la méthode des ondelettes (*wavelets* en anglais),[129,130] notamment pour la compression des spectres infrarouges en vue de l'utilisation de méthodes neuronales.[131] Cette transformation fournit des caractéristiques locales du signal qui peuvent s'avérer très utiles pour les problèmes de reconnaissance de motifs ou d'étalonnage multivarié.

---

[127] F. Despagne, D. L. Massart, *Chemometrics & Intelligent Laboratory Systems*, 40 (1998) p. 145
[128] L. Dolmatova, V. Tchistiakov, C. Ruckebusch, N. Dupuy, J. P. Huvenne, P. Legrand, *Journal of Chemical Information & Computer Science*, 39 (1999) p. 1027
[129] F. Ehrentreich, S. Nikolov, R. Wolkenstein, H. Hutter, *Mikrochimica Acta*, 123 (1998) p. 221
[130] B. Walczak, D. L. Massart, *Trends in Analytical Chemistry*, 16 (1997) p. 451
[131] M. Bos, J. A. M. Vrielink, *Chemometrics & Intelligent Laboratory Systems*, 23 (1994) p. 115

Analyse en composantes principales

L'analyse en composantes principales est une transformation des données, utilisée pour la réduction du nombre de caractéristiques. Cela s'effectue au moyen de combinaisons linéaires des variables initiales. Son principe consiste à exprimer l'information que contiennent les variables $\mathbf{X}=\{x_k,\ k=1...K\}$ par un nombre plus faible de variables $\mathbf{T}=\{t_a,\ a=1...A<K\}$ appelées composantes principales de **X** vérifiant $t_i'.t_j=0$ pour $i\neq j$. Cette correspondance s'effectue par l'intermédiaire de vecteurs qui maximisent la variance empirique contenue au sein de la matrice **X**. Ces vecteurs *loading* $\mathbf{P}=\{p_k,\ k=1...K\}$ vérifient de plus $p_i'.p_j=\delta_{ij}$. On extrait en général plus de facteurs qu'il n'en faut pour construire le modèle, éliminant finalement ceux qui ne sont pas significatifs ou sujets à caution.

Pour des variables centrées, on peut approcher **X** par **TP'** c'est-à-dire écrire **X=TP'+E** lorsque seules quelques composantes principales **T** sont calculées (A<K). Pour chaque composante principale, **P** et **T** sont calculés par les opérations suivantes :

$$\text{Équation 28} \quad \mathbf{P}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}$$

$$\text{Équation 29} \quad \mathbf{T} = \mathbf{XP}(\mathbf{P}'\mathbf{P})^{-1} = \mathbf{XP}$$

Ces formules sont très importantes pour l'écriture d'algorithmes efficaces, tels que celui proposé en annexe 1.

L'analyse en composantes principales est une méthode de projection linéaire décrivant toutes les contributions de la variance. Néanmoins, celles-ci présentent éventuellement des relations non-linéaires aussi bien entre-elles que du point de vue de la variable physico-chimique d'intérêt (Figure 11 A/ et B/ page 50). En ce sens, la PCA présente une méthode de compression efficace des données en vue de la construction de modèles empiriques dont les paramètres sont déterminés non-linéairement. Ainsi, la plupart des applications des RNA en analyse quantitative utilisent les facteurs *score* comme données d'entrée.[75,85,91,127]

Du point de vue des limitations de la méthode, une transformation linéaire n'assure évidemment pas la conservation parfaite de la structure du lot de données si celui-ci présente des contributions non-linéaires.[75] D'autre part, certaines des contributions extraites de la variance ne s'avèrent pas forcément informatives du point de vue de la qualité à prédire.

Présentation de la transformée des données par ondelettes

Les ondelettes sont des fonctions de base permettant la transformation des données et en particulier des données spectrales. Elles se distinguent des transformations plus classiques par leur caractère local, c'est-à-dire qu'elles diffèrent de zéro uniquement dans un domaine limité. Elles peuvent être localisées à différentes positions et sont construites de telle sorte qu'elles sont orthogonales entre elles.

Les ondelettes sont obtenues à partir d'une même fonction par effets de contraction / dilatation (*scaling* en anglais) et par translations (*shift* en anglais). Ces dernières sont nécessaires pour couvrir le domaine complet alors que le *scaling* assure l'analyse 'multi-résolution' du signal. Toutes les fonctions de base sont issues d'une même forme. Elles sont donc identiques les unes aux autres, ce qui leurs confère une structure fractale. Une forme mathématique en adéquation avec ces propriétés peut s'écrire :

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right)$$ où a est une variable de *scaling* et b une variable de *shift*.

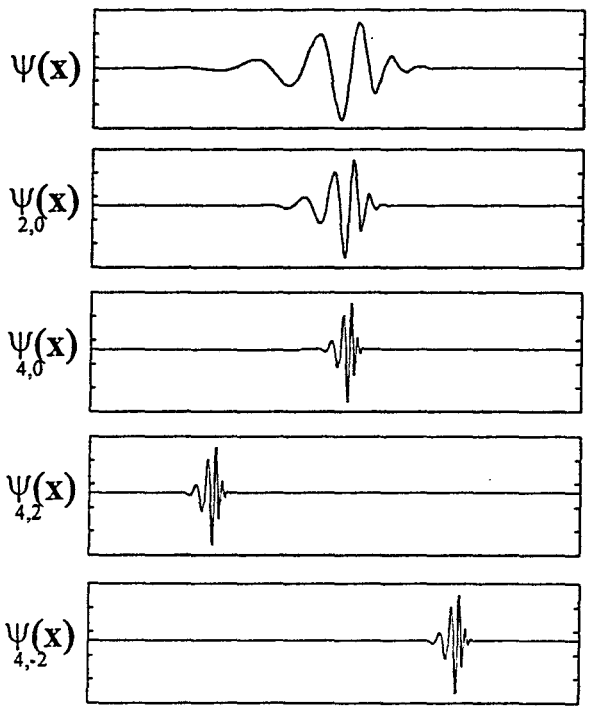


Figure 41 : Formes d'ondelettes issues d'une fonction de Daubechies d'ordre 8.

On peut ainsi utiliser n'importe laquelle ou plusieurs des fonctions pour représenter les caractéristiques d'un signal. De plus, en considérant des niveaux différents d'une même

famille, on peut approcher le signal avec le degré de précision voulu.

Il existe un grand nombre de types d'ondelettes, et chaque base de fonctions est indirectement spécifiée par un lot de coefficients. Une fois qu'une ondelette a été définie, les coefficients qui lui sont associés sont utilisés pour définir deux filtres, un filtre passe-bas et un filtre passe-haut. Le premier membre de la famille des ondelettes de *Daubechies*[132] est ainsi caractérisé par deux coefficients alors que le $n^{ième}$ comporte 2n coefficients. Les filtres se partagent les mêmes coefficients, mais avec des signes opposés et dans l'ordre inverse. L'algorithme hiérarchique de *Mallat*,[133] présenté Figure 42, est une approche très pédagogique de la méthode de transformation discrète des wavelets (DWT). Une fois les filtres construits, cet algorithme effectue une décomposition 'pyramidale' du signal et en propose une représentation hiérarchique et 'multi-résolution'.

Chaque signal traverse un filtre passe-haut et un filtre passe-bas et les sorties produites sont réduites d'un facteur 2 (*down sampling* en anglais). A chaque niveau, le lot de données issu du filtre passe-haut contient des coefficients caractéristiques des détails du signal, alors que celui issu du passe-bas contient les coefficients dits d'approximation. Ces derniers peuvent à nouveau être soumis à la même paire de filtres et la décomposition se poursuit. Plus le niveau de l'échelle est bas, moins il y a d'information dans le signal d'approximation. Autrement dit, c'est dans les coefficients de détail du niveau en question qu'est contenue la différence d'information significative entre deux niveaux d'approximation consécutifs.

[132] I. Daubechies, *Communications in Pure Applied Mathematics*, 41 (1988) p. 909
[133] I. Daubechies, S. Mallat, A. S. Willsky, *IEEE Transactions on Information Theory*, 38 (1992) p. 529

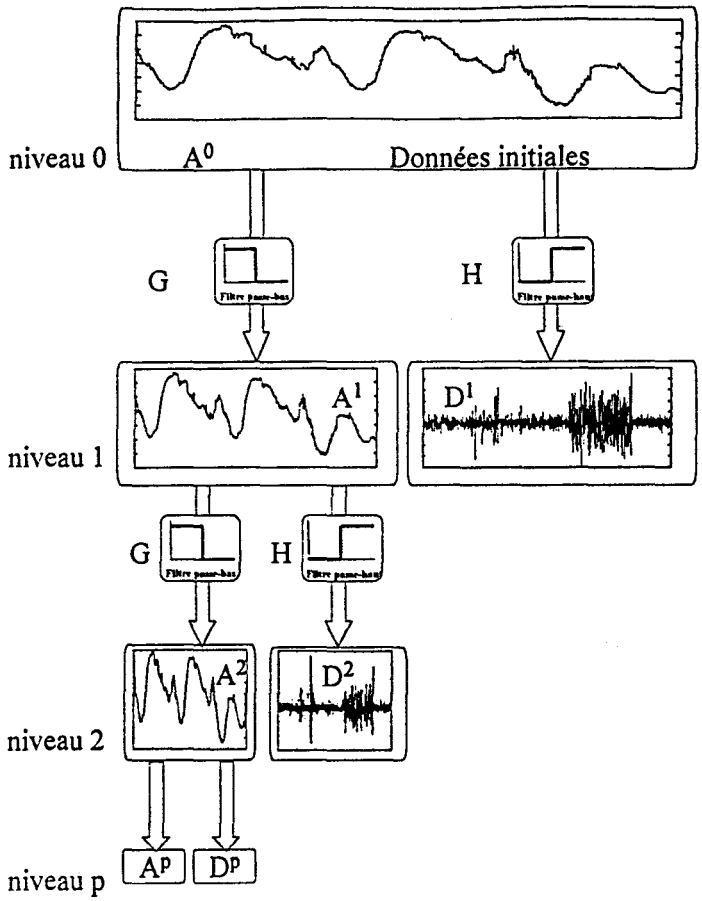


Figure 42 : Décomposition "pyramidale" de Mallat.

Entraîner et tester les réseaux

Des calculs d'erreurs quadratiques moyennes entre valeurs attendues et obtenues permettent de comparer les performances obtenues pour différents réseaux ou différents lots de données. Ce type d'estimateur n'est certes pas le meilleur indicateur de performance,[115] mais il en donne une idée globale suffisante pour la détermination du modèle le mieux adapté.

De manière générale, la construction d'un modèle d'étalonnage nécessite la répartition des données disponibles sur deux lots :

- un lot d'entraînement qui permet d'estimer les paramètres,
- un lot de test, dit aussi de validation, pour juger des capacités de généralisation du modèle construit.

Dans le cas des réseaux FF, il faut maîtriser l'entraînement, afin d'empêcher la sur-modélisation, par le biais d'un troisième lot de données :

- le lot de contrôle, représentatif de la population.

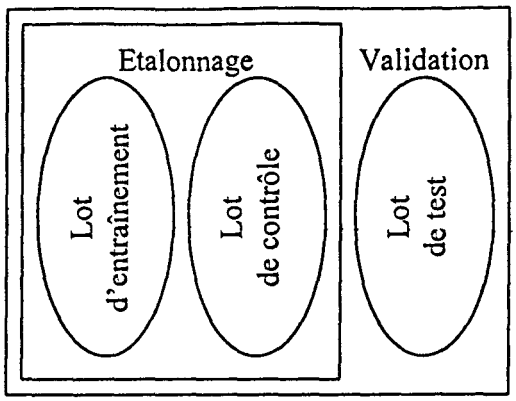


Figure 43 : Lots de données pour la construction d'un modèle de RNA-FF.

Dans le cas idéal, les trois ensembles de la Figure 43 sont indépendants. Cette condition n'est satisfaite que si leurs échantillons sont choisis au hasard. Mais il faut également s'assurer que les principales sources de variance sont incluses dans le lot d'entraînement, afin d'éviter toute extrapolation.

Le suivi de l'entraînement se fait par l'intermédiaire de la visualisation des erreurs produites par les lots utilisés pour l'étalonnage au cours des itérations. La Figure 44 idéalise les situations typiquement obtenues. Les courbes doivent présenter le même comportement, le minimum de la courbe de contrôle imposant l'arrêt de l'entraînement. Le modèle est ensuite validé par le lot de test, estimateur de l'erreur de généralisation. Cette méthode est la plus utilisée, particulièrement lorsque le nombre d'échantillons dont on dispose n'est pas idéalement élevé.[134]

[134] W. S. Sarle, *Proceedings of the 27th Symposium on the Interface of Computing Science and Statisitics*, (1995) p. 352 *ftp ://ftp.sas.com/pub/neural/inter95.ps.Z*

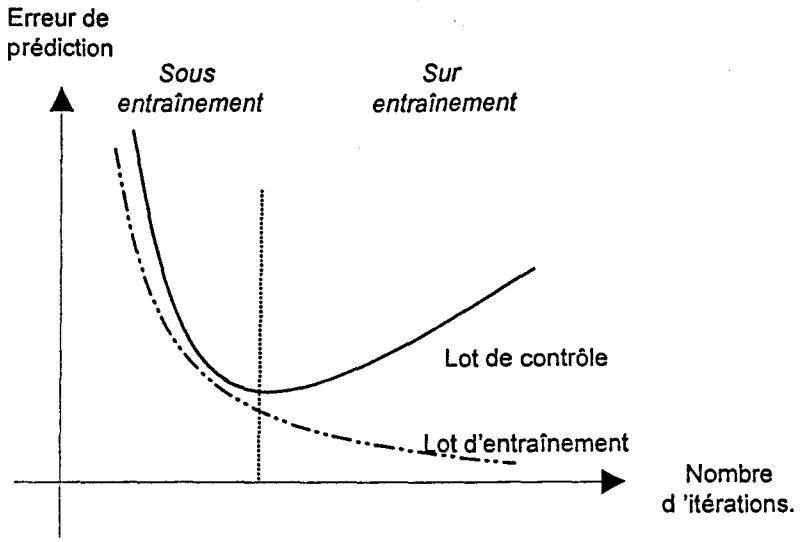


Figure 44 : Contrôle de l'apprentissage d'un réseau FF.

Il s'agit d'une méthode simple et rapide, utilisable pour des réseaux constitués d'un grand nombre de paramètres, décrite dans la littérature par l'anglicisme *d'early stopping*. Il reste à la charge de l'utilisateur de décider de la répartition des échantillons sur les lots utilisés pour l'apprentissage.

Etude expérimentale

Introduction

La cinétique d'apparition des peptides lors d'une hydrolyse pepsique d'hémoglobine bovine peut être caractérisée analytiquement. D'une répétition du procédé à l'autre, des peptides identiques sont élaborés dans le même ordre. Certains d'entre eux, intermédiaires, ont des propriétés ou des fonctions intéressantes et il est important de pouvoir les détecter dans le réacteur. Le degré d'hydrolyse est un bon indicateur de leur présence.

En pratique, chaque réaction évolue avec une cinétique propre, les procédés discontinus ne sont jamais exactement répétables. D'ailleurs, si c'était le cas, une mesure triviale du temps serait suffisante pour connaître l'avancement de la réaction. L'hydrolyse de l'hémoglobine bovine n'est donc pas un processus phénoménologique.[10] En conséquence, la prédiction du déroulement de la réaction n'est envisageable que par le biais d'un modèle expérimental, représentant la connaissance du système et qui permettra d'estimer des répétitions, inconnues du point de vue analytique, du procédé.

Une méthode physico-chimique rapide et non destructive pour la mesure du degré d'hydrolyse constitue un intérêt majeur pour les biologistes, généralement tributaires de mesures de référence longues et coûteuses. Il convient donc d'appréhender l'avancement de la réaction par une mesure globale de l'état du système et propre à ce dernier. La spectrométrie de vibration est une des méthodes analytiques les plus adaptées, mais elle ne peut cependant pas remplir à elle seule la fonction demandée. En effet, les spectres observés sur des systèmes réels sont complexes et rarement directement interprétables. C'est leur étude par des méthodes de chimiométrie qui permet éventuellement de pointer l'information significative. D'autre part, le procédé de contrôle doit être doté de réelles capacités d'adaptation[22] et les réseaux de neurones artificiels sont a priori adaptés à cette tâche. C'est néanmoins par l'examen de la structure des données que ce choix se justifie. Enfin, supervisé ou non, l'apprentissage des systèmes neuronaux requiert rigueur et connaissances.

Il est ainsi possible d'envisager le remplacement de mesures complexes et coûteuses par une instrumentation plus intelligente, bien qu'initialement moins adaptée, qui réalise le meilleur compromis possible. La chimiométrie est l'outil nécessaire pour permettre le passage de la mesure globale au capteur finalisé.

Toute étude de faisabilité, préalable au développement de méthodes dédiées sur le plan

instrumental, doit s'accommoder des conditions actuelles du procédé. Le chapitre 4 présente le contexte et évalue la possibilité d'un suivi de la réaction à partir des spectres de vibration d'échantillons prélevés. Divers niveaux de traitement sont envisagés. Indépendamment, nous considérons les effets de solvant sur la conformation, garante de l'activité des molécules biologiques. La combinaison des résultats mène finalement à proposer une interprétation de l'information spectrale.

Le contrôle en ligne, à proprement parler, est présenté dans le détail au cours du chapitre 5. Le spectromètre IRTF est, par l'intermédiaire d'une fibre optique, couplé à un réacteur de laboratoire qui fonctionne en cycles ouverts. Le travail consiste, en premier lieu, à rechercher une représentation de la connaissance du système par un modèle à partir des couples formés par les spectres et les degrés d'hydrolyse, mesurés lors de répétitions semblables (mais jamais identiques). Dans un deuxième temps, le but est bien entendu la prédiction du degré d'hydrolyse permettant la caractérisation des produits lors de répétitions qui n'ont pas participé au processus d'étalonnage. La méthodologie mise en place est détaillée dans son ensemble et les résultats sont commentés.

Chapitre 4

Faisabilité d'un suivi d'hydrolyse d'hémoglobine bovine par spectrométrie infrarouge

Une étude de faisabilité poursuit plusieurs objectifs, mais ne doit pas mettre en œuvre de modifications importantes des conditions opératoires avant d'en avoir démontré l'utilité. Ce type d'étude doit procurer, avec un délai raisonnable, un résultat tranché et justifié. Cependant, les conditions de mesure ne relèvent généralement pas d'un choix de l'analyste. Nous envisageons ici la faisabilité d'un suivi sur deux lots d'échantillons prélevés qui nous sont fournis ; nous nous accommodons d'échantillons contenant une quantité faible de protéine.

Dans un premier temps, la caractérisation des protéines permise par la spectrométrie de vibration est présentée. Ce sont les fréquences de vibration provenant des liaisons amide de l'hémoglobine qui sont responsables de l'allure des spectres observés. Ces vibrations sont par ailleurs intimement liées à la conformation de la molécule qui décide de son activité biologique.
La réaction d'hydrolyse d'hémoglobine est ensuite abordée du point de vue de ses mécanismes, de ses enjeux, des matériels et des méthodes.
Enfin, nous détaillons l'étude de faisabilité sur échantillons prélevés. Par la mise en œuvre de méthodes simples, on montre les potentialités de la chimiométrie pour le suivi et l'interprétation de la réaction.

4.1 Spectroscopie de vibration des protéines

La spectroscopie de vibration est une des techniques utilisées pour déterminer la structure et la dynamique des molécules biologiques. Elle permet entre autres de suivre les changements de conformation des peptides et des protéines dans leurs environnements naturels.
Les applications ont longtemps été considérées comme de véritables challenges,[135] l'eau étant le solvant de prédilection des molécules biologiques. Récemment, *Huhmer et al.*[136] ont énuméré une trentaine d'études, entre 1995 et 1997, utilisant la spectroscopie infrarouge à transformée de Fourier. Un tel développement est justifié par la quantité d'informations qu'il est possible d'inférer des spectres de vibration ainsi que l'avènement de nouvelles techniques instrumentales et de capacités performantes de calcul.

4.1.1 La liaison peptidique

Les peptides et les protéines sont des molécules biologiques de première importance composées d'acides aminés. La structure des acides aminés est présentée Figure 45. Il existe 20 acides aminés dont les groupes R latéraux peuvent être hydrophiles, hydrophobes ou même ioniques.

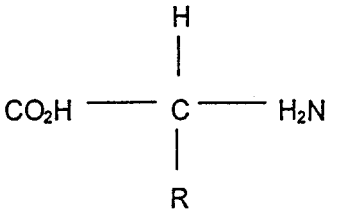


Figure 45 : Structure des acides aminés.

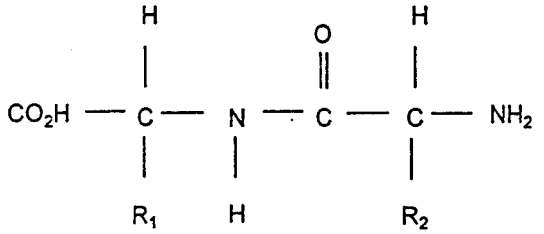


Figure 46 : La liaison peptidique.

Les protéines sont des polymères dont les acides aminés sont connectés par la liaison peptidique présentée Figure 46.

[135] M. Diem, *Introduction to Modern Vibrational Spectroscopy*, J. Wiley & Sons (1993)
[136] A. F. R. Huhmer, G. I. Aced, M. D. Perkins, R. N. Gursoy, D. S. S. Joys, C. Larive, T. J. Siahaan, C. Schoneich, *Analytical Chemistry*, 69 (1997) p. 29R

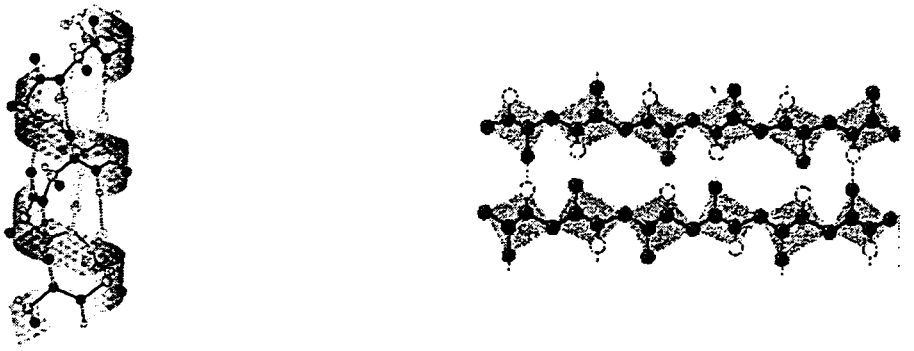
4.1.2 Structure d'une protéine

Structure primaire

Des centaines, voire des milliers d'acides aminés sont liés entre eux suivant une séquence propre à chaque macromolécule. Cet arrangement est déterminé par le code génétique et on s'y réfère sous le nom de structure primaire. Néanmoins, les propriétés fonctionnelles de la molécule sont déterminées par la forme tridimensionnelle prise par les acides aminés à l'issue du processus de repliement.[137]

Structure secondaire

Au sein de la forme repliée de protéines, il existe des motifs caractéristiques, appelés structures secondaires ou conformations. Lorsque le peptide est enroulé sous une forme hélicoïdale droite, stabilisée par des liaisons hydrogène, on parle d'hélice α. D'autres formes, comme les coudes, sont stabilisées par des liaisons covalentes, des interactions entre les résidus latéraux chargés ou des interactions hydrophobes ou hydrophiles avec le solvant. Les structures secondaires les plus connues sont représentées sur la Figure 47.



A/ B/

Figure 47 : Structures secondaires : a/ hélice α b/ feuillets β antiparalleles.

[137] T. E. Creighton, *Proteins – Structure and Molecular Properties*, Library of Congress Cataloguing (1993)

4.1.3 Modes de vibration

Lorsqu'on analyse les spectres de vibration provenant de molécules dont les masses avoisinent les 100 000 daltons, le nombre de degrés de liberté écarte naturellement les méthodes courantes d'interprétation de spectres. Il convient d'étudier des composés modèles auxquels il sera possible de se référer.

Vibrations amides

Le niveau énergétique des interactions mises en jeu par la structure secondaire des protéines provient de la stéréochimie de la liaison peptidique présentée Figure 48. Si des résidus successifs de la chaîne polypeptidique ont des orientations similaires, la molécule prend alors une structure symétrique.

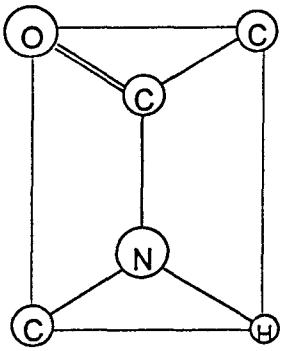


Figure 48 : Stéréochimie de la liaison peptidique dans les protéines.

Les fréquences de vibration correspondant à la liaison amide, observées dans toutes les protéines, ont été analysées en détail par les premiers calculs de modes normaux[138] et sont regroupées dans le Tableau 2.

Certaines de ces vibrations peuvent être affectées par les conformations.[139,140] Pour les analyses de structures secondaires, les vibrations amide I, II et III de la liaison peptidique ont un intérêt particulier. La vibration amide I est principalement attribuée à l'élongation C=O observée à 1653 cm^{-1}. Cette fréquence est très sensible à la conformation et peut varier de 50

[138] T. Miyazawa, T. Shimanouchi, S. Mizushima, *Journal of Chemical Physics*, 29 (1958) p. 621
[139] R. C. Loord, *Appied Spectroscopy*, 31 (1977) p. 187
[140] G. J. Thomas Jr, *Vibrational Spectra and Structure*, Elsevier Science (1975)

cm^{-1} entre les différentes structures. Pendant longtemps, cette contribution n'a été observée que dans les milieux non aqueux à cause de la contribution importante de l'eau située à 1620 cm^{-1}. Les vibrations amide II et III peuvent aussi être utilisées, mais leurs interprétations sont complexes.

Dénomination	Nombres d'onde (cm^{-1})	Description
Amide A	3250-3300	Elongation N-H
Amide I	1630-1700	Elongation C=O (C-N, C-N-H)
Amide II	1510-1570	Déformation C-N-H (C-C-N)
Amide III	1230-1330	Déformation C-N-H/C-N-C
Amide IV	630-750	Déformation O=C-N
Amide V	700-750	Déformation C-N-H hors du plan
Amide VI	~600	Déformation hors C=O-C du plan

Tableau 2 : Fréquences observées et contributions pour les vibrations amide.

Sensibilité à la conformation du mode amide I

Du fait du moment dipolaire relativement important associé à la liaison peptidique, les différents groupes carbonyles d'une protéine sont couplés. Il en résulte qu'en fonction de la structure secondaire, des absorptions différentes sont observées et résumées au sein du Tableau 3.

Structure Secondaire	Vibration amide I (cm^{-1})
Hélices α	1645-1660
Feuillets β	1665-1680
Coudes β	1640-1690

Tableau 3 : Fréquences de la vibration amide I pour différentes structures secondaires[135].

C'est donc la région amide I qui a été la plus utilisée en tant qu'indicateur qualitatif de la conformation des peptides et des protéines en solution. Le mode amide I est impliqué dans

toutes les structures secondaires par l'intermédiaire des liaisons hydrogènes. [141]
Néanmoins, les fréquences de vibration dépendent de la structure moléculaire exacte et ces valeurs ne peuvent être utilisées qu'à titre indicatif.

Sensibilité du mode amide I

Les biologistes sont intéressés par la composition d'une protéine en terme de structure secondaire. En effet, des changements de compositions structurales peuvent être corrélés à une perte d'activité biologique. La détermination des proportions relatives des diverses conformations[142] est donc un travail important. Ce problème doit néanmoins être abordé prudemment car des facteurs externes à la liaison peptidique, comme les groupes latéraux ou le solvant, affectent la vibration amide I.
Dans les protéines coexistent plusieurs structures secondaires ; des segments hélicoïdaux sont connectés la plupart du temps par des chaînes peptidiques désordonnées. Les spectres observés sont donc constitués de bandes larges résultant de la superposition de vibrations amide I spécifiques. En fait, une protéine présente une bande d'absorption d'environ 50 cm^{-1} de largeur à mi-hauteur, sur laquelle aucune contribution n'est directement perceptible. Même si les méthodes de déconvolution ont été largement employées,[143,144] les interprétations proposées sont souvent à considérer avec précaution.

4.1.4 L'hémoglobine bovine

Origine

Parmi les co-produits animaux, le sang issu des abattoirs représente une source qualitativement et quantitativement très riche en protéines (150 grammes de protéines par litre de sang). Celui-ci est composé d'une part du plasma, et d'autre part du cruor. Le plasma, de par ses propriétés coagulantes et émulsifiantes, représente la fraction traditionnellement la

[141] S. Krimm, *Biopolymers*, 22 (1983) p. 217
[142] K. Rahmelow, W. Hübner, *Analytical Biochemistry*, 241 (1996) p. 5
[143] R. W. Sarver, A. R. Frieman, T. J. Thamann, *Spectrochimica Acta*, Part A 53 (1997) p. 1889
[144] D. M. Byler, H. Susi, *Biopolymers*, 25 (1986) p. 469

mieux valorisée. C'est néanmoins le cruor, obtenu après centrifugation et principalement constitué d'hémoglobine, qui contient environ 70% des ressources protéiques.
L'hémoglobine bovine constitue une source importante de protéines de bonne qualité insuffisamment valorisée.

Composition

L'hémoglobine bovine, de masse moléculaire 64500 Dalton, est composée de 4 chaînes polypeptidiques. La partie protéique de la molécule est la globine, formée de deux chaînes α et deux chaînes β contenant respectivement 141 et 146 acides aminés. Les deux zones sombres de la Figure 49 correspondent aux chaînes β, les niveaux de gris plus faible représentant les chaînes α. Chacune de ces chaînes contient un groupement auquel l'oxygène peut se lier.

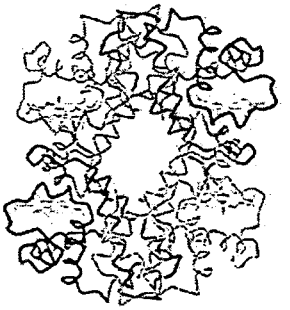


Figure 49 : Disposition des atomes de carbone d'une molécule d'hémoglobine.[145]

Absorption de l'hémoglobine bovine dans l'infrarouge

La Figure 50 représente un spectre caractéristique de l'hémoglobine bovine dans son solvant et un spectre du solvant seul. Les deux principales absorptions de la protéine, autour de 1660 cm^{-1} et 1550 cm^{-1}, correspondent respectivement aux bandes amide I et amide II dont l'interprétation chimique a été détaillée dans le Tableau 2.

[145] W. L. Nichols, G. D. Rose, J. Hopkins, L. F. Ten Eyck, B. H. Zimm *http ://www.sdsc.edu/IOTW/week26/iotw.html*

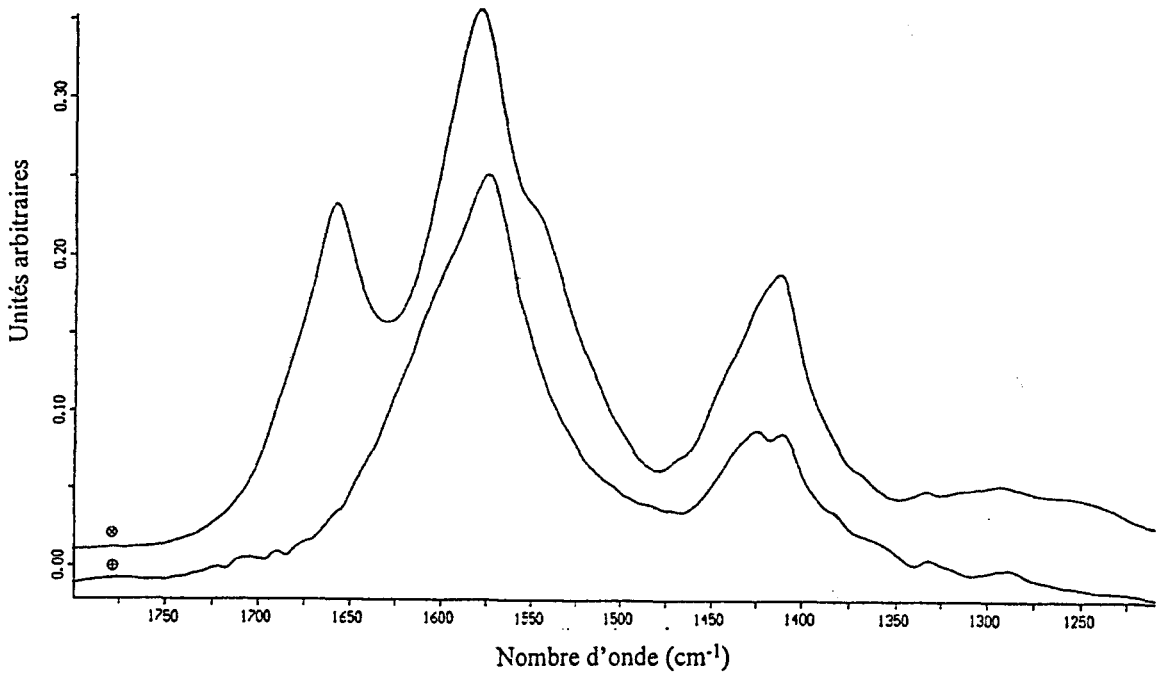


Figure 50 : Spectre de vibration de l'hémoglobine $^{\otimes}$, spectre du solvant acide acétique /acétate de sodium$^{\oplus}$.

La vibration amide I, la bande d'absorption centrée à 1656 cm-1, est caractérisée par la somme des contributions des structures secondaires de l'hémoglobine bovine et c'est l'hélice α qui est largement majoritaire. Ce résultat est en accord avec les études de structure à l'état solide.[146] La forme hélicoïdale représente jusqu'à 87% de la structure secondaire. En outre, le Tableau 3 rappelle que les composantes haute fréquence de la bande amide I sont associées aux coudes de la forme tridimensionnelle, alors que les nombres d'onde de l'intervalle 1620-1650 cm^{-1} sont corrélés à la proportion de structures désordonnées.

[146] J. S. Richardson, *Advances in Protein Chemistry*, 34 (1981) p. 167

4.2 L'hydrolyse de l'hémoglobine bovine

Les hydrolysats de protéines sont depuis longtemps valorisés dans le secteur agro-alimentaire pour leurs propriétés organoleptiques et fonctionnelles. Plus récemment, de nouveaux débouchés sont apparus nécessitant une meilleure définition et un meilleur contrôle des produits.

4.2.1 Enjeux

De nombreuses recherches ont fait de l'hémoglobine un des substrats protéiques les mieux connus. A cet intérêt académique s'ajoute le fait qu'il s'agisse d'un sous-produit industriel disponible en grande quantité. La tendance actuelle est donc à la production maîtrisée de peptides possédant des propriétés spécifiques pour une utilisation en nutrition clinique[147] ou dans les cultures cellulaires.[148] Néanmoins, des applications dans des domaines aussi sensibles nécessitent de disposer d'hydrolysats reproductibles et parfaitement définis du point de vue de la nature des peptides produits. L'extraction de ces peptides est effectuée par des techniques de fractionnement et de purification assez longues.
Ainsi, les méthodes enzymatiques représentent une voie naturelle, physiquement et chimiquement douce, de transformation de l'hémoglobine. Il est en général possible d'envisager des extrapolations à l'échelle pilote et des applications industrielles sont attendues sous peu.

4.2.2 La réaction d'hydrolyse

Les hydrolysats pepsiques de l'hémoglobine bovine sont, malgré leur complexité, totalement caractérisés du point de vue de la cinétique d'apparition des peptides et de la composition peptidique. Celle-ci pourrait ainsi être anticipée dans le cas, idéal, où la réaction est parfaitement reproductible. Les propriétés des produits sont alors toujours les mêmes à un

[147] A. Leke, J. M. Piot, F. Sannier, D. Guillochon, J. P. Ricard, J. P. Postel, J. P. Risbourg, J. P. Canarelli, *Nutrition Clinical Metabolism*, 4 (1990) p. 223
[148] D. Dive, J. M. Piot, F. Sannier, D. Guillochon, P. Charet, S. Lustrat, *Enzyme & Microbial Technology.*, 11 (1989) p. 165

instant donné.

Dans la pratique, des hydrolyses aux conditions initiales a priori semblables ne donnent jamais des cinétiques d'apparition des peptides identiques même si les états intermédiaires d'intérêts sont toujours produits. Ces changements sont probablement imputables aux matières premières naturelles utilisées, à des conditions d'hydrolyse ou à des conditions environnementales légèrement différentes. Il faut donc appréhender l'avancement de la réaction par une mesure globale de l'état du système et propre à ce dernier.

Degré d'hydrolyse

L'évolution des propriétés des peptides issus de la protéine hydrolysée est corrélée aux valeurs du degré d'hydrolyse désigné par la notation DH. Il s'agit d'un paramètre représentant le quotient du nombre de liaisons peptidiques clivées au nombre total de liaisons peptidiques.

L'intérêt de suivre une réaction d'hydrolyse par détermination du DH a été étudié en détails par *Adler-Nissen.*[149] En résumé, toute réaction d'hydrolyse fait intervenir un grand nombre de paramètres physico-chimiques. Hormis le pH, le degré d'hydrolyse permet à lui seul le contrôle de tous les autres paramètres. Cela signifie que le procédé est efficacement contrôlé par le suivi du DH, qu'il devient indispensable de mesurer rapidement et pratiquement.[150]

Cependant, il n'existe pas actuellement de mesure procurant des résultats concernant l'état d'avancement de la réaction dans un délai permettant une action sur le procédé. En outre, les protocoles existants décrits au paragraphe suivant, en plus d'être longs, nécessitent une quantité importante de manipulations techniques.

L'enzyme

La pepsine est une protéine d'origine animale. Le critère fondamental de choix d'une protéase est sa spécificité, mais d'autres facteurs interviennent dans la pratique comme son pH d'utilisation optimum, son prix ou sa disponibilité.

Les enzymes hydrolysent les protéines jusqu'à donner de petits peptides aux piètres propriétés fonctionnelles. Les peptides d'intérêt, eux, sont produits en cours de réaction. De fait, il est

[149] J. Adler-Nissen, *Journal of Chemical Technology and Biotechnology*, 32 (1982) p. 138
[150] J. M. Piot, *Thèse de doctorat d'état*, Université de Technologie de Compiègne (1989)

primordial de pouvoir limiter les protéolyses à la coupure de quelques liaisons peptidiques. Le contrôle du degré d'hydrolyse s'impose donc, en plus de la définition précise de la population peptidique.

La connaissance en temps réel de l'état réactionnel peut permettre le maintien de la réaction à un niveau où seuls les dérivés attendus sont produits.

4.2.3 Matériels et méthodes

L'hydrolyse enzymatique de l'hémoglobine bovine a fait l'objet de recherches fondamentales (isolements de peptides actifs) et appliquées (mise au point de procédés au stade pilote) par le Laboratoire de Technologie des Substances Naturelles (LTSN - Université des Sciences et Technologies de Lille) dirigé par le Professeur Guillochon. Les chercheurs de ce laboratoire ont quotidiennement en charge les préparations et les analyses qui sont décrites dans ce paragraphe. Des versions plus détaillées des modes opératoires sont d'ailleurs présentées dans divers travaux rédigés par les membres de cette équipe.[151,152]

Préparation de l'hémoglobine

Les solutions d'hémoglobine utilisées pour les hydrolyses sont réalisées par dilution d'une même solution-mère d'hémoglobine concentrée. Le contrôle de la concentration en hémoglobine de la solution préparée est effectué par une mesure d'absorbance UV. [152]

Hydrolyse de l'hémoglobine par la pepsine

On doit disposer de plusieurs solutions.

- Une solution de pepsine concentrée dont l'activité enzymatique est déterminée.
- Une solution-mère d'hémoglobine bovine.

Celle-ci sera diluée pour obtenir la concentration voulue. L'hémoglobine peut être préparée dans deux milieux distincts :

[151] N. Arroume-Nedjar, *Thèse de doctorat*, Université de Technologie de Compiègne (1991)
[152] B. Lignot, *Thèse de doctorat*, Université de Technologie de Compiègne (1998)

- le milieu "tampon", solution d'acide acétique / acétate de sodium assurant un pH égal à 4,5.
- le milieu "éthanol", 20% (v/v) d'éthanol sont ajoutés au milieu précédent.

• Une solution de borate de sodium 0,32 M à pH égal à 12,7 permettant l'arrêt de la réaction.

Le protocole est composé des étapes suivantes.

• Incuber à 23°C les préparations d'hémoglobine et de pepsine.

• Ajouter à 100 ml de la solution d'hémoglobine une quantité de pepsine permettant d'assurer un rapport nombre de moles d'enzymes sur nombre de moles de substrat constant.

• Effectuer des prélèvements de 1ml réguliers des hydrolysats.

Ajouter pour chacun d'entre-eux la solution d'arrêt qui inactive la pepsine par augmentation brutale du pH.

Détermination du degré d'hydrolyse

La détermination des DH des hydrolysats, basée sur la méthode proposée par *Adler-Nissen*,[153] consiste en un dosage par spectrophotométrie du chromophore formé après réaction de l'acide trinitrobenzène sulfonique (TNBS) avec les amines primaires, comme le montre la Figure 51. Il permet d'évaluer la concentration en fonctions NH_2 libérées dans le surnageant.

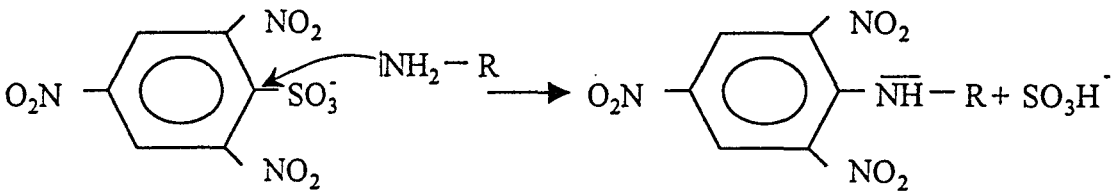


Figure 51 : Réaction du TNBS avec les groupements amines.

Le dosage est donné ici pour des prélèvements de 200 µl d'hydrolysats :

• Ajouter 50 µl d'une solution de HCl concentrée et 2 ml de tampon phosphate 0,2125 M de pH égal à 8,2.

[153] J. Adler-Nissen, *Journal of Agricultural & Food Chemistry*, 27 (1979) p. 1256

- Ajouter 2 ml de TNBS à 0,1%.
- Incuber 60 min à 50°C dans l'obscurité.
- Arrêter la réaction par ajout de 4 ml d'une solution de HCl 0,1N.
- Lire l'absorbance à 340 nm de chacun des tubes contre le blanc, en déduire le degré d'hydrolyse.

Méthodes d'analyse des peptides

La caractérisation exacte des produits de l'hydrolyse au cours du temps a été effectuée en deux étapes :[152]

- une première séparation des peptides contenus dans les hydrolysats d'hémoglobine est réalisée par chromatographie liquide haute performance de phase inverse.
- les constituants responsables de chacun des pics séparés par chromatographie sont identifiés en spectrométrie de masse par désorption laser, afin d'évaluer la pureté et la masse des peptides présents.

4.3 Etude de faisabilité du suivi d'hydrolyse sur échantillons prélevés

La maîtrise que le laboratoire LTSN de L'USTL possède des réactions d'hydrolyse d'hémoglobine bovine a permis d'accumuler, de caractériser et de conserver un certain nombre d'hydrolysats. Ces prélèvements ont été mis à notre disposition pour étudier la faisabilité d'un suivi d'hydrolyse d'hémoglobine bovine par spectrométrie infrarouge à transformée de Fourier. Certes, l'analyse d'échantillons prélevés s'écarte des objectifs initiaux en terme de mesure en ligne ou en temps réel. Néanmoins ce travail s'avère un investissement indispensable du point de vue de la justification voire de la crédibilité des travaux ultérieurs. Enfin, ce chapitre, présenté de manière volontairement succincte, peut être étoffé par la lecture des travaux auxquels il se réfère.[154,155]

4.3.1 Matériels et méthodes

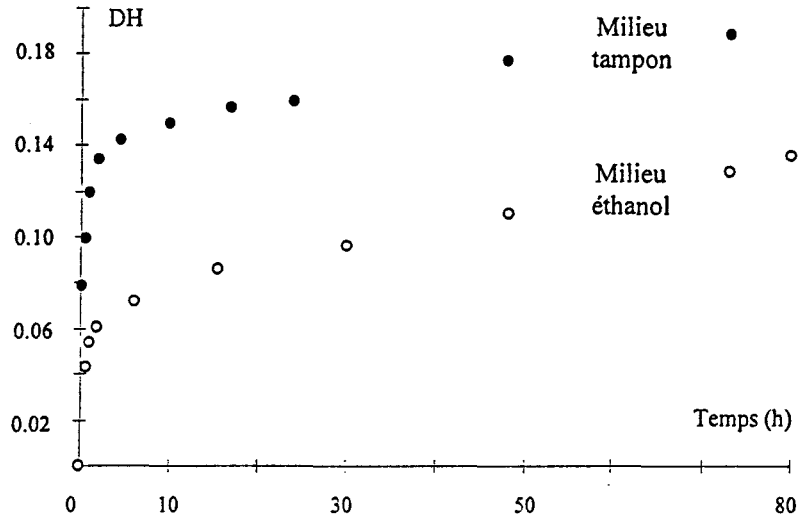


Figure 52 : Cinétiques d'hydrolyse.

Les échantillons analysés ont été prélevés lors de deux réactions distinctes d'hydrolyse d'hémoglobine à 0,2 % en masse se déroulant respectivement dans les milieux "tampon" et

[154] C. Ruckebusch, N. Nedjar-Arroume, S. Magazzeni, J. P. Huvenne, P. Legrand, *Journal of Molecular Structure*, 478 (1999) p. 185
[155] C. Ruckebusch, L. Duponchel, J. P. Huvenne, P. Legrand, N. Nedjar-Arroume, B. Lignot, P. Dhulster, D. Guillochon, *Analytica Chimica Acta*, 396 (1999) p. 241

"éthanol" décrits au paragraphe 2.3 de ce chapitre. Les conditions expérimentales sont fixées. Les méthodes de préparation des solutions d'hémoglobine bovine, de la pepsine, ainsi que l'estimation de l'activité enzymatique ne sont donc pas rappelées ici. La Figure 52 présente les cinétiques d'hydrolyse de l'hémoglobine bovine qui en résultent ; elles ne sont pas semblables bien que les mécanismes réactionnels soient identiques.

Echantillons

Il nous a été fourni 20 échantillons prélevés au cours de l'hydrolyse dans le milieu "tampon" (désignés par la lettre b) et 15 échantillons prélevés au cours de l'hydrolyse dans le milieu "éthanol" (désignés par la lettre e).

Hydrolyse d'hémoglobine "tampon"		**Hydrolyse d'hémoglobine "éthanol"**	
Echantillons	DH	Echantillons	DH
b0	0	e0	0
b1	0.006	e1	0.004
b2	0.012	e2	0.008
b3	0.020	e3	0.013
b4	0.032	***e4***	***0.020***
b5	***0.05***	e5	0.025
b6	0.052	***e6***	***0.03***
b7	***0.058***	***e7***	***0.05***
b8	***0.065***	e8	0.055
b9	0.078	e9	0.065
b10	***0.085***	e10	0.076
b11	0.1	***e11***	***0.098***
b12	***0.108***	e12	0.107
b13	0.115	***e13***	***0.110***
b15	***0.150***	e14	0.128
b16	0.156	e15	0.135
b17	***0.162***		
b18	0.180		
b20	***0.180***		

Tableau 4 : Caractéristiques des échantillons prélevés. (Les échantillons en italique gras correspondent à des valeurs interpolées de DH).

Le Tableau 4 regroupe les caractéristiques des échantillons prélevés. Les degrés d'hydrolyse de la plupart des échantillons ont été déterminés analytiquement. Ces DH sont corrélés à l'avancement de la réaction, c'est-à-dire à la population peptidique. Les valeurs rapportées

pour les prélèvements représentés en italique n'ont pas été mesurées mais interpolées.[154]

Echantillonnage

La technique d'échantillonnage consiste à laisser sécher 50 µl de solution sur une pastille de silicium de 13 mm de diamètre. Après évaporation du solvant non lié à la protéine, un dépôt mince et relativement uniforme peut être observé sur la pastille. Les spectres infrarouge sont alors enregistrés dans des conditions expérimentales de routine.
La normalisation utilisée procure aux variations d'absorbance de chaque échantillon une étendue unitaire.[154] Afin de respecter les conditions associées à l'emploi de fonctions de transfert sigmoïdes pour les réseaux, les valeurs numériques des DH ont été échelonnées dans l'intervalle [0,1 ; 0,9]. Elles sont ensuite reconverties dans le domaine initial pour l'exploitation des résultats.

Logiciels

On citera notamment :

- Le logiciel Unscrambler® (version 6.11b, CAMO) qui a été utilisé pour la construction des matrices d'échantillons, les pré-traitements ainsi que les calculs des analyses PCA et PLS.
- L'environnement Matlab™ (version 5.3, The Mathworks Inc.) qui permet l'écriture de routines ou la réalisation d'opérations matricielles.
- Le logiciel SNNS© (version 4.2), pour les calculs basés sur des techniques neuronales, fonctionnant sur plate-forme UNIX. Il a été développé par l'Institut pour les systèmes parallèles et distribués de hautes performances (IPVR) de l'Université de Stuttgart.[156]

[156] A. Zell, G. Mamier, M. Vogt, J. Wieland, *SNNS User Manual*, Report n°6-95 (1995)

4.3.2 Discussion : spectres et techniques d'échantillonnage

Techniques d'échantillonnage

La spectrométrie IRTF des protéines utilise majoritairement l'enregistrement de spectres de transmission[157] de liquide, malgré un certain nombre de contraintes techniques qui rendent difficile l'obtention de résultats de bonne qualité. En particulier, pour éviter une absorption importante des molécules d'eau, il faut disposer de solutions suffisamment concentrées en protéines et utiliser des chemins optiques inférieurs à 10 µm. Lorsque la concentration en protéine est si faible qu'elle interdit ce type de mesure, le séchage de la solution, dont résulte un film, peut constituer une alternative. En effet, le dépôt en question possède une concentration en protéine élevée et permet généralement d'obtenir un rapport signal sur bruit correct. Des spectres de qualité acceptable peuvent donc être enregistrés pour de faibles quantités de protéines.

Pour des solutions contenant peu d'hémoglobine, la technique de dépôt et séchage sur pastille de silicium est relativement pratique, et constitue le meilleur compromis que nous ayons observé. Celui-ci tient compte du rapport signal sur bruit et de la répétabilité des spectres. De plus, cette méthode ne semble pas dénaturer l'hémoglobine ; l'information chimique de structure est conservée, ce qui a été aussi observé par ailleurs.[158] La formation de films ne détruit généralement pas la structure native des protéines puisque celle-ci reste hydraté lors du séchage.[159] Dans des conditions expérimentales beaucoup plus dures, certaines études rapportent néanmoins l'observation de dénaturations.[160]

Spectres obtenus

Afin d'accentuer les variations spectrales et de s'affranchir de contributions éventuelles de ligne de base, les dérivées secondes des spectres ont été calculées. La Figure 53 présente quelques spectres obtenus au cours d'hydrolyses dans les milieux "tampon" et "éthanol". Les

---

[157] K. A. Oberg, A. L. Fink, *Analytical Biochemistry*, 256 (1998) p. 92
[158] J. Sajid, A. Elhaddaoui, S. Turrel, *Journal of Raman Spectroscopy*, 28 (1997) p. 165
[159] J. Safars, P. P. Roller, G. C. Ruben, D. C. Gajdusek, C. J. Gibbs, *Biopolymers*, 33 (1993) p. 1461
[160] R. W. Sarver, A. R. Friedman, T. J. Thaman, *Spectrochimica Acta*, Part A 53 (1997) p. 1889

deux régions spectrales (centrées sur 1575 cm^{-1} et 1400 cm^{-1}), où l'absorption est maximale, correspondent à des contributions du milieu. Il s'agit en effet des vibrations d'élongation symétrique et antisymétrique du groupe acétate résonnant (COO^-). Le large pic d'absorption de faible intensité qui s'étend autour de 1660 cm^{-1} correspond aux vibrations amide I de l'hémoglobine bovine. Les vibrations amide II ne sont quant à elles pas observables.

En outre, l'observation ne permet pas de distinguer les spectres enregistrés sur des échantillons provenant de l'hydrolyse dans le milieu "tampon" et de ceux concernant la réaction dans le milieu "éthanol". Ils sont visuellement semblables et cela reste vrai quel que soit l'état d'avancement de la protéolyse. Bien que la bande Amide I constitue la région contenant l'information, les modifications ne sont pas détectables sur le spectre. Cela peut s'expliquer par l'étroitesse du domaine des variations possibles du DH. Même lorsque la réaction est terminée, 80% des liaisons peptidiques sont intactes.

Figure 53 : Spectres d'échantillons d'hydrolyse en dérivée seconde.

Effets de solvants

Une étude effectuée parallèlement à celle de la protéolyse de l'hémoglobine a montré l'influence d'une dénaturation avancée de l'hémoglobine bovine sur le spectre infrarouge.[108] Il s'agit de caractériser les effets produits lors de l'ajout d'un volume d'éthanol croissant au

milieu "tampon" (acide acétique / acétate de sodium). Les spectres obtenus pour des proportions en éthanol variant de 0 % à 50 % sont regroupés Figure 54.

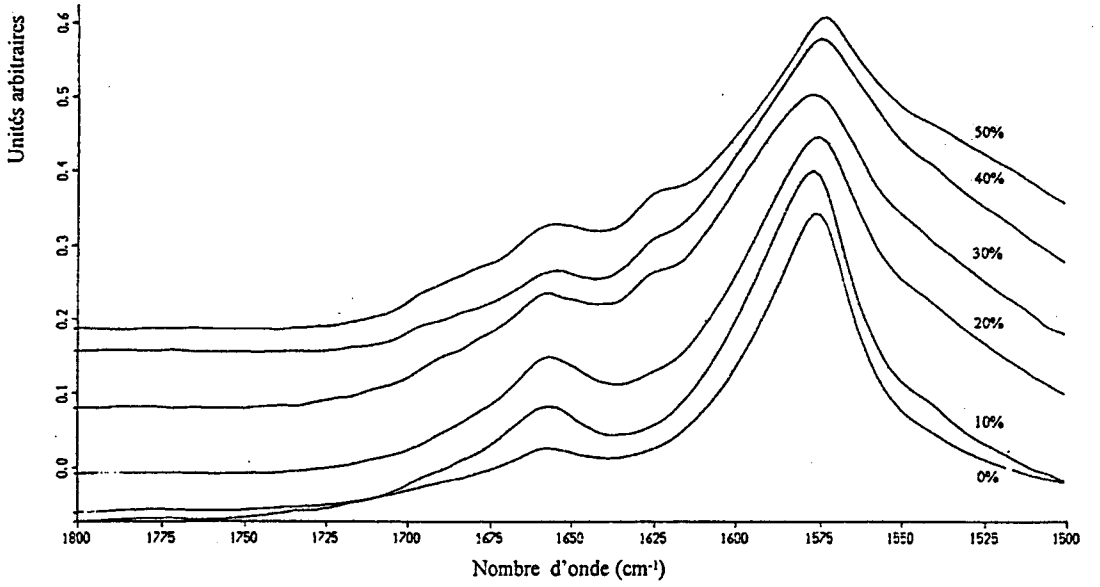


Figure 54 : Spectres de l'hémoglobine dans les milieux concentrés en éthanol.

La dénaturation se traduit par l'apparition d'un épaulement à 1625 cm^{-1} pour des concentrations en éthanol supérieures à 20 % en volume.

L'estimation quantitative de ces changements structuraux, a été effectuée par mesure d'aires sous les pics obtenus en dérivée seconde dans la région spectrale de la vibration amide I. Ces résultats sont résumés par le graphique de la Figure 55. On constate que l'ajout d'éthanol au milieu "tampon" n'est pas dénaturant pour l'hémoglobine bovine tant que la proportion d'alcool ne dépasse pas environ 20 % en volume. En effet, la proportion d'hélice α décroît au fur et à mesure que la concentration en éthanol augmente et finit par se stabiliser au-delà d'une proportion de 30 % d'alcool en volume. Cette perte de conformation en hélice s'accompagne principalement d'une augmentation de la proportion de structure désordonnée et, dans une moindre mesure, de la quantité de coudes.

La proportion d'hélice α permet ainsi de juger de l'état, natif ou dénaturé, de l'hémoglobine bovine. Ces résultats sont en accord avec d'autres études qui ont montré qu'une proportion de l'ordre de 20 % d'éthanol correspond à une stabilisation maximale des molécules d'hémoglobine soumises à des conditions potentiellement dénaturantes[161] ou encore à une

[161] T. Asukara, K. Adachi, E. Schwartz, *Journal of Biological Chemistry*, 253 (1978) p. 6423

hydratation complète des molécules d'alcool, entraînant à une stabilisation des liaisons hydrogènes. [162]

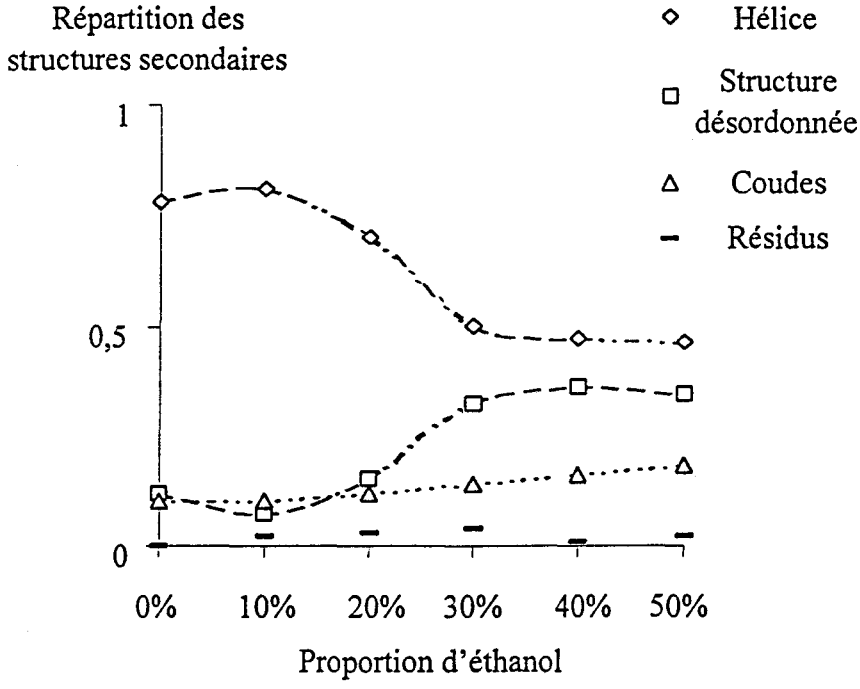


Figure 55 : Changements de structure secondaire (estimée à partir des aires des contributions individuelles discernables sur les spectres en dérivée seconde) de l'hémoglobine bovine observés en fonction de la proportion en éthanol.

La vibration amide I de la liaison peptidique peut être considérée comme un indicateur qualitatif de la structure secondaire d'une protéine, en particulier de l'hémoglobine bovine.

4.3.3 Vers le suivi de la réaction

Lorsque les caractéristiques spectrales des échantillons ne sont pas directement discernables, le premier réflexe de l'analyste consiste bien souvent à effectuer une analyse en composantes principales, afin d'obtenir la localisation des échantillons sur les directions de plus grande variance.

[162] N. Nishi, K. Koga, C. Ohshima, K. Yamamoto, U. Nagashima, K. Nagami, *Journal of the American Chemical Society*, 109 (1987) p. 7353

Analyse en composantes principales

Pour l'analyse en composantes principales, seul l'intervalle spectral 1600-1700 cm^{-1} a été utilisé. En effet, la qualité relative des spectres, imputable à la répétabilité de la technique d'échantillonnage utilisée, rend impossible l'utilisation du domaine complet. Lorsque l'intervalle 1200-1800 cm^{-1} est analysé dans son ensemble, la variance qui oriente le modèle provient des contributions dues au solvant acétate. Une étape de sélection du domaine spectral est donc nécessaire.

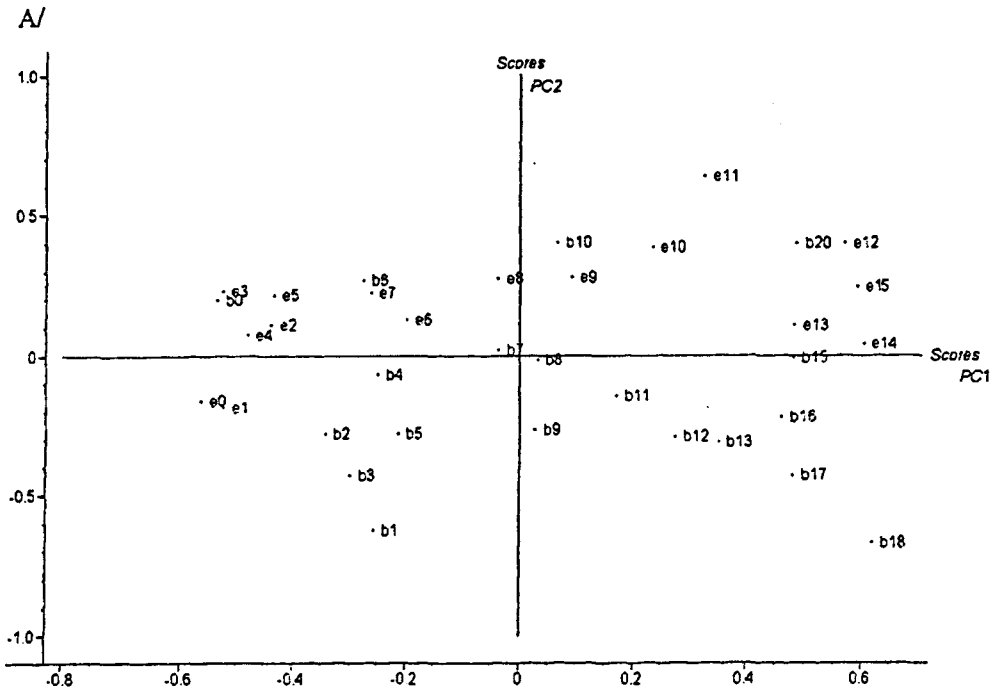


Figure 56 : Analyse en composantes principales du lot d'échantillons – Représentation du plan engendré par les deux premiers axes principaux. *(Les échantillons sont représentés par une lettre correspondant au milieu et un numéro d'ordre propre à chaque hydrolyse, Tableau 4).*

La projection du lot d'échantillons sur le plan formé par les deux premiers axes principaux correspondant aux valeurs propres les plus élevées est présentée ci dessus Figure 56. Une observation rapide de la représentation confirme une constatation faite à partir des spectres. Quel que soit le milieu de provenance des échantillons, milieu "éthanol" ou "tampon", ils sont représentés de manière homogène dans le plan. On peut donc considérer qu'ils appartiennent à un même lot et qu'il n'est pas possible de distinguer, à partir des spectres, de classification sur le milieu dans lequel la protéolyse est effectuée. Cela reste vrai pour les plans engendrés par

des combinaisons d'axes principaux d'ordres supérieurs. Par contre, on peut remarquer un arrangement chronologique des représentations des échantillons le long de la première composante principale. Cet axe représente par ailleurs 48% de la variance totale contenue dans le lot de spectres.

La Figure 57 est une représentation des degrés d'hydrolyse des échantillons en fonction leurs projections respectives (appelées *score*) sur le premier axe principal qui s'accorde avec la chronologie. Cette règle générale doit être néanmoins nuancée. Certains échantillons la respectent mal ou ne s'y soumettent pas, et les relations d'ordre entre échantillons voisins ne sont pas toujours convenables. C'est par exemple le cas, visible Figure 56, pour les prélèvements b1 ou e2. Enfin, que ce soit sur le deuxième axe principal ou sur les axes principaux d'ordre supérieur, on ne distingue pas d'organisation particulière ou de classification des échantillons.

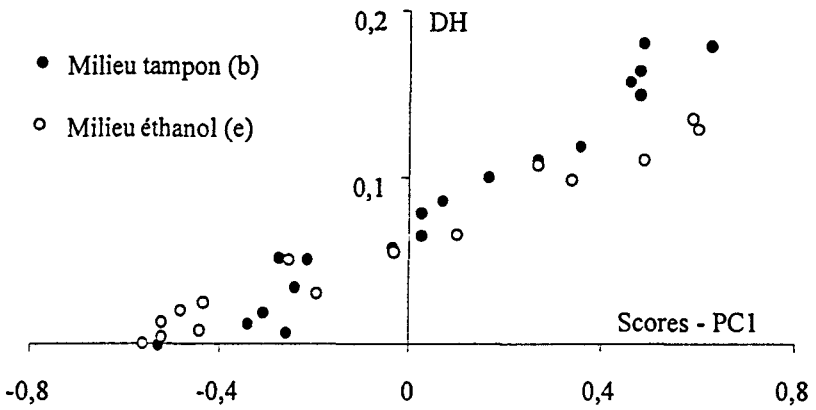


Figure 57 : Degré d'hydrolyse représenté en fonction des scores du premier axe principal.

La PCA associe à chaque axe principal un vecteur normalisé appelé vecteur *loading* qui maximise la variance expérimentale. La Figure 58 présente le facteur *loading* associé au premier axe principal. Il montre un maximum assez étroit, autour de 1660 cm^{-1}, qui sera identifié dans la partie interprétation de ce paragraphe. C'est cette composante spectrale qui a le plus influencé la construction de la direction principale de variance.

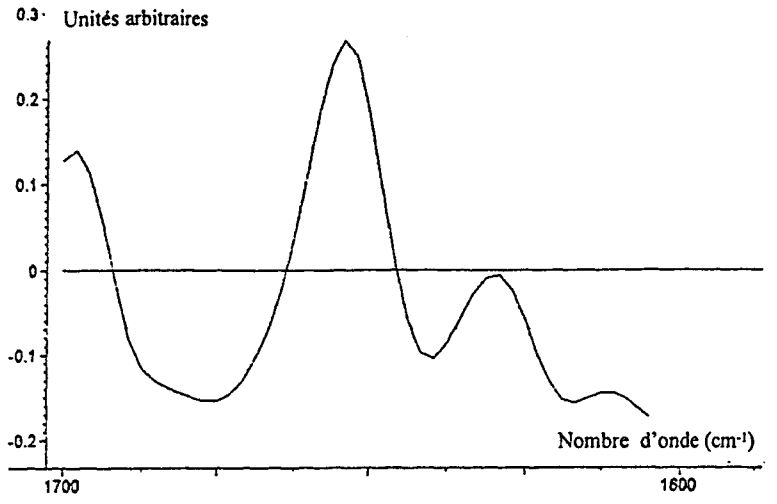


Figure 58 : Vecteur *loading* associé à la première composante principale.

Estimation de l'avancement de la réaction

Ce paragraphe synthétise les résultats obtenus lors de la construction d'un modèle étalonnage multivarié, par la régression PLS ou les réseaux multicouches FF, pour la prédiction du DH. Nous n'entrons pas dans le détail des méthodes. Notons cependant qu'une attention particulière a été apportée à la procédure d'étalonnage pour des réseaux *feed-forward*, compte tenu du nombre relativement faible d'échantillons dont on dispose pour réaliser cette étude de faisabilité.[155]

La qualité des modèles construits a été estimée par un paramètre appelé RMSEP (*Root Mean Squared Error Prediction* en anglais). Il s'agit d'un paramètre de validation expérimentale explicité en annexe 2.

Les erreurs standards obtenues en prédiction lors de cette étude,[155] aussi bien pour les méthodes neuronales que pour la régression PLS, couvrent une gamme qui s'étend autour de 0,02. Ces valeurs numériques ne sont significatives que si elles sont comparées à l'ordre de grandeur de la variable estimée. Il s'agit ici du degré d'hydrolyse dont les valeurs typiques sont comprises entre 0 et 0,2. Le manque de répétabilité des données spectrales semble interdire de meilleures capacités prédictives. En effet, quelle que soit la méthode utilisée, la qualité des données est une limitation incontournable de la précision et des capacités prédictives du modèle.

Par exemple, les projections des échantillons b20, b17 et b15 dans le plan composé des deux premières composantes principales sont très voisines (Figure 56). Cela signifie que les

spectres sont quasiment identiques du point de vue des contributions à la variance totale prises en compte par ces deux axes propres. En conséquence, comme le montre la Figure 59 sur laquelle les échantillons en question sont entourés, le modèle prédit naturellement des valeurs de DH équivalentes et sous-estimées pour ces trois échantillons.

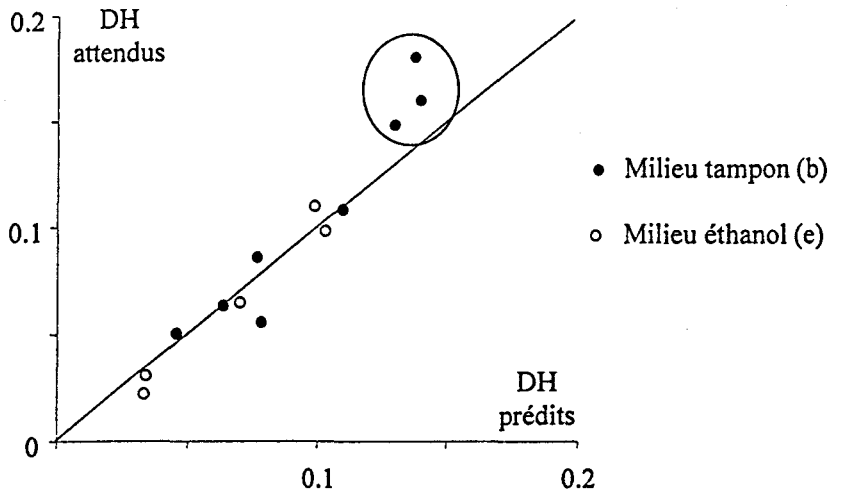
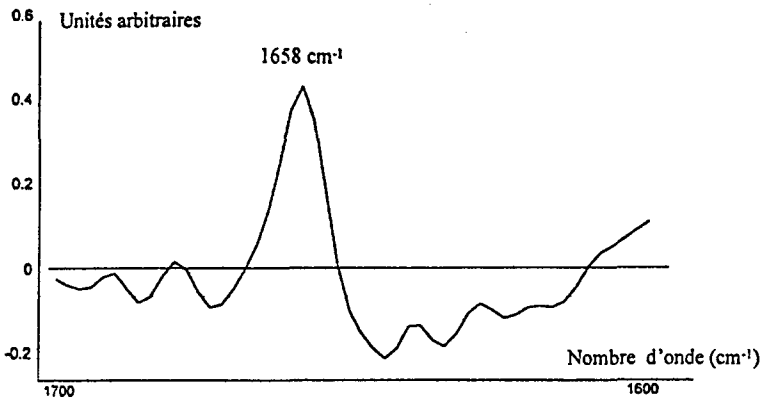
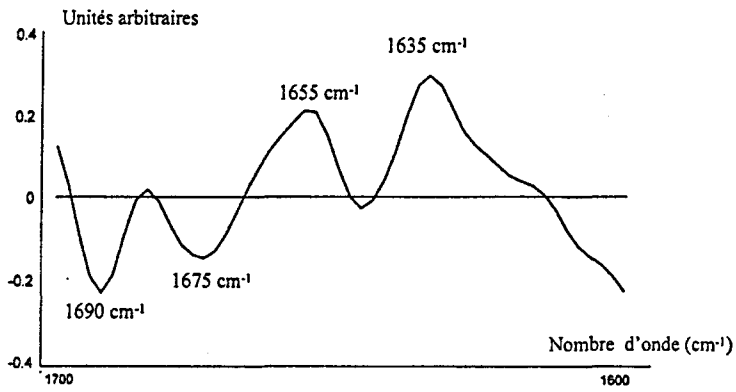


Figure 59 : Représentation des capacités de prédiction d'un des modèles construits.

Un des avantages indéniables de la régression PLS sur les méthodes basées sur les RNA provient de la possibilité d'interprétation immédiate des résultats par le physico-chimiste. Les composantes *loading* de la régression PLS expriment la manière dont l'information contenue au sein de chaque variable dans l'espace des nombres d'onde est reliée à l'évolution du degré d'hydrolyse. Les *loading* correspondant aux deux premiers axes principaux prennent en compte presque 70% de la contribution totale de la variance au modèle. Ils sont présentés Figure 60 et leur interprétation est détaillée dans le paragraphe qui suit.



A/



B/

Figure 60 : Premier (A/) et second (B/) facteurs *loading* extraits pour la prédiction du DH.

Interprétation

Le mécanisme moléculaire associé à l'hydrolyse d'hémoglobine bovine dans les milieux considérés ici est appelé "one by one".[163,164] Trivialement, cela signifie que les molécules sont clivées l'une après l'autre. Ce mécanisme implique la présence, tout au long de la réaction, d'hémoglobine sous forme native (conformation en hélice dominante) et de l'ensemble des peptides finaux, en proportion variable en fonction de l'avancement de la réaction. Pour être potentiellement hydrolysable, chaque molécule d'hémoglobine doit subir une dénaturation qui a pour effet de rendre accessible à l'enzyme les sites potentiels de coupure de la protéine. Sur le spectre, les effets produits sont qualitativement comparables à ceux mis en avant lors de

[163] J. A. Rupley, *Methods in Enzymology*, 11 (1967) p. 905
[164] K. Linderstrom-Lang, *Proteins and Enzymes*, Stanford university Press (1952) p. 53

l'étude de l'influence de la proportion d'alcool dans le solvant. Cette analyse, détaillée au paragraphe 3.2 de ce chapitre, permet donc une compréhension plus aisée des phénomènes physico-chimiques qui sont traduits par la bande d'absorption amide I. D'autre part, le clivage lui-même modifie forcément les propriétés de structure, que ce soit des molécules d'hémoglobine ou des peptides résultants.

Le vecteur *loading*, Figure 58 page 117, associé à la construction du premier axe de la PCA, pointe les variations spectrales les plus importantes. Il s'agit essentiellement de la région autour de 1600 cm^{-1} mais il existe également d'autres contributions.

Pour la régression PLS, l'analyse de la première composante *loading* présentée Figure 60A/ page 119 permet de mettre en valeur les variables les plus importantes pour la prédiction de l'avancement de la réaction, situées autour de 1658 cm^{-1}. Cette information est corrélée à la conformation en hélice α de la protéine, dont la proportion par rapport à l'ensemble de la structure évolue au cours du temps, au profit de formes désordonnées de structure.

D'autre part, le vecteur *loading* associé au deuxième axe principal, présenté Figure 60 B/ page 119, laisse apparaître deux contributions positives à 1635 cm^{-1} et 1655 cm^{-1}, ainsi que deux contributions négatives à 1675 cm^{-1} et 1690 cm^{-1}.

Les coudes de la structure, observables autour 1670 cm^{-1}, traduisent une certaine structuration de la molécule. Ils sont ainsi corrélés négativement aux contributions des structures désordonnées, qui s'étendent de 1620 cm^{-1} à 1650 cm^{-1}. En outre, ce facteur semble comporter des contributions résiduelles du signal correspondant à la conformation en hélice, qui peuvent être considérée comme des compensations.

4.3.4 Conclusion et perspectives

Ces étapes préliminaires laissent envisager la faisabilité du suivi des protéolyses d'hémoglobine bovine à partir des spectres infrarouges enregistrés sur des échantillons prélevés. En effet, la prédiction du DH peut être effectuée avec une erreur standard de l'ordre de 0,02 pour une gamme analytique [0 ; 0,2]. Ce résultat est convenable compte tenu des difficultés. En outre, elles mettent en évidence que ce sont principalement les phénomènes de changement de structure secondaire, qu'ils soient liés à la dénaturation de l'hémoglobine bovine ou/et aux coupures de celles-ci voire aux peptides présents en solution, qui sont mesurés par spectrométrie infrarouge au cours de l'hydrolyse. Chaque entité structurale de la protéine contribue au spectre infrarouge et les contours de la bande amide I, constitués de

recouvrement de bandes, sont complexes. Si la sensibilité de la conformation des bandes infrarouge est reconnue, l'aspect quantitatif de l'analyse en terme de structure secondaire est cependant moins immédiat.

Néanmoins, la méthode d'échantillonnage utilisée est relativement exigeante et coûteuse. La qualité spectrale est inévitablement limitée par le manque d'uniformité du séchage. Cette technique ne permet pas d'envisager le suivi en temps réel bien qu'elle représente une méthode de substitution intéressante aux protocoles de détermination analytique du degré d'hydrolyse. En dépit de cela, il s'agit quasiment de la seule méthode envisageable lorsqu'on s'intéresse à de si faibles concentrations.

Parallèlement, la question de la dénaturation a été abordée et le séchage semble respecter la structure. Certains auteurs élargissent d'ailleurs ce problème aux études cristallographiques de référence[165] ; jusqu'à quel point la structure "statique" des cristaux représente correctement la conformation de la protéine dans l'environnement complexe et dynamique des cellules vivantes ?

L'instrumentation spectroscopique infrarouge, combinées à des méthodes numériques plus ou moins sophistiquées d'analyses, permet d'obtenir des résultats informatifs du degré d'hydrolyse, estimateur de l'avancement de la réaction. Plus précisément, l'état de la réaction peut être atteint par le biais de la corrélation entre les spectres infrarouge du milieu contenant protéines et peptides, et du degré d'hydrolyse.

Si l'on souhaite effectuer une mesure en ligne du degré d'hydrolyse, cela impose en plus des contraintes liées à la dimension temporelle, de pouvoir analyser les produits sous leur forme brute. Or, la spectrométrie infrarouge n'autorise pas de mesure sur des solutions aussi faiblement concentrées. La poursuite de ce travail implique donc nécessairement une modification des conditions expérimentales et notamment une augmentation de la concentration en hémoglobine. Heureusement, ceci peut être effectué sans modifier la production, à condition de maintenir le rapport entre enzyme et substrat constant. Les spectres infrarouge des liquides sont alors le résultat de recouvrement d'informations spectrales et d'interférences éventuelles dues au solvant. Les traitements chimiométriques sont encore plus nécessaires.

Les conditions de la mesure se dégradant, les méthodes neuronales sont les plus à même de gérer ce type de situations réelles à l'environnement changeant ou mal défini. Ce choix

[165] W. K. Surewicz, H. H. Mantsch, D. Chapman, *Biochemistry*, 32 (1993) p. 83

technologique sera validé par le biais de l'observation de la structure des données. Néanmoins, des méthodes plus communes comme la régression PLS ont montré, dans ce chapitre, leur intérêt pour l'interprétation et la justification de la mesure.

Chapitre 5

Suivi d'hydrolyse d'hémoglobine bovine en réacteur

Afin de permettre une certaine flexibilité de la production, les réactions biotechnologiques sont couramment réalisées au sein de réacteurs fonctionnant en cycles ouverts. L'optimisation de ces réacteurs est corrélée, soit à la connaissance de la cinétique, soit à l'évaluation des produits du procédé. Le nombre important de liaisons peptidiques susceptibles d'être clivées, en parallèle et en série, lors d'une hydrolyse enzymatique est une limitation naturelle à la possibilité d'estimer des paramètres cinétiques. Certaines études rapportent l'utilisation d'équations empiriques simples.[166] Elles constatent néanmoins que les résultats sont peu satisfaisants du fait des différences significatives entre les courbes d'hydrolyse obtenues pour l'hémoglobine bovine, malgré des conditions expérimentales les plus semblables possible.
L'information produite par les observations par spectrométrie infrarouge est très représentative, bien que d'interprétation difficile. En effet, la sélectivité spectrale n'est généralement pas suffisante pour permettre une approche par des traitements simples. La modélisation empirique d'un système pour un suivi quantitatif *in situ* implique en conséquences plusieurs difficultés potentielles.

Après avoir présenté les méthodes mises en place pour l'observation du système et discuté de la validité des enregistrements, nous caractérisons les données pour justifier l'utilisation de techniques neuronales. Diverses procédures de transformation des données sont ensuite envisagées. L'analyse en composantes principales est la méthode retenue, en dépit de certaines limitations théoriques à propos des non-linéarités, afin d'assurer une aptitude à la généralisation des modèles. Nous détaillons enfin la construction de modèles valides et efficaces pour la prédiction de répétitions inconnues du procédé.

---

[166] M. C. Marquez, M. A. Vazquez, *Process Biochemistry*, 35 (1999) p. 111

5.1 Méthodologie instrumentale

Au-delà de la présentation des procédures et des installations expérimentales, nous commentons ici les choix et les compromis que nous avons été amenés à réaliser. Il a fallu, en particulier, estimer et maîtriser la stabilité du signal analytique par le biais de mesures de répétabilité. Une certaine variabilité du signal se justifie du point de vue instrumental. Néanmoins, l'observation d'un procédé doit procurer une information qualitativement et quantitativement distincte. Même si la chimiométrie peut s'accommoder d'observations relativement médiocres, c'est d'abord l'instrumentation qui fournit le résultat analytique.

5.1.1 Pré-requis

Matières premières

Concernant les procédés d'hydrolyse analysés dans ce chapitre, des solutions "tampon" concentrées à 5 % en hémoglobine bovine* sont utilisées. Rappelons que le milieu "tampon" est un mélange acide acétique / acétate de sodium (0,1 M, pH = 4,5).
Une solution de pepsine* permet d'initier chaque réaction. Son activité enzymatique est ajustée de telle sorte que le rapport enzyme sur substrat soit le même pour toutes les répétitions du procédé.
En ce qui concerne les échantillons prélevés, l'arrêt de la production peptidique s'effectue à l'aide d'une solution de borate de sodium (0,32 M, pH=12,7) qui inactive la pepsine, par augmentation brutale du pH.

Analyses de référence

La détermination des degrés d'hydrolyse est réalisée sur des hydrolysats prélevés selon la méthode décrite au paragraphe IV.2.3. Ces valeurs peuvent éventuellement être corrélées à la

* Sigma-Aldrich®, St Quentin Fallavier, France

caractérisation exacte des peptides obtenue par chromatographie liquide haute performance en phase inverse et spectrométrie de masse par désorption laser.

5.1.2 Conditions opératoires

Montage expérimental

Un schéma de l'installation expérimentale, mise en place au laboratoire, est présenté Figure 61.

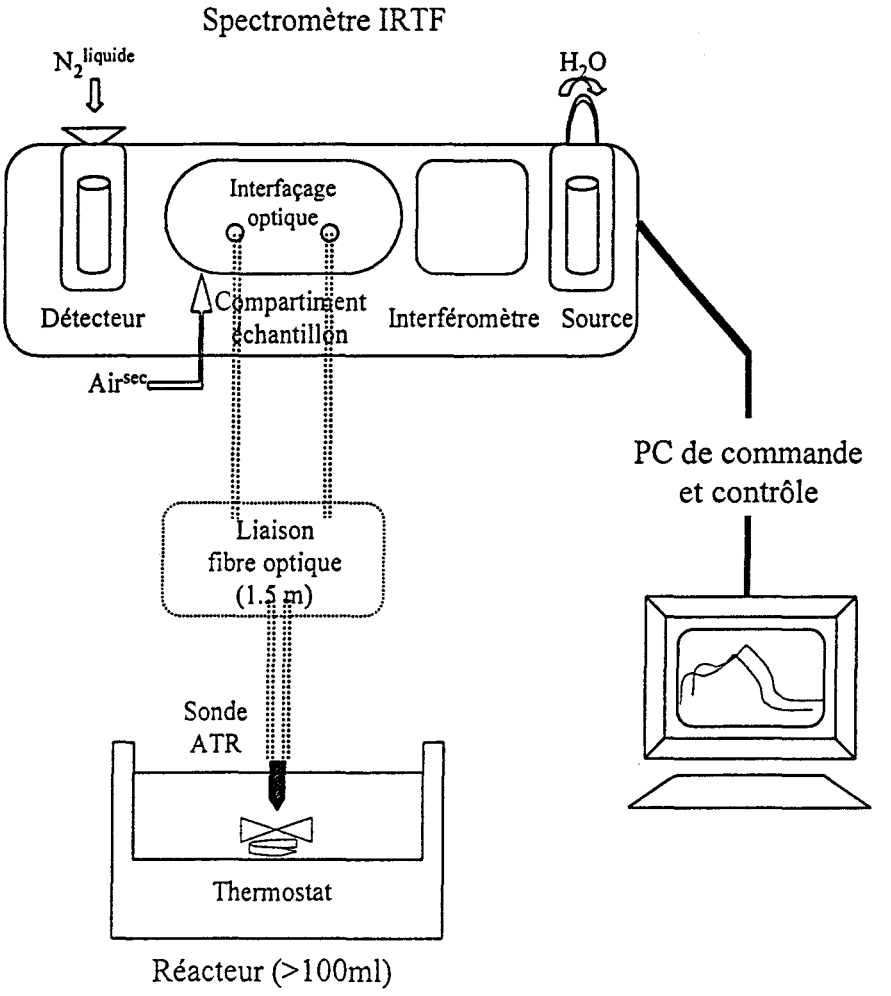


Figure 61 : Dispositif expérimental pour l'enregistrement en ligne de spectres de vibration.

La source infrarouge est un bâtonnet de carbure de silicium (*Globar*) dont la température est maintenue constante par le biais d'une circulation d'eau. Le détecteur photovoltaïque (Mercure Cadmium Tellure), équipant le spectromètre IRTF Bruker* IFS-48, doit quant à lui être en permanence maintenu à la température de l'azote liquide. Le réservoir utilisé assure une autonomie que nous avons estimée à environ 6 heures. Si cela s'avère nécessaire, il peut être réapprovisionné en cours d'expérimentation. Par ailleurs, le compartiment échantillon est purgé des éventuelles traces de vapeur d'eau et de dioxyde de carbone dont les contributions spectrales sont perturbatrices.

Le spectromètre est couplé au réacteur par l'intermédiaire d'une liaison par fibres optiques** en verres de chalcogénures. L'interfaçage s'effectue à l'intérieur du compartiment échantillon. Ce couplage est optimisé par l'intermédiaire des réglages de l'orientation des miroirs, de la hauteur du support, et du positionnement focal des fibres d'entrée et de sortie. Une fois cette opération effectuée, le compartiment ne nécessite plus d'être ouvert et une purge efficace est assurée. A l'autre extrémité de la fibre optique, l'énergie infrarouge est interfacée par un cristal ATR de séléniure de zinc (ZnSe) à deux réflexions. Le diamètre de cette sonde, qui assure le couplage de l'énergie avec le milieu réactionnel, est de 12 mm. Sécurité et simplicité de manipulation et de mise en place sont les avantages principaux d'un tel système, garantissant une reproductibilité satisfaisante des mesures. De plus, si la concentration en analyte est suffisante, les solutions opaques peuvent être analysées. Par contre, les pertes d'énergie associées à ce dispositif constituent un inconvénient incontournable. Celles-ci se produisent non seulement au sein du matériau constituant la fibre optique mais aussi, et surtout, aux interfaces. Globalement, l'énergie utilisable en sortie représente, selon les dispositifs instrumentaux, de 10 % à 1 % de l'énergie initiale. Nous avons mesuré un signal optimum de 1800 coups (unités arbitraires) sur le détecteur contre 32000 en l'absence de sonde, soit un résultat proche de 6 %.

L'ensemble du dispositif expérimental est placé dans une pièce climatisée dont la température est maintenue autour de 20° C.

Enfin, l'enregistrement, ainsi que certains pré-traitements des spectres, sont effectués au niveau de l'ordinateur qui pilote le spectromètre. A terme, celui-ci est supposé intégrer les traitements de chimiométrie, et ainsi, gérer la caractérisation l'avancement de la réaction pour permettre une éventuelle action sur le système.

---

* Bruker, Wissembourg, France
** Graseby Specac Limited™, Orpington, Royaume-Uni

Procédure

Les différentes répétitions de l'hydrolyse, réalisées en réacteur de laboratoire, sont analysées sur une durée qui peut atteindre 8 heures. Cette période est suffisante pour couvrir le domaine de production des peptides intéressants, qui apparaissent plutôt au cours des premières heures du procédé. La sonde ATR n'est, à aucun moment, extraite du réacteur. Cela rend impossible tout enregistrement de spectres de référence en cours de manipulation.
Ainsi, la procédure suivie pour chaque répétition du procédé est décrite ci dessous.

- Le spectre de la référence, un milieu "tampon" provenant de la même solution mère que celle utilisée comme substrat de l'hémoglobine, est enregistré à 23°C.
- Le réacteur est ensuite chargé avec la solution d'hémoglobine.

La température du système est contrôlée et maintenue à 23°C. Avant que la réaction ne soit initiée, le spectre de l'hémoglobine dans son milieu est enregistré. Ce spectre constitue en quelque sorte le zéro de la mesure.

- L'ajout de l'enzyme dans le réacteur marque le début de la réaction d'hydrolyse.

Des spectres sont alors enregistrés à des intervalles qui ne sont pas forcément réguliers, toutes les 10 à 30 minutes. Un procédé complet se trouve ainsi typiquement caractérisé par une quarantaine de spectres. En cours de manipulation, les prélèvements effectués sont stockés au réfrigérateur avant d'être soumis aux analyses produisant les résultats de référence pour le degré d'hydrolyse.

Acquisition des spectres

Une observation nécessite d'accumuler 500 fois (on parle de *scans* en anglais) l'interférogramme, pour une résolution instrumentale de 4 cm^{-1}. Le résultat est ensuite convolué par une fonction d'apodisation triangulaire. Le signal de référence étant enregistré dans les mêmes conditions, le spectre obtenu subit alors une transformation de Fourier discrète avec un facteur de "*zero-filling*" égal à 2. La résolution apparente est ainsi de 1,928 cm^{-1}.

Tous les spectres sont soumis à une correction de ligne de base. Pour chaque valeur d'absorbance, celle-ci consiste à soustraire la moyenne des absorbances de l'intervalle compris entre 1800 cm^{-1} et 1900 cm^{-1}, où il n'existe pas d'absorption caractéristique. Par ailleurs, les spectres sont tronqués et seule la zone spectrale significative est conservée. Finalement, les

résultats sont donc présentés sous la forme de spectres d'absorbance décrits par 141 valeurs numériques, les intensités aux nombres d'onde compris dans la gamme entre 1452,135 cm^{-1} et 1722,121 cm^{-1}.

En cours de manipulation, les spectres typiquement obtenus sont semblables à celui présenté Figure 62.

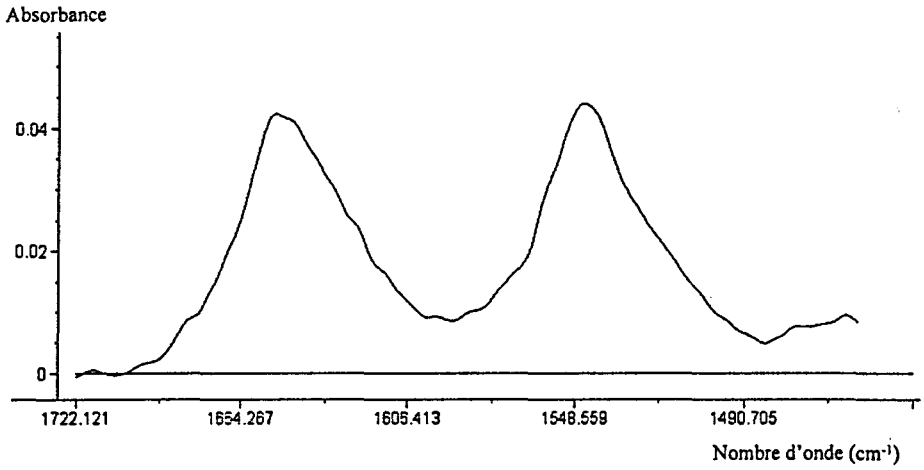


Figure 62 : Spectre obtenu en cours d'hydrolyse de l'hémoglobine.

Les bandes amides I et II, centrées respectivement autour de 1655 cm^{-1} et 1545 cm^{-1}, sont clairement observées. Celles-ci ont une étendue spectrale importante représentant la somme de plusieurs contributions, que la résolution ne permet pas de distinguer. D'autre part, le faible niveau d'absorbance est imputable au rendement énergétique médiocre de la liaison par fibres optiques.

Sur les spectres, nous avons constaté un niveau de bruit de l'ordre de 10^{-4} unités d'absorbance. Il en résulte un rapport signal sur bruit de l'ordre de 400 pour un spectre d'hémoglobine, enregistré par le système ATR-fibre optique. Ce rapport acceptable, bien qu'en deçà des valeurs typiquement observées en spectrométrie dans l'infrarouge moyen, est plus la conséquence du faible niveau de signal, que du niveau de bruit.

Commentaires

Les études des systèmes biologiques et physiologiques nécessitent d'utiliser des solutions salines comme tampon, afin de conserver la substance dans l'état désiré. Effectuer le spectre de l'eau seule comme mesure de référence, ne constitue alors pas une solution suffisante. En

effet, les ions acétates CH_3COO^- et sodium Na^+ du milieu "tampon", dans lequel l'hémoglobine est dissoute, perturbent l'organisation des molécules d'eau et induisent des décalages de nombres d'onde. La technique de réflexion totale atténuée (ATR) facilite la soustraction et l'échantillonnage des solutions aqueuses. Des spectres de bonne qualité sont rapidement obtenus, même pour des solutés dissous dans des solvants très absorbants.

Nous avons envisagé une correction de ligne de base très simple utilisée dans la littérature pour la correction des variations d'indice de réfraction[167] ou de température.[168] Nous avons constaté que la ligne de base est principalement soumise à des effets de dérive dus au détecteur photovoltaïque. Cela se caractérise, avant tout, par un *offset* constant[169] mais les détecteurs de type MCT montrent aussi des comportements non-linéaires.

Par ailleurs, nous effectuons un suivi par spectrométrie sur un intervalle de temps qui excède l'autonomie du détecteur infrarouge. De fait, il est nécessaire d'effectuer le réapprovisionnement en azote liquide de son réservoir. La conséquence majeure, visible sur le spectre, est l'apparition d'un décalage global de celui-ci ; heureusement pris en compte par la correction de ligne de base. Néanmoins, nous nous imposons une attente de quelques minutes avant l'enregistrement d'un nouveau spectre, pour garantir la stabilité du signal.

Concernant les 500 accumulations, ce nombre a été choisi à l'issu de plusieurs essais. Combiné à une résolution instrumentale de 4 cm^{-1}, il permet d'obtenir un rapport signal sur bruit satisfaisant. C'est la relative faiblesse du signal retourné après transport par le système optique qui nous oblige à accumuler un nombre important d'interférogrammes. Néanmoins, cela ne constitue pas une contrainte, le temps nécessaire pour effectuer une mesure restant inférieur à 5 minutes.

Enfin, la fonction d'apodisation permet de réaliser une troncation de l'interférogramme plus douce que dans le cas où il serait uniquement borné par la limite de déplacement du miroir mobile. L'apparition de lobes latéraux sur les bandes obtenues après la transformée de Fourier est ainsi évitée.[170]

La méthode de remplissage connue sous le nom de "*zero-filling*" consiste, quant à elle, en l'ajout de zéros aux extrémités de l'interférogramme. Cela limite les pertes éventuelles liées à la discrétisation du spectre précédant la transformation de Fourier.[170] Un facteur 2 permet de doubler le nombre de points échantillonnés sans modifier le niveau de bruit. La résolution

---

[167] E. Furusjo, L. G. Danielson, E. Konberg, M. Rentsch-Jonas, B. Skagerberg, *Analytical Chemistry*, 70 (1998) p. 1726
[168] P. Fayolle, D. Pique, B. Perret, E. Latrille, G. Corrieu, *Applied Spectroscopy*, 50 (1996) p. 1325
[169] K. Rahmelow, W. Hubner, *Applied Spectroscopy*, 51 (1997) p. 160
[170] W. Herres, *HRGC-FTIR Theory and applications*, Hüthig Verlag (1987)

apparente du spectre se trouve donc, artificiellement, accrue. Enfin, la précision du pas d'échantillonnage du spectre (1,928 cm^{-1}) tient aux positions d'échantillonnage de l'interférogramme qui dérivent des zéros de la figure d'interférence produite par un laser hélium-néon.

5.1.3 Discussion : validité des enregistrements

Stabilité du signal

Les observations, pour les analyses de procédés par spectrométrie infrarouge à transformée de Fourier,[167,171] impliquent l'enregistrement d'une série de spectres sur une période relativement longue. Cependant, l'unicité de la mesure de référence ne permet de prendre en compte que les perturbations initiales. De fait, la justesse des observations subit les conséquences de l'existence d'instabilités, à long terme, de la réponse du spectromètre.

Les résultats présentés concernent des tests effectués sur une période de 6 à 8 heures, en respectant certaines conditions opératoires simples et quelques précautions évidentes assurant un isolement relatif du montage. D'ailleurs, les tests de stabilité effectués par les constructeurs sont toujours réalisés durant la nuit. En outre, il est recommandé de ne pas tirer d'air sec sur la colonne d'alimentation utilisée pour le spectromètre.[172]

La stabilité des enregistrements est estimée par des calculs de répétabilité spectrale, effectués sur les spectres après correction de la ligne de base. Les perturbations d'origine instrumentale sont ainsi prises en compte le mieux possible. La répétabilité mesure la capacité à donner le même résultat lorsque plusieurs observations successives d'un même échantillon sont effectuées. Cet estimateur est quantifié par le paramètre RSD (pour *Relative Standard Deviation* en anglais) qui est calculé par l'Équation 30.

$$\text{Équation 30} \quad RSD(\%) = \frac{s}{\bar{a}} \times 100$$ avec s l'écart type des mesures et $\bar{a}$ l'absorbance moyenne au nombre d'onde considéré.

[171] Y. Wang, B. Kowalski, *Applied Spectroscopy*, 46 (1992) p. 764
[172] C. A. Young, K. Knutson, J. D. Muller, *Applied Spectroscopy*, 47 (1993) p. 7

Le Tableau 5 regroupe les résultats des calculs de répétabilité obtenus sur 3 fenêtres centrées sur les maxima d'absorbance des bandes amide I et II (notées respectivement Ω_I et Ω_{II}) ainsi qu'à 1625 cm^{-1} (Ω_{1625}). Cette valeur particulière est généralement associée à la répétabilité la plus médiocre, résultant de la soustraction de spectres de solutions aqueuses.[173] Chacune de ces trois fenêtres s'étend sur 5 nombres d'onde successifs. C'est la valeur moyenne des résultats sur ces 5 positions qui est proposée dans le tableau.

Afin de caractériser les dérives instrumentales, trois solutions de protéines, a priori stables physiquement et chimiquement, sont suivies.

	Ω_I RSD (%)	Ω_{II} RSD (%)	Ω_{1625} RSD (%)
Solution d'acétamide dans l'eau *Référence eau*	**0,9 %**	**1,2 %**	**2,1 %**
Solution d'hémoglobine dans l'eau *Référence eau*	1,4 %	1,1 %	2,3 %
Solution d'hémoglobine dans le milieu "tampon" *Référence solution d'acétate de sodium*	1,9 %	2,1 %	3,5 %

Tableau 5 : Coefficients RSD (%) de solutions inertes.

Notons tout d'abord, même s'ils ne sont pas rapportés ici, que des résultats identiques ont été obtenus que le réacteur soit couvert ou à l'air libre. D'éventuels effets dus à l'évaporation ou à l'oxydation de la solution ne sont pas donc observés.

Concernant les solutions d'acétamide et d'hémoglobine dans l'eau, il s'agit de mélanges inertes qui ne subissent aucune évolution au cours du temps. Les valeurs de répétabilité calculées sont d'ailleurs satisfaisantes et en accord avec les résultats usuels. Nous observons un coefficient RSD de l'ordre de 1 % sur les bandes amides, de 2 % autour de 1625 cm^{-1}. Globalement, de tels niveaux de variation sont attribués, dans la littérature, à des changements locaux de température.[174] Effectivement, des perturbations dont l'amplitude peut atteindre 4 %, par degré Celsius, du signal spectral de la réponse ont été observées.[175] Par ailleurs, il est rapporté qu'une précision spectroscopique de l'ordre du pour-cent nécessiterait un contrôle au

[173] K. P. Ishida, P. R. Griffiths, *Applied Spectroscopy*, 47 (1993) p. 584
[174] J. A. De Haseth, *Applied Spectroscopy*, 36 (1982) p. 544
[175] D. B. Macbride, C. G. Malone, J. P. Hebb, E. G. Cravallo, *Applied Spectroscopy*, 51 (1997) p. 43

dixième de degré de la température de l'air ambiant.[176] Néanmoins, les mécanismes par lesquels la température affecte le spectre n'ont pas été clairement identifiés et d'autres auteurs incriminent plutôt des écarts incontrôlés de température entre le spectre de référence et les observations.[168]

Concernant l'hémoglobine dans son milieu réactionnel, une dégradation notable des résultats est observée. Les valeurs de répétabilité calculées sur les bandes amide sont, globalement, deux fois plus importantes tandis que le coefficient RSD atteint 3,5 % autour de 1625 cm^{-1}. Cela traduit, tout d'abord, la détérioration de la stabilité des mesures lorsque le mélange acide acétique / acétate de sodium constitue la référence. D'autre part, on constate que la compensation du signal attribué à l'eau est plus délicate, ce dont témoigne la valeur moyenne pour le RSD à 1625 cm^{-1}.

Nous avons également envisagé un autre critère, présenté Figure 63, pour apprécier la stabilité d'une mesure. Il s'agit de l'observation, en fonction du temps, des facteurs *score* résultant de l'analyse en composantes principales d'une matrice de données spectrales issue de l'échantillonnage d'une solution d'hémoglobine dans le milieu "tampon" en l'absence d'enzyme. Les résultats rendent compte de la dynamique et sont visuellement convaincants. La projection des spectres sur les axes principaux Cp1 et Cp3, exprimant respectivement 58% et 10% de la variance extraite, ne laissent pas entrevoir de disposition particulière au cours de l'avancement. Concernant l'axe Cp2, au contraire, il semble que l'on puisse déceler une certaine tendance au sein du nuage de points. Cette remarque est confirmée par le calcul d'une corrélation linéaire, faible mais significative au regard de la variance sur la pente. L'analyse des vecteurs *loading*, présentées sur la partie droite de la figure, est délicate. La contribution principale à la variance concerne les profils amide I et II dans leur ensemble. Néanmoins, les variations sont relativement faibles comparativement au niveau de bruit. De plus, un facteur uniquement positif traduit un effet global mais qui n'évolue apparemment pas de manière cohérente au cours du temps. Cet effet est probablement imputable à des variations de ligne de base de deuxième ordre, puisqu'il peut être gommé en observant les dérivées secondes des spectres. Le deuxième *loading* quant à lui caractérise le profil de la bande amide I. Il s'explique certainement par le manque relatif de stabilité de l'hémoglobine dans son milieu, même en l'absence d'enzyme.

[176] K. Rahmelow, W. Hubner, *Applied Spectroscopy*, 51 (1997) p. 160

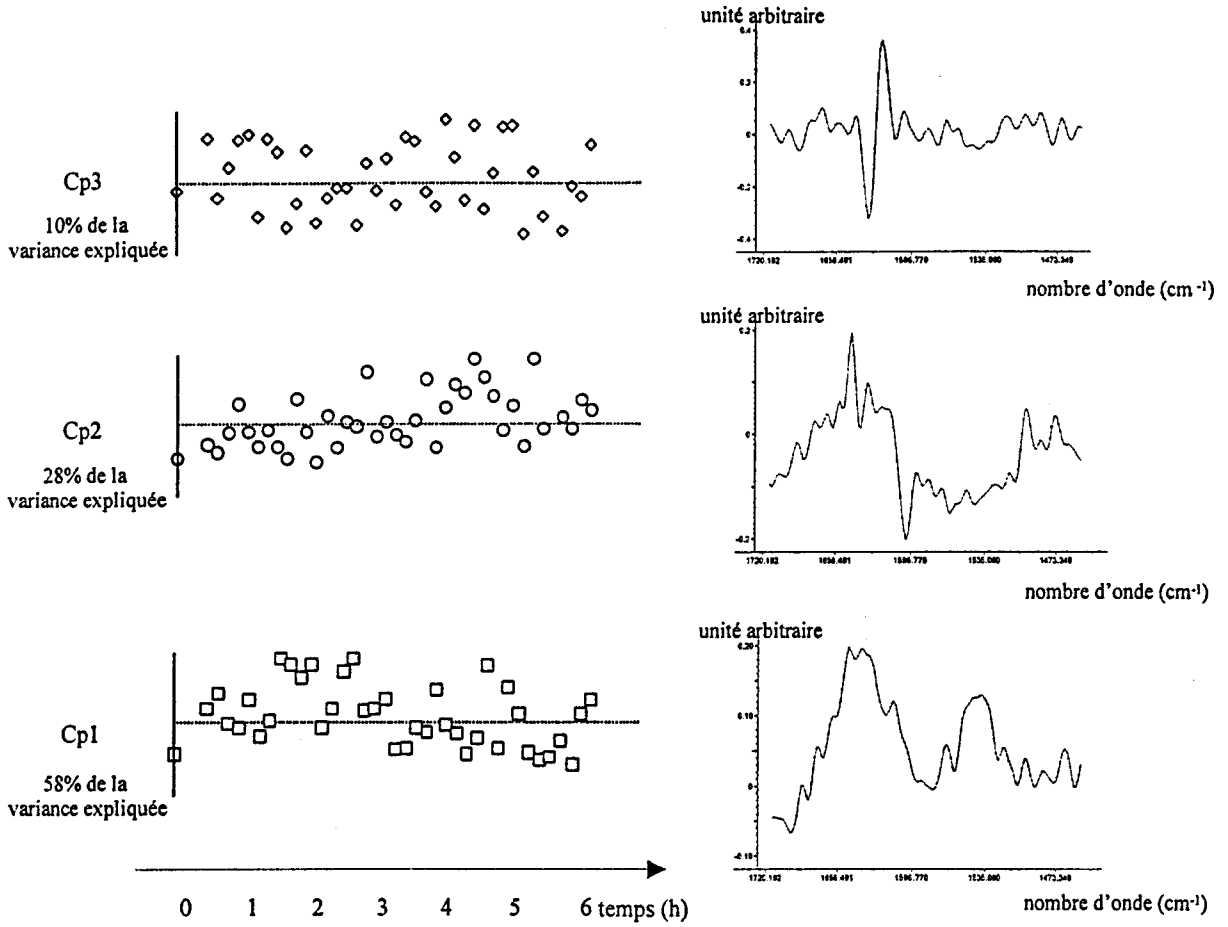


Figure 63 : Représentation dynamique des scores.

Observation qualitative d'une hydrolyse

Les calculs de répétabilité permettent de caractériser la stabilité. Néanmoins, ils peuvent également servir de référence pour quantifier la variabilité du signal, imputable au procédé d'hydrolyse.

Le Tableau 6 rapporte la moyenne des résultats obtenus sur les spectres enregistrés en cours d'hydrolyse d'hémoglobine bovine, c'est à dire en présence de pepsine. Les valeurs du Tableau 5, calculées pour une solution inerte dans les mêmes conditions, sont rappelées à titre de comparaison.

	Ω_I RSD(%)	Ω_{II} RSD(%)	Ω_{1625} RSD(%)
Solution d'hémoglobine dans le milieu 'tampon' Référence solution d'acétate de sodium	1,9 %	2,1 %	3,5 %
Hydrolyse d'hémoglobine dans le milieu 'tampon' Référence solution d'acétate de sodium	**4,2 %**	**4,7 %**	**9,3 %**

Tableau 6 : Coefficients RSD (%) pour une hydrolyse d'hémoglobine bovine.

Il apparaît que les variations spectrales, traduite par le coefficient RSD (%), sont deux à trois fois plus importantes lorsque les spectres sont enregistrés au cours d'une hydrolyse. Il existe donc une information propre au procédé, même si la Figure 64 montre que l'appréciation visuelle de celle-ci est discutable.

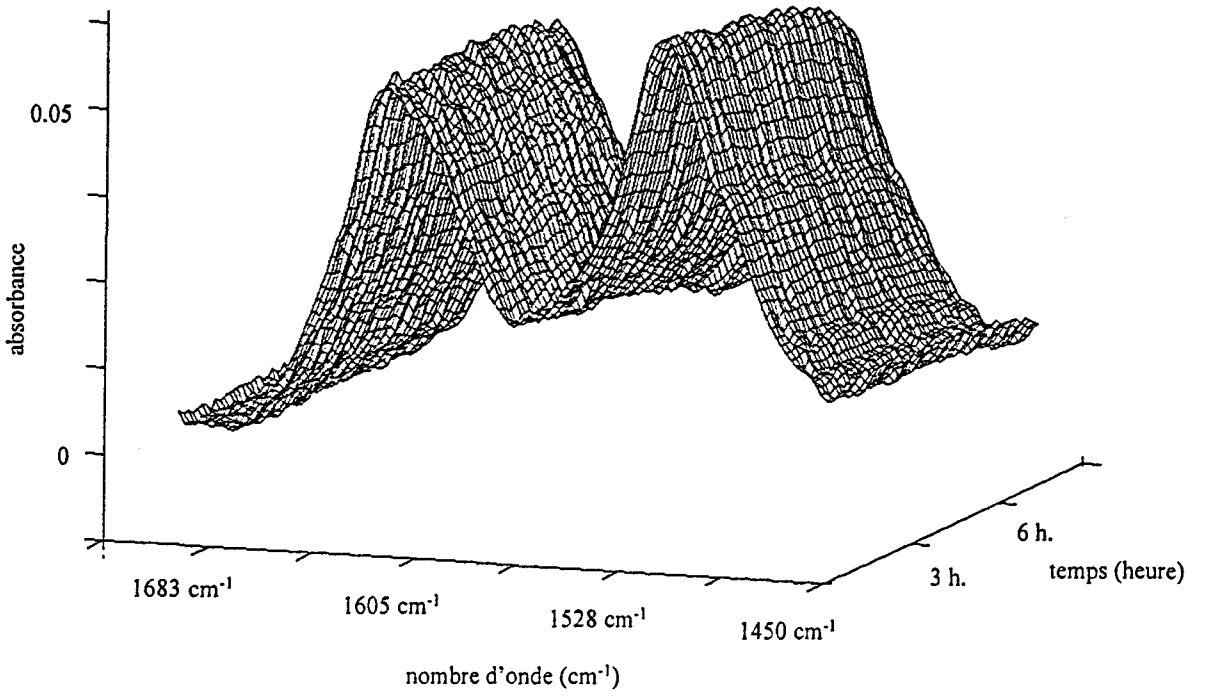


Figure 64 : Allures des spectres observés au cours du procédé.

La différence de signal entre un spectre enregistré au début de l'hydrolyse et un spectre correspondant à un état avancé du procédé est faible. Plusieurs justifications peuvent être

avancées notamment liées au mécanisme d'hydrolyse dont il est question. D'une part, seule une partie des molécules d'hémoglobine subit des modifications. D'autre part, aussi bien l'hémoglobine native que la gamme complète des peptides sont présents dans le milieu du début à la fin du procédé. Enfin, les variations observées sont probablement liées à des changements de conformation qui peuvent s'avérer très faibles comparativement à la quantité de protéine adsorbée sur le cristal.[177]

Les écart-types des spectres, enregistrés lors des expériences présentées dans le Tableau 6, sont illustrés Figure 65 et confirment les observations précédentes. Etant donné que les niveaux d'absorbance des spectres de la solution inerte et des spectres de l'hémoglobine sont équivalents, une comparaison semi-quantitative directe de ces profils est envisageable. Ils ne sont pas homothétiques, le spectre de déviation standard du signal analytique observé en cours d'hydrolyse se distinguant de celui représentant une solution inerte.

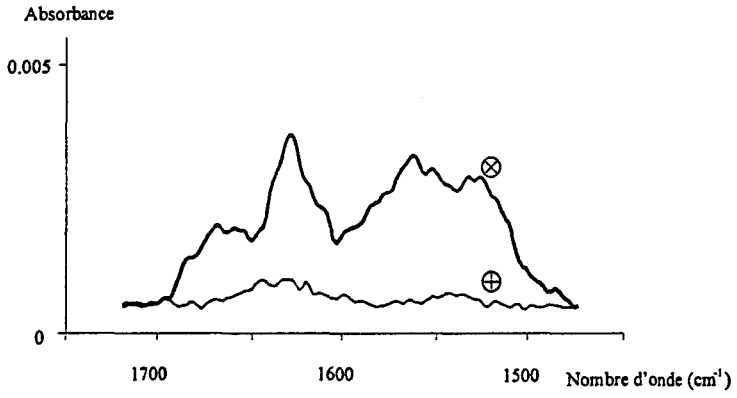


Figure 65 : Ecarts-types des données spectrales *(⊕ pour une solution inerte ⊗ pour une hydrolyse d'hémoglobine).*

Finalement, c'est la projection des spectres sur la première composante produite par la PCA de la matrice spectrale qui caractérise le plus nettement une certaine évolution en fonction de l'avancement de la réaction.

A ce propos, la Figure 66 regroupe les facteurs *score* et les vecteurs *loading* correspondant aux trois axes principaux extraits préférentiellement lors de la nouvelle analyse réalisée. Le premier *loading* s'interprète comme une variation d'ensemble des spectres, représentant 91 % de la variance, qui est dans une certaine mesure suffisante pour exprimer le classement des échantillons. Il regroupe d'ailleurs la plupart des informations spectrales d'intérêt, une certaine importance est donnée au signal à 1625 cm^{-1} ainsi qu'à la bande amide II.

[177] K. Fahny, *Biophysical Journal*, 75 (1998) p. 1306.

Ces résultats permettent de pointer le signal analytique caractéristique du déroulement du procédé, le distinguant clairement des variations de fond associées à l'instrumentation. Il s'agit d'une étape longue et laborieuse, mais essentielle. Certes, la faiblesse des niveaux de signal observés est lié à l'emploi de fibres optiques dans le moyen infrarouge. Par contre, le manque apparent de spécificité du signal est plutôt imputable à l'état physique de l'échantillon, au grand nombre de liaisons peptidiques présentes, ainsi qu'au faible pourcentage de ces liaisons qui subissent un clivage.

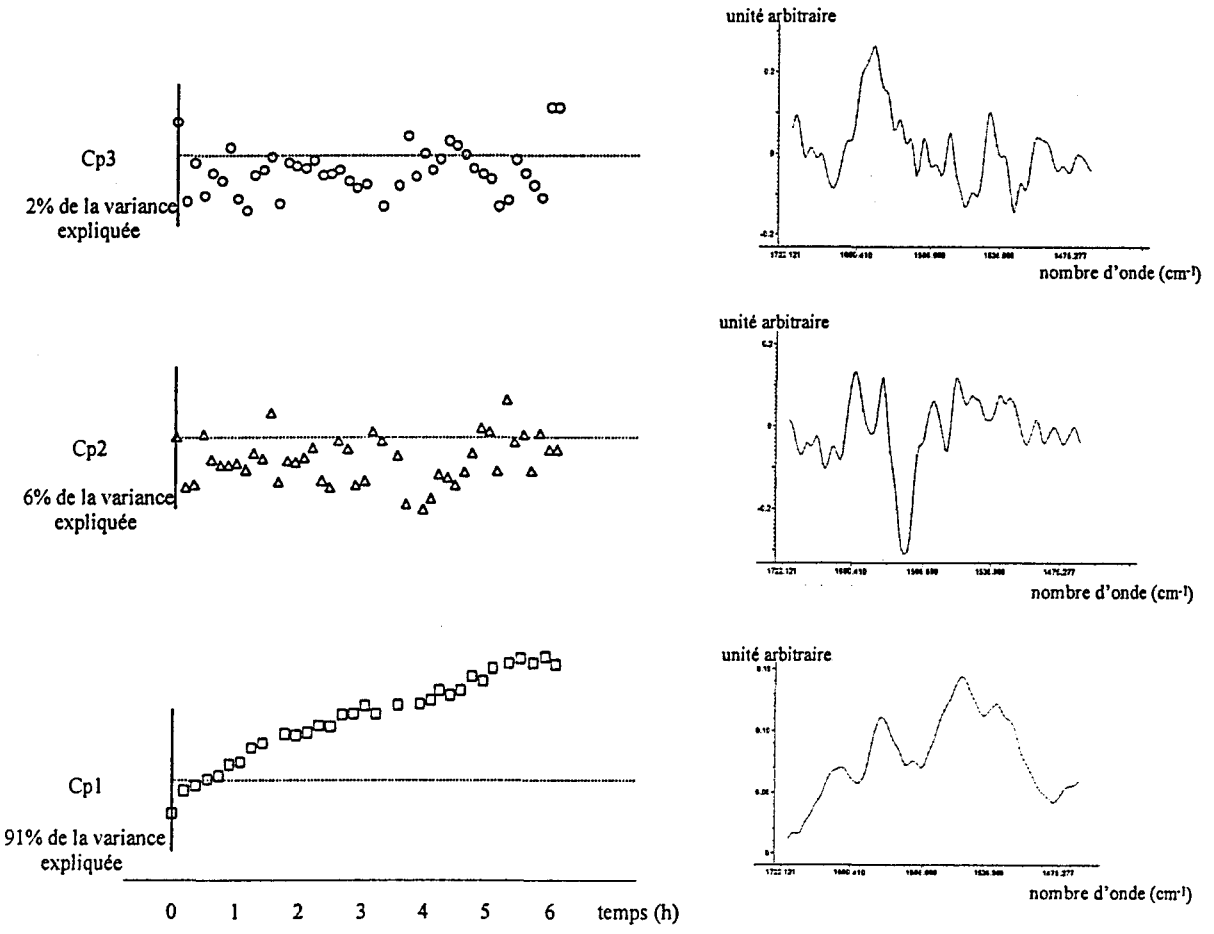


Figure 66 : Dynamique des scores pour une hydrolyse d'hémoglobine bovine.

5.2 Caractérisation des données

Nous caractérisons tout d'abord la reproductibilité de l'expérience. Afin d'accumuler un nombre intéressant de données pour la construction d'un modèle, la manipulation est renouvelée dans des conditions expérimentales les plus semblables possible.
Concernant les analyses de référence sur échantillons prélevés, des comportements propres à chaque répétition peuvent être décelés. Cette caractéristique des procédés enzymatiques induit un certain nombre de difficultés. Néanmoins, elle constitue la motivation principale pour la recherche de nouveaux capteurs ou de méthodes instrumentales innovantes.
D'autre part, la visualisation de l'ensemble des spectres par des méthodes de projection révèle l'existence de groupes de données (*clusters* en anglais) ainsi que la présence de non-linéarités. Nous serons donc amenés à construire des modèles capables de gérer ces disparités.

5.2.1 Mise en place de l'étalonnage

Les matrices de données, utilisées pour la construction des modèles, se composent de données spectrales auxquelles sont associées les valeurs de DH mesurées.

Répétitions du procédé

Nous exploitons les échantillonnages réalisés sur six manipulations d'hydrolyse de l'hémoglobine bovine dans le milieu "tampon". Celles-ci sont des répétitions du même procédé, le réacteur de laboratoire étant vidé puis rechargé en matières premières pour chacune d'entre-elles. Ces expériences sont regroupées dans le Tableau 7. Chaque répétition r_i est décrite par sa durée, le nombre de spectres observés et le nombre de prélèvements réalisés. Exception faite de l'expérience r_3 qui a été stoppée précocement, toutes les répétitions sont analysées sur une durée équivalente, autour de 8 heures. Par ailleurs, le nombre de prélèvements est sensiblement le même pour toutes les hydrolyses.
Les spectres aberrants, présentant des caractéristiques hors du commun, ne sont pas considérés, et sont à soustraire de la comptabilité du Tableau 7. La matrice complète comporte finalement 238 spectres.

Répétition	Nombre de spectres observés	*Nombre de spectres écartés*	Durée	Nombres de prélèvements
r_1	42	*1*	8h30	25
r_2	45	*2*	7h30	29
r_3	23	*0*	4h	14
r_4	45	*2*	8h	26
r_5	49	*5*	8h	22
r_6	44	*0*	8h	23

Tableau 7 : Répétitions du procédé.

Mesures de référence

La valeur du degré d'hydrolyse est estimée, par la méthode analytique de référence, pour chaque prélèvement. La Figure 67 présente les résultats obtenus pour les répétitions r_2 et r_4 du procédé.

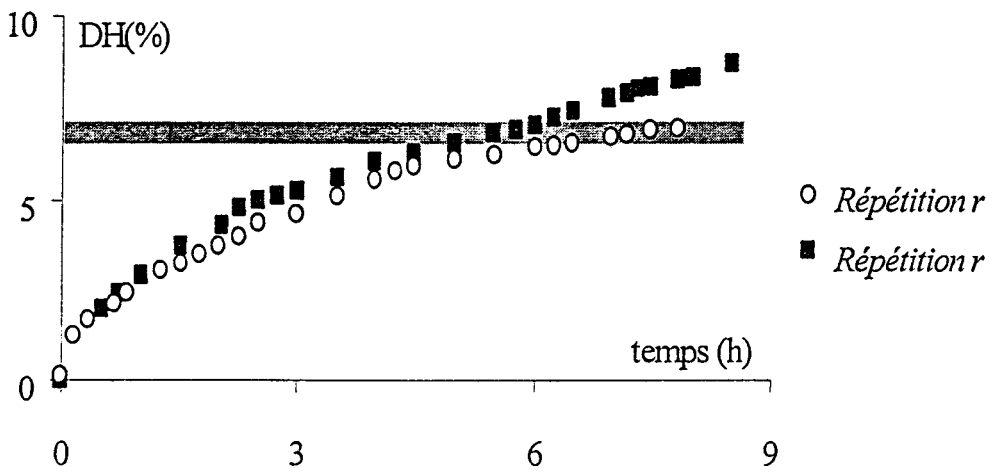


Figure 67 : Degrés d'hydrolyse de référence.

On constate que d'une répétition à l'autre, les valeurs mesurées sont, à un instant donné, notablement différentes. C'est pour cette raison qu'une mesure élémentaire du temps écoulé n'est pas suffisante pour caractériser l'avancement de la réaction, en terme de peptides

produits. Contrôler le procédé consiste à le maintenir autour d'un niveau donné, matérialisé par la bande grisée sur le graphe ; le paramètre pertinent est donc le DH. Par ailleurs, aucun écart important n'est observé entre échantillons voisins, c'est à dire lorsque les variations sur les abscisses sont faibles. Quant à la répétabilité des mesures de référence, une déviation standard de l'ordre de 0,2 % dans la gamme des DH peut être observée ; ce qui confirme de manière plus rationnelle cette constatation.
La lourdeur des analyses de référence empêche la détermination analytique d'un prélèvement pour chaque spectre. La construction des matrices d'étalonnage a donc nécessité l'interpolation linéaire de certaines valeurs du degré d'hydrolyse. Cette étape n'a pas présenté de difficultés particulières. Une description complète, échantillon par échantillon, pour toutes les répétitions est disponible en annexe 3.

Construction des lots d'échantillons

La construction des lots d'échantillons, qu'il s'agisse de ceux utilisés pour la phase d'apprentissage ou de celui permettant de tester le modèle, peut s'avérer délicate. Il peut être relativement simple et bon marché d'obtenir des données d'entrée telles que des spectres. Néanmoins, le problème réside souvent dans l'obtention des mesures de référence pour tous les prélèvements correspondant. Deux types de distributions des données sont alors envisageables[5] et le contenu des lots d'apprentissage, d'entraînement et de test est détaillé dans le Tableau 8 page 140.

• Les répartitions R_I, R_{II}, R_{III} des échantillons couvrent l'ensemble des variations possibles associées aux six répétitions r_i du procédé. Chaque lot est représentatif de toutes les classes de données présentes et les données sont distribuées de manière homogène sur les lots d'entraînement, de contrôle et de test (ce que représente la notation $r_1\ r_2\ r_3\ r_4\ r_5\ r_6$ du Tableau 8). Ces répartitions sont donc utilisées lors d'une stratégie prudente d'approche des données, lorsqu'il est avéré que celles-ci ne sont pas homogènes, pour construire un modèle global. Néanmoins, ce type de distribution ne convient que pour la prédiction a posteriori, à partir des spectres provenant d'une manipulation enregistrée dans sa globalité et intégrée dans les lots d'entraînement et de contrôle. Enfin, l'utilisation de trois répartitions distinctes des données, contenant les mêmes échantillons mais répartis différemment, permet d'une part d'éviter l'apparition de biais et, d'autre part, de moyenner les résultats obtenus.

• Les répartitions R_V et R_{VI} seront utilisés, au paragraphe V.4, pour estimer les capacités des modèles à généraliser lors de la présentation de lots inconnus. Pour ce faire, toutes les données descriptives d'une répétition sont exclues des lots d'apprentissage et consacrées uniquement au lot de test, bâti exclusivement sur r_5 (respectivement r_6) pour la répartition R_V (respectivement la répartition R_{VI}). Cette approche par les facteurs représentatifs considère ici que cinq répétitions sur six sont représentatives de l'ensemble des variations utiles.

Répartition	**Lot d'entraînement** Nombre et *répétition de provenance* des échantillons	**Lot de contrôle** Nombre et *répétition de provenance* des échantillons	**Lot de test** Nombre et *répétition de provenance* des échantillons
R_I	80 $r_1 r_2 r_3 r_4 r_5 r_6$	79 $r_1 r_2 r_3 r_4 r_5 r_6$	79 $r_1 r_2 r_3 r_4 r_5 r_6$
R_{II}	81 $r_1 r_2 r_3 r_4 r_5 r_6$	79 $r_1 r_2 r_3 r_4 r_5 r_6$	78 $r_1 r_2 r_3 r_4 r_5 r_6$
R_{III}	81 $r_1 r_2 r_3 r_4 r_5 r_6$	78 $r_1 r_2 r_3 r_4 r_5 r_6$	79 $r_1 r_2 r_3 r_4 r_5 r_6$
R_V	98 $r_1 r_2 r_3 r_4 r_6$	94 $r_1 r_2 r_3 r_4 r_6$	44 r_5
R_{VI}	97 $r_1 r_2 r_3 r_4 r_5$	97 $r_1 r_2 r_3 r_4 r_5$	44 r_6

Tableau 8 : Répartition des échantillons sur les lots *(les r_i représentent les répétitions du procédé qui sont intégrées dans chaque lot – Par exemple, seule la répétition r_5 du procédé est utilisée en phase de test pour la répartition R_V des échantillons).*

La répartition des échantillons sur les lots d'entraînement, de contrôle et de test s'effectue au hasard. Néanmoins, nous imposons la contrainte d'utiliser les données dont les DH ont été interpolés pour construire en priorité les lots de test, puis de contrôle, mais jamais d'entraînement. Chaque échantillon n'est évidemment utilisé que dans un seul lot et chaque lot contient un nombre à peu près semblable de données sauf lorsque les dimensions des lots de test sont imposées par l'échantillonnage des répétitions utilisées (pour R_V et R_{VI}). Il s'agit, lorsqu'on dispose de suffisamment de données, de la situation la plus favorable.

5.2.2 Visualisation des données

Accéder à la structure des données, c'est observer l'ensemble des spectres simultanément. Chaque donnée étant décrite par 141 nombres d'onde, il nous faut envisager une transformation de la représentation des données. Les méthodes de projection, associées généralement à une réduction de l'information, procurent des alternatives visuellement acceptables. Nous présentons les résultats obtenus par deux techniques, les analyses en composantes principales et les cartes de Kohonen. Ceux-ci sont suivis d'une discussion consacrée à la complémentarité, pour l'analyste, de leurs interprétations respectives.

Représentations issues de l'analyse en composantes principales

On effectue l'analyse en composantes principales sur la matrice **X**(238 ; 141) des spectres. La Figure 68 représente la projection des données dans le sous-espace engendré par les deux premiers axes principaux (Cp1 et Cp2), représentant respectivement 71 % et 12 % de la variance extraite.

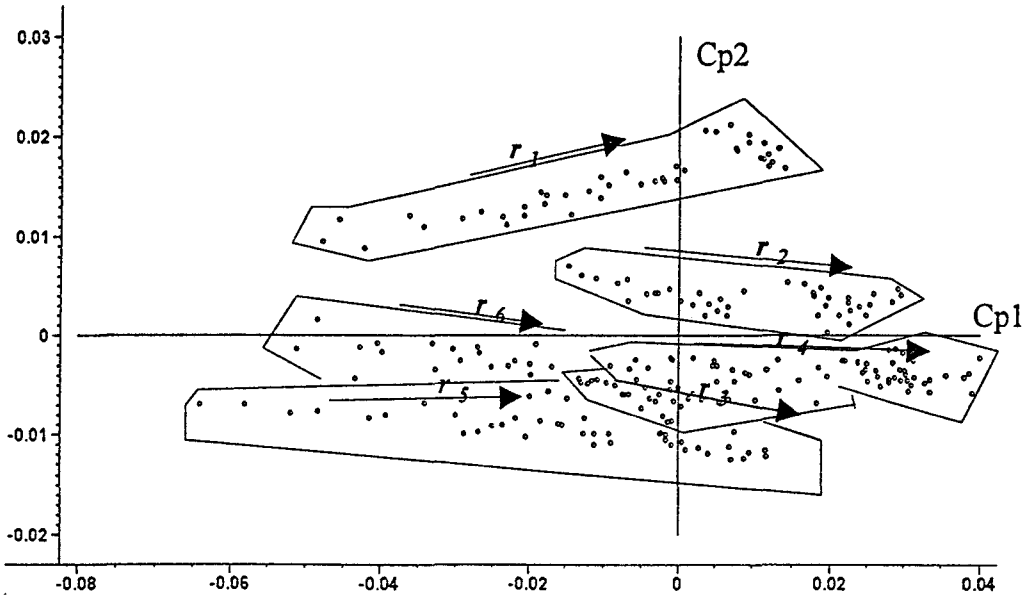


Figure 68 : Projection des données spectrales dans le plan (Cp1 ; Cp2).

Les projections représentant les spectres échantillonnés au cours d'une même répétition ont été encadrées et étiquetées. La flèche indique la direction dans laquelle les représentations des spectres ont tendance à se déplacer, au cours de l'avancement de la manipulation.
La Figure 69, quant à elle, concerne la projection des données sur le sous-espace engendré par les axes principaux (Cp2 et Cp3). La troisième composante concerne 7 % de la variance totale.

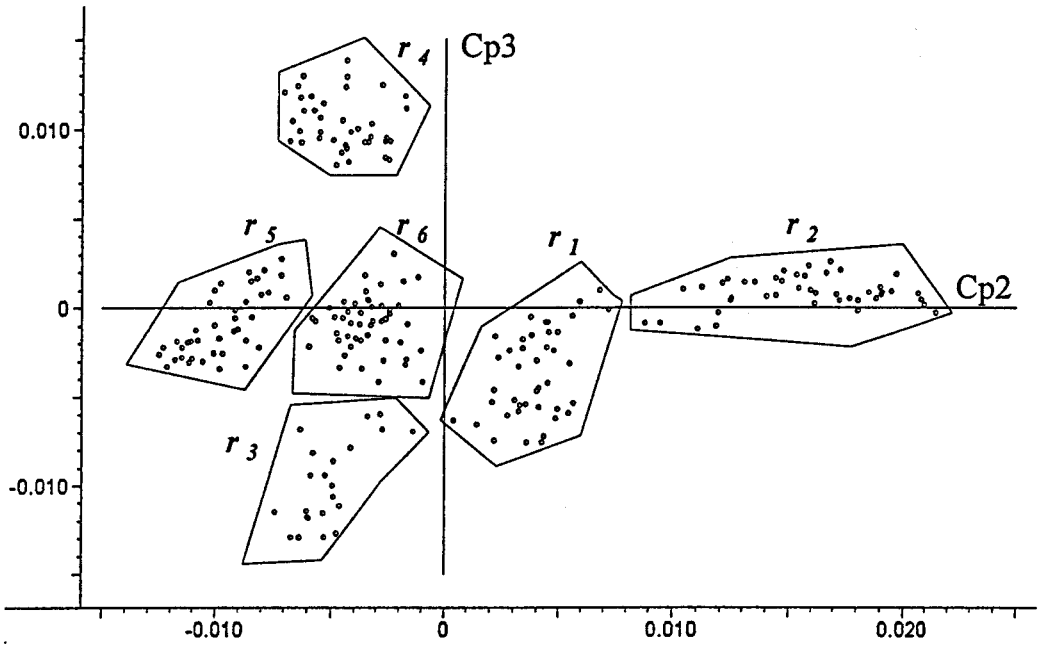


Figure 69 : Projection des données spectrales dans le plan (Cp2 ; Cp3).

L'analyse des vecteurs *loading*, dont résulte la construction de ces trois axes sera effectuée ultérieurement, dans le paragraphe consacré à la discussion.
Dans la pratique, ce type de représentation permet, comme présenté schématiquement au chapitre 3 paragraphe 1.1, la détection de points aberrants éventuels ainsi qu'une construction interactive et intelligente des lots de données.

Cartes de Kohonen

Nous utilisons également l'apprentissage non supervisé afin d'obtenir une représentation de la structure des données sur une carte à deux dimensions, dite de Kohonen. Les caractéristiques de chaque catégorie éventuelle des données sont trouvées de manière autonome par

l'algorithme d'apprentissage décrit au paragraphe 3.3.2, la représentation étant organisée de telle sorte que la topologie des entrées soit conservée. La Figure 70 et la Figure 71 illustrent cette façon de faire.

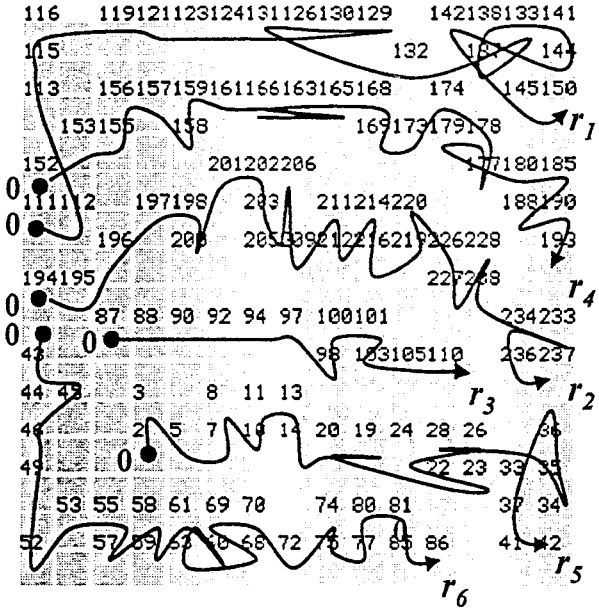


Figure 70 : Carte des caractéristiques.

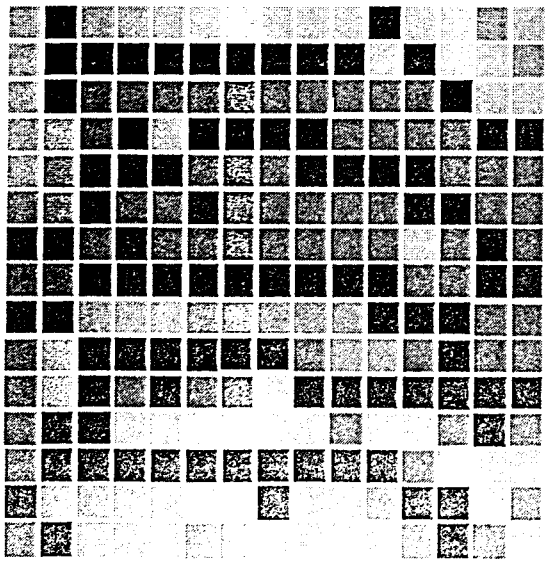


Figure 71 : Carte des activités.

Le choix des dimensions de la carte dépend des objectifs de la visualisation. Nous proposons ici les résultats obtenus pour une architecture produisant une carte carrée de côté 15 c'est à dire représentée par 15×15 neurones. En effet, observer la structure des données nécessite une carte composée de suffisamment de positions neuronales pour permettre une représentation des 238 vecteurs d'entrée qui ne soit pas trop contrainte. D'autre part, la version de l'algorithme d'entraînement de Kohonen utilisé impose la spécification de plusieurs paramètres.[156] Hormis la dimension de la carte, le rayon initial du voisinage dans lequel a lieu l'apprentissage et le taux d'apprentissage de départ doivent être choisis. Pour le premier de ces paramètres, nous préconisons une valeur équivalente aux dimensions de la carte, c'est à dire 15 dans le cas présent. Le taux d'apprentissage, quant à lui, est initialisé à une valeur très proche de 1. La modération de ces deux paramètres, après chaque itération, assure la convergence. Cela nécessite de spécifier un facteur de décroissance à un niveau proche de 1 (typiquement 0,999). En effet, une diminution lente du taux d'apprentissage garantit des représentations qui ne dépendent pas de la distribution initiale des poids. L'apprentissage est finalement stoppé lorsque le rayon du voisinage est inférieur à 1 ou lorsque le taux d'apprentissage atteint des valeurs proches de zéro, arbitrairement choisies.

La Figure 70 représente la carte des caractéristiques que nous obtenons. Les neurones vainqueurs, correspondant à chaque donnée d'entrée, sont étiquetés et les trajectoires représentant les parcours liés aux six répétitions sont grossièrement manuellement dessinées, puis orientées pour faciliter la lecture. Les unités de la carte repérées par **0** correspondent aux positions que l'apprentissage assigne au spectre initial de chaque répétition.

La carte des activités est présentée Figure 71. La distance spécifiée pour les calculs d'adéquation des vecteurs d'entrée avec les vecteurs poids est la distance euclidienne. Les niveaux de gris peuvent donc être interprétés comme des indications de la similarité entre les différentes unités. D'autre part, nous avons noirci les unités qui ne se voient assigner aucun échantillon.

Discussion

L'information majeure que partagent ces représentations est la présence de groupes au sein du lot de données. La Figure 68 et la Figure 69 montrent clairement que les projections des données des différentes répétitions échantillonnées occupent des zones distinctes. De plus, concernant la première représentation, il est possible d'extraire au sein de chacune des régions

une information à propos de la répartition des données les unes par rapport aux autres. En effet, la représentation des facteurs *score* des données spectrales sur l'axe horizontal Cp1 croit presque systématiquement avec l'évolution du procédé. Ainsi, pour chaque répétition, la projection de plus faible *score* correspond à l'enregistrement de l'hémoglobine bovine dans son état initial. Notons d'ailleurs que si les conditions initiales des répétitions étaient parfaitement identiques, ces points seraient superposés. En outre, les étendues des projections sur cet axe des différentes manipulations sont significativement variables. Cela laisse penser que la composante Cp1 n'est probablement pas, à elle seule, représentative de la corrélation au DH. Néanmoins, en l'absence de Cp1 comme par exemple dans le plan (Cp2 ; Cp3), il est très difficile d'extraire une tendance de l'observation des placements des projections les unes par rapport aux autres. Enfin, il convient de rappeler que chaque visualisation n'est pas représentative de la totalité de la variance, ce qui est indubitablement un inconvénient.

Les cartes de Kohonen sont de construction et d'interprétation plus techniques que les représentations issues de la PCA. Par contre, elles permettent un traitement visuel de la totalité de l'information qui peut s'avérer plus représentatif. C'est ainsi que, Figure 70, les manipulations semblables représentent des étendues comparables. De plus, les neurones vainqueurs correspondant aux premiers spectres de chaque répétition, marqués 0, sont trouvés dans un même voisinage. Cela semble indiquer qu'un état initial commun à toutes les répétitions peut être inféré des spectres. Une explication de cette aptitude remarquable tient probablement au fait que les valeurs initiales des poids de la carte sont soumises à une contrainte numérique du type de celle de l'Équation 14. Il a été démontré qu'imposer une telle interdépendance, en éliminant un degré de liberté aux poids, peut supprimer la variance associée à des décalages de ligne de base.[178] Enfin, les neurones vainqueurs ont tendance à se répartir de la gauche vers la droite, lorsque le procédé se déroule. Cela confirme l'observation faite dans le cas de la PCA concernant une répartition en fonction du degré d'avancement de la réaction. Néanmoins, il n'est pas possible de révéler de représentation unique de la cinétique. Les données se groupent en priorité selon la répétition considérée, avant d'être ordonnées. L'observation de la carte des activités, Figure 71, confirme cette constatation. Certains neurones ne sont jamais déclarés vainqueurs. Ceux-ci, représentés en noir, dessinent de véritables frontières[179] entre les classes formées par les données d'entrée.

---

[178] R. Difoggio, *Applied Spectroscopy*, 54 (2000) p. 94A
[179] L. Chen, J. Gasteiger, *Journal of the American Chemical Society*, 119 (1997) p. 4033

Bien que les données d'entrée présentent des groupes, il faut envisager la construction d'un modèle capable de les décrire dans leur ensemble. C'est typiquement dans ces conditions délicates que les RNA peuvent donner des résultats satisfaisants. Néanmoins, la prédiction d'une manipulation inconnue, que l'on envisagera à l'aide des répartitions R_V et R_{VI} de données, demandera une certaine forme d'extrapolation sur les données d'entrée qui devra être considérée avec précaution.

5.2.3 Observation de relations non-linéaires

La recherche d'éventuelles non-linéarités est partie intégrante du processus de caractérisation de la structure des données. L'existence de non-linéarités apparentes[75] signifie, la plupart du temps, l'abandon des méthodes linéaires pour le développement de l'étalonnage.

Entre les données d'entrée

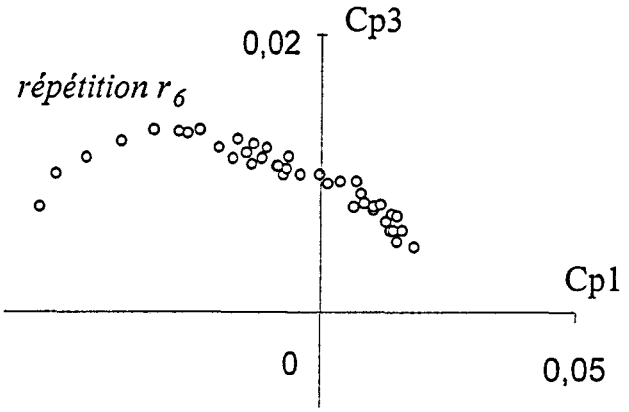


Figure 72 : Observation dans le plan (CP_1 ; CP_3).

Le degré de linéarité des variations existant au sein des données spectrales est révélé par la présentation des projections des échantillons dans les différents sous-espaces engendrés par la PCA. A titre d'exemple, la Figure 72 propose une visualisation des spectres caractérisant la répétition r_6 dans le plan engendré par le premier et le troisième axe principal.

Entre les données d'entrée et de sortie

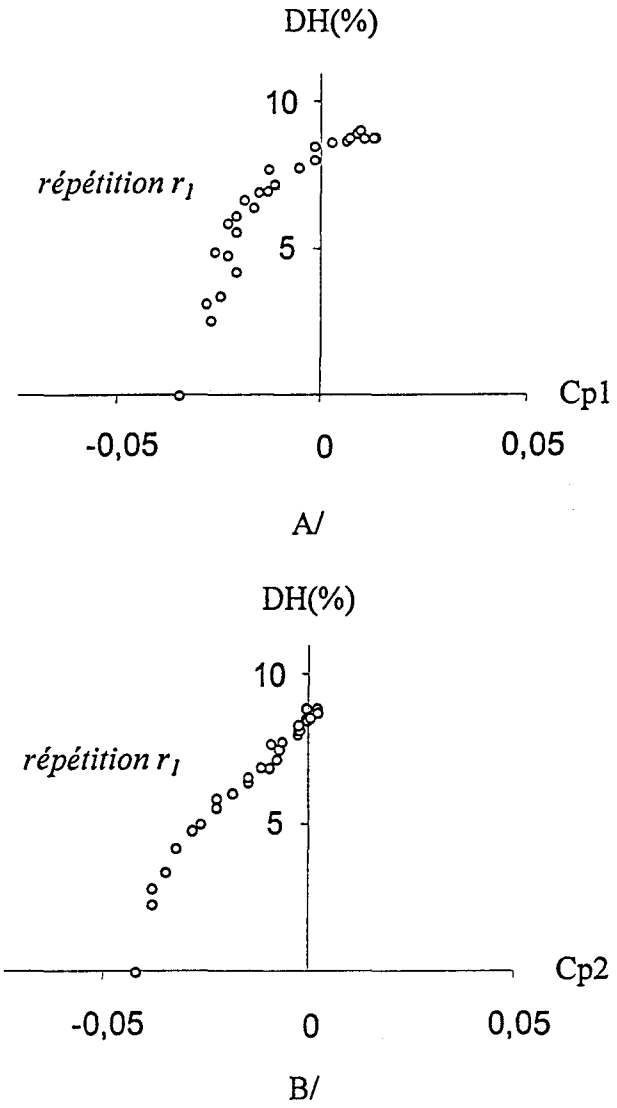


Figure 73 : Représentation du DH en fonction des scores de la Cp 1 (A/)

et de la Cp2 (B/).

La relation entre les données de sortie et les données d'entrée, dont la forme analytique est inconnue, peut également présenter un caractère non-linéaire. Nous représentons Figure 73 les valeurs analytiques du DH en fonction des facteurs *score* des échantillons de la répétition r_1.

Commentaires

De manière générale, les formes des nuages de points présentés ci-dessus montrent des écarts à un comportement purement linéaire. La Figure 72 présente les relations entre les données d'entrée sur des directions perpendiculaires de variance. Il n'est pas possible de décrire les données par une combinaison linéaires des deux variables (en l'occurrence Cp1 et Cp3). Un niveau intrinsèque de non-linéarités relativement important est ainsi constaté.

D'autre part, la Figure 73 de la page précédente atteste que les relations entre les valeurs de DH de référence et les scores sur les premiers axes principaux ne sont pas forcément linéaires, en particulier pour les valeurs de DH faibles c'est à dire en début de réaction. Nous sommes en présence de non-linéarités, certes modérées, mais avérées. Elles sont dues aux recouvrements de signaux, aux comportements du détecteur mais aussi probablement en majeur partie à la relation analytique elle-même. Quoi qu'il en soit, ces observations confortent le choix de techniques de modélisation non-linéaires, en particulier les RNA.

5.3 Construction et interprétation d'un modèle global

Le paragraphe précédent a justifié le choix de l'utilisation des RNA-FF pour l'étalonnage multivarié de la relation spectres / DH. Au-delà de la présence de corrélations non-linéaires, plus ou moins prononcées, c'est l'identification de classes au sein des données d'entrée et de sortie qui a motivé cette décision.

Nous construisons ici un modèle global, étalonné sur toutes les répétitions du procédé. Des données représentatives de toutes les classes, c'est à dire les répartitions R_I, R_{II} et R_{III} des données, sont donc utilisées et l'échantillonnage de l'espace d'entrée est le plus uniforme possible. Le développement d'un modèle performant est un travail délicat et interactif. Il nécessite d'optimiser à la fois l'apprentissage et la topologie du réseau, mais aussi la présentation des données. Ainsi, nous montrons que l'analyse en composantes principales, même si elle induit intrinsèquement une limitation concernant la prise en compte des non-linéarités, peut être une méthode de réduction adéquate des données.

5.3.1 Apprentissage

La phase d'apprentissage des RNA-FF est délicate, dépendante du problème et des données, et doit être abordée avec rigueur.

Estimation de l'erreur

Au cours de l'étalonnage, il faut pouvoir estimer à tout moment l'écart entre les sorties produites par le réseau et celles attendues, fournies par l'analyse de référence. La plupart des études utilisent des estimateurs basés sur l'erreur MSE[180] (*Mean Squared Error* en anglais), qui correspond au critère calculé par l'Équation 27. C'est également la grandeur qui est calculée par le logiciel SNNS.[156] La MSE est ici calculée à partir de données expérimentales n'ayant pas participé à la construction de l'étalonnage, on parle alors de méthode de validation externe.

[180] D. Svozil, V. Kvasnicka, J. Pospichal, *Chemometrics & Intelligent Laboratory Systems*, 39 (1997) p. 43

Les résultats que nous présentons concernent la racine carrée de la MSE, notée RMSE. Ce critère présente l'avantage d'être exprimé dans l'unité de la valeur analytique de référence. Par ailleurs, les valeurs numériques obtenues correspondent à des moyennes, lorsque les étalonnages sont répétés à partir de distributions initiales différentes des poids d'un réseau. En effet, les réseaux de neurones sont des outils sensibles aux conditions initiales.[181] Des valeurs particulières des paramètres peuvent ainsi éventuellement induire des situations ponctuellement favorables mais qui ne peuvent être en aucun cas considérées comme générales. Enfin, notons que les erreurs quadratiques donnent, de manière générale, plus de poids aux résidus importants ainsi qu'aux types de données dominants.

Paramétrage de l'algorithme

L'Équation 26 formalise l'utilisation d'un terme d'affaiblissement des poids pour améliorer la qualité de l'apprentissage. Elle résume l'originalité de l'algorithme RPROP. En effet, des valeurs excessives pour les paramètres ajustables engendrent une variance importante des données de sortie.[82] Certains résultats montrent d'ailleurs que l'amplitude des valeurs des poids peut jouer un rôle plus important que leur nombre, en ce qui concerne les capacités de généralisation.[182]

L'implémentation de RPROP dans le logiciel SNNS laisse trois degrés de liberté à l'utilisateur.[156] L'un d'entre-eux concerne l'amplitude initiale des modifications, notées Δ_{ji} dans l'Équation 24, qui se voit, par défaut lorsque les poids initiaux sont distribués de manière aléatoire, attribuer la valeur 0,1. Un autre paramètre spécifie la borne supérieure permise pour la mise à jour des poids. Fixé par défaut à une valeur arbitrairement élévée, la convergence n'y est sensible que dans certaines conditions particulières.[156] Ces deux premiers paramètres peuvent être modifiés en cours d'apprentissage, sans conséquence sur la qualité de celui-ci. Par contre, il appartient à l'expérimentateur de choisir correctement le troisième paramètre, noté α. Celui-ci détermine la relation entre l'erreur de sortie et l'importance de la réduction des valeurs des poids. Nous avons constaté des résultats satisfaisants pour des valeurs comprises entre 3 et 5.

[181] I. V. Tetko, D. J. Livingstone, A. I. Luik, Journal of *Chemical Information and Computer Science*, 35 (1995) p. 826
[182] P. L. Bartlett, M. C. Mozer, M. I. Jordan, T. Petsche, *Advances in Neural Information Processing Systems*, The MIT Press (1997)

Contrôle

L'apprentissage d'un RNA-FF est avant tout un problème d'optimisation qui nécessite deux lots de données indépendants entre-eux, mais également indépendants du lot de test. Le lot d'entraînement sert à déterminer les poids du modèle tandis que le lot de contrôle permet de diriger l'entraînement. Le sur-entraînement du réseau, c'est à dire typiquement la modélisation du bruit, est ainsi évité au mieux.

D'un point de vue pratique, la méthode d'*early stopping* permet, par le biais de l'observation de l'évolution de l'erreur RMSE de contrôle, l'arrêt de l'apprentissage. En effet, si les données sont distribuées de manière satisfaisante sur chacun des lots, l'erreur d'entraînement et l'erreur de contrôle montrent, dans un premier temps, un comportement homothétique, décroissant. Néanmoins, la courbe représentant l'erreur de contrôle s'incurve inévitablement, présente un minimum, puis croît de nouveau alors que l'erreur d'entraînement ne montre pas de changement de pente. Ce minimum correspond aux valeurs de poids les mieux adaptés pour l'apprentissage de la relation entrée / sortie à partir des données courantes.

Le nombre d'itérations nécessaire à l'obtention des résultats est variable. Nous avons observé qu'il dépend légèrement du nombre de poids et donc du taux de compression des données d'entrée. Néanmoins, il semble plus sensible à la forme utilisée pour la représentation des données, en particulier au degré de signification ou encore à la qualité de celles-ci. Enfin, les données de chaque lot sont, au cours de chaque itération, présentées dans un ordre variable et déterminé par le hasard.

5.3.2 Transformation des données

L'objectif des différents pré-traitements des données est d'améliorer la description numérique du problème, afin de rendre le processus d'apprentissage plus performant. Celui-ci étant supervisé, cette étape préliminaire de préparation concerne à la fois les données d'entrée et de sortie.

Transformation des données de sortie

La fonction de transfert utilisée pour le neurone de sortie est la fonction sigmoïde présentée Figure 24 page 64, qui produit des valeurs dans l'intervalle]0 ; 1[. Les valeurs cibles, de sortie, doivent donc nécessairement être comprises à l'intérieur de cette gamme. De plus, il est préférable, pour profiter des capacités intrinsèques de la fonction sigmoïde, qu'elles couvrent la plus grande partie de celle-ci.

En conséquence, les valeurs de DH ont systématiquement été redistribuées dans un intervalle borné par les valeurs 0,1 et 0,9. L'Équation 31, adaptée de la forme générale,[75] présente cette mise à l'échelle (*scaling* en anglais).

$$\text{Équation 31} \quad x_i' = \frac{x_i - x^{min}}{x^{max} - x^{min}} \times 0{,}8 + 0{,}1 .$$

La notation x_i désigne la valeur de l'intervalle de départ, dont les bornes supérieures et inférieures sont désignées respectivement par x^{max} et x^{min}. La notation x_i' désigne DH exprimé dans le nouvel intervalle [0,1 ; 0,9].

Les erreurs RMSE calculées sur le lot de test (RMSEP) correspondent donc à des unités dans cette gamme. Lorsque cela présente un intérêt, ces erreurs seront exprimées dans la gamme de DH des résultats analytiques [0 % ; 8,7 %]. Cela nécessite de calculer à nouveau le critère sur les valeurs de sortie converties.

Transformation des données spectrales

La normalisation des vecteurs d'entrée présente une influence moindre dans le cadre des RNA-FF, sauf si cette transformation a pour motivation une propriété physico-chimique. En effet, cette transformation peut être annihilée par des modifications des paramètres de poids et de biais, les variables d'entrée étant combinées linéairement sur la couche cachée. Notons que ce n'est pas le cas pour les réseaux de Kohonen ; les données d'entrée subissent donc une transformation préalable.[50]

Dans cette étude, nous n'avons pas constaté d'exigence particulière en ce qui concerne d'éventuelles normalisations des données d'entrée. Par contre, la différentiation des données spectrales peut présenter de nombreux avantages, que ce soit concernant la correction d'effets

de ligne de base, du premier ou du deuxième ordre ou la décomposition de massifs, qui facilite l'interprétation.

Nous présentons donc les résultats obtenus lorsque les données d'entrées sont exploitées sous leur forme initiale, en dérivée d'ordre 1 et en dérivée d'ordre 2. A cet effet, nous avons utilisé la méthode de Savitzky et Golay[49], avec un polynôme de degré deux et une fenêtre spectrale de 11 points. Le Tableau 9 regroupe les erreurs RMSE obtenues en apprentissage et en test pour une architecture 141×4×1, construite sur la base des spectres complets. L'ordre de grandeur du nombre d'itérations est de l'ordre de 300 à 500. Il faut noter que le nombre de paramètres ajustables, associé à la topologie du modèle, ne répond pas strictement au critère assurant la détermination du problème.

Forme des spectres	Entraînement RMSE (%)	Contrôle RMSE (%)	Test RMSEP (%)
Non dérivés	**0,05**	**0,05**	**0,10**
Dérivée première *Savitsky et Golay*[49]	0,07	0,08	0,11
Dérivée seconde *Savitsky et Golay*[49]	0,06	0,07	0,13

Tableau 9 : Erreurs obtenues sur les spectres complets en fonction l'ordre de la dérivée
(moyenne des valeurs obtenues pour les répartitions R_I, R_{II} et R_{III} dans la gamme [0,1 ; 0,9]).

Le Tableau 10 montre le même type de tendance, lorsque l'on répète l'expérience en utilisant des lots d'échantillons construits à partir des facteurs *score* de la PCA. L'architecture proposée a priori est alors 10×5×1 et l'apprentissage nécessite de l'ordre de 200 itérations.

Ces résultats concernent des réseaux qui n'ont pas été optimisés, ni du point de vue de la réduction du nombre de neurones d'entrée, ni du point de vue du nombre d'unités intermédiaires. Ils montrent toutefois, sans ambiguïté, que l'utilisation d'une forme dérivée des spectres ne présente aucun intérêt en terme de qualité de l'apprentissage ou de précision de la prédiction. Par ailleurs, ils illustrent très simplement l'effet d'un meilleur respect du critère concernant le rapport entre le nombre d'échantillons et le nombre de liens composant le modèle. Les valeurs de RMSEP les plus faibles sont en effet systématiquement observées lorsque les scores sont utilisés, quel que soit l'ordre de la dérivée des spectres initiaux.

Forme des spectres	Entraînement RMSE (%)	Contrôle RMSE (%)	Test RMSEP (%)
Non dérivés	**0,04**	**0,05**	**0,07**
Dérivée première *Savitsky et Golay*[49]	0,06	0,08	0,10
Dérivée seconde *Savitsky et Golay*[49]	0,07	0,07	0,09

Tableau 10 : Erreurs obtenues sur les scores en fonction l'ordre de la dérivée *(moyenne des valeurs obtenues pour les répartitions R_I, R_{II} et R_{III} dans la gamme [0,1 ; 0,9]).*

La différentiation des données spectrales présente de nombreux avantages détaillés au chapitre 2. Néanmoins, la dégradation du rapport signal sur bruit qu'elle engendre inévitablement, même lorsque la différentiation de *Savitsky* et *Golay* est employée, peut s'avérer rédhibitoire. Il semble que ce soit le cas ici et, plus généralement, lorsque la qualité des données spectrales utilisées n'est pas aussi satisfaisante que celle associée aux spectres infrarouge enregistrés dans des conditions plus adéquates, en transmission par exemple.
C'est le spectre complet qui contient en principe le maximum d'informations pouvant être obtenues.[87] En effet, comme toute opération de pré-traitement modifie le signal, elle n'est utile que si l'information négligée n'est pas significative. Par contre, si cette dernière s'avère importante, une dégradation notable des résultats peut survenir.

5.3.3 Compression des données, topologie

Quel que soit le modèle construit, le rapport du nombre d'échantillons au nombre de poids du réseau doit être maximum. C'est le principe de parcimonie, dont nous avons discuté au chapitre 3, qui impose de diminuer le plus possible le nombre de paramètres ajustables du réseau. Cela rend le modèle plus facile à entraîner, plus robuste et plus lisible. Il convient donc, et nous avons déjà succinctement abordé le sujet, de réduire la dimension du vecteur représentant les données d'entrée en projetant celles-ci dans un sous-espace représentatif. Deux possibilités ont été envisagées pour la transformation des données, l'analyse en composantes principales et la compression par ondelettes. Après optimisation de la dimension

de la couche intermédiaire, nous comparons les topologies respectives des réseaux construits ainsi que leurs capacités prédictives.

Méthodologie

Nous avons précédemment montré que le lot de donnée transformé présente un caractère non linéaire aussi bien du point de vue des variables latentes (Figure 72) que concernant la relation entre les données d'entrée et de sortie (Figure 73).

L'utilisation de l'analyse en composantes principales pour la transformation d'un lot de données où des non-linéarités sont suspectées a été discutée au Chapitre 3. En outre, il faut noter que cette transformation n'est pas, intrinsèquement, une méthode de compression. Elle ne le devient que lorsque les composantes, qui ne sont pas considérées comme significatives, sont négligées. Ainsi, toute la difficulté réside dans le choix de cette coupure qui engendre inévitablement une perte d'information.

La méthodologie est différente dans le cas des transformées en ondelettes. En effet, la dimension des données est réduite d'un facteur deux lors de chaque étape de filtrage qui produit les coefficients d'approximation et de détails. Par ailleurs, les composantes principales ont un caractère plutôt global alors que les coefficients issus de la compression par ondelettes possèdent des caractéristiques locales.[130]

Quelle que soit la méthode de compression choisie, l'indépendance des lots de données d'entraînement, de contrôle et de test reste une priorité. Ainsi, les sous-espaces sur lesquels sont décrits les lots d'échantillons sont d'abord déterminés à partir du lot de données d'entraînement. Ensuite, les échantillons de contrôle et de test sont projetés dans ce sous-espace, et leurs nouvelles coordonnées sont calculées.

Concernant l'optimisation de la topologie des RNA, la méthodologie mise en place ne dépend pas de la transformation choisie. Tout d'abord, il faut préciser que la couche de sortie, composée d'un neurone unique retournant la valeur analytique du DH, n'est évidemment pas concernée. A ce propos, nous avons constaté, de manière plus générale, que les modèles consacrés à la prédiction d'une seule réponse à la fois sont plus performants,[88] sauf lorsque les variables de sortie sont corrélées.[101] En outre, notre intérêt se porte exclusivement à des réseaux composés d'une seule couche intermédiaire qui garantissent, sous certaines conditions, la propriété d'approximation universelle.[77,116] Il n'existe malheureusement pas de règle théorique concernant le choix du nombre de neurones cachés, relié au compromis biais-

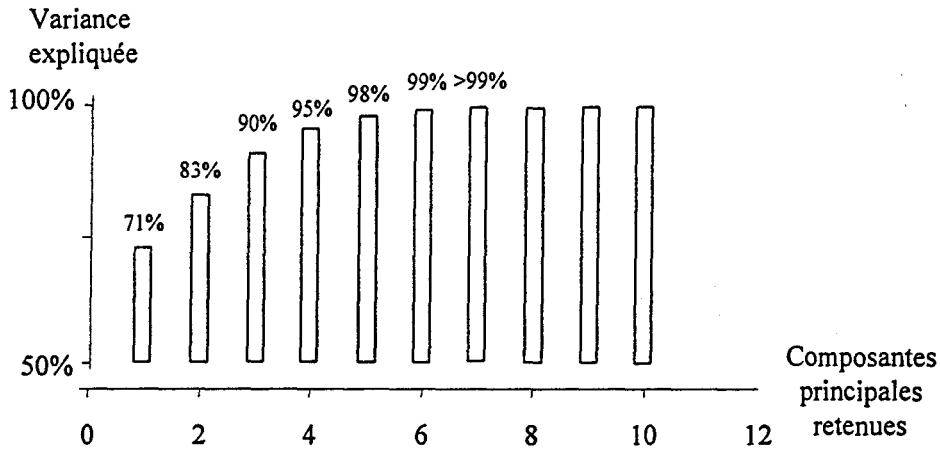
variance[120] dont nous avons discuté au chapitre 3 et particulièrement à propos de la Figure 44. Les critères empiriques proposés pour ce choix devraient tous tenir compte de la complexité de la fonction (inconnue) modélisée, du nombre de données pour l'apprentissage ou du niveau de bruit des données de sortie. Ainsi, il n'y a pas réellement de meilleur moyen que de procéder par essais successifs et d'estimer indépendamment à chaque fois l'erreur produite par le lot de test. Par la suite, la minimisation du nombre d'unités intermédiaires sera systématiquement recherchée. Concernant la couche d'entrée des RNA-FF, une méthode très simple permet d'optimiser le nombre de facteurs *score* de la PCA ou le nombre de coefficients pour les ondelettes. En effet, une approche systématique peut être envisagée, basée sur l'observation de l'erreur associée à la présentation des échantillons du lot de test, lorsque le nombre de neurones d'entrée est progressivement réduit à partir d'une taille délibérément large. La courbe décrite par l'erreur de prédiction en fonction de la complexité décrit alors inévitablement un minimum. Cette approche est très sûre mais elle impose de réitérer la phase d'apprentissage pour chaque topologie construite. Il s'agit d'un inconvénient et des algorithmes permettant une sélection plus automatique ont d'ailleurs été étudiés.[183,184]
Néanmoins, pour un nombre donné de variables d'entrée, la performance reste en partie fonction du nombre de neurones cachés et la difficulté est d'effectuer les deux optimisations conjointement. Cette approche peut être facilitée par l'observation des poids des liens, et des activités des neurones, lorsque l'analyse de l'architecture du réseau est visuellement acceptable.

Facteurs scores de l'analyse en composantes principales

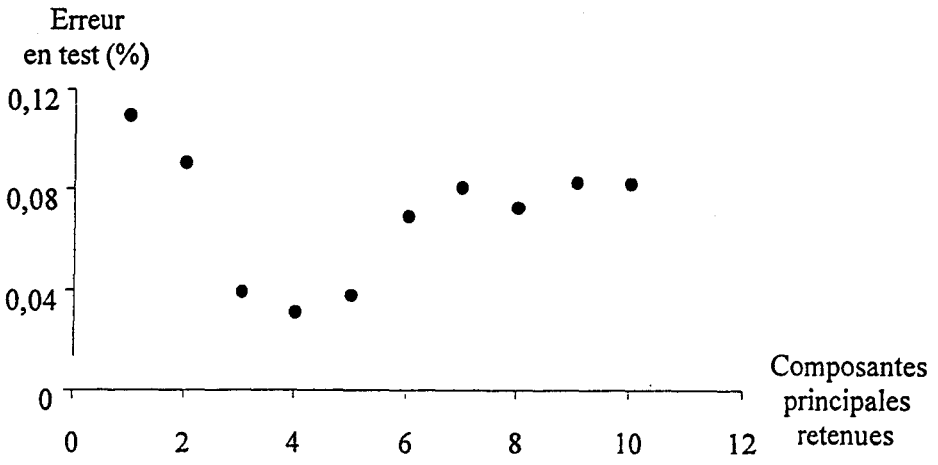
Les facteurs *score* représentent les projections des données spectrales sur les axes principaux générés par l'ACP. A ce propos, les dix premières composantes contiennent du point de vue de la variance quasiment toute l'information disponible dans la matrice des données (soit plus de 99 % de la variance extraite). C'est le sens de la Figure 74A/.
Même si une partie de cette information n'est probablement pas significative, la couche d'entrée est initialement constituée de 10 neurones. Cette valeur a été choisie comme la dimension limite pour garantir la parcimonie et éviter la sur-modélisation du problème. Ce

[183] F. Despagne, D. L. Massart, *Chemometrics & Intelligent Laboratory Systems*, 40 (1998) p. 145
[184] L. G. Weyer, S. D. Brown, *Journal of NIR spectroscopy*, 4 (1996) p. 163

nombre d'unités est ensuite réduit pas à pas, jusqu'à un seul neurone, tout en observant les capacités de prédiction du réseau. Une représentation graphique des résultats est proposée Figure 74B/. Les répartitions d'échantillons R_I, R_{II} et R_{III} ont là encore été testées à chaque fois. Il n'a jamais été observé de comportement particulier de l'un ou l'autre des lots, ce qui confirme que les distributions des échantillons y sont équivalentes.



A/ Variance expliquée



B/ Optimisation de la couche d'entrée

Figure 74 A/ Variance expliquée ET B/ : Influence du nombre d'axes principaux retenus *(pour 5 neurones intermédiaires, moyennés pour les répartitions R_I, R_{II} et R_{III}, dans la gamme [0,1 ; 0,9]).*

En ce qui concerne les paramètres, les meilleurs comportements lors de l'apprentissage ont été obtenus pour un coefficient α de RPROP égal à 5 ; seule une faible modification des poids est permise.

Le profil de la Figure 74B/ est typique de l'évolution de l'erreur de prédiction en fonction du nombre de composantes principales retenues. Il montre que le meilleur modèle, du point de vue du critère RMSEP, peut être construit à partir de la considération des scores des quatre premiers axes principaux. Comme le précise la Figure 74A/, ces quatre directions propres prennent en compte 95 % de la variance présente dans le lot de données. Ainsi, le plus faible critère RMSEP obtenu est 0,03, pour des DH exprimés dans l'intervalle [0,1 ; 0,9],. Cela correspond à une erreur de 0,3 % dans la gamme des valeurs initiales de DH (%), soit [0 % ; 8,7 %]. D'autre part, on constate qu'incorporer uniquement les deux premiers axes, soit 83 % de la variance totale, ne permet pas d'atteindre des niveaux d'erreur corrects lors de la prédiction. A l'opposé, la considération de plus de 5 neurones entraîne une dégradation de la qualité de prédiction, autour de 0,08. Néanmoins, il y a une tendance à la stabilisation de l'erreur à partir de cette sixième composante. La prise en compte du signal indubitablement associé à un bruit ne semble donc pas détériorer outre mesure les aptitudes du modèle. Il présente ainsi un caractère relativement robuste vis à vis d'un certain niveau de bruit, lorsque celui-ci est également présent au sein des données d'entraînement.

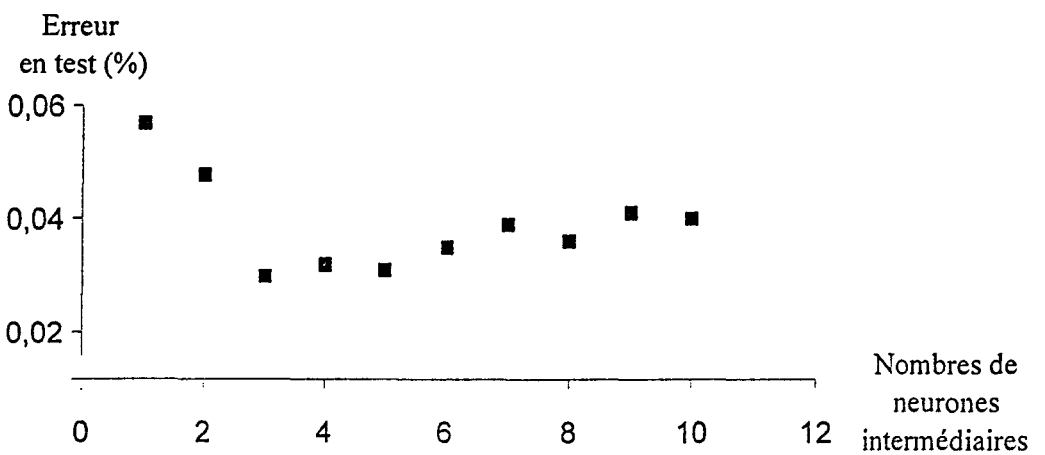


Figure 75 : Optimisation de la couche intermédiaire (pour 4 neurones d'entrée, moyennés pour les répartitions R_I, R_{II} et R_{III}, dans la gamme [0,1 ; 0,9]).

L'analyse de la Figure 75 précise la plus faible erreur obtenue lorsque différents nombres de neurones intermédiaires sont testés, pour une couche d'entrée constituée des quatre premiers neurones. Pour trois neurones cachés, l'erreur RMSEP affiche 0,03, soit 0,3 % lorsqu'elle est convertie dans la gamme des valeurs analytiques. Au-delà de trois neurones cachés, le niveau d'erreur croît lentement puis a plutôt tendance à se stabiliser. Cette observation est probablement liée au critère d'arrêt sur la validation que nous avons choisie. Effectivement, l'emploi de la méthode d'*early stopping* réduit l'influence du nombre de neurones cachés, puisqu'il préserve du phénomène de sur-entraînement.[185] Néanmoins, il peut parfois s'avérer utile de conserver une dimension relativement importante pour la couche intermédiaire afin d'éviter l'écueil que constituent les mauvais optima locaux.[134]

Lorsque les facteurs *score* tiennent lieu de données d'entrée, la topologie du réseau a été optimisée et peut être notée 4×3×1. L'erreur associée à la prédiction du DH à l'aide de ce modèle global est alors proche de 0,3 % (0,27 % exactement) dans la gamme [0 % ; 8,7 %]

Coefficients des ondelettes

L'approche de compression des spectres par ondelettes ne permet pas de démarrer l'optimisation à partir d'un nombre déterminé de neurones d'entrée. En effet, pour chaque famille de fonctions choisie (nous utiliserons uniquement les ondelettes de *Daubechies*), plusieurs niveaux de compression sont possibles et un nombre donné de coefficients d'approximation et de détails correspond à chaque niveau. Cet aspect a été présenté dans le chapitre 3. De plus, il est envisageable d'utiliser l'un ou l'autre de ces types de coefficients, voire les deux.

Les spectres ont été tronqués de telle sorte qu'ils soient décrits par 128 valeurs d'absorbance. Ils présentent ainsi un nombre de descripteurs égal à une puissance de deux. Nous avons entraîné des réseaux construits sur les lots de coefficients d'approximation produits par les familles de fonctions de *Daubechies* 4, 7 et 10 (notées respectivement DB4, DB7, DB10) pour les niveaux 1 à 5 en approximation, et pour 5 neurones cachés. La dimension de la couche d'entrée en entrée prend ainsi des valeurs allant de 64 à 4 unités. L'utilisation des

[185] D. J. Livingston, D. T. Manallack, I. V. Tetko, *Journal of Computer-aided and Molecular Design*, 11 (1997) p. 135

coefficients caractéristiques des détails a certes été envisagée, mais elle n'a pas montré de possibilités.

Le Tableau 11 rapporte ainsi uniquement les critères RMSEP obtenus en test pour les coefficients d'approximation. Nous avons constaté avec intérêt que la convergence des réseaux construits sur la base des données fournies par ces coefficients nécessite un nombre d'itération important, de l'ordre de 1000 à 10000, quel que soit le niveau de compression incriminé.

Niveau d'approximation– *dimension des données d'entrée*	**DB4** RMSEP (%)	**DB7** RMSEP (%)	**DB10** RMSEP (%)
1 – *64*	0,05	0,05	0,06
2 – *32*	**0,03**	**0,03**	0,04
3 – *16*	0,05	**0,04**	0,04
4 – *8*	0,09	0,06	0,07
5 – *4*	0,15	0,10	0,09

Tableau 11 : Erreurs de prédiction pour différents niveaux de compression (5 neurones intermédiaires, moyennés pour les répartitions R_I, R_{II} et R_{III}, dans la gamme [0,1 ; 0,9]).

Le minimum d'erreur correspond à un critère égal à 0,03, pour le niveau 2 dans le cas des familles d'ondelettes DB4 et DB7. Cela correspond à un nombre de neurones d'entrée de 32 qui, pour un nombre de neurones cachés ultérieurement optimisé à 4, engendre une architecture comportant 132 paramètres ajustables. Ce nombre est élevé si l'on considère que le modèle final devra montrer d'importantes capacités de généralisation, pour un nombre de données d'entrée équivalent. Il semble donc préférable, quitte à engendrer une faible dégradation de la prédiction, de choisir le modèle construit à partir du niveau 3, pour la DB7. Une architecture 16×4×1, plus satisfaisante, qui produit un critère RMSEP de 0,04 dans la gamme [0,1 ; 0,9].

Discussion

Les ordres de grandeur des erreurs obtenues par le biais des deux méthodes de réduction des données sont équivalents, autour de 0,03 dans la gamme [0,1 ; 0,9]. Des architectures

différentes, pour lesquelles par exemple la représentation des données n'est pas la même, peuvent converger vers des solutions similaires, pour l'estimateur choisi. Cela s'effectue même si la répartition de l'information, c'est à dire des valeurs des poids, est différente et cela caractérise la flexibilité intrinsèque des méthodes neuronales. D'autre part, pour un nombre fixé de neurones d'entrée, la performance reste fonction du nombre d'unités intermédiaires. Pour un lot d'apprentissage donné, le nombre idéal de neurones cachés du point de vue de la prédiction ne varie apparemment pas significativement en fonction de la représentation choisie des données d'entrée. Le nombre de neurones cachés dépend plutôt de la complexité de la tâche d'apprentissage.[186]

Enfin, l'analyse de la Figure 76 représentant les données prédites en fonction des valeurs cibles semble montrer que l'on dispose, sur le plan de la qualité des prédictions, de deux modèles équivalents. Cette représentation illustre également les difficultés des modèles à prédire avec précision le début de la gamme. Effectivement, les points correspondant à des DH inférieurs à 0,3 s'écartent plus franchement de la première bissectrice. Il s'agit de l'intervalle où la dynamique est la plus importante et qui, de plus, présente un fort aspect non-linéaire du point de vue de la relation entre les données d'entrée et les données de sortie.

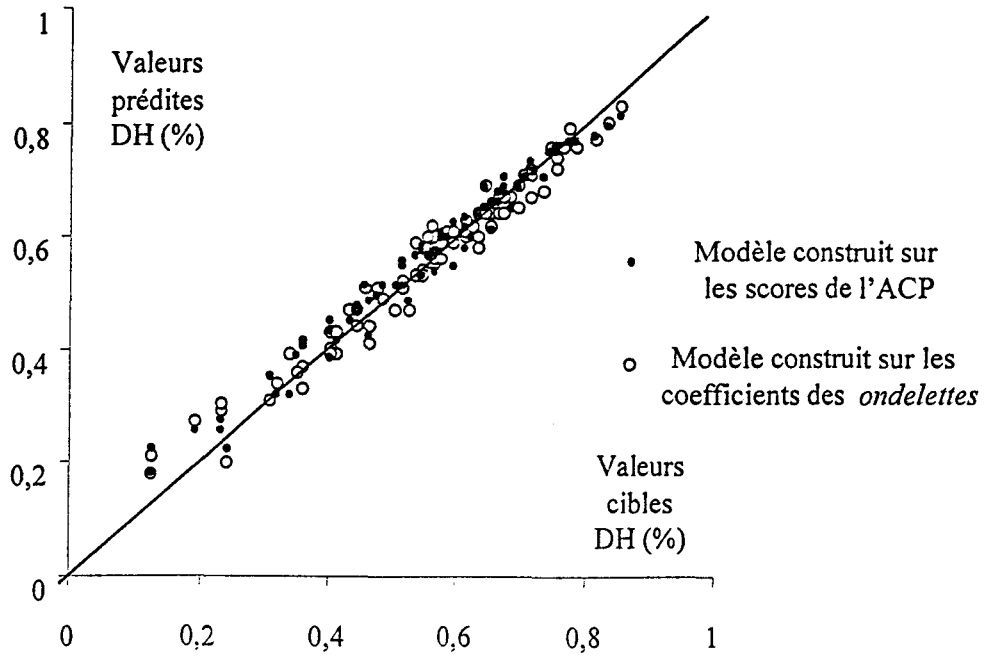


Figure 76 : Prédiction des modèles *(exprimées dans la gamme [0,1 ; 0,9]).*

[186] O. Fujita, *Neural Networks*, 11 (1998) p. 851

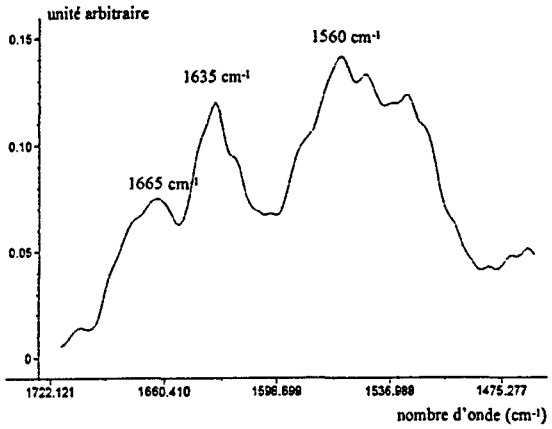
Le choix d'une méthode de compression n'est donc pas seulement basé sur la valeur RMSEP. Ici, le modèle retenu est celui construit sur la base des facteurs *score* issus de l'analyse en composantes principales de la matrice spectrale. En effet, il est décrit par un nombre de paramètres plus acceptable, ce qui est un pré-requis essentiel pour la poursuite du travail. De plus, nous prendrons soin de montrer dans le paragraphe suivant qu'il est bien adapté à une identification, à partir des spectres enregistrés dans l'infrarouge moyen, des phénomènes physico-chimiques.

5.3.3 Interprétation et analyse

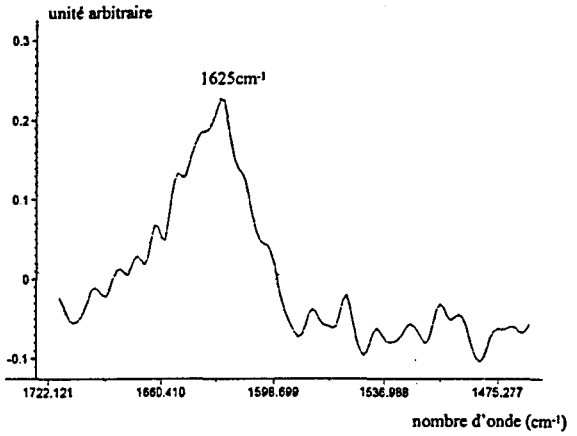
L'utilisation des variables latentes de l'analyse en composantes principales comme données d'entrée des RNA-FF permet de développer un modèle d'étalonnage global pour la prédiction du degré d'hydrolyse. Restent à caractériser les informations exploitées et à en proposer une éventuelle attribution pour permettre une meilleure compréhension des mécanismes de la réaction.

Analyse qualitative des vecteurs *loading*

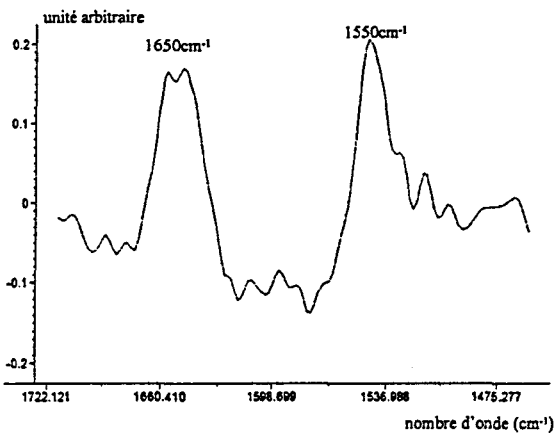
La Figure 77 de la page suivante propose une représentation des quatre vecteurs *loading* formant la base vectorielle sur laquelle les données d'entrée ont été projetées.
Le vecteur *loading* illustrant le premier axe principal est une composante globale. Il s'agit d'une contribution positive pour chaque nombre d'onde. La plupart des informations spectrales d'intérêt que nous avons détaillées au chapitre précédent y sont regroupées. En effet, ce vecteur décompose le profil de la bande amide I détaillé Tableau 3. Les contributions à 1665 cm^{-1} et 1635 cm^{-1} sont probablement imputables à des modifications de la structure tridimensionnelle des molécules d'hémoglobine, puisqu'ils concernent des nombres d'ondes caractéristiques des proportions de coudes et de structures désordonnées.[155] D'autre part, ces signaux sont également corrélés à l'information concernant le profil de la bande amide II, caractérisée à 1560 cm^{-1}. Enfin, le mécanisme d'hydrolyse de l'hémoglobine en question, connu sous le nom de *one by one*,[163,164] implique la présence simultanée du signal de l'hémoglobine et de l'information correspondant à l'ensemble des peptides tous présents tout au long de la réaction ; cela peut être le sens de cette composante globale.



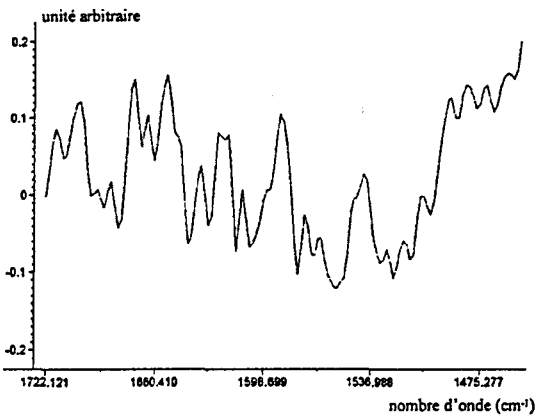
A/ *Loading* de l'axe 1.



B/ *Loading* de l'axe 2.



C/ *Loading* de l'axe 3.



D/ *Loading* de l'axe 4.

Figure 77 : Coefficients *loading* correspondant aux axes principaux pris en compte par le modèle.

L'information spectrale contenue au sein du vecteur *loading* du troisième axe possède deux contributions, à 1650 cm^{-1} à 1550 cm^{-1}, nombres d'onde caractéristiques des bandes amide I et amide II. Il a été rapporté que des variations d'intensité relatives de ces bandes peuvent être associées à des modifications de structure secondaire ou tertiaire.[173]

Les deux axes précédemment étudiés sont en accord avec l'explication avancée à la fin du chapitre 4 concernant l'observation de l'avancement du procédé par spectrométrie infrarouge. Le mécanisme fournissant une information spectrale est probablement la dénaturation de

protéines, qu'elle soit préalable à la fragmentation peptidique dans un milieu "tampon" ou engendrée par les cassures des liaisons peptidiques.

Concernant le vecteur associé au deuxième axe, il ne tient compte que de l'information à 1625 cm^{-1}. Nous avons déjà discuté du fait que celle-ci soit due à la prise en compte de spectres de solutions aqueuses[187] comme référence. En outre, les contributions spectrales du solvant évoluent certainement en cours d'hydrolyse.

Enfin, le signal qui concerne le quatrième axe correspond à des contributions de l'information réparties sur tout le domaine spectral et qui ne sont a priori pas directement interprétables.

Importance relative des informations

Bien que les axes principaux soient ordonnés (du premier vers le quatrième axe) de part la proportion de la variance globale qu'ils prennent en compte, cette information est transparente en ce qui concerne le modèle neuronal. En effet, l'importance de chaque axe n'est déterminée que par la qualité de sa contribution à la modélisation.

Les informations amenées par les facteurs *score* des composantes principales 1 et 3 semblent posséder une signification chimique dans le sens où elles concernent des variations de structure. Néanmoins, le premier axe principal ne permet pas, à lui seul, de prédire le paramètre descriptif de la réaction avec une précision satisfaisante. L'erreur RMSEP est de l'ordre de 0,11 (Figure 74). Elle correspond à un niveau d'erreur de 1,3 % dans la gamme [0 ; 8,7 %] ce qui est considérable.

Construire un modèle prenant en compte les scores sur le quatrième axe principal permet d'améliorer légèrement le critère RMSE de prédiction, de 0,04 à 0,03 dans la gamme [0,1 ; 0,9]. L'information spectrale exprimée par ce vecteur *loading* ne possède pourtant pas de signification physico-chimique particulière ; elle peut résulter de phénomènes de compensations ou de contributions résiduelles des signaux attribués aux axes précédemment construits. Il pourrait donc être légitime de préconiser de limiter la couche d'entrée du réseau au trois premiers neurones.

Enfin, la deuxième source de variations dans la matrice des données spectrales est probablement imputable aux molécules d'eau ou à des modifications de la structure du solvant

[187] K. P. Ishida, P. R. Griffiths, *Applied Spectroscopy*, 47 (1993) p. 584

en cours de d'hydrolyse. La prise en compte de cette information peut également être critiquée, elle n'apporte que peu d'amélioration du point de vue de la prédiction (0,09 au lieu de 0,11) par rapport au cas où seule les composantes sur le premier axe sont considérées. Néanmoins, c'est sur cette composante que la présence de classes au sein du lot de données est la plus marquée et cette constatation est un argument pour l'intégration de cette information au modèle.

5.4 Modèles robustes

Par définition, les modèles robustes ne sont pas sensibles à certaines déviations par rapport à la distribution initiale.[188] Leurs propriétés sont nécessaires pour permettre la prédiction de sorties associées à des données d'entrée situées aux limites du domaine d'étalonnage.
Suivre le procédé en temps réel impose de prédire le degré d'hydrolyse à partir d'observations sur des répétitions nouvelles, c'est à dire impossibles à intégrer aux lots d'entraînement et de validation. A ce titre, nous envisageons l'analyse des répartitions R_V et R_{VI} (Tableau 8) qui dédient respectivement l'ensemble des répétitions r_5 et r_6 au lot de test, et uniquement à celui-ci. Il faut remarquer que l'erreur de prédiction calculée alors est un estimateur biaisé de l'erreur de généralisation, puisque les lots d'échantillons ne sont pas construits au hasard.[189] La méthode la plus efficace consisterait à inclure dans les lots d'apprentissage, artificiellement ou non, toutes les sources possibles de variation. Cependant, lorsqu'il n'est pas envisageable de les identifier toutes, nous sommes contraints de supposer que les perturbations éventuelles du procédé et des données spectrales sont prises en compte au sein des répétitions utilisées pour l'apprentissage. Par prudence, il est donc conseillé de répéter l'opération pour d'autres répartitions d'échantillons. Le choix des répartitions R_V et R_{VI} est intéressant du point de vue de la représentativité de leurs échantillons par rapport aux capacités de généralisation des modèles.
Notons enfin que l'optimisation de la couche d'entrée, constituée des scores des échantillons sur les variables latentes produites par une nouvelle analyse en composantes principales, doit être réitérée en tenant compte des précautions déjà mentionnées au paragraphe 3.3 de ce chapitre.

5.4.1 Construction des modèles

Prédire une répétition nouvelle, analytiquement inconnue, est d'autant plus délicat qu'il existe une classification forte des données en fonction de leur répétition d'origine. L'observation de la structure des données dans l'espace de départ, engendré par les axes principaux (Figure 68 et Figure 69), révèle que la difficulté est sensiblement différente pour les deux répartitions

[188] I. E. Frank, R. Todeschini, *The Data Analysis Handbook*, Elsevier (1994)
[189] S. M. Weiss, C. A. Kulikowski, *Computer Systems that learns*, M. Kaufman (1991)

considérées. En effet, la prédiction de la répétition r_5 nécessite une certaine forme d'extrapolation, en particulier sur la deuxième composante principale. Par contre, la répétition r_6 occupe une position plus centrale au sein du nuage de données ce qui devrait faciliter la prédiction.

Optimisation de la couche d'entrée

Le Tableau 12 regroupe, pour les deux répartitions d'échantillons dont il est question, certaines des valeurs obtenues pour le critère RMSEP (%) lors de l'optimisation du nombre de neurones de la couche d'entrée. Celui-ci a été décrémenté de 10 à 2, par pas unitaires. Le nombre d'itérations et le paramétrage de l'apprentissage, équivalents à ceux précisés au paragraphe précédent, ne sont pas explicitement rappelés.

Répartition	**10×5×1** RMSEP (%)	**5×5×1** RMSEP (%)	**4×5×1** RMSEP (%)	**3×5×1** RMSEP (%)
R_V	0,08	0,04	**0,04**	0,05
R_{VI}	0,07	0,03	**0,03**	0,04

Tableau 12 : Optimisation de la couche d'entrée *(RMSEP dans la gamme [0,1 ; 0,9]).*

L'architecture la plus adéquate est, dans le cas de R_V comme dans le cas de R_{VI}, basée sur la considération de quatre unités d'entrée. Cela correspond à la dimensionalité de l'espace d'entrée, après optimisation, dans le cas du modèle global. L'architecture 4×5×1 donne alors lieu à des erreurs de 0,03 et 0,04 dans la gamme [0,1 ; 0,9], respectivement pour R_{VI} et R_V. Les résultats obtenus pour R_V sont, numériquement, légèrement moins bons, bien que d'un niveau tout à fait acceptable. Cela s'explique probablement par la difficulté du travail de généralisation dans le cas de la répartition R_V, puisque l'extrapolation des données dans l'espace d'entrée est requise.

Optimisation de la couche intermédiaire

Le Tableau 13 propose les résultats de l'optimisation du nombre de neurones de la couche cachée

Répétition	4×5×1 RMSEP (%)	4×4×1 RMSEP (%)	4×3×1 RMSEP (%)	4×2×1 RMSEP (%)	4×1×1 RMSEP (%)
R_V	0,04	**0,04**	0,04	0,06	0,09
R_{VI}	0,03	**0,03**	0,04	0,05	0,11

Tableau 13 : Optimisation de la couche intermédiaire *(RMSEP dans la gamme [0,1 ; 0,9]).*

Le meilleur critère RMSEP (%) obtenu, exprimé dans la gamme [0,1 ; 0,9], correspond à la valeur 0,03 pour la répartition R_{VI} et à la valeur 0,04 pour la répartition R_V.
La topologie optimisée se résume donc à quatre neurones d'entrée, quatre neurones cachés et une unité de sortie, soit 4×4×1. Convertis dans la gamme d'origine [0 ; 8,7 %], les niveaux d'erreur sont alors de l'ordre de 0,4 % pour la répartition R_V (0,37 % pour être exact) et 0,3 % pour la répartition R_{VI} (0,31 %). Ces valeurs représentent des moyennes quadratiques légèrement supérieures, mais comparables, au résultat obtenu concernant le modèle développé au paragraphe précédent. Il est également instructif d'observer qualitativement les prédictions, en fonction des valeurs attendues, pour juger de l'efficacité de l'étalonnage.

Résultats

L'observation de la Figure 78 et de la Figure 79, obtenues toutes deux à partir des tableaux de résultats proposés en annexe 4, permet d'analyser les qualités prédictives des deux modèles, tout au long de la cinétique de la réaction, et de les comparer.

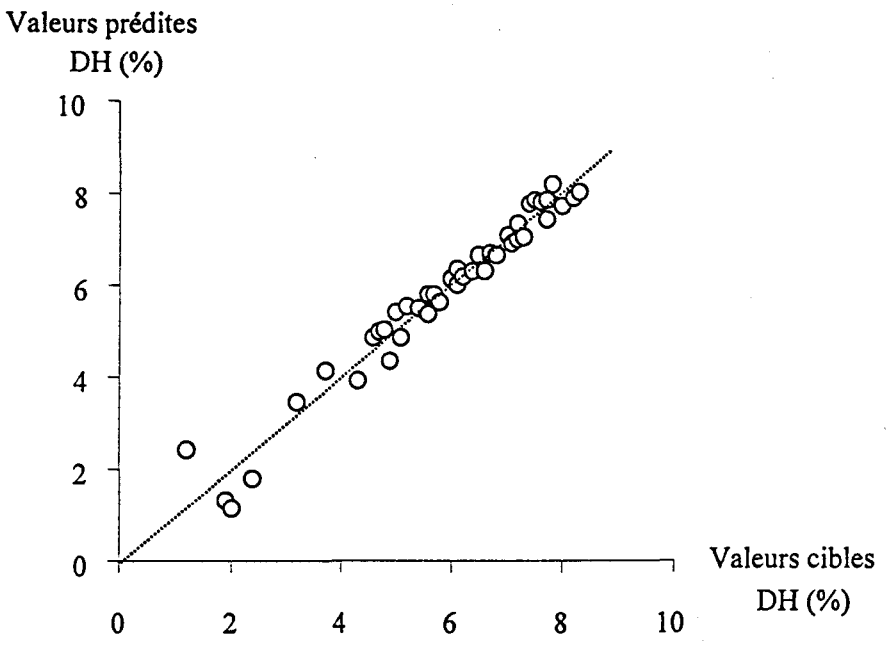


Figure 78 : Prédiction de la répétition r_5 *(dans la gamme [0 ; 8,7 %], RMSEP=0,37 %).*

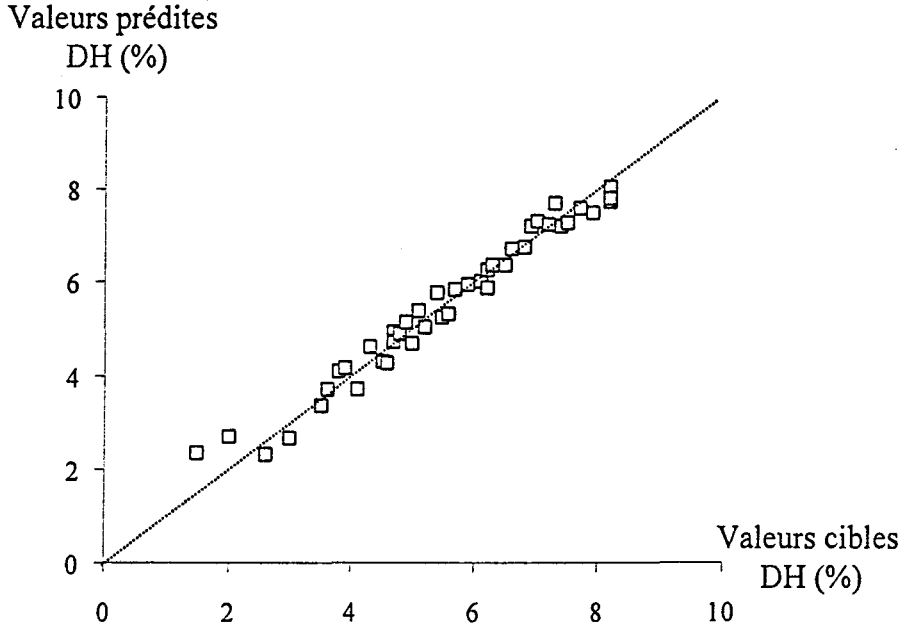


Figure 79 : Prédiction de la répétition r_6 *(dans la gamme [0 ; 8,7 %], RMSEP=0,31 %).*

Tout comme lors de l'analyse des prédictions du modèle global, la première constatation concerne la plus grande difficulté du modèle à prédire le début de la manipulation. Au contraire, lorsque les DH sont plus importants, les valeurs se rapprochent de la première bissectrice, oscillant de part et d'autre de celle-ci au cours de l'avancement de la manipulation.

Enfin, les valeurs de DH correspondant à la fin de la réaction sont les plus correctement prédites, avec une précision équivalente à celle observée dans le cas des meilleurs modèles globaux. D'ailleurs, de manière générale la perte de précision associée à la prédiction d'une répétition d'hydrolyse complètement inconnue n'est jamais dramatique, ce qui est un atout incontestable des RNA-FF et une preuve de leurs capacités de généralisation.

5.4.2 Interprétation des modèles

La capacité des réseaux neuronaux à représenter des relations complexes entre des données d'entrée et des données de sortie a été démontrée et des capacités prédictives au moins équivalentes à celles obtenues par d'autres méthodes sont attendues, lorsque les applications sont adéquates.[85] Nous en avons tiré avantage pour la prédiction du DH en vue du suivi d'hydrolyse d'hémoglobine bovine. Les RNA-FF donnent des résultats très satisfaisants même lors de la généralisation à des manipulations inconnues, nécessitant dans le cas de la répartition R_V des facultés d'extrapolation sur les données d'entrée. Néanmoins, le principal frein à l'utilisation des méthodes neuronales en chimie analytique des procédés provient de leur réputation d'herméticité ou encore de "boîte noire". Nous appliquons ici certaines méthodes d'observation à l'interprétation des modèles neuronaux. Celle-ci n'est pas tout à fait immédiate mais elle s'avère possible et utile. Il existe des possibilités de déduction ou d'interprétation, même s'il est vrai que la forme mathématique exacte du modèle reste inconnue.[75,80]

La Figure 80 présente un étiquetage des neurones du réseau auquel nous nous référons par la suite.

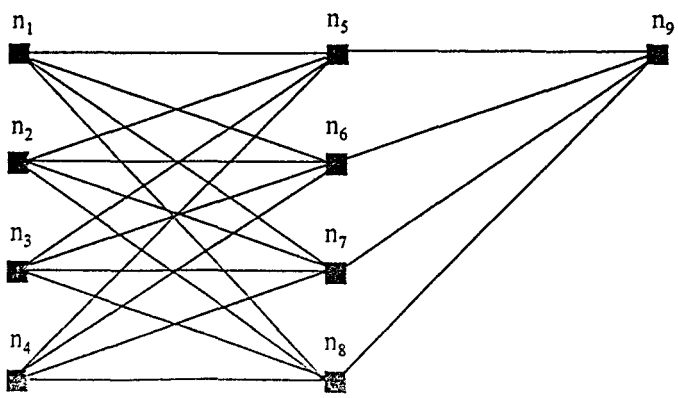


Figure 80 : Etiquetage des neurones du réseau.

Sélection des neurones d'entrée

L'optimisation de la topologie est une étape déterminante, longue et délicate, définie et contrôlée par l'expérimentateur. Le nombres de neurones, donc de poids, doit être aussi faible que possible. C'est pour cela que les facteurs *score* issus de l'analyse en composantes principales ont été utilisés en dépit de l'aspect mathématique linéaire de cette transformation. Cependant l'étape de réduction peut être poussée plus loin. Des méthodes de sélection efficaces concernent, entre autres, l'évaluation des contributions des variables d'entrée à la variance des réponses prédites[183] ou le calcul des effets de la suppression de chacun des poids sur l'erreur globale.[190]

Etant donné le faible nombre de paramètres associés à l'architecture qui concerne le problème courant, nous envisageons simplement la suppression des unités d'entrée, une à une, puis l'observation de l'erreur. Le Tableau 14 regroupe donc les niveaux d'erreur en prédiction correspondant. Ce tableau permet de discuter de l'influence relative des entrées du réseau sur la valeur RMSEP (%). Le fait que l'observation des poids entre la couche d'entrée et la couche de sortie ne soit pas directement représentative de leurs importances respectives est ainsi contourné. En effet, à cause des opérations de produits et de sommes qui interviennent lors du traitement, les neurones les plus importants ne sont pas forcément associés aux poids les plus grands.

Unités d'entrée	$n_1\ n_2\ n_3\ n_4$	$.\ n_2\ n_3\ n_4$	$n_1\ .\ n_3\ n_4$	$n_1\ n_2\ .\ n_4$	$n_1\ n_2\ n_3\ .$
Répartition	RMSEP (%)	RMSEP (%)	RMSEP (%)	RMSEP (%)	RMSEP (%)
R_V	**0,04**	0,08	0,07	0,10	0,04
R_{VI}	**0,03**	0,07	0,09	0,12	0,04

Tableau 14 : Erreurs en prédiction pour différentes architectures *(4 neurones intermédiaires, dans la gamme [0,1 ; 0,9]).*

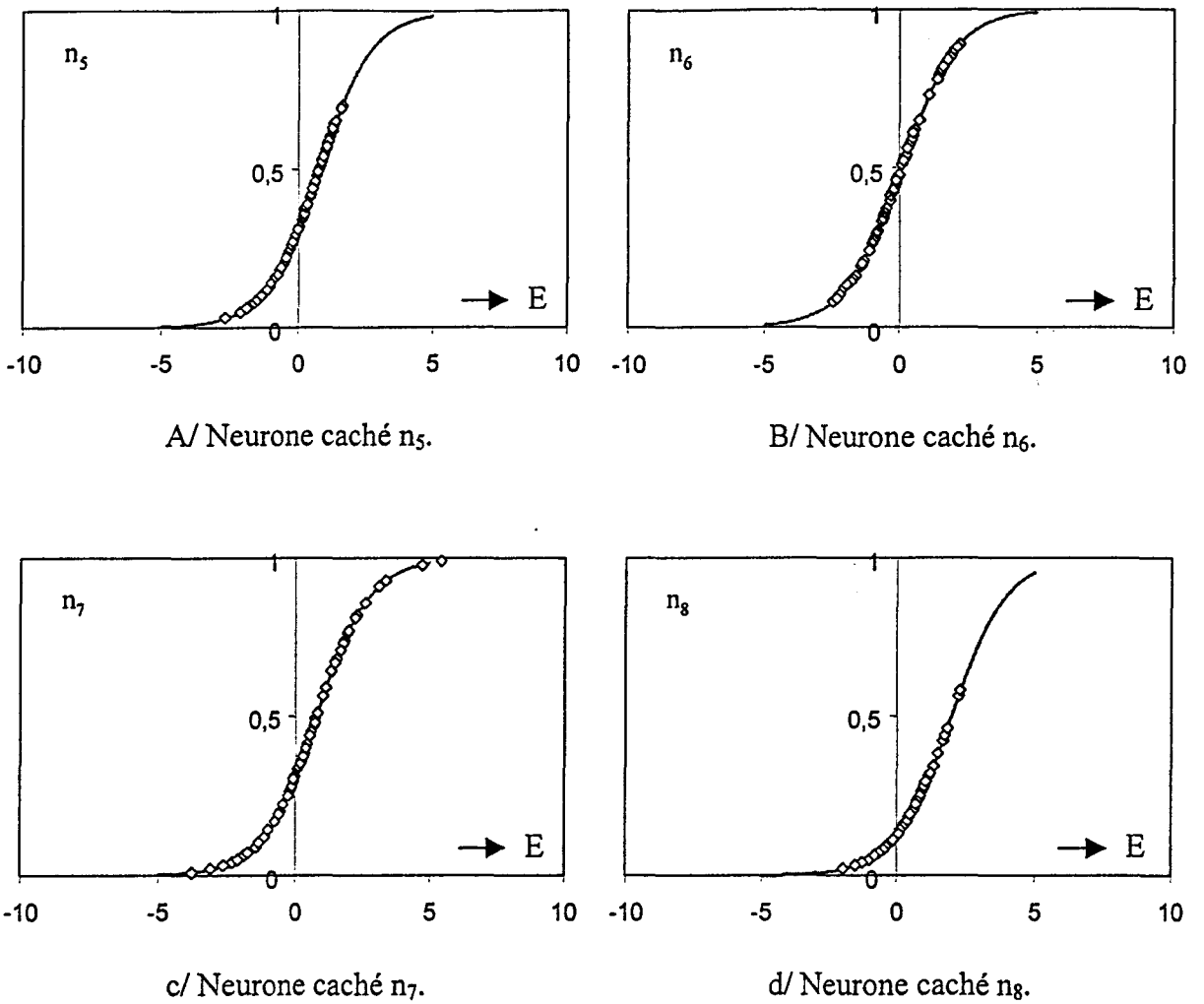
Les résultats optimaux, représentés en gras, correspondent aux architectures construites sur les quatre premières unités. On constate, par ailleurs, que la dégradation la plus prononcée par

[190] R. Fletcher, *Practical Methods of Optimization*, Wiley (1987)

rapport à ces valeurs est associée à la suppression des liens provenant du troisième neurone d'entrée (0,10 % et 0,12 % pour n_1 n_2 . n_4). Cette unité reçoit les contributions de la troisième composante principale qui présente une allure similaire à celle présentée Figure 77C/ page 163. Elle corrèle les intensités des bandes amide I et amide II. A l'opposé, les contributions les moins flagrantes sont celles associées à la quatrième composante principale, difficilement caractérisée sur le plan physico-chimique. Celle-ci est néanmoins conservée. En effet, pour le modèle global présenté précédemment, l'amélioration produite par cette information était significative du point de vue de l'erreur de prédiction. Enfin, on constate que les composantes associées aux deux premiers neurones sont indispensables à l'établissement d'un modèle correct. Leur omission, décrite par les résultats des colonnes notées . n_2 n_3 n_4 et n_1 . n_3 n_4, double le niveau d'erreur en prédiction.
Ainsi, cette méthode permet une interprétation "semi-quantitative" des vecteurs *loading* qui peuvent être ordonnés en fonction de leur importance du point de vue de la précision du modèle. On contourne ainsi une limitation mentionnée en conclusion de l'étude de faisabilité.

Rôle des neurones intermédiaires

Les neurones de la couche intermédiaire, auxquels sont attribuées des fonctions de transfert sigmoïde, permettent la modélisation de relations non-linéaires, mais également linéaires. Cet aspect a été détaillé au chapitre 3. Ainsi, l'étude de l'activation des neurones cachés peut, éventuellement, donner des indications concernant le degré de non-linéarité du lot de données. La Figure 81 représente, pour les quatre neurones cachés, l'ensemble des positions d'activation de la fonction sigmoïde en fonction des valeurs d'entrée E, pour les données du lot d'entraînement de la répartition R_V, à l'issue de l'apprentissage.



A/ Neurone caché n_5.

B/ Neurone caché n_6.

c/ Neurone caché n_7.

d/ Neurone caché n_8.

Figure 81 : Activations des neurones intermédiaires (unités arbitraires) *(architecture4*4*1, lot d'entraînement, répartition R_1).*

Constatons en premier lieu que les unités intermédiaires n_5 et n_8 révèlent un comportement équivalent. Leurs fonctions de transfert sont activées, à la fois, dans leur zone fortement non linéaire pour des données correspondant au début de la manipulation et dans leur domaine linéaire pour des données associées au déroulement du procédé. La même information semble donc partagée par deux unités, ce qui pourrait expliquer l'équivalence des modèles à trois et quatre neurones intermédiaires. D'autre part, l'information transmise par le neurone n_6 peut, quant à elle, être considérée comme linéaire puisque seule la zone centrale de la fonction est activée. Enfin, les données se répartissent sur le neurone n_7 en utilisant toute la dynamique de la fonction, mettant en avant une complémentarité des deux types d'informations.

A titre indicatif, tous les poids sont rassemblés dans le Tableau 15. Certes, l'observation des poids entre les neurones de la couche d'entrée et ceux de la couche intermédiaire n'est pas forcément significative de l'importance des variables d'entrée. Par contre, à condition que le réseau ne soit constitué que d'une seule unité de sortie, l'observation des poids issus de la couche intermédiaire apporte une information sur l'importance relative des neurones cachés.

neurones	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8
n_5	**3,76**	-2,01	1,03	**-13,95**				
n_6	-0,06	-0,19	**18,07**	-0,7				
n_7	**4,35**	**26,2**	0,68	0,24				
n_8	**-4,47**	**-3,70**	-0,68	**-10,56**				
n_9					**1,85**	**-0,84**	**0,47**	**-3,36**

Tableau 15 : Valeurs numériques des poids *(répartition R_V, les poids en gras correspondent aux liens représentés Figure 82).*

Pour le réseau entraîné sur la répartition R_V, la Figure 82 associe à chaque neurone intermédiaire la valeur du poids associé au lien vers l'unité de sortie.

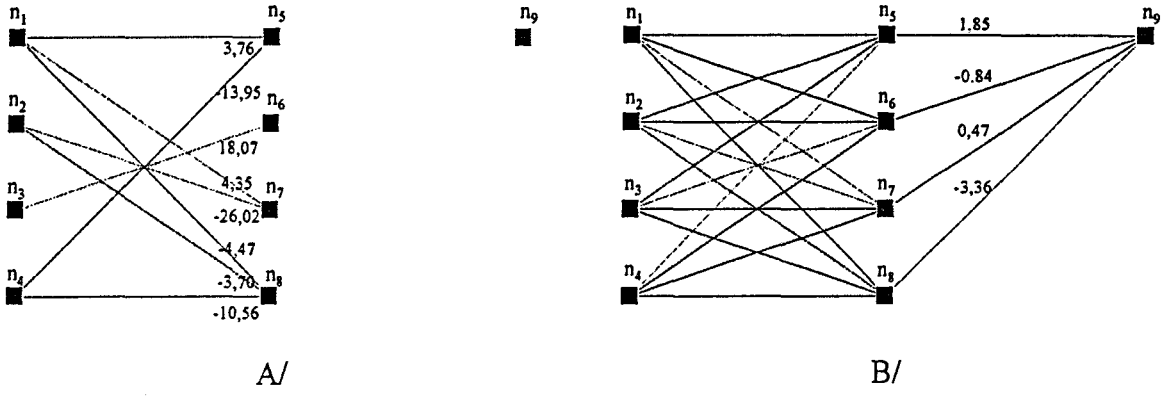
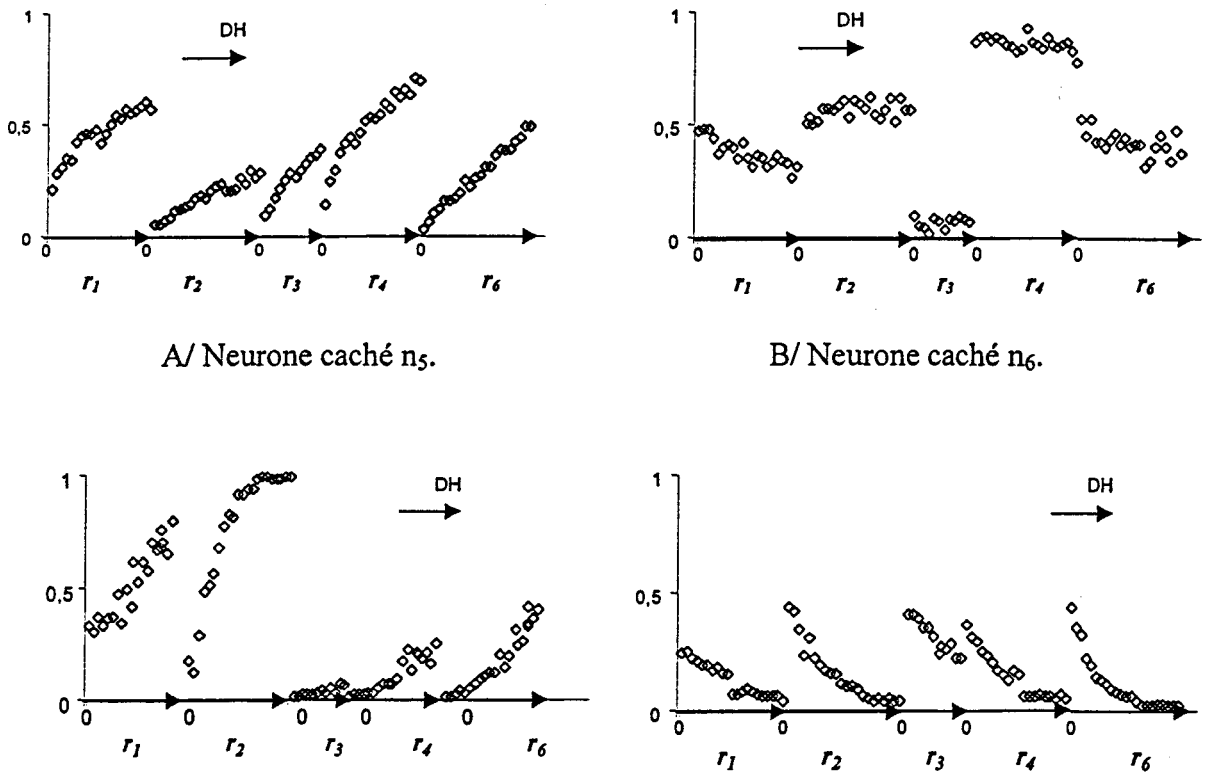


Figure 82 : Valeurs des poids associés aux liens d'entrée (A/) et de sortie (B/) *(répartition R_V).*

L'information prépondérante, correspondant à la valeur numérique -3,36, est associée au neurone caché n_8 qui sollicite plutôt la partie basse de la fonction de transfert sigmoïde et en particulier la zone non-linéaire. Parallèlement, l'observation des activations du neurone intermédiaire n_5 traduit une information équivalente, mais pondérée par une valeur positive,

1,85, attribuée probablement une certaine forme de compensation. Aux neurones n_6, dont les contributions sont principalement linéaires, et n_7, dont les contributions sont à la fois linéaires et non-linéaires linéaires, correspondent des valeurs de poids plus faibles ; respectivement -0,84 et 0,47.



A/ Neurone caché n_5. B/ Neurone caché n_6.

C/ Neurone caché n_7. D/ Neurone caché n_8.

Figure 83 : Observation des activations des neurones intermédiaires en fonction du DH (unités arbitraires) *(architecture4*4*1, lot d'entraînement, répartition R_1).*

Pour terminer, la Figure 83 peut apporter une confirmation de cette interprétation. Les valeurs d'activation des neurones cachés y sont représentées en fonction du degré d'hydrolyse pour chaque répétition du lot d'entraînement. Les graphes A/ et D/ (pour n_5 et n_8) montrent ainsi des tendances équivalentes, au signe près. Le graphe C/ est plus surprenant et plus difficilement interprétable. En effet, le neurone caché n_7 accorde une plus grande dynamique, aux répétitions r_1 et r_2. Le graphe B/, associé au neurone n_6 au caractère linéaire, peut être interprété comme une "mise à niveau" des données d'entrée de chaque répartition. Pour chaque répétition, les valeurs des activations se maintiennent à un niveau constant. C'est la

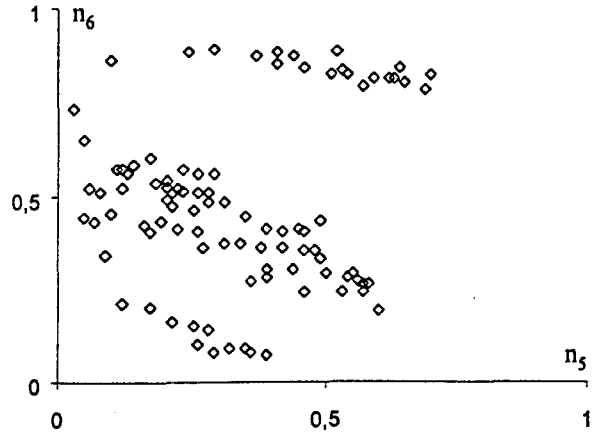
phase d'apprentissage du modèle qui est à l'origine cette "normalisation" entre les répétitions. Celle-ci est probablement nécessaire pour permettre des prédictions à partir de données présentant des classes.

Etude des sorties produites par les neurones intermédiaires

Pour être complet, il est également instructif d'observer les activations des neurones cachés représentées les unes en fonction des autres.[75] Cela permet la visualisation des données après l'étalonnage pour, par exemple, la détection a posteriori de points aberrants ou de classes de données. En ce sens, l'information produite est en quelque sorte redondante avec les représentations des scores de la PCA.

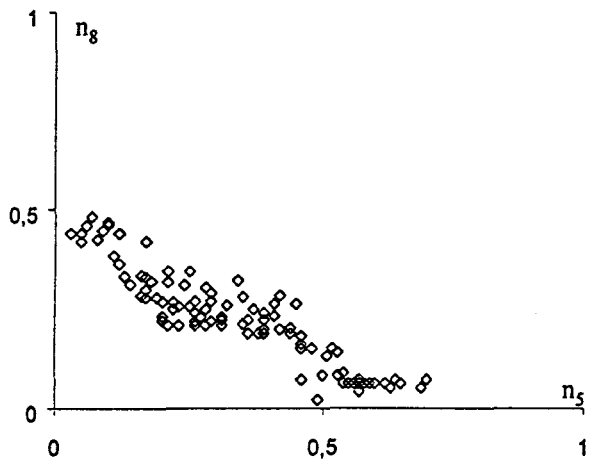
Les différents graphes de la Figure 84, excepté le C/, montrent tous des classes d'échantillons. Néanmoins, les groupes de données formant les classes sont moins identifiés que dans le cas de l'observation des représentations des scores.

Une autre interprétation possible de ces graphes peut être envisagée en termes de corrélation des données spectrales. Ainsi, les lots de données distribués horizontalement ou verticalement ne traduisent que peu de corrélation entre les sorties associées aux deux neurones cachés étudiés. Par exemple, sur les graphes A/ et D/, les sorties de l'unité n_6 ne montrent, au sein d'un même groupe, pratiquement pas de corrélation avec celles de n_5 et n_7. Par ailleurs, de fortes non-linéarités peuvent être facilement détectées sur certains graphes, le graphe B/ par exemple. Enfin, le graphe C/ est particulier puisque les données y sont distribuées de manière homogène. Il concerne les neurones intermédiaires n_5 et n_8 dont les comportements similaires ont déjà fait l'objet de plusieurs remarques.

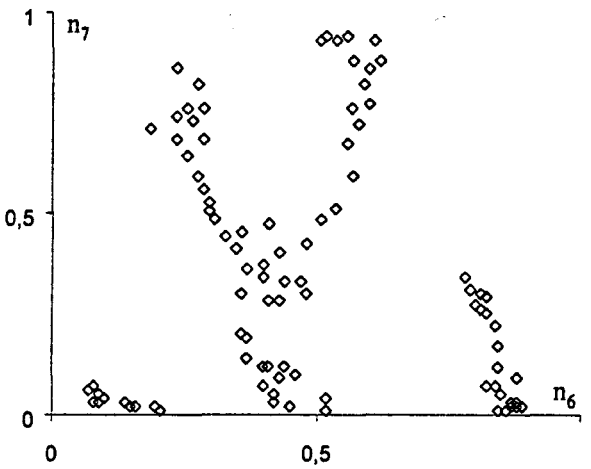


A/ Graphe (n_5 ; n_6).

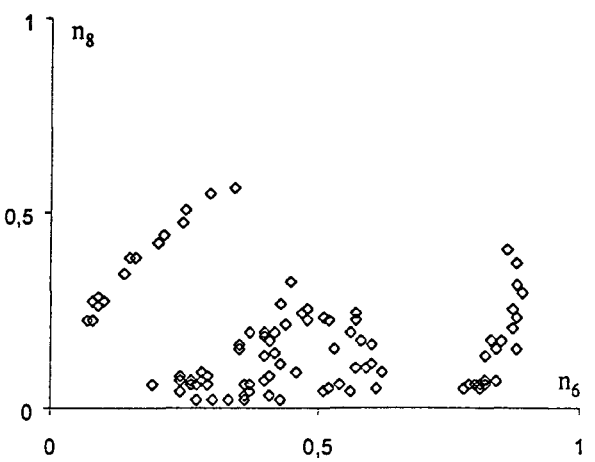
B/ Graphe (n_5 ; n_7).



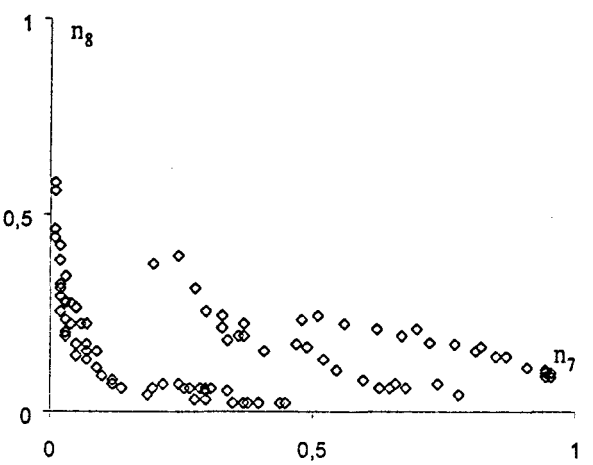
C/ Graphe (n_5 ; n_8).



D/ Graphe (n_6 ; n_7).



E/ Graphe (n_6 ; n_8).



E/ Graphe (n_7 ; n_8).

Figure 84 : Répartitions des valeurs d'activations sur les neurones intermédiaires

(architecture4×4×1, lot d'entraînement, répartition R_V).

5.5 Bilan de l'étude

C'est l'application physico-chimique qui doit être la motivation de la chimiométrie. Nous avons mesuré le degré d'hydrolyse enzymatique de l'hémoglobine bovine en vue de contrôler la production peptidique en réacteur de laboratoire. Certains peptides possèdent des propriétés organoleptiques et fonctionnelles intéressantes mais il n'existe cependant pas de méthode transférable sur le réacteur pour le contrôle de la cinétique. Les résultats résumés dans ce paragraphe montrent que le recours à la modélisation empirique est une alternative intéressante en vue d'une meilleure connaissance, plus rapide et plus directe, permettant éventuellement une action ou une prise de décision.

Après avoir présenté les méthodes instrumentales mises en place pour l'observation du système, nous avons d'abord discuté de la validité des enregistrements. Ainsi, des coefficients RSD variant de 1 % à 3% pour des manipulations d'essai sur des solutions inertes ont été obtenus autour des fréquences de vibration caractéristiques. En comparaison, les mêmes critères, calculés pour une hydrolyse d'hémoglobine dans des conditions environnementales et instrumentales les plus semblables possibles, sont de l'ordre de 4 % à 10 %. Il existe donc bien une information spectrale propre au procédé.

Par la suite, nous avons caractérisé les données pour justifier l'utilisation de techniques neuronales. Ce choix a été guidé certes par l'existence de non-linéarités plus ou moins franches, mais surtout suite à la détection de classes au sein des données d'entrée et de sortie. En effet, l'utilisation des cartes de Kohonen et des représentations des scores issus de la PCA ont permis de déceler des comportements propres à chaque répétition du procédé. Cette difficulté justifie à elle seule la recherche d'un modèle expérimental à l'aide des méthodes neuronales et en particulier des réseaux multicouches.

Les méthodes neuronales pour l'étalonnage multivarié exigent une attention et une expérience particulière de la part de l'utilisateur. Entre autres, l'architecture doit être optimisée et les modèles doivent être construits sur une base respectant les conditions élémentaires assurant la généralisation. La PCA est alors, malgré le caractère linéaire de la transformation mathématique, une méthode adéquate de réduction des données pour améliorer la description numérique du problème, en particulier vis à vis de l'apprentissage.

Ainsi, des répartitions différentes des données ont été envisagées. D'une part, celles utilisées pour l'élaboration d'un modèle global. Elles couvrent l'ensemble des variations possibles et contiennent, en entraînement, en validation et en test, toutes les répétitions du procédé

disponibles. Le niveau atteint pour l'erreur de prédiction est alors estimé par un RMSEP inférieur à 0,3 % dans la gamme de DH [0 ; 8,7 %] pour une architecture 4×3×1. D'autre part, il est possible de construire des répartitions des données dédiées à l'évaluation des capacités de généralisation ; une répétition est exclue de la procédure d'apprentissage et est conservée à cet effet. Les résultats obtenus sont alors, dans le cas le plus difficile c'est à dire pour une répartition nécessitant des facultés d'extrapolation sur les données d'entrée, autour de 0,4 % dans la même gamme. De telles erreurs sont tout à fait tolérables compte tenu des niveaux observés pour les mesures de référence du degré d'hydrolyse et de la précision recherchée pour le contrôle.

A terme, on cherchera probablement à se placer dans une optique où l'on envisagera un outil permettant la stabilisation du procédé dans une zone particulière de la dynamique. Les peptides d'intérêts y seraient produits, disons à titre d'exemple, pour des valeurs de DH dans l'intervalle entre 6 % et 7 %. Le Tableau 16 extrait les résultats de l'annexe 4 et une erreur RMSEP inférieure à 0,2 % (0,16 % exactement) peut être calculée.

Ordre de l'échantillon	**DH cible** (en % dans la gamme [0 ; 8,7])	**DH prédits** (en % dans la gamme [0 ; 8,7])
23	6	6,11
24	6,1	6,02
25	6,1	6,32
26	6,2	6,15
27	6,4	6,28
28	6,5	6,63
29	6,6	6,29
30	6,7	6,67
31	6,8	6,63
32	7	7,06

Tableau 16 : Valeurs prédites et valeurs cibles dans une gamme de DH d'intérêt *(échantillons extraits du lot de prédiction, répartition R_V, architecture 4×4×1).*

Certes, remettre en cause les objectifs d'un modèle d'étalonnage, et en particulier la gamme prospectée, ne doit pas se faire indépendamment de la technique utilisée. Même si le passage de la modélisation de l'ensemble de la cinétique à un système plus local peut être supporté ici,

d'autres techniques dont nous avons discuté au paragraphe 3.1.2, seraient probablement également adaptées.

Enfin, l'intérêt des méthodes de chimiométrie tient indiscutablement à l'attention particulière qui peut être portée à l'interprétation des modèles construits. Du point de vue de l'information physico-chimique révélée par les vecteurs *loading*, ce sont les profils et les intensités relatives des bandes amide I et amide II qui sont informatifs. Par ailleurs, concernant les possibilités qu'engendre une meilleure caractérisation numérique du problème, la faible dimension des réseaux de neurones finalement exploités peut être mise en avant. Nous avons souligné qu'il est certainement injuste de restreindre les réseaux neuronaux à des outils de modélisation abstraits. En effet, l'observation des poids entre le niveau intermédiaire et le niveau de sortie a ainsi permis de caractériser l'importance de chacun des neurones cachés. Simultanément, l'analyse des domaines sur lesquels leurs fonctions sigmoïdes sont activées a aidé à déterminer leurs rôles respectifs et le degré de non-linéarité des traitements effectués.

Conclusion

Un système instrumental capable d'engendrer indirectement une mesure pertinente de l'avancement de la protéolyse pepsique d'hémoglobine bovine a pu être élaboré. Il combine les observations du système par spectrométrie infrarouge et leur exploitation par des outils de chimiométrie. Il est ainsi possible d'envisager la construction d'un capteur dédié à la mesure du degré d'hydrolyse, indicateur à lui seul de la composition peptidique du milieu. Le recours aux modèles empiriques peut être une alternative efficace et intéressante en vue d'une connaissance du procédé plus rapide et permettant à terme de s'affranchir de mesures de référence coûteuses. Le choix de méthodes basées sur les réseaux de neurones artificiels a été motivé par la complexité du problème et la nature des observations. Ils permettent, selon l'architecture choisie, aussi bien la visualisation des données par l'intermédiaire des cartes de Kohonen, que la modélisation supervisée via les réseaux multicouches.

Une étude de faisabilité a tout d'abord été effectuée hors-ligne sur des échantillons prélevés. Elle a montré la possibilité d'estimer le degré d'hydrolyse à partir des spectres infrarouge enregistrés en transmission, traités par des modèles neuronaux ou des régressions des moindres carrés partiels. Par ailleurs, nous avons pu mettre en évidence que ce sont principalement des changements de structure de la molécule, imputables au clivage ou à la dénaturation, qui constituent l'information physico-chimique significative. Néanmoins, pour envisager le suivi en ligne, il faut effectuer l'observation spectroscopique du milieu sous sa forme liquide. La mesure est réalisée par l'intermédiaire d'une liaison fibre optique entre le compartiment échantillon du spectromètre et le réacteur de laboratoire. L'interface est un cristal qui assure la réflexion totale atténuée de l'énergie infrarouge. Les spectres observés, très ressemblants, présentent un niveau d'absorbance faible mais leur qualité est suffisante pour être informatifs.

L'étude des données spectrales a, d'une part, révélé une classification en fonction de la répétition, d'autre part, établit la présence de non-linéarités confortant de fait le choix des méthodes neuronales pour la modélisation. Pour satisfaire au principe de parcimonie, l'analyse en composantes principales des données spectrales a été réalisée. La topologie optimisée, le modèle permet alors la prédiction, avec une erreur moyenne satisfaisante, des données enregistrées sur des répétitions analytiquement inconnues du procédé.

La chimiométrie est certes impliquée dans la production des données mais également dans l'extraction d'informations perspicaces, de déductions ou d'inductions que l'on associe généralement au concept d'intelligence. Les modèles construits ont, autant que possible, été interprétés ; du point de vue de l'information physico-chimique que contiennent les

observations comme du point de vue des paramètres du modèle, poids et activations des neurones. Finalement, l'étude interactive de ces mécanismes permet d'aboutir à une mesure construite plus sûre, voire validée. Toute interprétation, même incomplète, est une valeur ajoutée aux résultats de prédiction en termes de confiance et de connaissance des dépendances entre variables ou entre facteurs.

Enfin, les traitements décrits pourront (devront) être localisés sur le procédé pour permettre une action en retour, déterminée par la réponse fournie par le modèle. Ce transfert n'est délicat, du point de vue des modèles, qu'en ce qui concerne l'adaptation de l'étalonnage à une gamme plus ouverte de procédés. En effet, s'il y a une difficulté à utiliser les méthodes neuronales, celle-ci réside dans la construction de leur architecture et dans leur apprentissage. Une fois le modèle construit, il se résume à un certain nombre de valeurs numériques, qu'il est possible de transférer aisément, sur une carte à puces par exemple. Cette solution instrumentale matérielle, plutôt que logicielle, convient à la mise en œuvre d'une instrumentation peu coûteuse, dédiée et impliquée dans un système global.

Annexes

Algorithme NIPALS :[5]

L'algorithme extrait un facteur à la fois. Pour les facteurs a =1, 2, ..., A, on calcule $\mathbf{t_a}$ et $\mathbf{p_a}$ à partir de $\mathbf{X_{a-1}}$.

0• Sélectionner les valeurs de départ c'est à dire $\mathbf{t_a}$, la colonne de $\mathbf{X_{a-1}}$ dont la somme des carrés est la plus élevée. Répéter alors les étapes suivantes jusqu'à convergence.

1• Améliorer l'estimation du facteur *loading* $\mathbf{p_a}$ en projetant la matrice $\mathbf{X_{a-1}}$ sur $\mathbf{t_a}$

$$\mathbf{p_a'} = (\mathbf{t_a't_a})^{-1}\mathbf{t_a'X_{a-1}}$$

2• Normaliser $\mathbf{p_a}$ à l'unité

$$\mathbf{p_a} = \mathbf{p_a}\left(\mathbf{p_a'p_a}\right)^{-0.5}$$

3• Améliorer l'estimation $\mathbf{t_a}$ en projetant la matrice $\mathbf{X_{a-1}}$ sur $\mathbf{p_a}$

$$\mathbf{t_a'} = \mathbf{X_{a-1}p_a}(\mathbf{p_a'p_a})^{-1}$$

4• Améliorer l'estimation des valeurs propres $\tau a = \mathbf{t_a}'\,\mathbf{t_a}$.

5• Contrôler la convergence.

Si la différence entre les valeurs propres obtenues lors de deux itérations successives est inférieure à une limite définie, alors la méthode a convergé. On effectue alors la soustraction

$$\mathbf{X_a} = \mathbf{X_{a-1}} - \mathbf{t_a p_a'}$$

et reprendre l'étape 0• pour le facteur suivant.

Sinon, reprendre l'étape 1•.

Algorithme de PLS (Wold) :[5]

Les variables **X** et y sont centrées et on détermine un entier A supérieur au nombre de variables à extraire de X. Pour chaque facteur a=1, 2, ..., A, les étapes suivantes sont effectuées pour réaliser l'étalonnage.

1• Déterminer le vecteur *loading* w_a qui rend maximale la covariance X_{a-1} et y_{a-1} en respectant $w_a'.\ w_a=1$.

wa est un vecteur unitaire solution de $X_{a-1} = y_{a-1} w_a' + E$.

2• Construire la composante *score* t_a, par projection de X_{a-1} sur w_a.

Résoudre $X_{a-1} = t_a w_a' + E$ ou $t_a = X_{a-1} w_a$.

3• Régresser X_{a-1} sur t_a pour obtenir la composante *loading* p_a'

$$X_{a-1} = t_a p_a' + E \text{ a pour solution } p_a = X_{a-1}' \frac{t_a}{t_a' . t_a}.$$

4• Calculer les composantes q_a en déterminant les solutions de

$$y_{a-1} = t_a q_a + f \text{ qui s'écrivent } q_a = y_{a-1}' \frac{t_a}{t_a' . t_a}.$$

5• Calculer de nouvelles erreurs résiduelles de **X** et y par soustraction des effets modélisés pour le facteur a courant.

$$E = X_{a-1} - t_a p_a'$$

$$f = y_{a-1} - t_a q_a$$

Méthodes de validation.

Une des méthodes permettant de quantifier et d'optimiser les qualités prédictives d'un modèle consiste à réaliser une estimation des données fournies par celui-ci sur un lot de M échantillons indépendants. Le paramètre appelé RMSEP (*Root Mean Square Error of Prediction* en anglais) est un indicateur des capacités prédictives de l'étalonnage construit.

$$RMSEP = \sqrt{\frac{\sum_{j=1}^{M} (\hat{y}_j - y_j)^2}{M}}$$

avec $\hat{y}_j$ la valeur prédite pour $j^{ième}$ échantillon pour lequel on attend y_j.[74]

Répétition r_1		Répétition r_2		Répétition r_3	
Echantillon	DH (%)	Echantillon	DH (%)	Echantillon	DH (%)
1	**0**	**1**	**0**	**1**	**0**
2	**3**	**2**	**1,4**	**2**	**1,2**
3	**3,3**	**3**	**2**	**3**	**1,7**
4	**3,7**	**4**	**2,4**	**4**	**1,9**
5	**4,5**	**5**	**2,6**	**5**	**2,5**
6	**4,8**	**6**	**3,3**	**6**	**3**
7	**4,9**	**7**	**3,5**	**7**	**3,2**
8	**5,1**	**8**	**3,7**	**8**	**3,4**
9	**5,2**	**9**	**3,8**	**9**	**3,6**
10	**5,4**	**10**	**4**	**10**	**3,7**
11	**5,6**	**11**	**4**	**11**	**3,8**
12	**5,7**	**12**	**4,1**	**12**	**3,9**
13	**5,9**	**13**	**4,2**	**13**	**4**
14	**6**	**14**	**4,4**	**14**	**4,1**
15	**6,1**	**15**	**4,5**	**15**	**4,3**
16	**6,2**	**16**	**4,8**	**16**	**4,4**
17	**6,3**	**17**	**5**	**17**	**4,5**
18	**6,3**	**18**	**5,1**	**18**	**4,7**
19	**6,5**	**19**	**5,2**	**19**	**4,8**
20	**6,5**	**20**	**5,3**	**20**	**5**
21	**6,6**	**21**	**5,4**	**21**	**5,2**
22	**6,6**	**22**	**5,4**	**22**	**5,3**
23	**6,7**	**23**	**5,5**	**23**	**5,4**
24	**6,8**	**24**	**5,5**		
25	**7**	**25**	**5,5**		
26	**7**	**26**	**5,6**		
27	**7,2**	**27**	**5,6**		
28	**7,3**	**28**	**5,6**		
29	**7,5**	**29**	**5,7**		
30	**7,5**	**30**	**5,8**		
31	**7,7**	**31**	**5,8**		
32	**7,8**	**32**	**5,9**		
33	**7,9**	**33**	**5,9**		
34	**8**	**34**	**6**		
35	**8,1**	**35**	**6,1**		
36	**8,1**	**36**	**6,2**		
37	**8,2**	**37**	**6,2**		
38	**8,3**	**38**	**6,3**		
39	**8,3**	**39**	**6,3**		
40	**8,4**	**40**	**6,4**		
41	**8,5**	**41**	**6,5**		
42	**8,7**	**42**	**6,5**		
		43	**6,6**		
		44	**6.7**		
		45	**6,7**		

Les échantillons prélevés et analysés par la méthode de référence sont représentés en gras. Les autres sont le résultat d'interpolations linéaires.

Répétition r_4		Répétition r_5		Répétition r_6	
Echantillon	**DH (%)**	**Echantillon**	**DH (%)**	**Echantillon**	**DH (%)**
1	**0**	**1**	**0**	**1**	**0**
2	**1,2**	**2**	**1,2**	**2**	**1,5**
3	**1,8**	**3**	**1,9**	**3**	**2**
4	**2,7**	**4**	**2**	**4**	**2,6**
5	**3**	**5**	**2,4**	**5**	**3**
6	**3,2**	**6**	**2,8**	**6**	**3,5**
7	**3,3**	**7**	**3,2**	**7**	**3,6**
8	**3,5**	**8**	**3,7**	**8**	**3,8**
9	**3,6**	**9**	**4,3**	**9**	**3,9**
10	**3,7**	**10**	**4,6**	**10**	**4,1**
11	**3,8**	**11**	**4,7**	**11**	**4,3**
12	**4**	**12**	**4,8**	**12**	**4,5**
13	**4,1**	**13**	**4,9**	**13**	**4,6**
14	**4,1**	**14**	**4,9**	**14**	**4,7**
15	**4,2**	**15**	**5**	**15**	**4,7**
16	**4,2**	**16**	**5,1**	**16**	**4,8**
17	**4,6**	**17**	**5,2**	**17**	**4,8**
18	**4,8**	**18**	**5,4**	**18**	**4,9**
19	**5**	**19**	**5,6**	**19**	**5**
20	**5,5**	**20**	**5,6**	**20**	**5,1**
21	**5,6**	**21**	**5,7**	**21**	**5,2**
22	**6**	**22**	**5,8**	**22**	**5,4**
23	**6,1**	**23**	**6**	**23**	**5,5**
24	**6,2**	**24**	**6,1**	**24**	**5,6**
25	**6,2**	**25**	**6,1**	**25**	**5,7**
26	**6,3**	**26**	**6,2**	**26**	**5,9**
27	**6,4**	**27**	**6,4**	**27**	**6,1**
28	**6,5**	**28**	**6,5**	**28**	**6,2**
29	**6,5**	**29**	**6,6**	**29**	**6,2**
30	**6,6**	**30**	**6,7**	**30**	**6,3**
31	**6,6**	**31**	**6,8**	**31**	**6,5**
32	**6,7**	**32**	**7**	**32**	**6,6**
33	**6,8**	**33**	**7,1**	**33**	**6,8**
34	**6,8**	**34**	**7,2**	**34**	**6,9**
35	**6,9**	**35**	**7,2**	**35**	**7**
36	**6,9**	**36**	**7,3**	**36**	**7,2**
37	**7**	**37**	**7,3**	**37**	**7,3**
38	**7,2**	**38**	**7,4**	**38**	**7,4**
39	**7,2**	**39**	**7,5**	**39**	**7,5**
40	**7,4**	**40**	**7,5**	**40**	**7,7**
41	**7,5**	**41**	**7,6**	**41**	**7,9**
42	**7,7**	**42**	**7,6**	**42**	**8,2**
43	**7,8**	**43**	**7,7**	**43**	**8,2**
44	**8,1**	**44**	**7,7**	**44**	**8,2**
45	**8,3**	**45**	**7,8**		
		46	**8**		
		47	**8**		
		48	**8,2**		
		49	**8,3**		

Les échantillons prélevés et analysés par la méthode de référence sont représentés en gras.

Répétition r_5 Echantillon	DH (%) Valeurs cibles	DH (%) Valeurs prédites	Répétition r_6 Echantillon	DH (%) Valeurs cibles	DH (%) Valeurs prédites
2	1,2	2,41	1	1,5	2,36
3	1,9	1.32	2	2	2,68
4	2	1,13	3	2,6	2,32
5	2,4	1,77	4	3	2,67
7	3,2	3,45	5	3,5	3,34
8	3,7	4,12	6	3,6	3,72
9	4,3	3,9	7	3,8	4,09
10	4,6	4,83	8	3,9	4,17
11	4,7	4,98	9	4,1	3,69
12	4,8	5,04	10	4,3	4,62
13	4,9	4,32	11	4,5	4,29
15	5	5,42	12	4,6	4,25
16	5,1	4,86	13	4,7	4,92
17	5,2	5,52	14	4,7	4,73
18	5,4	5,47	15	4,8	4,92
19	5,6	5,37	16	4,8	4,89
20	5,6	5,8	17	4,9	5,14
21	5,7	5,78	18	5	4,68
22	5,8	5,62	19	5,1	5,37
23	6	6,11	20	5,2	5,04
24	6,1	6,02	21	5,4	5,78
25	6,1	6,32	22	5,5	5,23
26	6,2	6,15	23	5,6	5,32
27	6,4	6,28	24	5,7	5,83
28	6,5	6,63	25	5,9	5,96
29	6,6	6,29	26	6,1	6,01
30	6,7	6,67	27	6,2	5,86
31	6,8	6,63	28	6,2	6,27
32	7	7,06	29	6,3	6,35
33	7,1	6,91	30	6,5	6,37
34	7,2	6,98	31	6,6	6,72
35	7,2	7,32	32	6,8	6,76
37	7,3	7,01	33	6,9	7,19
38	7,4	7,73	34	7	7,32
39	7,5	7,83	35	7,2	7,24
41	7,6	7,77	36	7,3	7,68
43	7,7	7,85	37	7,4	7,21
44	7,7	7,42	38	7,5	7,27
45	7,8	8,18	39	7,7	7,59
46	8	8,2	40	7,9	7,47
47	8	7,89	41	8,2	7,78
48	8,2	8,39	42	8,2	8,03
49	8,3	8.11	43	8,2	7,79

Index des tables et figures

Glossaire

Activation : fonction des données d'entrée et des paramètres.

Adaptation : modification automatique de paramètres, en général pour optimiser une mesure de performance.

Algorithme : séquence cyclique d'instructions.

Apprentissage : changement de comportement dépendant de l'expérience et améliorant la performance d'un système. Désigne à la fois l'entraînement et le contrôle.

Apprentissage non-supervisé : schéma d'apprentissage sans connaissance à priori à propos des données utilisées.

Apprentissage supervisé : schéma d'apprentissage au cours duquel la différence entre la valeur produite et la valeur attendue diminue.

Architecture : organisation globale du transfert d'informations, du stockage et des traitements.

Backpropagation : méthode d'optimisation des vecteurs poids qui propage l'erreur des niveaux de sortie vers l'entrée et utilisée pour les réseaux *feedforward* multicouches.

Caractéristique : information élémentaire qui représente une ou plusieurs propriétés d'un objet.

Carte auto-organisatrice : réseau compétitif qui produit une représentation topologique de l'espace d'entrée sur les unités de la carte de sortie.

Classe : ensemble de points voisins les uns des autres dans l'espace des données.

Classification : méthode consistant à répartir les échantillons dans des catégories en fonction des valeurs des variables observées.

Cognitif : lié au traitement de l'information par l'être humain.

Compétitif : choix de l'élément portant l'optimum d'un critère, implique le calcul de ce critère pour tous les éléments.

Compression : méthode de réduction des données qui préserve leur spécificité pour l'application.

Connexion : lien entre les neurones, couplant leurs signaux en tenant compte du paramètre poids.

Couche d'entrée : lieu des neurones recevant directement les signaux d'entrée.

Couche cachée : couche de neurones intermédiaires dans un réseau multicouche.

Curse of dimensionality : augmentation rapide voire intolérable des exigences de calcul lorsque la dimension des vecteurs d'entrée croît.

Déterministe : liés aux conséquences inévitables des conditions de départ.

Distance : mesure spécifique de similitude entre des ensembles de valeurs.

Early-stopping : technique pour l'apprentissage permettant d'assurer une modélisation optimale du point de vue du compromis biais-variance et engendrant ainsi de meilleures capacités de généralisation.

Etalonnage : utilisation des données empiriques et de la connaissance disponible a priori pour prédire quantitativement une information inconnue. Le but est la substitution d'une mesure par une méthode moins coûteuse et suffisamment précise.

Etalonnage multivarié : développement d'un modèle quantitatif pour la prédiction sûre des propriétés recherchées à partir d'un certain nombre de variables prédictives.

Entrée : les neurones d'entrée sont des distributeurs de signaux et ne jouent aucun rôle actif dans leur modification.

Entraînement : enseignement forcé.

Feedback : signaux propagés de la sortie vers l'entrée.

Feedforward : signaux propagés de l'entrée vers la sortie, sans retour vers le neurone d'origine.

Généralisation : capacité à répondre d'une même manière à un ensemble d'entrées, certaines d'entre-elles n'appartenant pas aux lots d'apprentissage.

Heuristique : choix de l'expérimentateur, basés sur l'expérience et l'intuition.

Hiérarchique : relatif à une organisation sur plusieurs niveaux, subordonnés.

Intelligence artificielle : capacité d'un système à réaliser des tâches qui requièrent habituellement une certaine forme d'intelligence.

Label : symbole décrivant une classe de représentations.

Lien : connexion entre les neurones, couplant le signal et le poids associé.

Loading : vecteurs maximisant la variance empirique contenue au sein de la matrice des données lors d'une analyse en composantes principales.

Lot de contrôle : ensemble d'exemple permettant la variation des paramètres du modèle pour son optimisation.

Lot d'entraînement : ensemble d'exemples utilisé pour calculer les paramètres du modèle.

Lot de test : ensemble d'exemple utilisé pour caractériser la performance d'un modèle entraîné.

Métrique : propriété déterminant une distance symétrique entre les éléments.

Modèle : description simplifiée et approximative d'un système ou d'un procédé.

Modèle non paramétrique : méthode qui ne se base sur aucune fonction mathématique pour

la description des données, mais qui se réfère directement aux exemples disponibles.

Modèle paramétrique : méthode basée sur des fonctions mathématiques définies, incluant des paramètres libres, pour la description des données.

Modélisation : recherche d'une fonction analytique ou d'une procédure qui produit une sortie spécifique pour un type d'entrées défini.

Moindres carrés : méthode consistant à minimiser la somme des carrés des résidus.

Multivarié : les éléments individuels sont décrits par au moins deux variables.

Neurone : unité spécialisée qui transmet et traite les signaux.

Ondelettes : fonctions de base permettant la transformation des données spectrales.

Poids : paramètre réel associé au lien entre deux neurones artificiels.

Point aberrant : point aux coordonnées atypiques qui peut masquer la structure des autres données.

Pré-traitement : ensemble d'opérations sur les signaux, précédant leur utilisation.

Principe de parcimonie : consiste à accorder une préférence au modèle le plus simple.

Répétition (*batch*) : ensemble de tâches réalisées lors d'une même procédure.

Représentation : codage de l'information pour le calcul.

Régression : analyse des relations entre deux variables ou plus. Cette relation est exprimée sous la forme d'une fonction mathématique qui peut être utilisée pour la prédiction. Une application importante est l'étalonnage.

Réseau de neurones artificiels : réseau massivement parallèle d'éléments simples, interconnectés et organisés hiérarchiquement.

Réseau multicouche ***feedforward*** : architecture composée de couches de neurones successives qui reçoivent leurs entrées du niveau précédent.

Résidu : différence entre la valeur prédite et la valeur de référence.

Score : valeur des projections des données sur les axes propres générés par l'analyse en composante principale.

Sigmoïde : fonction non-linéaire utilisée pour le transfert des données dans un neurone artificiel.

Sous-modélisation : modélisation ne détectant pas toute la complexité du signal.

Sur-modélisation : modélisation considérant le bruit présent au sein du lot de données.

Topologie : configuration physique et logique des connections d'un réseau.

Vainqueur : neurone qui, lors de l'apprentissage compétitif, se voit attribué l'activation la plus importante.

Validation externe : estimation de l'erreur de prédiction à partir d'un nombre limité de

données hors du lot d'étalonnage.

Validation interne : estimation de l'erreur de prédiction à partir d'un nombre limité de données provenant du lot d'étalonnage.

Variable latente : combinaison linéaire des variables de l'espace d'origine.

Vecteur d'entrée : vecteur formé par les données d'entrée.

Voisinage : ensemble de neurones localisés topologiquement au plus près d'un neurone central.

Weight decay : Méthode de régularisation des poids basée sur l'ajout d'un terme de pénalité dans l'expression de la fonction d'erreur.

Index bibliographique

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Y

Z