Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Chimie 9 (2006) 364–373
COMPTES RENDUS CHIMIE
http://france.elsevier.com/direct/CRAS2C/

Application du traitement par entropie maximale aux données RMN multidimensionnelles ; cas de l'échantillonnage partiel

Marc-André Delsuc [a,*], Dominique Tramesel [b]

[a] *Centre de biochimie structurale, Inserm UMR 554–CNRS UMR 5048, 29, rue de Navacelles, 34090 Montpellier, France*
[b] *NMRtec, 288, rue d'Uppsala, Les Allées du Bois, 64/2, 34080 Montpellier, France*

Reçu le 7 avril 2005 ; accepté le 8 juin 2005

Disponible sur internet le 26 août 2005

Résumé

La méthode d'analyse par entropie maximale permet de réaliser l'analyse spectrale des données de RMN multidimensionnelles. Cette méthode permet l'analyse spectrale de données mesurées avec un échantillonnage partiel, qui ne pourraient pas être analysées par transformée de Fourier. Cette possibilité ouvre la porte à la mise en place d'échantillonnages optimisant en même temps les durées d'acquisition, la sensibilité et la résolution des spectres finaux. Dans cette étude, nous présentons la mise en œuvre de ce principe dans le logiciel Gifa. Nous explorons les échantillonnages aléatoires, et radiaux, et montrons, sur des exemples simulés et réels, que l'analyse par entropie maximale retrouve l'ensemble des caractéristiques de spectres, tout en permettant un gain de temps d'acquisition allant entre 4 et 10. ***Pour citer cet article : M.-A. Delsuc, D. Tramesel, C. R. Chimie 9 (2006)***.

Abstract

Application of Maximum-Entropy processing to NMR multidimensional datasets, partial sampling case. Multidimensional NMR experiments can be analysed with the Maximum Entropy method. This method permits the analysis of partially sampled data, which would not be analysed by the Fourier transform technique. This allows the set-up of alternative samplings, which optimize the acquisition time, as well as the sensitivity, and the resolution of the final spectra. In this study we show the implementation of this principle in the Gifa software. We consider random and radial samplings, and show, on simulated as well as real data, that the Maximum-Entropy method is able to recover all the features of the spectra, and permits in the same time gains in acquisition times ranging from 4 to 10. ***To cite this article: M.-A. Delsuc, D. Tramesel, C. R. Chimie 9 (2006)***.

Mots-clés : RMN 3D ; Entropie maximale ; Gifa ; Échantillonnage partiel ; Échantillonnage aléatoire

Keywords: 3D NMR; Maximum entropy; MaxEnt; Gifa; Partial sampling; Random sampling

* Auteur correspondant.
Adresses e-mail : ma.delsuc@cbs.cnrs.fr (M.-A. Delsuc), dominique@nmrtec.com (D. Tramesel).

1631-0748/$ - see front matter © 2005 Académie des sciences. Publié par Elsevier SAS. Tous droits réservés.
doi:10.1016/j.crci.2005.06.013

1. Introduction

L'application du principe de l'entropie maximum dans le cadre du traitement de données permet de proposer une approche générale de l'analyse de donnée : MaxEnt. Cette approche permet d'analyser d'une manière optimale tous les problèmes pouvant s'exprimer sous forme linéaire. Elle est utilisée très couramment dans le domaine du traitement de l'image, en particulier en astronomie et en astrophysique. Mais ses applications sont extrêmement diverses.

1.1. Position du problème

Cette grande flexibilité a permis que le traitement par entropie maximum soit appliqué à de nombreux aspects du traitement des spectres de RMN. Ainsi, cette approche a été proposée pour augmenter la résolution des spectres et optimiser le rapport signal sur bruit [1–5], analyser et mesurer les structures de couplage [6–8], analyser les DOSY par transformée de Laplace Inverse [9], etc. Cette flexibilité permet aussi de se libérer des contraintes des protocoles rigides d'échantillonnage des temps dans les expériences multidimensionnelles. C'est pourquoi ont été présenté dans la littérature des protocoles d'acquisition et de traitement, s'appuyant sur MaxEnt et permettant de raccourcir les temps de mesure d'expériences 2D et 3D [10–13].

La sensibilité des spectromètres actuels permet d'obtenir des rapports signal à bruit suffisants en des temps d'accumulation très courts. Cependant le protocole régulier d'acquisition des expériences multidimensionnelles impose un nombre de points minimum sur chacun des axes spectraux, et donc un temps minimum d'acquisition. Pour une expérience de 3D, même en réduisant le nombre d'accumulation au minimum, ce temps est de l'ordre de plusieurs heures, souvent au-delà de ce qui est nécessaire pour un rapport signal sur bruit suffisant. Une solution pour réduire le temps d'acquisition consiste à modifier le protocole d'échantillonnage des axes temporels, en abandonnant l'échantillonnage régulier et en choisissant un échantillonnage qui permette de réduire le temps total de mesure. Cependant, l'abandon de l'échantillonnage régulier ne permet plus d'utiliser la transformée de Fourier rapide, qui impose un échantillonnage régulier. D'autres techniques de traitement doivent donc être envisagées. Dans ce travail, nous présentons l'utilisation du traitement par entropie maximale pour réaliser cette analyse.

1.2. La transformée de Fourier et les contraintes d'acquisition

L'usage est maintenant universel d'utiliser la transformée de Fourier pour réaliser l'analyse de données temporelles issues de la RMN par impulsion. Pour des raisons de rapidité et d'efficacité, c'est l'algorithme de transformée de Fourier rapide (*Fast Fourier Transform: FFT*) qui est largement utilisé [14].

Lors de la mesure temporelle de la RMN, une impulsion est envoyée à l'échantillon. Cette impulsion induit une réponse du système de spin, sous la forme d'une émission d'un signal à la fréquence de résonance : le FID. Ce signal électrique doit être numérisé (transformation de la valeur analogique en valeur numérique) et échantillonné (mesure faite à une série d'instants particuliers) pour pouvoir être manipulé sous forme informatique. Nous sommes aujourd'hui habitués à régler de manière empirique de nombreux paramètres pour cette mesure temporelle : fréquence d'échantillonnage, durée d'échantillonnage, nombre de points, etc. et il y a de nombreuses contraintes sur ces paramètres, dont les sources sont variables. Nous allons passer en revue l'origine de quelques-unes de ces contraintes, et examiner comment ces contraintes sont modifiées ou conservées quand on modifie le protocole d'échantillonnage [14].

Le premier paramètre est la durée d'échantillonnage $T_{\max}$, c'est-à-dire le temps pendant lequel la mesure est effectuée. Cette durée détermine la résolution spectrale maximale de la mesure. En effet, pour discerner deux signaux séparés en fréquence, il faut attendre que cette différence de fréquence s'exprime et que des battements s'installent. Si l'on pose qu'il faut avoir une période complète de battements pour séparer les deux signaux, alors une résolution de $\Delta\nu$ impose une durée liée par la relation suivante :

$$T_{max} = 1/\Delta\nu \tag{1}$$

Il faut remarquer que cette limite de résolution est directement liée au choix de l'observation d'une période de battement complète. Si une méthode d'analyse permet de séparer les fréquences sur une évolution plus courte, la résolution spectrale $\Delta\nu$ en sera augmentée.

Les algorithmes de transformée de Fourier nous imposent un échantillonnage régulier de l'évolution temporelle, et ainsi la notion de fréquence d'échantillonnage $F_{\text{échant}}$ apparaît naturellement. Le nombre de

points de la mesure N et la durée d'échantillonnage sont directement liés :

$$N = F_{\text{échant}}\ T_{\max} \tag{2}$$

Cette fréquence d'échantillonnage introduit aussi une limite à la fréquence détectable la plus élevée $F_{\max}$. En effet, entre deux points successifs de l'échantillon, une fréquence parcourant plus d'une demi-période est indiscernable de celle parcourant moins d'une demi-période en sens opposé ; en effet, modulo 2 π, $\pi + \alpha$ est indiscernable de $-\pi + \alpha$. Cette limite s'exprime par la relation de Nyquist–Shannon :

$$F_{\max} = F_{\text{échant}}/2 \tag{3–1}$$

Cette dernière peut s'écrire en reliant la largeur spectrale LS mesurée à la période d'échantillonnage $\Delta t = 1/F_{\text{échant}}$:

$$LS = 1/\Delta t \tag{3–2}$$

dans le cas d'un échantillonnage complexe (dans ce cas, le signe de la fréquence est détecté et $LS = 2\,F_{\max}$).

Pour finir, l'algorithme de FFT est plus rapide dans le cas où le nombre de points N est une puissance de 2. Cette contrainte est cependant assez faible, et il existe des algorithmes permettant de faire ce calcul sur un nombre quelconque de points.

2. Le traitement des données par entropie maximale, et l'échantillonnage partiel

2.1. Principes généraux

Nous allons présenter ici rapidement les grands principes de l'analyse spectrale de données RMN par entropie maximale (MaxEnt). Pour une présentation plus complète, nous invitons le lecteur à se tourner vers la littérature [1–16].

Le principe de l'analyse spectrale par MaxEnt est de chercher un spectre dont la transformée de Fourier inverse corresponde le mieux aux données brutes (le FID). C'est une approche inverse, c'est-à-dire que l'on va du spectre vers les données et non pas l'inverse. L'adéquation du FID calculé à partir du spectre avec les données expérimentales se fait par l'évaluation du χ^2 :

$$\chi^2 = \sum \left(\frac{D_i^{\text{exp}} - D_i^{\text{calc}}}{\sigma_i} \right)^2 \tag{4}$$

Cependant, un simple critère de distance entre les données mesurées et les données reconstruites à partir du spectre est insuffisant. En effet, le bruit expérimental induit une incertitude σ_i sur chaque point de mesure. Cette incertitude fait que la valeur optimum du χ^2 n'est pas la valeur minimum, mais suit le niveau d'incertitude donné, généralement $\sum\sigma_i$. Il existe donc une infinité de spectres compatibles avec les données expérimentales parmi lesquels il faut choisir. C'est pourquoi l'on choisit de maximiser conjointement l'entropie au sens de Shannon [15] S présente dans le spectre :

$$S = -\sum p_j \log\,(p_j) \text{ avec } p_j = \frac{f_j}{\sum f_j} \tag{5}$$

où f_j est un point du spectre reconstruit.

Autrement dit, on choisit le spectre qui minimise la quantité d'information au sens de Shannon présente dans le spectre reconstruit.

Cette optimisation est largement non linéaire, et c'est ce qui pose le problème numérique principal de cette approche. Cependant, la partie la plus lourde du calcul est prise par des transformées de Fourier.

L'avantage de cette approche est de pouvoir introduire très facilement toutes les informations que l'on possède a priori sur les données : bruit, données tronquées, points aberrants, mais aussi formes de raies, structures fines. La possibilité de retirer les effets introduits par ces perturbations permet d'effacer (de déconvoluer) les structures associées dans les spectres. On va ainsi se débarrasser, non seulement des artefacts de troncature, mais aussi du bruit, et même des raies larges, des structures fines, etc.

En réalité, on peut considérer la méthode MaxEnt comme une méthode générale de traitement des données, et c'est ce qui explique pourquoi elle est utilisée dans des domaines scientifiques très divers. Malheureusement, cette généralité se paye au prix de la lourdeur algorithmique de l'appareil mathématique. Ainsi, l'algorithme est relativement difficile à développer et à maîtriser. Des travaux récents, réalisés pour l'analyse de DOSY, ont beaucoup simplifié le paramétrage de

l'algorithmique disponible dans Gifa [16]. Pour finir, les temps de calcul, bien qu'importants, sont devenus tout à fait compatibles avec une utilisation régulière.

2.2. L'échantillonnage partiel

Nous avons vu que l'algorithme de transformée de Fourier impose de faire un échantillonnage régulier des axes temporels. L'approche inverse du calcul par MaxEnt permet d'introduire facilement des échantillonnages alternatifs. En effet, en approche inverse, la transformée de Fourier n'est pas appliquée sur les données, mais une transformée inverse est appliquée sur le spectre reconstruit. La contrainte n'est donc plus sur les données, mais sur le spectre, et l'échantillonnage temporel n'a plus à être régulier.

En revanche, la transformée de Fourier rapide est toujours utilisée. Et donc le FID recalculé par le calcul inverse est forcément régulièrement échantillonné. Ainsi, pour pouvoir calculer un écart entre le FID expérimental et le FID reconstruit, il faut donc que les points du FID expérimental soient placés sur la grille régulière imposée par la FFT. En revanche, tous les points n'ont pas besoin d'être présents. C'est pourquoi on parle d'échantillonnage partiel.

Le principe du calcul par MaxEnt du spectre à partir de données partiellement échantillonnées est présenté sur la Fig. 1.

2.3. Différents types d'échantillonnages

L'échantillonnage partiel n'est utile que pour accélérer la mesure du spectre de RMN ; il n'a donc pas beaucoup d'utilité en RMN 1D ; en revanche, en RMN 2D, on peut appliquer un échantillonnage le long de l'axe non classique $t1$; en 3D le long des deux axes non classiques $t1$ et $t2$; etc.

De nombreux types d'échantillonnages peuvent être réalisés, et il n'existe pas de « meilleur » échantillonnage : celui-ci doit être optimisé en fonction de l'expérience réalisée. Dans le cas de l'échantillonnage le long d'un axe (axe $t1$ d'une 2D), l'idéal est de choisir des échantillons dont la densité le long de l'axe temporel évolue suivant le profil du signal d'intérêt [11]. Ainsi, dans l'exemple de la Fig. 1, les points d'échantillons sont choisis aléatoirement, avec une loi d'amortissement exponentiel de 2 Hz. Dans certains cas, on peut choisir un échantillonnage purement aléatoire, ou même une loi plus exotique (par exemple dans le cas de l'expérience COSY [12]).

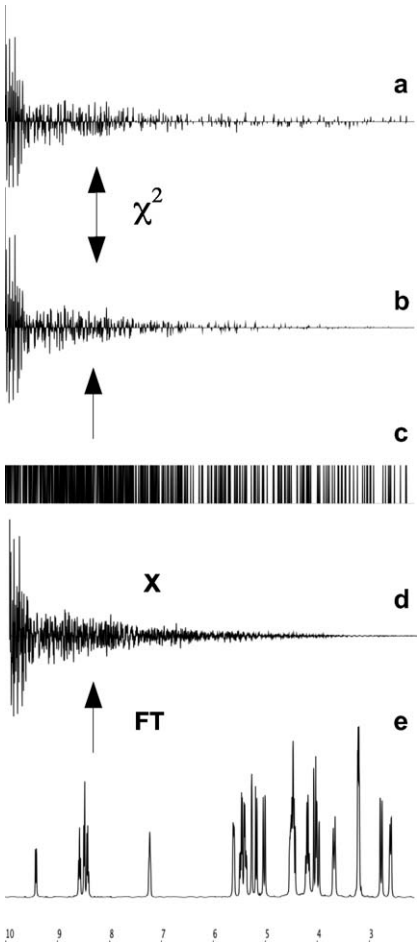

Fig. 1. Cette figure présente le principe de l'analyse par MaxEnt d'une mesure avec un échantillonnage partiel. Le FID expérimental (**a**), échantillonné partiellement peut être vu comme le résultat du produit d'un FID complet par une fonction d'échantillonnage (**c**). Au cours de l'analyse par MaxEnt, on calcule un FID reconstruit, (**d**) complet à partir de l'estimation du spectre (**e**), mais la comparaison entre le FID reconstruit et le FID expérimental n'est faite que sur les points mesurés (**b**). La fonction d'échantillonnage présentée ici correspond à un échantillonnage de 1024 points dans 4096, avec une densité de points qui suit une loi exponentielle décroissante de 2 Hz.

Dans le cas de l'échantillonnage 2D (plan $t1 \times t2$ d'une 3D), plusieurs protocoles sont possibles. Les différents types d'échantillonnages sont présentés sur la Fig. 2. Chaque axe peut être échantillonné indépendamment, résultant en un échantillonnage du plan par bloc (Fig. 2a). Il est possible d'échantillonner le plan 2D de manière complètement aléatoire (Fig. 2b).

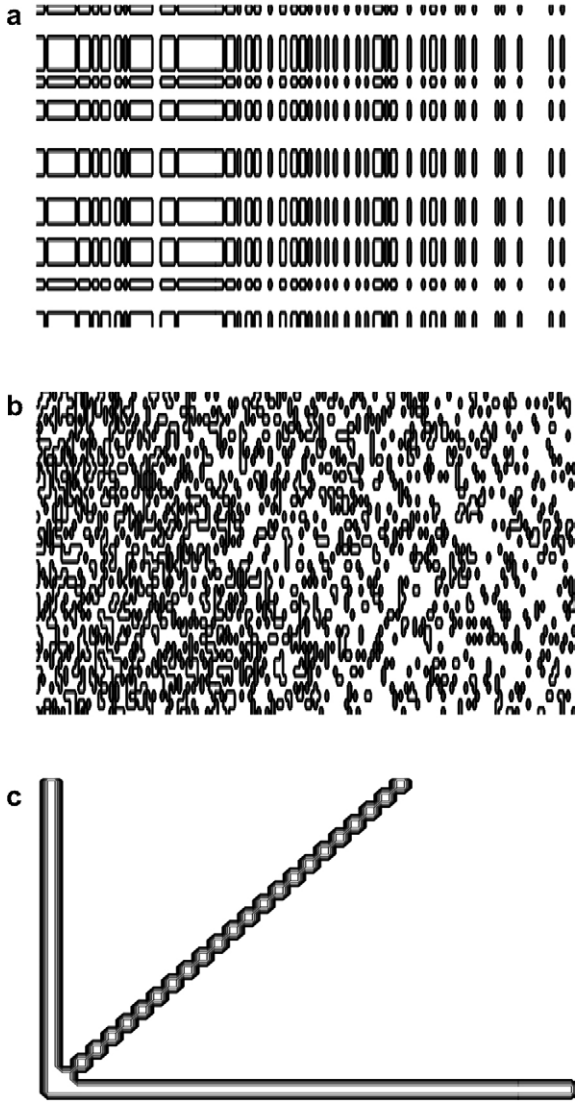

Fig. 2. Différents exemples d'échantillonnages.
(**a**) Échantillonnage par bloc, l'axe $t1$ (vertical – 40 points complexes), est échantillonné régulièrement, l'axe $t2$ (horizontal – 190 points complexes), est échantillonné avec un amortissement de 5 Hz. Le taux d'échantillonnage moyen est de 25 %.
(**b**) Échantillonnage aléatoire; l'axe $t1$ de (40 points complexes) est échantillonné régulièrement, l'axe $t2$ (190 points complexes) est échantillonné avec un amortissement de 5 Hz. Le taux d'échantillonnage moyen est de 25 %.
(**c**) Échantillonnage diagonal, seulement les axes $t1 = 0$, $t2 = 0$ et t1 et t2 incrémentés ensembles ont été échantillonnés. L'axe t1 fait 24 points complexes, l'axe t2 fait 35 points complexes. Le taux d'échantillonnage est de 9,65 %.
Tous les échantillonnages présentés sont hypercomplexes.

Ces deux modes réalisent un échantillonnage régulier du plan ; en revanche, il peut être relativement difficile de mettre en place le protocole d'acquisition. Par ailleurs, le traitement ne peut être réalisé qu'avec le protocole MaxEnt (ou un protocole équivalent), et il peut-être difficile de vérifier rapidement la qualité de l'expérience. Pour finir, un certain nombre d'expériences peuvent être envisagées comme des expériences produites par un échantillonnage particulier. Par exemple, la mesure de différents plans 2D issus d'une 3D, tels que proposé initialement par Brutscher et al. [17] et repris par plusieurs auteurs [18], peut se représenter sous la forme d'un échantillonnage (Fig. 2c). Ces auteurs ont proposé de réaliser un traitement 2D de ces expériences issues d'une 3D, et de remonter à l'information 3D en analysant les positions des pics dans les différentes expériences 2D. Par ailleurs, Kupče et Freeman [19] ont proposé de reconstruire la 3D ainsi sous-échantillonnée par différents algorithmes de projection–reconstruction, reconstruisant ainsi la 3D tout entière. Ce travail de reconstruction peut aussi être réalisé par MaxEnt, comme nous le verrons plus loin.

Quelques remarques supplémentaires peuvent être faites. Dans tous ces échantillonnages, il faut faire attention, pour chaque valeur de $t1 \times t2$, de mesurer les quatre points de la tétrade hypercomplexe. Bien que, à notre connaissance, rien ne justifie théoriquement cette contrainte, nous avons pu vérifier que les reconstructions MaxEnt sont plus stables dans ce cas. Ensuite, on peut se demander comment les différentes relations présentées plus haut sont conservées. L'eq. (1), qui relie résolution et temps d'acquisition, est vraiment liée au mode d'analyse. L'hypothèse selon laquelle il faut mesurer une période complète de battement pour discerner deux fréquences proches ne tient plus dans une approche où l'incertitude des données est prise en compte. Dans l'analyse par MaxEnt, la résolution ultime que l'on peut espérer obtenir est aussi liée à la qualité de la mesure. De même, le rapport entre le nombre de points d'échantillon et la résolution disparaît. Seule la relation de Nyquist–Shannon est conservée, car le phénomène de repliement de fréquence est une propriété de l'échantillonnage périodique, qui est conservée dans cette approche.

De nombreux types d'échantillonnages peuvent être réalisés ; cependant, tous ne sont pas utiles. En regardant la Fig. 1, on réalise que l'action d'échantillonnage est équivalente à une multiplication dans l'espace de

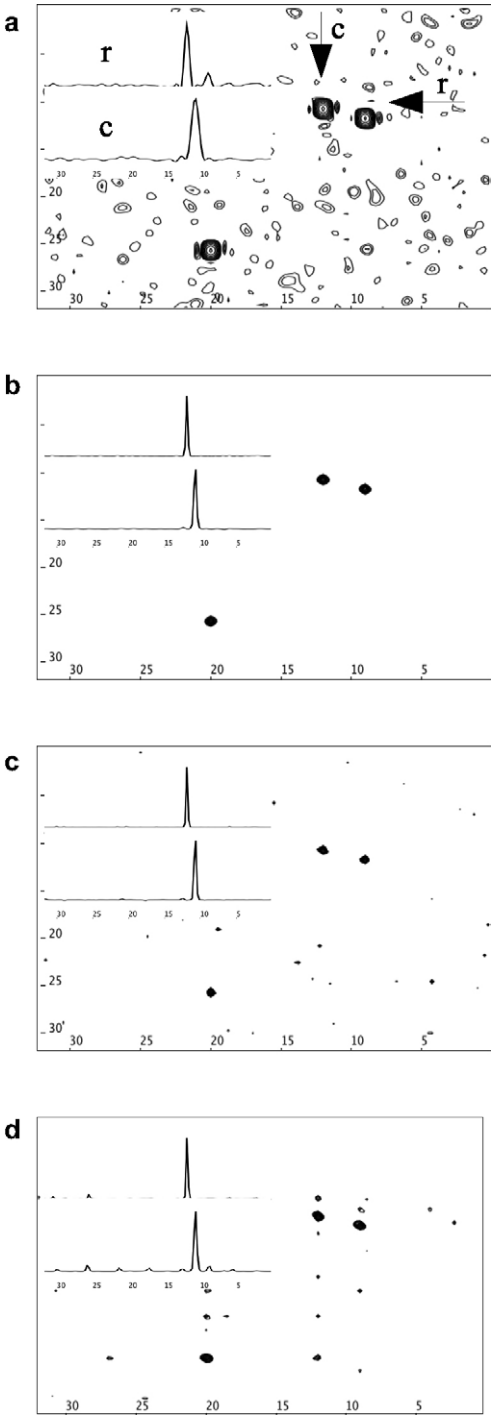

Fig. 3. Exemples de traitement sur un plan simulé d'une expérience 3D-HNCA.
Trois signaux, d'une intensité de 100 ont été simulés sur un plan de 24 points complexes (F1) par 35 points complexes (F2). Un bruit gaussien d'une intensité de 30 dans le domaine des temps a été rajouté aux données. Ces données ont été calculées dans différentes conditions, pour obtenir un spectre de 128 × 128 points. Chaque tracé est montré avec le plus bas niveau tracé à 1/32^{e} du point le plus haut du spectre. Sont aussi montrées une ligne et une colonne extraites au niveau du pic situé à (F1 = 10,5 ppm/F2 = 12 ppm).
(**a**) Traitement par FFT, apodisation avec une arche de sinus non shiftée.
(**b**) Traitement par MaxEnt, sans déconvolution particulière.
(**c**) Traitement par MaxEnt des données échantillonnées partiellement par 840 points répartis aléatoirement sur le plan. Le taux d'échantillonnage est de 25 %.
(**d**) Traitement par MaxEnt des données échantillonnées par les trois axes $t1 = 0$; $t2 = 0$; $t1 = t2$; protocole présenté sur la Fig. 3c. Le taux d'échantillonnage est de 9,65 %.

temps. À ce titre, le spectre est donc convolué par la transformée de Fourier de la fonction d'échantillonnage. Bien que l'analyse par MaxEnt réalise la déconvolution des données, certaines fonctions d'échantillonnage sont à éviter, car réalisant une perte définitive d'information. Par exemple, un sous-échantillonnage périodique (ne gardant qu'un points sur n) a pour effet de réduire la largeur spectrale de détection. De manière générale, les échantillonnages aléatoires sont préférables, car la perturbation sur le signal est minimum, et dispersée au travers de toute la largeur spectrale.

3. Exemples de traitement par MaxEnt

3.1. Données simulées

La Fig. 3 présente des exemples de traitement par MaxEnt de données simulées en utilisant des paramètres de tailles et de largeurs spectrales utilisées pour une expérience 3D-HNCA. L'axe F1 simule l'axe ^{15}N et F2 simule le ^{13}C. Trois raies d'égales intensités et de largeur 10 Hz sur chaque axe ont été simulées, et un bruit gaussien égal à 30 % du signal temporel a été rajouté. Ces mêmes données ont été utilisées pour explorer les différents traitement de données possibles. Tous les spectres présentés ont été calculés avec au final 128 × 128 points dans le spectre, et toutes les figures sont présentées avec exactement les mêmes paramètres graphiques. En particulier, le niveau le plus bas est tracé à 1/32^{e} de l'intensité du plus grand pic. Tous les calculs ont été réalisés avec un ordinateur Macintosh G4 à 1,5 GHz sous MacOs 10.3.8. Les traitements par MaxEnt sont arrêtés, soit quand le critère de χ^2 est

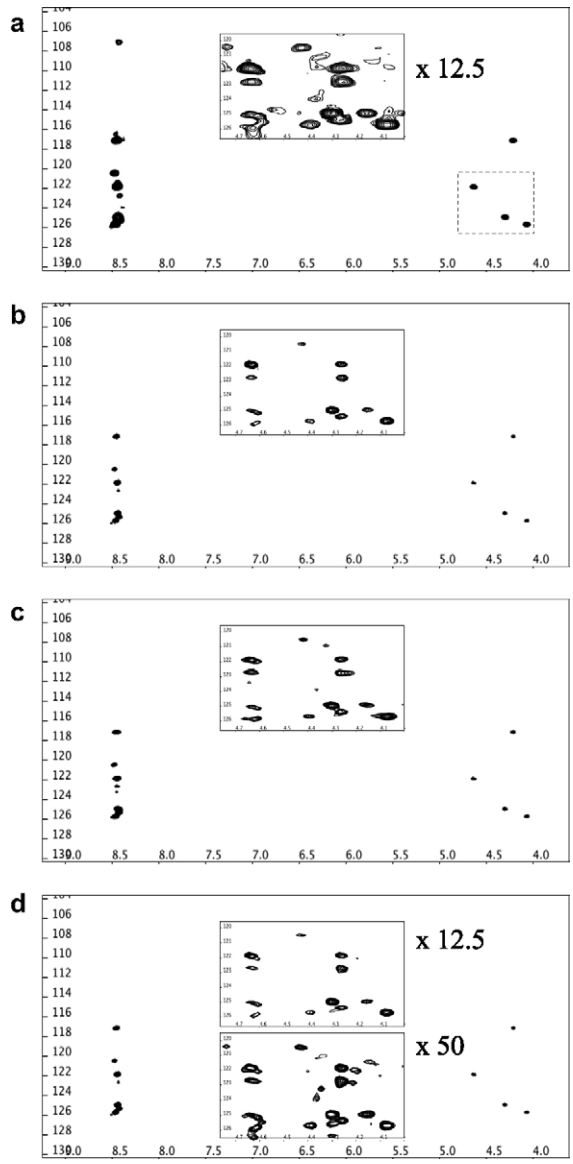

Fig. 4. Exemples de traitement par MaxEnt d'un plan ^{15}N–^{1}H issu d'un 3D 15 N-HSQC NOESY réalisée sur la protéine RpfBc de 108 résidus (données généreusement communiquées par C. Roumestand).
L'axe ^{15}N a été échantillonné sur 40 points complexes, l'axe ^{1}H a été échantillonné sur 190 points complexes. L'expérience complètement échantillonnée a duré 85 h.
(**a**) Traitement par FT, l'axe ^{15}N a été étendu à 80 points complexes par prédiction linéaire, le spectre final contient 256 × 512 points spectraux. La zone en encart a été multipliée par 12,5 en échelle verticale.
(**b**) Traitement par MaxEnt des mêmes données, avec une déconvolution de 5 Hz le long de l'axe de F2.
(**c**) Traitement par MaxEnt des mêmes données échantillonnées en utilisant le schéma présenté sur la Fig. 2a.
(**d**) Traitement par MaxEnt des mêmes données échantillonnées en utilisant le schéma présenté sur la Fig. 2b. Un deuxième encart présente la même zone multipliée par 50 en échelle verticale.

atteint, soit après 100 itérations en cas de stagnation de la convergence, pour un temps de calcul maximum de 5 secondes environ.

La Fig. 3a présente le spectre obtenu avec ces données par transformée de Fourier. On observe la largeur et la forme des raies dominées par l'effet de troncature de la mesure temporelle, qui est ici très sévère. De plus, on observe très facilement le niveau de bruit injecté dans les données simulées. La Fig. 3b présente les même données calculées par MaxEnt. On observe un fort affinement des signaux et une réjection du bruit. L'affinement est dû à l'approche inverse : en effet, dans cette approche, l'effet de la troncature du FID est très largement minimisé, et l'on s'approche de la largeur intrinsèque de la raie simulée. La réjection du bruit est due principalement à l'approche statistique de l'incertitude sur les données, permise par le calcul du χ^2.

La Fig. 3c présente le calcul par MaxEnt réalisé sur des données où seulement 25 % des points du plan $t1 \times t2$ ont été échantillonnés, avec un protocole équivalent à celui présenté sur la Fig. 2b. Dans ce cas, l'échantillonnage a été fait avec une statistique constante et non pas exponentielle, car ici l'amortissement dû à la relaxation est quasiment insensible. On observe une légère apparition d'un bruit de fond, dû d'une part à la diminution de la quantité de signal mesuré et d'autre part à la convolution du bruit d'échantillonnage dans le signal. Enfin, la Fig. 3d présente les mêmes données échantillonnées avec le protocole présenté sur la Fig. 2c. Dans ce cas, on observe une montée du bruit assez importante, mais du même ordre de grandeur que celui observé dans le spectre de la Fig. 3a, et n'empêchant pas l'interprétation des données. Par exemple, sur la colonne extraite, on voit que les artéfacts les plus intenses sont de l'ordre de 10 % du pic le plus intense. Ces artéfacts viennent principalement de la convolution des données par la structure de l'échantillonnage et peuvent être assimilés aux artéfacts que l'on observerait si on traitait les mêmes données par l'algorithme de projection–reconstruction proposé par Kupče et Freeman [19].

3.2. Données réelles

La Fig. 4 présente un traitement de données réalisé sur des données réelles. Est présenté ici un plan F1 × F2 extrait d'une expérience de 3D-HSQC-NOESY mesurée sur la protéine RpfBc de 108 résidus [20]. L'expérience a été réalisée en 85 h, l'axe ^{15}N (F1) étant échantillonné sur 40 points complexes, l'axe ^{1}H (F2) étant échantillonné sur 190 points complexes, avec huit scans par incrément. Cette expérience présente des difficultés supplémentaires en termes d'analyse par rapport aux données présentées sur la Fig. 3. En premier lieu, la protéine étudiée présente une zone de plus de dix résidus non structurés en solution, les résidus dans cette zone présentent des signaux intenses et fins, très différents des signaux présentés par les résidus structurés, qui sont moins intenses et nettement plus larges. Cette répartition de signaux différents pose une difficulté supplémentaire, du fait de la grand dynamique présente dans une telle expérience. En effet, le ratio entre l'intensité du signal diagonal le plus intense et celle du signal NOESY présenté par un des résidus structurés est ici supérieur à 100. Ensuite, ces mêmes signaux intenses et fins sont la source d'un bruit $t1$ d'autant plus marqué que l'accumulation a été très longue (près de quatre jours). Les données présentées correspondent à un plan extrait à $F3 = 8{,}5$ ppm, qui présente une forte superposition de signaux intenses issus des zones non structurées de la protéine et de signaux NOE issus des zones structurées.

La Fig. 4a présente ce plan traité de manière optimum par prédiction linéaire et transformée de Fourier, le plan étant calculé sur 256 × 512 points spectraux, ainsi que dans tous les calculs suivants. Les raies intenses vers $F2 = 8{,}5$ ppm correspondent aux signaux diagonaux issus de la zone non structurée de la protéine. L'encart présente un zoom de la zone vers $F1 = 123$ ppm × $F2 = 4$ ppm, cette zone contenant les informations pertinentes pour la détermination de la structure de la protéine étudiée. L'échelle verticale ayant été multipliée par 12,5 dans cet encart, le ratio entre la raie la plus intense et le premier niveau tracé dans cet encart est de 400. On remarque sur cette figure le fort niveau de bruit $t1$ présent vers 8,5 ppm.

La Fig. 4b présente les mêmes données traitées par MaxEnt. Dans ce calcul ainsi que dans les suivants, nous avons appliqué une déconvolution de 5 Hz le long de l'axe F2, le nombre d'itérations ayant été limité à 200, pour un temps de calcul de 85 s. On observe dans cette figure que la forme de raie a été optimisée, du fait de la déconvolution en F2 et de l'absence de sensibilité à la troncature en F1. On vérifie aussi que les signaux les plus faibles sont conservés. On observe que les intensités relatives des différents signaux sont légèrement modifiées dans ce spectre. La première raison en est la déconvolution, qui, en affinant les raies les plus larges, modifie les intensités relatives, mais en gardant l'intégrale du pic relativement constante. Une deuxième raison en est l'algorithme de MaxEnt lui-même, qui a tendance à minimiser les signaux les moins intenses. Bien que présent ici, ce phénomène est relativement peu important, comme on peut le vérifier dans la zone en encart. La Fig. 4c présente les mêmes données, où nous avons simulé l'effet d'un échantillonnage partiel réalisé avec le protocole de la Fig. 2a. La Fig. 4d présente les mêmes données, dans lesquelles l'échantillonnage présenté à la Fig. 2b a été simulé. Cet échantillonnage présente l'avantage de répartir plus largement le bruit d'échantillonnage, et donc de moins distordre les données ; il présente en revanche l'inconvénient d'être le plus difficile à implanter sur le spectromètre. Dans les deux cas (Fig. 4c et d), la mesure de ces quantités n'aurait duré que 22 h. On voit que les signaux sont relativement peu modifiés, ce qui indique bien que le temps d'acquisition était nécessité par la résolution plus que par le rapport signal sur bruit. Un examen attentif des spectres ainsi obtenus permet de vérifier que les signaux faibles, présents dans le spectre initial (Fig. 4a) sont retrouvés dans les spectres des Figs. 4c et d, mais que les intensités ne sont pas complètement conservées. Dans tous les cas, les spectres présentent un niveau d'artefacts assez intenses, dus principalement au bruit $t1$ généré par les signaux intenses de la zone déstructurée. Il faut remarquer, à ce titre, qu'il s'agit ici de simulation, alors que, dans le cas réel, on peut s'attendre à une diminution des artéfacts de bruit $t1$ dans les spectres échantillonnés partiellement, du fait de la diminution du temps total d'expérience.

4. Discussion

4.1. Le traitement par MaxEnt

Nous avons vu que le traitement par MaxEnt est une approche très générale. La même méthode, et le même

interface utilisateur, peut être utilisé pour traiter des situations très différentes. Cependant cette technique, comme toute technique, possède des forces et des faiblesses que nous allons passer en revue.

C'est une approche inverse, dans laquelle le bruit est statistiquement pris en charge. Cela assure une grande immunité face au bruit dans les données et aux problèmes de convolution des données par des causes extérieures telles que troncature, élargissement et, bien sûr, échantillonnage. C'est ce dernier aspect qui permet d'obtenir les résultats présentés ici. Nous avons ainsi montré que, suivant les expériences, et suivant les échantillonnages partiels utilisés, les gains en temps de mesure peuvent aller jusqu'à des facteurs de 4 à 10. Ces chiffres sont en accord avec les différentes approches déjà publiées. De plus, le principe consiste à minimiser la quantité d'information présentée dans les spectres ; à ce titre, cette approche permet des gains plus élevés dans le cas de spectres pauvres en information, c'est-à-dire contenant peu de raies, et surtout une dynamique faible. C'est ce qui explique les gains qui peuvent être obtenus sur des expériences de type HNCA, qui répondent mieux à ces critères que les expériences de type de HSQC-NOESY.

La première difficulté est la complexité de l'algorithmique. Celle-ci impose des temps de calcul importants, mais cet aspect est minimisé par la rapidité obtenue par les ordinateurs aujourd'hui. Par exemple, avec les paramètres présentés ici, le traitement complet par MaxEnt de la 3D HSQC-NOESY avec échantillonnage aléatoire a pris 1 h 30 min, à comparer au temps de calcul par FT et prédiction linéaire, qui a pris de 4 à 20 min sur le même système, selon le protocole utilisé.

En outre, du fait de la forte non-linéarité du problème d'optimisation, l'algorithme est sensible à l'accumulation des erreurs d'arrondi, inhérentes aux calculs informatiques. Cela pose des difficultés pour l'analyse de données de très grandes tailles. C'est ce qui explique que le traitement ne soit pas réalisé en MaxEnt pour la 3D entière, ce qui serait concevable mais impraticable, mais qu'il soit réalisé plan par plan, après un traitement classique le long de l'axe $F3$. Pour finir, un des problèmes de l'analyse par MaxEnt est la conservation des signaux les moins intenses face à des signaux très intenses. Nous avons pu vérifier que les signaux les plus faibles sont parfaitement conservés, même si la dynamique globale n'est pas complètement conservée.

4.2. Implantation dans Gifa

Il est clair qu'une implémentation efficace et facile d'emploi de MaxEnt est indispensable pour pouvoir être utilisée d'une manière régulière au laboratoire. Le logiciel Gifa [21] est disponible depuis longtemps, et contient depuis la première heure la possibilité de traitement par entropie maximale [4] ainsi que l'analyse de données avec un échantillonnage partiel [11]. Une nouvelle version du logiciel Gifa a été développée dans le but dans facilité l'utilisation. Outre, les fonctionnalités augmentées, et la vitesse de traitement optimisée, la principale nouveauté de cette version consiste en une modélisation très fine de ce qu'est un traitement de données en RMN. Cette modélisation permet de complètement séparer l'action à effectuer et le paramétrage de cette action. Il en résulte une interface très facile d'emploi pour l'utilisateur, où l'ensemble des paramètres ne se déploient pas dans une combinatoire infinie, car la modélisation permet de cadrer les différents scénarios ayant une certaine utilité, et d'éliminer tous les autres. Cette nouvelle version implémente tous les nouveaux protocoles d'acquisition partielle, et permet de facilement les mettre en œuvre.

Le travail effectué lors de la création de la base de données binaire RMN NMRb [22] nous a fourni une modélisation fine de l'ensemble des données utilisées (FID, spectres, traitements, pics, intégrales...). Cette modélisation nous permet aujourd'hui de produire un logiciel possédant une architecture très structurée et modulaire. Cette architecture est composée de plusieurs couches clairement définies, chacune de ces couches étant dédiée à une fonction précise.

Nous trouvons ainsi trois couches distinctes, chacune s'appuyant sur les outils définis dans les couches précédentes :

- un noyau mathématique implantant les fonctions telles que FFT, MaxEnt, transformée de Hilbert, transformée de Laplace, etc ;
- un ensemble de boîtes à outils dédiées au traitement du signal et à la RMN :
  - apodisation ;
  - causalisation des données ;
  - correction de ligne de base, etc.
- une interface de haut niveau permettant de définir de manière complète tout les calculs à réaliser.

Pour finir, l'interface graphique n'est plus prise en charge par ce logiciel, mais de nombreuses facilités sont

introduites, permettant de réaliser facilement l'interface graphique adaptée à son usage propre.

Cette répartition stricte des fonctions permet de développer chacune de ces parties le plus efficacement possible. En particulier, chaque couche est développée dans le langage le plus adapté au problème (respectivement : Fortran pour le noyau, un langage de macro spécifiquement dédié au traitement du signal pour les boîtes à outils ; XML et Java pour la définition des traitements à réaliser). De plus, le développement étant réalisé en équipe, chaque partie est gérée par une personne spécialiste de ce domaine.

De plus lors du développement de Gifa, nous avons voulu éviter le phénomène de « boîte noire » propre aux logiciels commerciaux ; en effet, tous les paramètres de traitement, les données produites, et les macro de traitement sont stockés dans des formats ouverts (XML, textes...). Le format des données s'appuie sur la norme *Open-Document* [23], qui normalise le stockage de données hiérarchiques telles qu'on les trouve en RMN, la description des contenus, les liens entre les différents éléments, ainsi que la compression des données. De plus un gros effort de documentation du code a été réalisé.

Comme il est dit ci-dessus, l'interface graphique n'est plus prise en charge par Gifa, mais celui-ci présente une interface extrêmement simple pour être interfacée par le logiciel utilisateur de son choix. Ainsi, nous sommes en train de développer une interface de type web permettant de piloter Gifa à distance. Cette possibilité va permettre d'ajouter des capacités de traitement au site NMRb [22], mais aussi permettre, par exemple, de mettre en place les calculs lourds de type MaxEnt sur un cluster dédié pour le calcul.

Pour finir, le logiciel commercial NMRnotebook utilise cette fonctionnalité de Gifa, et propose une interface graphique confortable, écrite en java et OpenGL, native sur toutes les architectures informatiques actuelles.

5. Conclusion

Nous avons montré comment le traitement de données par entropie maximale permet d'analyser des données mesurées avec un échantillonnage partiel. Et cela, qu'il s'agisse d'échantillonnages partiels aléatoires du plan $t1 \times t2$ aussi bien que de parcours favorisant certains plans particuliers. Il est possible, par cette approche, non seulement de traiter des données de type RMN 2D et 3D, mais aussi d'optimiser les temps d'acquisition, en permettant, dans le cas où la sensibilité est suffisante, de gagner systématiquement près d'un facteur 4 dans les temps d'acquisition à résolution constante. De manière alternative, ce gain permet d'optimiser résolution et sensibilité pour un temps de mesure donné.

Références

[1] F. Ni, G.C. Levy, H.A. Scheraga, J. Magn. Reson. 66 (1986) 385.
[2] E.D. Laue, M.R. Mayger, J. Skilling, J. Staunton, J. Magn. Reson. 68 (1986) 14.
[3] E.D. Laue, J. Skilling, J. Staunton, S. Sibisi, R.G. Brereton, J. Magn. Reson. 62 (1985) 437.
[4] M.-A. Delsuc In, in: J. Skilling (Ed.), Maximum Entropy and Bayesian Methods, Kluwer Academic, Dordrecht, 1989, pp. 285.
[5] J.A. Jones, P.J. Hore, J. Magn. Reson. 92 (1991) 276.
[6] M.A. Delsuc, G.C. Levy, J. Magn. Reson. 76 (1988) 306.
[7] J.A. Jones, D.S. Grainger, P.J. Hore, G.J. Daniell, J. Magn. Reson, Ser A 101 (1993) 162.
[8] V. Stoven, J.-P. Annereau, M.-A. Delsuc, J.-Y. Lallemand, J. Chem. Inf. Comput. Sci. 37 (1997) 265.
[9] M.-A. Delsuc, T.E. Malliavin, Anal. Chem. 70 (1998) 2146.
[10] J.C.J. Barna, E.D. Laue, M.R. Mayger, J. Skilling, S.J.P. Worral, J. Magn. Reson. 73 (1987) 69.
[11] M. Robin, M.A. Delsuc, E. Guittet, J.-Y. Lallemand, J. Magn. Reson. 92 (1991) 645.
[12] P. Schmieder, A.S. Stern, G. Wagner, J.C. Hoch, J. Biomol. NMR 3 (1993) 569.
[13] D. Rovnyak, D.P. Frueh, M. Sastry, Z.-Y.J. Sun, A.S. Stern, J.C. Hoch, G.J. Wagner, Magn. Reson. 170 (2004) 15.
[14] R.N. Bracewell, The Fourier Transform and its Applications, McGraw-Hill, New York, 1986 (2e éd.).
[15] C. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois, Urbana, USA, 1964.
[16] T. Gostand, thèse, université de Montpellier, 2004.
[17] (a) J.-P. Simorre, B. Brutscher, M.S. Caffrey, D. Marion, J. Biomol. NMR 4 (1994) 325; B. Brutscher, J.-P. Simorre, M.S. Caffrey, D. Marion, J. Magn. Reson. B. 105 (1994) 77.
[18] (a) K. Ding, A. Gronenborn, J. Magn. Reson. 156 (2002) 262; (b) S. Kim, T. Szyperski, J. Am. Chem. Soc. 125 (2003) 1385; (c) W. Kozminski, I. Zhukov, J. Biomol. NMR 26 (2003) 157.
[19] (a) E. Kupče, R. Freeman, J. Biomol. NMR 27 (2003) 383; (b) E. Kupče, R, Freeman, J Magn. Reson. 2 (2005) 317.
[20] M. Cohen-Gonsaud, P. Barthe, F. Pommier, R. Harris, P.C. Driscoll, N.H. Keep, C. Roumestand, J. Biomol. NMR 30 (2004) 373.
[21] J.-L. Pons, T.E. Malliavin, M.-A. Delsuc, J. Biomol. NMR 8 (1996) 445.
[22] J.-L. Pons, T.E. Malliavin, D. Tramesel, M.-A. Delsuc, Bioinformatics 20 (2004) 3707.
[23] http://www.oasis-open.org, 2003.