Quantification moléculaire par spectrométrie de masse à base de NEMS : modélisation et inversion du problème.

Rémi PÉRENON[1], Ali MOHAMMAD-DJAFARI[2], Laurent DURAFFOURG[1], Pierre GRANGEAT[1],

[1]CEA Leti,
MINATEC Campus, 17 Rue des Martyrs, F38054 Grenoble Cedex 9, France

[2]Laboratoire des Signaux et Systèmes, UMR 8506 (CNRS - Supélec - Université ParisSud)
Supélec, Plateau de Moulon, 3 Rue Joliot Curie, 91192 Gif-sur-Yvette, France

remi.perenon@cea.fr, djafari@lss.supelec.fr
laurent.duraffourg@cea.fr, pierre.grangeat@cea.fr

Résumé – Cet article présente une méthode de traitement adaptée à la spectrométrie de masse à base de NEMS (Nano Electro Mechanical Systems). Les NEMS sont des nano-capteurs suffisamment sensibles pour détecter la masse d'une molécule unique. Il est ainsi possible d'effectuer un traitement en mode comptage afin d'estimer le spectre de la solution analysée par le spectromètre de masse à partir de son contenu moléculaire. Dans cet article, nous proposons une modélisation de la chaîne d'acquisition et nous effectuons ce traitement en nous basant sur des méthodes MCMC (Monte-Carlo Markov Chain) pour résoudre le problème inverse. Des premiers résultats sur données simulées sont présentés.

Abstract – This article exposes a processing method adapted to NEMS (Nano Electro Mechanical Systems) Mass-Spectrometry. NEMS are nanometric sensors which are sensitive enough to detect the mass of a single molecule. Thus, it is possible to perform a counting mode processing in order to estimate the mass spectrum of the analyzed solution. In this article, we introduce a model of the acquisition chain and we describe the associated processing based on MCMC (Monte-Carlo Markov Chain) methods to solve the inverse problem. Preliminary results on simulated data are presented.

- $\underline{x}$ désigne un vecteur temporel discrétisé ;
- x_i désigne la i-ème composante du vecteur $\underline{x}$;
- $*$ désigne la convolution ;
- I_N désigne la matrice identité dans un espace de dimension N ;
- $\mathcal{N}\left(\underline{x}, \underline{\mu}, \Sigma\right)$, $\Gamma\left(x, k, \theta\right)$ et $B\left(x, a, b\right)$ désignent respectivement les lois normale, Gamma et Bêta relatives au vecteur $\underline{x}$ ou à la variable x.

1 Introduction

La protéomique (l'étude des protéines), est une science présente dans de nombreux domaines de la médecine, comme la recherche contre le cancer par exemple. Les protéines sont des cas particulièrement intéressants de biomarqueurs. En effet, ces dernières décrivent à la fois l'expression des gènes et l'influence de l'environnement, et sont de plus accessibles dans un bon nombre de fluides organiques (sang, sérum, urine ...).

Un outil d'analyse particulièrement privilégié pour la protéomique est la spectrométrie de masse [1]. Celle-ci a pour objectif de fournir le spectre de masse d'une solution, puis d'identifier et de quantifier les protéines de la solution à partir de ce spectre. Les spectromètres de masse actuels comportent généralement plusieurs fonctions, à savoir une source d'ions (source MALDI ou électrospray, par exemple) permettant de vaporiser et d'ioniser les molécules et un analyseur de spectre de masse à proprement parler.

Dans ce document, nous nous intéressons à l'analyseur de spectre. Typiquement, cet analyseur effectue une étape d'analyse où les molécules ionisées sont séparées en fonction de leur masse (*via* un système de temps de vol ou de trappe à ions, par exemple), puis une étape de détection où le flux d'ions est transformé en un flux d'électrons. Ce flux d'électrons est analysé et traité afin de reconstituer le profil moléculaire de la solution. Ces dernières années, de nombreux efforts ont été faits afin de concevoir des systèmes de plus en plus performants et des méthodes de traitement de plus en plus précises.

Parallèlement à ces développements, les nanotechnologies sont en plein essor, en particulier les systèmes nanométriques électromécaniques, ou NEMS [2] [3] [4]. Les NEMS sont de très petits capteurs sur lesquels s'adsorbent les molécules et qui, de par leur petite taille, sont sensibles à l'ajout de très faibles masses pouvant aller jusqu'à la masse de molécules uniques (100 Dalton). Ces développements conduisent à une idée nouvelle : l'utilisation de NEMS afin d'effectuer une tâche d'analyseur de spectre de masse en mode comptage.

Dans le travail proposé ici, nous nous inspirons du schéma expérimental proposé par Naik et al. [2]. Nous proposons une modélisation statistique du système ainsi qu'une méthode d'inversion bayésienne [5] afin de pouvoir retrouver la quantité de molécules adsorbées ainsi que leurs masses. Cette méthode d'inversion sera testée sur des données simulées.

2 Modèle direct

Naik et al. [2] proposent un système constitué d'un électrospray et d'une électronique de focalisation (un hexapôle) permettant d'ioniser la matière et de guider les ions jusqu'à un détecteur NEMS comprenant l'électronique de lecture. Les capteurs utilisés ici sont des nano structures mécaniques en forme de poutre. Ces capteurs se comportent mécaniquement comme un système masse-ressort, ce qui implique que leur fréquence de résonance varie avec la masse effective de la poutre. Celle-ci est bombardée par les molécules ionisées qu'elle adsorbe séquentiellement. Chaque adsorption d'une molécule entraîne une chute de la fréquence de résonance de la poutre par effet gravimétrique pouvant être reliée à la masse de la molécule. Il est donc possible d'effectuer une détection et une estimation de la masse au niveau de la molécule unique et de combiner les mesures élémentaires associées à chaque événement pour estimer le spectre de masse des molécules incidentes. Le capteur est relié à une électronique de lecture qui a pour objectif de lire la fréquence de résonance de la poutre. Cette électronique est constituée d'un premier dispositif qui stimule la poutre et d'un deuxième qui lit l'amplitude du déplacement de celle-ci. Le tout est intégré dans un système en boucle fermée qui asservit la sortie du système à la fréquence de résonance du NEMS.

Nous considérons que l'électrospray et l'électronique de focalisation se comportent tous deux comme un gain indépendant de la molécule considérée, ce qui revient à dire que l'électronique de focalisation et l'électrospray n'induisent que des pertes non discriminatoires vis-à-vis du type de molécule. Cette hypothèse se justifie par le fait que toutes les molécules observées sont de même nature (dans le cas présent, toutes les molécules sont des protéines). Ceci nous permet dans cet article de nous pencher uniquement sur les molécules détectées par la poutre : nous supposerons qu'il suffirait de multiplier le nombre de molécules détectées au niveau de la poutre par une constante pour estimer le nombre de molécules de la solution.

Pour modéliser le comportement de la poutre, nous considérons que cette dernière est soumise à un train temporel d'impulsions modulées en amplitude $m(t) = \sum_{i=1}^{I} m_i \delta(t_i)$. L'adsorption d'une molécule au temps t_i est décrite par une distribution de Dirac $\delta(t_i)$, m_i est la masse de cette molécule et I est le nombre d'adsorbtions pendant le temps d'observation.

Dans cette logique, chaque molécule bombardée sur la poutre fait varier la fréquence de résonance de cette dernière. La chute en fréquence de résonance qui en découle est supposée immédiate et parfaite, ce qui signifie que l'électronique en aval du dispositif présente des temps de réponse plus longs que celui de la poutre. L'amplitude de cette chute de fréquence va dépendre à la fois de la masse m_i adsorbée et du « gain » de la poutre. Ce gain est fonction d'une part des propriétés intrinsèques de la poutre et d'autre part du lieu z_i où la molécule s'est adsorbée sur la poutre. En effet, la fréquence de résonance dépend de la masse effective de la poutre, laquelle dépend non pas uniquement de la masse totale mais aussi de la distribution de masse le long de la poutre. Ainsi, lors de l'adsorption d'une molécule de masse m_i en z_i, la fréquence de résonance chute de $c(m_i, z_i)$.

L'électronique de lecture, quant à elle, vient lire la fréquence de résonance de la poutre en fonction du temps. On suppose que l'électronique se comporte comme un filtre linéaire invariant dans le temps. La juxtaposition de la poutre et de l'électronique de lecture conduit à l'expression de $g(t)$, la sortie non bruitée du système :

$$g(t) = g_0 + \sum_{i=1}^{I} c(m_i, z_i) h(t - t_i) \tag{1}$$

où g_0 est la valeur de $g(t)$ en $t = 0$ et $h(t)$ est la réponse impulsionnelle (infinie) du système de lecture. Cela signifie que le filtre $h(t)$ transforme une chute de fréquence en un signal de sortie de la boucle de lecture.

Par la suite, afin de modéliser le problème sous l'angle d'un problème statistique, nous faisons un certain nombre d'hypothèses simplificatrices. Nous discrétisons le problème et observons le signal sur un nombre fini d'échantilons T.

Les molécules sont envoyées sur la poutre par un processus sans mémoire. À chaque instant, l'adsorption d'une molécule est un processus de Bernoulli de paramètre π (inconnu). De ce fait, on introduit $\underline{q}$, une variable de Bernoulli symbolisant la présence d'un événement à chaque instant.

Dans un premier temps, nous considérons qu'un seul type de molécule arrive dans le système : ces molécules sont de masse m_0 (inconnue). De plus, pour cette première inversion, nous fixons $\underline{z} = z_0\, \underline{q}$, z_0 étant connu, ce qui revient à supposer que la poutre a été traitée (fonctionnalisée) de sorte que l'adsorption des molécules ne se produise que sur une zone que l'on considère suffisamment étroite pour que la variabilité du lieu d'adsorption n'ait pas d'influence sur la mesure.

Le problème s'écrit donc sous l'angle de la déconvolution impulsionnelle en introduisant les quantités suivantes :

$$\underline{m} = m_0\, \underline{q}, \quad \underline{z} = z_0\, \underline{q}, \quad \underline{f} = \underline{c}(\underline{m}, \underline{z}), \quad \underline{g} = \underline{f} * \underline{h} \tag{2}$$

Nous modélisons le signal observé $\underline{y}$ comme un signal aléatoire gaussien de moyenne $\underline{g}$ et de matrice de covariance $\sigma^2 I_T$, la puissance du bruit σ^2 étant supposée connue. Avec ces définitions, nous pouvons modéliser le système de mesure d'une manière hiérarchique comme indiqué sur la figure 1, les inconnues recherchées étant en rouge, les autres inconnues en gris/noir, les grandeurs connues en vert et l'observation en bleu.

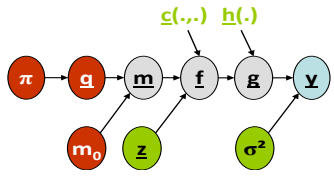

FIGURE 1 – Modélisation baysienne hiérarchique

3 Inversion

Afin d'inverser le problème, c'est-à-dire afin de retrouver la masse m_0 et le nombre d'événements I à partir des données $\underline{y}$, nous mettons en place une méthode d'inversion découlant d'une démarche bayésienne. L'estimateur de la solution retenu est celui de l'espérance *a posteriori* (EAP). Le calcul de l'EAP est réalisé en mettant en œuvre un échantillonneur de Gibbs et l'algorithme SMLR (Single Most Likely Replacement) [6].

Dans un premier temps, nous écrivons la loi jointe de l'ensemble des paramètres. Il est possible de marginaliser facilement certaines variables. C'est notamment le cas lorsque certaines variables présentent un lien déterministe. Ainsi, en utilisant (2) nous marginalisons les variables $\underline{\mathrm{m}}$, $\underline{\mathrm{f}}$ et $\underline{\mathrm{g}}$.

Il est ensuite possible d'expliciter la loi jointe pour les variables restantes, à savoir le signal observé $\underline{y}$, la masse des molécules m_0, le vecteur des événements $\underline{q}$ et le paramètre π.

$$p\left(\underline{y}, m_0, \underline{q}, \pi\right) \propto p\left(\underline{y}|m_0, \underline{q}\right) p\left(m_0\right) p\left(\underline{q}|\pi\right) p\left(\pi\right) \tag{3}$$

Nous allons désormais définir des *a priori* sur les paramètres recherchés : m_0 suit une loi Gamma de paramètres k_M et θ_M, $\underline{q}$ est une variable de Bernoulli de paramètre π, et π suit une loi Bêta de paramètres a_π et b_π.
Compte-tenu de ces *a priori*, la loi jointe (3) devient :

$$\begin{array}{c} \mathcal{N}\left(\underline{y}, \underline{c}\left(m_0\ \underline{q}, z_0\ \underline{q}\right) * \underline{h}\right) \Gamma\left(m_0, k_M, \theta_M\right) \\ \left(\pi^{\sum_{i=1}^{T} q_i} + (1-\pi)^{\sum_{i=1}^{T}(1-q_i)}\right) \beta\left(p, a_\pi, b_\pi\right) \end{array} \tag{4}$$

En écrivant $\mathcal{L}\left(\underline{y}\right)$ la quantité $\mathcal{N}\left(\underline{y}, \underline{c}\left(m_0\ \underline{q}, z_0\ \underline{q}\right) * \underline{h}, \sigma^2\right)$, les lois *a posteriori* conditionnelles pour chacun des paramètres s'expriment comme suit :

$$p\left(m_0|\underline{y}, \underline{q}, \pi\right) \propto \mathcal{L}\left(\underline{y}\right) \Gamma\left(m_0, k_M, \theta_M\right) \tag{5}$$

$$p\left(\underline{q}|\underline{y}, m_0, \pi\right) \propto \mathcal{L}\left(\underline{y}\right) \left(\pi^{\sum_{i=1}^{T} q_i} + (1-\pi)^{\sum_{i=1}^{T}(1-q_i)}\right) \tag{6}$$

$$p\left(\pi|\underline{y}, \underline{q}, m_0\right) \propto \beta\left(\pi, a_\pi + \sum_{i=1}^{T} q_i, b_\pi + \sum_{i=1}^{T}(1-q_i)\right) \tag{7}$$

Pour mettre en œuvre l'inversion, nous nous basons sur une méthode MCMC et l'échantillonneur de Gibbs. Chacun des paramètres inconnus m_0, $\underline{q}$ et π va être itérativement échantillonné selon sa loi *a posteriori* conditionnelle. Les valeurs numériques des paramètres inconnus utilisées pour calculer les lois *a posteriori* conditionnelles seront les valeurs courantes de ces paramètres. Au bout d'un certain temps (le « temps de chauffe »), le tirage d'un paramètre peut être considéré comme étant un tirage selon la « vraie » loi *a posteriori* du paramètre.

Avec ces valeurs échantillonnées selon la « vraie » loi, il est possible de moyenner les échantillons obtenus après le temps de chauffe et de disposer ainsi d'un estimateur EAP.

La loi *a posteriori* conditionnelle pour π (7) ayant la forme d'une loi bêta, π peut être aisément échantillonnée. Les lois *a posteriori* conditionnelles pour m_0 (5) et $\underline{q}$ (6), quant à elles, n'ont pas une forme permettant un échantillonnage simple, notamment en raison de l'expression de $\mathcal{L}\left(\underline{y}\right)$. Pour échantillonner m_0 et $\underline{q}$, nous proposons d'estimer la valeur de la loi *a posteriori* conditionnelle de m_0 (respectivement $\underline{q}$) dans un voisinage de la valeur courante. L'échantillon suivant est tiré selon la méthode de la fonction de répartition inverse.

Si la définition de la notion de voisinage est ici immédiate pour m_0, à savoir une gamme allant de $m_0 - V_M$ à $m_0 + V_M$ par pas de P_M, la notion de voisinage pour $\underline{\mathrm{q}}$ est plus complexe à définir. Idier et al. [6] proposent, dans leur description de l'algorithme SMLR, qu'un vecteur $\underline{\mathrm{q}}$ ait pour voisins les vecteurs auxquels un événement a été ajouté ou un événement a été retiré. Nous reprenons cette définition en ajoutant dans la liste des voisins d'un vecteur $\underline{\mathrm{q}}$ les vecteurs tels qu'un événement ait été déplacé et que deux événements aient été fusionnés. De plus, lorsqu'un nouveau vecteur $\underline{\mathrm{q}}$ du voisinage est proposé, l'algorithme peut choisir (avec une probabilité p_1), de proposer en même temps une modification de la masse de sorte que $m_0 \sum_{i=1}^{T} q_i$ soit inchangée.

Il s'avère qu'en l'état, l'algorithme ne converge pas vers la bonne solution. En effet, ce dernier peut rester bloqué dans des *optima* locaux. Ceci s'explique par le fait que les lois *a posteriori* conditionnelles sont très piquées, et présentent généralement un support trop étroit pour que les plages de valeurs proches de la « vraie » valeur soient explorées. Ainsi, le caractère « exploratoire » des méthodes MCMC est perdu. Il est donc nécessaire de modifier l'algorithme afin que ce dernier puisse explorer correctement l'espace des possibilités.

Dans notre cas, ceci s'effectue très simplement par une procédure dite de remise en cause. A chaque itération, il est possible, avec la probabilité p_2 de supprimer, d'ajouter un événement, de fusionner deux à deux les événements du vecteur $\underline{\mathrm{q}}$, ou encore de ré-échantillonner la masse avec la loi *a priori*. S'en suivent N_1 itérations (où aucune remise en cause n'est effectuée). Si la valeur de la loi jointe a été augmentée pendant ces N_1 itérations, les valeurs courantes sont conservées. Sinon, les valeurs antérieures à la remise en cause sont restaurées. Cette étape permet à l'algorithme d'éviter les problèmes numériques dus aux trop faibles valeurs de loi *a posteriori* (zones de l'espace jamais explorées), d'explorer bien plus rapidement l'espace des solutions possibles et de sortir des *optima* locaux.

Une fois les N_i itérations effectuées, m_0 (respectivement I) est estimée en moyennant les L_M valeurs de m_0 (respectivement les nombres d'événements des L_M valeurs de $\underline{q}$) qui ont permis de générer les valeurs les plus élevées de la loi jointe.

4 Résultats

Afin de mener à bien une simulation permettant de tester l'algorithme, nous mettons en place un générateur de signaux tel que décrit dans la section « Modèle direct ». En terme de simulation, nous prenons comme valeur numériques : pour T, 1000, pour $\underline{c}(\underline{m}, \underline{z})$, les valeurs données dans [3], pour z_0, 1, pour g_0, 0 et pour $\underline{h}$, $1 - e^{-\underline{t}/10}$. Avec ces valeurs, l'amplitude d'une chute δg sur le signal $\underline{g}$ induite par l'adsorption d'une

molécule de masse m_0 est d'environ $7.4\, m_0$.

Nous exécutons l'algorithme pour plusieurs valeurs de masse et de nombre de molécules. Nous prenons comme paramètres de l'algorithme d'inversion $N_i = 1000$, $V_M = 100$, $P_M = 1$, $k_M = 2$, $\theta_M = 300$, $a_\pi = 1$, $b_\pi = 10$, $p_1 = 0.5$, $N_1 = 25$, $p_2 = 0.95$ et $L_M = 100$. m_0 et π sont intialisées selon leur loi *a priori* et $\underline{q}$ est initialisé à un vecteur nul.

Le tableau 1 donne les performances de la méthode décrite ci-dessus pour différents niveaux de bruit sur le signal observé. La figure 2 illustre le signal à traiter et le signal reconstruit avec le meilleur jeu de paramètres trouvé par l'algorithme.

Jusqu'à un certain niveau de bruit, les paramètres sont très bien estimés. Cependant, pour un niveau de bruit trop élevé ($\sigma = 5000$, $\delta g \approx 6000$), l'algorithme ne parvient pas à converger et donne des solutions éloignées des vrais paramètres.

TABLE 1 – Erreur en valeur absolue des erreurs d'estimation de masse (Δ_M) et de nombre d'événements détectés (Δ_I)

Vraie valeur	$m_0 = 800$ $I = 10$ $\delta g \approx 6000$	$m_0 = 350$ $I = 10$ $\delta g \approx 2600$	$m_0 = 350$ $I = 15$ $\delta g \approx 2600$
$\sigma = 500$	$\Delta_M = 0.0$ $\Delta_I = 0$	$\Delta_M = 0.0$ $\Delta_I = 0$	$\Delta_M = 0.1$ $\Delta_I = 0$
$\sigma = 1000$	$\Delta_M = 0.4$ $\Delta_I = 0$	$\Delta_M = 1.1$ $\Delta_I = 0$	$\Delta_M = 1.6$ $\Delta_I = 0$
$\sigma = 2000$	$\Delta_M = 2.2$ $\Delta_I = 0$	$\Delta_M = 0.4$ $\Delta_I = 0$	$\Delta_M = 2.2$ $\Delta_I = 0$
$\sigma = 5000$	$\Delta_M = 0.6$ $\Delta_I = 0$	$\Delta_M = 36.1$ $\Delta_I = 1$	$\Delta_M = 18.2$ $\Delta_I = 1$

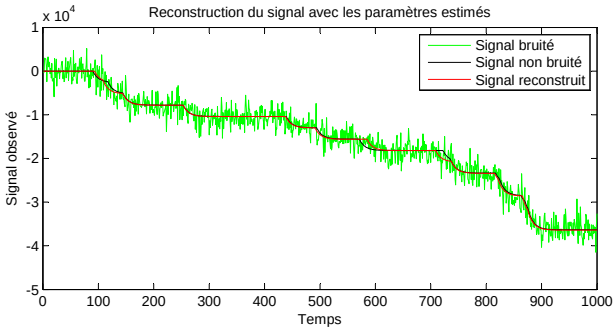

FIGURE 2 – Illustration de signaux non-bruité, bruité et reconstruit pour $m_0 = 350$, $I = 15$ et $\sigma = 2000$

5 Extension au cas multi-masse

Pour cette inversion, nous avons pris un certain nombre d'hypothèses fortes, notamment celle de la présence d'une seule valeur de masse dans le système. Afin de lever cette hypothèse, il est possible de ne plus considérer que les masses sont uniformes mais que les molécules appartiennent à des « classes », toutes les molécules d'une classe donnée ayant la même masse. Nous traduisons ces connaissances en changeant le modèle. Ainsi, nous remplaçons m_0 par c_M, vecteur donnant la masse de chaque classe, qui suit une loi Gamma pour chacune de ses composantes, et également π par c_π, vecteur donnant la probabilité de l'absence de molécule plus la probabilité d'occurrence de chaque classe, qui suit une loi de Dirichlet. Dans ce cas, $\underline{q}$ est une variable non plus binaire mais discrète prenant une valeur entière entre 0 et le nombre de classes, telle que $p(q_i = k) = c_{\pi\, k-1}$. La modélisation hierarchique devient celle de la figure 3. Une démarche similaire à celle présentée dans la section « Inversion » permet de résoudre le problème.

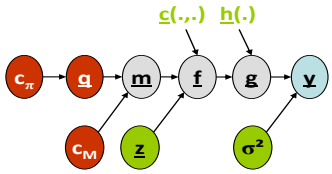

FIGURE 3 – Modélisation dans le cas multi-masse

6 Conclusion

Nous avons présenté dans cette communication une première modélisation statistique du problème de spectrométrie de masse à base de NEMS. En prenant un certain nombre d'hypothèses, nous sommes parvenus à inverser le problème et à estimer les paramètres moléculaires de masse et de nombre d'une espèce moléculaire donnée, en fonctionnant en mode comptage. Si les estimations semblent bonnes dans le cadre de cette étude, il reste néanmoins nécessaire de lever les deux hypothèses prises, à savoir celle de la masse unique, pour laquelle nous avons proposé une méthode de résolution, et celle de la variabilité de la position de la molécule sur le capteur.

Références

[1] R. Aebersold, M. Mann. *Mass spectrometry-based proteomics* Nature, 2003, 422, 198-207.

[2] A.K. Naik, M.S. Hanay, W.K. Hiebert, X.L. Feng, M.L. Roukes. *Towards single-molecule nanomechanical mass spectrometry* Nature Nanotechnology, 2009, 4, 445-450.

[3] S. Dohn, W. Svendsen, A. Boisen, O. Hansen. *Mass and position determination of attached particles on cantilever based mass sensors* Review of Scientific Instruments, 2007, 78

[4] E. Mile, G. Jourdan, I. Bargatin, S. Labarthe, C. Marcoux, P. Andreucci, S. Hentz, C. Kharrat, E. Colinet, L. Duraffourg. *In-plane nanoelectromechanical resonators based on silicon nanowire piezoresistive detection* Nanotechnology, 2010, 21(16)

[5] A. Mohammad-Djafari. *Bayesian inference for inverse problems in signal and image processing and applications* International Journal of Imaging Systems and Technology, 2006, 16, 209-214.

[6] J. Idier, Ed., *Approche bayésienne pour les problèmes inverses - Chapitre 5 Déconvolution impulsionnelle* Hermès Science Publications, Paris, nov. 2001.