# Questions d'ajustement

Hervé This

**Résumé** En recherche scientifique, les ajustements de courbes expérimentales sont permanents. On se propose, dans cet article, de montrer que, sans modèle théorique, des ajustements arbitraires sont une mauvaise, pratique, mais que des essais peuvent être justifiés par des considérations de « simplicité ». Deux mesures de cette dernière sont considérées : la complexité au sens de Kolmogorov et l'entropie de Shannon.

**Mots-clés** **Ajustement, simplicité, complexité, entropie, Kolmogorov, Shannon.**

**Abstract** **Fitting chemical data**

For scientific research, fitting experimental data is ubiquitous because the scientific activity is based on the following sequences of activities: (1) identification of a phenomenon; (2) quantitative characterization of the phenomenon; (3) grouping the data into synthetic laws; (4) looking for the mechanisms of the phenomenon, making a "model", a "theory"; (5) looking for testable consequences of the theory; (6) experimental tests in view of refuting the predicted consequences. Moving from step 2 to step 3 needs "fitting": when experimental data are acquired, grouping them in synthetic laws means traditionally looking for an analytical function going through the data. How can this work be done? The question being too general, it is proposed in this article to observe that without *a priori* theoretical model, arbitrary fitting is a poor practice (what many of us know), but some tests can be justified by questions of "simplicity". Two quantitative measurements of such a notion are discussed here: Kolmogorov complexity and Shannon entropy.

**Keywords** **Fitting, simplicity, complexity, entropy, Kolmogorov, Shannon.**

En recherche scientifique, les ajustements de courbes expérimentales sont permanents, car l'activité scientifique est fondée sur la séquence de travaux suivante : (1) identification d'un phénomène ; (2) caractérisation quantitative du phénomène ; (3) réunion des données en lois synthétiques ; (4) recherche de mécanismes, établissement d'un « modèle », d'une « théorie » ; (5) recherche de conséquences testables de la théorie ; (6) test expérimental de la théorie, en vue de l'améliorer [1].

Or le passage de l'étape 2 à l'étape 3 nécessite une opération d'« ajustement » : quand on a des points expérimentaux, la réunion en lois synthétiques revient classiquement à la recherche d'une fonction analytique qui passe par les points [2]. Comment effectuer ce travail ?

La question étant excessivement générale, on se propose ici de montrer que sans modèle théorique, des ajustements arbitraires sont une mauvaise pratique (ce que beaucoup d'entre nous savent), mais que des essais peuvent être justifiés par des considérations de « simplicité » [3]. Deux mesures de cette dernière seront considérées ici… arbitrairement : la complexité au sens de Kolmogorov et l'entropie de Shannon.

## Les ajustements s'imposent dans des circonstances variées

Partons d'un exemple : la production d'un bouillon de carotte par « cuisson de carottes dans de l'eau » (pour une publication, on dirait « traitement thermique de tissu racinaire de carottes, *Daucus carota* L., en solution aqueuse »). Avec le temps, l'eau se charge de divers solutés, qui sont essentiellement des saccharides (glucose, fructose, saccharose) et des acides aminés. Au début du traitement, la variation de la concentration en solutés est rapide, parce qu'il y a une forte différence de concentrations entre l'intérieur du tissu végétal et la solution (forte différence de potentiel chimique) ; puis, progressivement, le tissu s'appauvrit en composés qui vont se dissoudre dans la solution, de sorte que la concentration en ces composés augmente dans la solution, jusqu'à un « équilibre » (en réalité, il peut y avoir des complications, telle l'hydrolyse du saccharose, mais nous ne considérerons pas ici cette éventualité). Bref, si l'on mesure la concentration en l'un des solutés, on obtient le diagramme *concentration = f(temps)* de la *figure 1*.

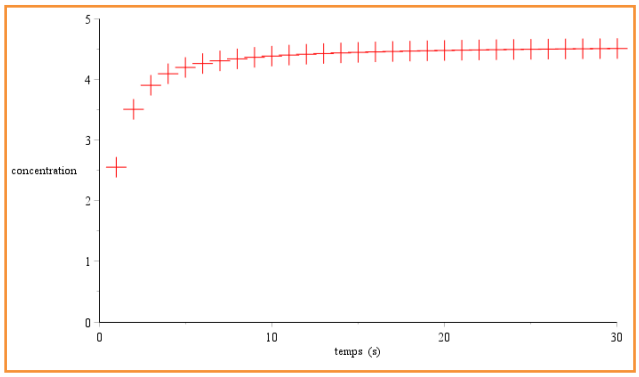


Figure 1 - Une courbe fréquemment observée quand on place un système en contact avec un environnement dont la composition diffère.

Un tel cas est fréquent en physico-chimie, et notamment en science des aliments ; on peut le rencontrer quand on place un système physico-chimique dans un nouvel environnement : lors de la mise en contact, la forte différence entre le système et son environnement est à l'origine de modifications rapides, mais ensuite, la différence se réduisant, le « moteur » de l'échange est ralenti [4].

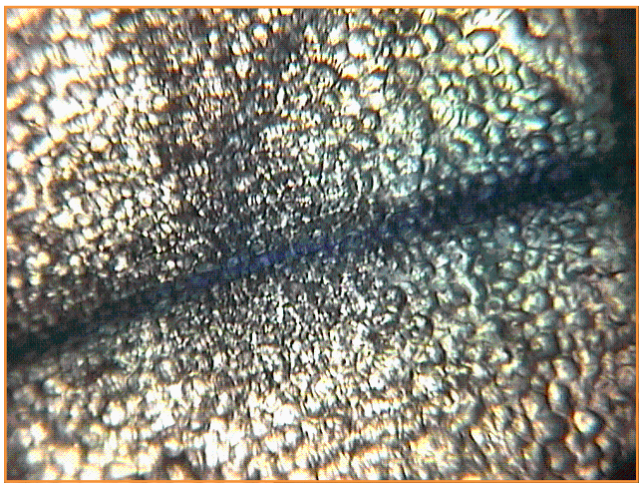

Figure 2 - Dans un tissu végétal, les tissus conducteurs (xylème, phloème) peuvent échanger des composés avec un environnement aqueux. Ici, on a placé un échantillon de bulbe d'oignon, *Allium cepa* L., dans une solution de bleu de méthylène, lequel a migré dans les canaux du tissu.

Quelle « loi » correspond à une telle répartition ? On dispose parfois d'un modèle théorique : par exemple, pour le traitement thermique de tissus végétaux en solution aqueuse, on peut initialement supposer que les composés sapides (saccharides, acides aminés) s'échangent à partir des tissus conducteurs que sont les « canaux » du xylème et du phloème, qui, respectivement, montent la sève brute des racines vers la partie aérienne de la plante et descendent la sève élaborée en sens inverse [5] (*figure 2*).

Dans un tel système, si les composés étaient échangés par simple diffusion, la résolution des équations de diffusion conduirait à un « modèle » (en pratique, il s'agit de lois analytiques, avec des paramètres inconnus qui dépendraient des particularités du système : coefficient de diffusion, particularités géométriques du système, caractéristiques physiques telle la viscosité, etc.), et l'on devrait alors « ajuster » la courbe aux points expérimentaux, afin d'identifier les paramètres inconnus du modèle.

Nous reviendrons plus loin sur la méthode à mettre en œuvre pour un tel ajustement, mais il faut d'abord observer que l'on ne dispose pas toujours d'un modèle physique au début d'une étude scientifique, de sorte que lorsque l'on a fait une expérience pour explorer un phénomène et que l'on a obtenu des points expérimentaux, on est tenté de tester un ajustement des points expérimentaux par des fonctions analytiques que l'on construit pour reproduire la forme dessinée par les points expérimentaux. Dans le cas du bouillon de carotte précédemment exposé, si l'on ignorait la possibilité d'une libération des composés par diffusion à partir des tissus conducteurs, on pourrait chercher une fonction qui aurait une tangente verticale à l'origine (libération rapide au début du traitement thermique) et une asymptote horizontale à l'infini (équilibre entre le tissu végétal et la solution).

Parfois aussi, les études font apparaître le besoin d'une interpolation qui ne soit pas linéaire, et il est commode d'obtenir un ajustement, que l'on utilise ensuite pour des calculs variés (*figure 3*).

## Comment éviter l'arbitraire ?

Dans le cas d'une absence de modèle comme d'une interpolation, on est mis en position de choisir une fonction d'ajustement sans avoir de modèle. Comment s'y prendre ? Quelle fonction analytique choisir ? De très nombreux ouvrages et articles ont été consacrés à cette question, mais de nombreux étudiants continuent de faire la faute qui consiste à élaborer arbitrairement des fonctions convenables, avec une pléthore de paramètres [6]. On gagne à enseigner que les opérations ne doivent pas être arbitraires, mais justifiées, d'une part, et même si le principe de simplicité (le « rasoir d'Ockham », du nom du moine Guillaume d'Ockham qui vivait au XIV$^{e}$ siècle) n'est pas une panacée (par exemple, l'orbite des planètes n'est pas un cercle mais une ellipse), il est au moins une justification raisonnable ; et Louis-Joseph Gay-Lussac montra que tous les acides ne contiennent pas de l'oxygène, comme Antoine-Laurent de Lavoisier l'avait trop rapidement conjecturé) [7].

Par exemple, dans le cas évoqué précédemment (voir *figure 1*), une combinaison de fonctions analytiques « simples » (cela signifie généralement « connues de l'utilisateur ») conduit à une fonction « raisonnable ». Ainsi, puisque l'évolution a une asymptote horizontale à des temps infinis, on peut considérer que la fonction est égale à la valeur de l'asymptote à laquelle on soustrait une fonction monotone décroissante et tendant vers 0. Quelle fonction ? Ayant suivi les enseignements de collège ou de lycée, on sait que la fonction exponentielle part de la valeur 1 pour x = 0, et tend vers + ∞ quand l'abscisse tend vers + ∞ ; la fonction exp(-x), de ce fait, part de l'ordonnée 1 en x = 0, mais tend vers 0 quand l'abscisse tend vers + ∞, de sorte que, si l'on soustrait cette fonction exp(-x) à une constante, on peut avoir, moyennant des normalisations élémentaires, un comportement qui pourrait décrire l'évolution expérimentale observée.

Autrement dit, on pourrait « raisonnablement » utiliser la fonction d'ajustement $f(x)=\left(a_1\cdot\left(1-\exp\left(-b_1\cdot x^{c_1}\right)\right)\right)$ pour décrire les données expérimentales ; ici, on introduit *a priori* les trois paramètres ajustables $a_1$, $b_1$, $c_1$, et l'emploi d'une commande d'ajustement (des logiciels de calcul formel tels que Maple en proposent plusieurs : Arrayinterpolation, BSpline, Polynomialinterpolation, LeastSquares...) conduit à déterminer ces paramètres ajustables, de sorte que l'on obtient finalement une courbe qui « passe très bien » par les points expérimentaux (*figure 4*).

Évidemment, l'ajustement est d'autant plus facile que les incertitudes de mesure sont notables. Toutefois, ce serait une erreur de débutant que d'utiliser un adjectif (« bien », « mal ») ou un adverbe (« très ») en science, et c'est un bon conseil que de recommander aux étudiants de remplacer

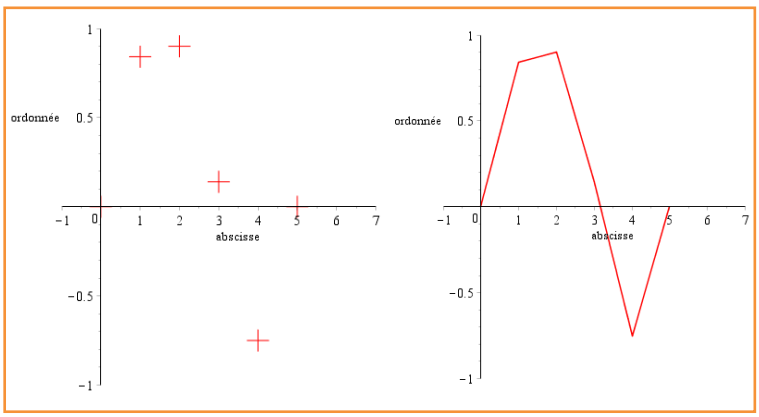


Figure 3 - Comment interpoler entre les points expérimentaux de gauche ? Manifestement l'interpolation linéaire est peu satisfaisante.

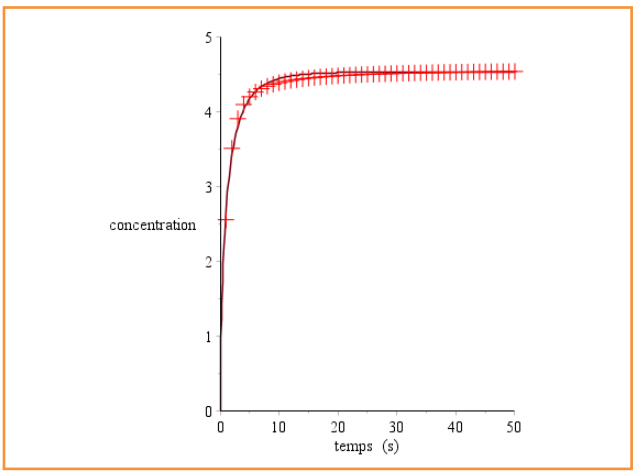


Figure 4 - Un ajustement possible des points expérimentaux de la *figure 1*.

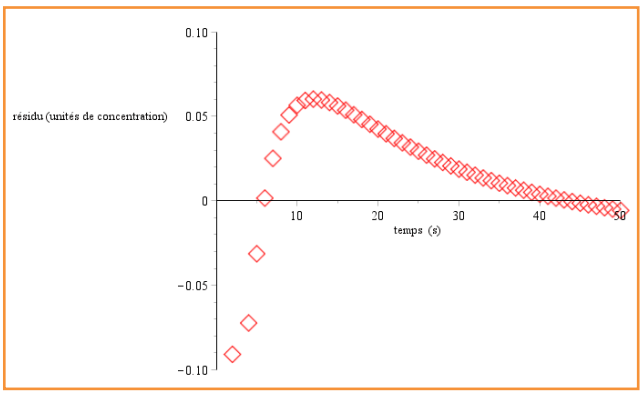


Figure 5 - Les résidus pour le premier ajustement.

ces mots par la réponse à la question « combien ? ». C'est la raison pour laquelle un ajustement ne se soutient qu'avec une estimation quantitative de la qualité de l'ajustement, telle la somme des carrés des différences entre l'ordonnée de la fonction d'ajustement et le point expérimental à la même abscisse, ou l'affichage des résidus. Par exemple, pour l'ajustement précédent, voir les résidus sur la *figure 5.*

Pour autant, la fonction choisie pour l'ajustement précédent était arbitraire, tout comme le nombre de paramètres (pris arbitrairement égal à 3 dans le cas précédent). Pourquoi ne pas choisir plutôt une fonction décroissante de façon monotone de type $\frac{1}{x}$, que l'on soustrairait encore à une valeur constante associée à l'asymptote ? Cette fois, il faut régler des questions de valeur à l'origine, mais on parvient très rapidement à imaginer la forme $a_2 \cdot \left(1 - \left(\frac{1}{1 + b_2 \cdot x^{c_2}}\right)\right)$, avec les trois paramètres ajustables $a_2, b_2, c_2$.

Dans ce nouveau cas, on obtient encore une courbe qui s'ajuste également « très bien », tout comme ce serait d'ailleurs le cas pour une fonction $a_3.\arctan(b_3.x + c_3)$, encore avec trois paramètres ajustables $a_3$, $b_3$, $c_3$, ou pour d'autres fonctions que nous pourrions nous amuser à construire.

Nous n'affichons pas ici les résultats des ajustements que nous venons de proposer, parce que les courbes sont quasi superposables à celle de la première fonction d'ajustement. Évidemment, on pourrait grossir des parties pour faire apparaître des différences... mais cela n'aurait pas de sens, car l'incertitude sur les mesures est bien supérieure, dans notre exemple, à la différence entre les courbes d'ajustement. Bien sûr, dans tous les cas, la somme des carrés des résidus est différente, mais si peu.

Bref, il manque une raison non arbitraire de choisir entre les différentes possibilités d'ajustement, entre les diverses formes analytiques à retenir pour effectuer ces ajustements.

## Le plus simple d'abord, à l'aide d'une mesure de la complexité

Selon le principe de parcimonie, nous aurions intérêt à choisir en priorité la fonction d'ajustement la plus simple, avant de passer éventuellement à des fonctions plus compliquées, au cas où cette fonction la plus simple ne conviendrait pas. Toutefois, « simple » est un adjectif, qui mérite donc d'être rendu quantitatif.

La « complexité de Kolmogorov » [8], du nom du mathématicien russe Andrei Kolmogorov (1903-1987), est un moyen d'y parvenir, puisque la complexité peut être considérée comme l'inverse de la simplicité.

Plus précisément, la complexité de Kolmogorov – d'une fonction en l'occurrence – est la taille du plus petit programme (les algorithmiciens parlent « plus proprement » de machines de Turing) avec lequel on peut définir la fonction.

Comment calculer cette complexité ? Il suffit d'utiliser un programme de compression sans pertes tels ceux qui équipent tous les ordinateurs (pour comprimer un fichier, avec des systèmes d'exploitation tels que Linux ou Windows, on se contente souvent de faire un clic droit, puis on coche la case « compresser »).

Commençons par examiner la notion de complexité de Kolmogorov sur un exemple plus simple qu'une fonction : une suite de lettres. On pressent qu'une série de dix lettres « a » est plus « complexe » qu'une seule lettre « a », et que la série de lettres « abcdefghij » est plus « complexe » que « aaaaaaaaaa ». C'est ce que confirme le calcul proposé pour l'estimation de la complexité de Kolmogorov : à l'aide de l'opération de compression de fichiers, un fichier texte qui contient un enchaînement de 1 000 000 lettres « a » se réduit à 1 148 octets (soit un facteur de compression de presque 1 000) ; un fichier de 100 000 répétitions de « abcdefghij » se réduit à 2 137 octets (la complexité est presque deux fois supérieure) ; un fichier de 10 000 répétitions d'une séquence aléatoire de cent caractères se ramène à 3 697 octets.

Cette méthode, qui utilise un algorithme de compression, donne une approximation « raisonnable » de la complexité de Kolmogorov, si l'on utilise des fichiers assez gros, avec des régularités assez apparentes [9].

Comment appliquer la complexité de Kolmogorov au choix de fonctions d'ajustement ? Une manière consiste à développer d'abord les fonctions en série avant d'appliquer l'algorithme de compression aux fichiers où sont inscrits les coefficients du développement. Par exemple, pour la fonction $1 - \frac{1}{1+x}$, le développement en $x = 0$ (jusqu'à une puissance 19 de x) est :

$$x - x^2 + x^3 - x^4 + x^5 - x^6 + x^7 - x^8 + x^9 - x^{10} + x^{11} - x^{12} + x^{13} - x^{14} + x^{15} - x^{16} + x^{17} - x^{18} + x^{19} + O(x^{20})$$

De sorte que l'on chercherait à déterminer par compression la séquence « 1, -1, 1, -1, 1, -1.....1 ». Et c'est ainsi,

par exemple, que l'on mesure que, par cette manière, la fonction 1/(1+x) est plus simple que la fonction sin(x).

Bien sûr, il demeure l'arbitraire d'avoir choisi un développement au voisinage d'un point plutôt qu'un autre développement (un développement en série entière, qui serait plus naturel, une développement asymptotique, un développement en fractions continues...), et bien sûr, il demeure l'arbitraire du nombre de termes utilisés, mais si nous voulons faire mieux, rien n'empêche d'accumuler les manières, de les comparer, de les explorer (par exemple, si l'on conserve le développement en x = 0, on pourrait étudier la variation de la complexité de Kolmogorov en fonction du nombre de termes du développement).

Et pour les fonctions examinées plus haut : comment sont-elles ordonnées ? Voici un exercice que l'on peut proposer à des étudiants, surtout s'ils utilisent des logiciels de calcul formel, qui déterminent les développements en série. Par exemple, avec le logiciel Maple, le développement de la fonction $1 - \frac{1}{x+1}$ s'obtient en une commande :

$$series\left(1 - \frac{1}{x+1}, x = 0\right)$$

Cette commande procure le résultat instantanément. D'ailleurs, on observera que, puisque la fonction en inverse ne peut être décrite en totalité (son rayon de convergence est limité), l'utilisation d'un développement en série entière doit être abandonnée au profit d'une autre méthode de description des fonctions.

## Une seconde méthode : le calcul de l'entropie de Shannon

L'entropie de Shannon, due au mathématicien américain Claude Shannon (1916-2001), est une autre possibilité [10]. Cette fonction, qui correspond à la quantité d'information contenue ou délivrée par une source d'informations, peut être définie par l'expression suivante :

$$H_b(X) = \sum_{i=}^{n} P_i \log_b\left(\frac{1}{P_i}\right)$$

Ici on considère une source, qui est une variable aléatoire discrète X comportant *n* symboles, chaque symbole $x_i$ ayant une probabilité $P_i$ d'apparaître. La fonction $H_b$ désigne l'entropie en base *b*, mais, souvent en informatique, *b* est égale à 2.

Une telle définition s'accorde avec l'idée thermodynamique de l'entropie. Par exemple, si la source n'émet qu'un signe (par ex. la lettre « a »), alors après *n* émissions, on obtient la chaîne de caractères « aa...a » : il n'y a pas d'incertitude sur le résultat émis. En revanche, si la source peut émettre 26 signes possibles, tirés au hasard, alors la chaîne devient bien plus incertaine.

Là encore, les progrès de l'informatique ont considérablement facilité le travail : ce calcul de l'entropie de Shannon est codé dans la plupart des logiciels de calcul formel, tel Maple. Par exemple, pour ce dernier, la détermination de l'entropie d'une chaîne de caractères s'obtient simplement par l'opération *Entropy(« chaine_de_caractères »)*. Et c'est ainsi que, pour une chaîne de 1 000 lettres « a », on calcule une entropie nulle. Pour une chaîne où l'on répète 100 fois « ab », l'entropie est calculée égale à 1 ; pour 100 fois « abc », on arrive à 1,58, et l'on passe à 2,32 pour 100 répétitions de « abcde ».

Comment utiliser l'entropie pour classer les fonctions d'ajustement par ordre de complexité croissante ? Là encore, on peut utiliser les développements en série des fonctions analytiques proposées... À nouveau, je laisse ici « en exercice » le soin de comparer les trois fonctions considérées en début d'article pour l'ajustement des points expérimentaux de la *figure 1*.

Et c'est ainsi que, à l'aide de l'entropie de Shannon ou de la complexité de Kolmogorov, on peut indiquer une procédure qui, bien qu'arbitraire elle-même, n'est pas injustifiée... Car c'est bien cela dont il s'agit : ne pas donner la première solution qui nous passe par la tête, et donner des justifications de nos méthodes. Avec les deux mesures de la complexité indiquées ci-dessus, nous pouvons choisir une « fonction plus simple selon une procédure particulière », quitte, si nous avons le temps, l'envie, le goût, à entrer plus dans les détails. Mais en tous cas, on aura conservé l'idée de mettre toujours le simple avant le compliqué : à défaut d'être justifié, le principe de parcimonie est ainsi un principe explicite, non arbitraire, que l'on peut exposer à nos examinateurs ou rapporteurs.

Finalement, cette question des ajustements arbitraires, résolue par l'emploi de paramètres quantitatifs de la complexité, est surtout un paradigme pour le travail scientifique. De même que les méthodes mises en œuvre doivent être validées, les calculs ne peuvent être arbitraires, et, plus généralement, nous devons être en mesure de justifier de nos pratiques à nos « rapporteurs » d'articles, tout comme à nous. Cela va du premier des gestes expérimentaux à notre stratégie scientifique, ou plutôt, dans l'ordre inverse, de notre stratégie à notre pratique. Quelles sont nos questions scientifiques, et pourquoi les mettons-nous en œuvre ? Quelles questions scientifiques étudions-nous et pourquoi ? Quelles techniques expérimentales utilisons-nous et pourquoi ? Quels calculs faisons-nous et pourquoi ? Ces questions sont évidemment terribles, mais la production de savoir mérite bien que nous nous interrogions, n'est-ce pas ?

## Références


[1] This H., *Cours de gastronomie moléculaire N° 1*, Belin, **2011**.
[2] Ducauze C., *Chimie analytique, analyse chimique et chimiométrie*, Lavoisier, Tec & Doc, **2014**.
[3] McCall J.J., Induction: from Kolmogorov and Solomonoff to De Finetti and back to Kolmogorov, *Metroeconomica*, **2004**, *55*, p. 195.
[4] Cazor A., Deborde C., Moing A., Rolin D., This H., Sucrose, glucose and fructose extraction in aqueous carrot root extracts prepared at different temperatures by means of direct NMR measurements, *J. Agric. Food Chem.*, **2006**, *54*, p. 4681.
[5] This H., Solutions are solutions, and gels are almost solutions, *Pure Appl. Chem.*, **2012**, *85*, p. 257.
[6] *Numerical Methods for Nonlinear Engineering Models*, J.R. Hauser, Springer Verlag, **2009**.
[7] Biard J., *Guillaume d'Ockham, Logique et Philosophie*, PUF, **1997**.
[8] Shen A., *Kolmogorov Complexity*, Uppsala University Lecture, **2007** ; Delahaye J.-P., Zenil H., On the Kolmogorov-Chaitin complexity for short sequences, in *Randomness and Complexity, From Chaitin To Leibniz*, C.S. Calude (ed.), **2007**, World Scientific, p. 123-130.
[9] Belabbes S., Richard G., On using SVM and Kolmogorov complexity for spam filtering, *Int. conference on artificial intelligence* (FLAIRS, mai **2008**), Miami (FL), AAAI Press, **2008**, p. 130.
[10] Shannon C.E., A mathematical theory of communication, *The Bell System Technical Journal*, **1948**, *27*, p. 379 et 623.



**Hervé This**
est directeur de l'AgroParisTech-Inra International Centre for Molecular Gastronomy*.

* International Centre for Molecular Gastronomy AgroParisTech-INRA, Groupe de Gastronomie moléculaire (Laboratoire de chimie analytique, UMR 1145 Ingénierie Procédés Aliment GENIAL), 16 rue Claude Bernard, F-75005 Paris.
Courriel : herve.this@agroparistech.fr