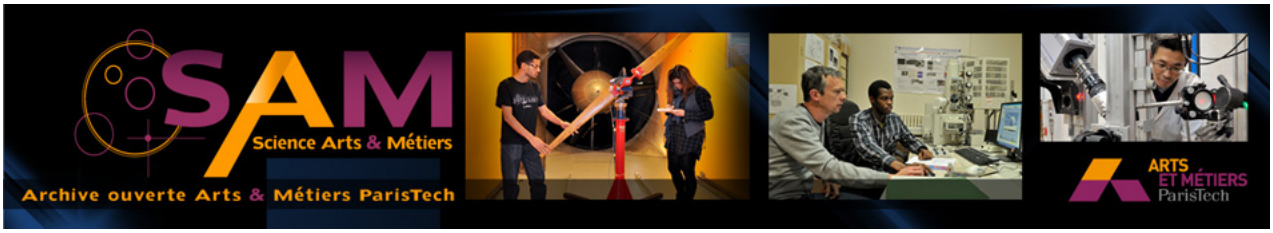


Relations entre l'entropie physique, le codage de l'information et l'énergie de simulation

M. Bigerelle et A. Iost

Résumé : Dans cette note, nous analysons l'entropie de mélange d'un système physique à l'équilibre par la théorie algorithmique de l'information. Nous montrons l'existence d'un isomorphisme entre cette entropie et la taille du programme informatique qui simule ce système physique. Cet isomorphisme doit être construit en respectant certaines règles, et le meilleur résultat est obtenu en combinant les algorithmes de compression RLE « run length encoding » et Huffman. Si le système physique est codé par la composition de ces deux isomorphismes basés sur le dénombrement des séquences identiques puis de leur codage en dictionnaire, nous pouvons quantifier l'entropie de systèmes binaires ou ternaires à l'état d'équilibre. De plus, il existe une relation affine entre l'énergie de simulation du système physique et l'énergie libre du système.

PACS N° : 89.70.+c

Abstract: It is shown that an isomorphism exists between the mixing entropy and the size of a computer program that simulates the physical system. This isomorphism must be constructed with respect to some theorems, and it is shown that the composition of two isomorphisms, one based on a run length encoding and another by encoding sequences in a dictionary allows us to quantify the entropy of binary and ternary systems at the equilibrium. Finally, it is shown that the energy consumed by a physical system encoded by this system and executed on a Turing machine is proportional to the free energy of the thermodynamic system.

1. Introduction

Depuis les années 1950, de nombreux travaux abordent le lien entre l'entropie physique et l'entropie informationnelle [1, 2]. Brillouin [3] fut l'un des pionniers dans l'étude de cette relation mentionnée pour la première fois par Shannon [4]. Shannon s'intéresse principalement au débit informatique d'un message exprimé par exemple en bits par seconde, qu'il appelle la capacité (C) du canal. Il postule

que l'entropie H du message à transmettre est analogue à l'entropie de Boltzmann à une constante près. Il montre sous réserve d'ergodicité du message que H correspond à l'état le plus probable et que le canal ne peut dépasser une vitesse de traitement supérieure à C pour une longueur de chaîne infinie [4]. Ce théorème implique une vitesse de transmission liée au message lui-même — donc à son entropie — et la possibilité d'augmenter le débit de transmission en codant le message initial suivant une taille minimale. Cette taille minimale est bornée par l'entropie, mais la méthode de codage (qui dépend de l'entropie du message à encoder) n'est pas précisée. Les deux exemples donnés [4] pour coder de manière optimale l'information peuvent être considérés comme des précurseurs des codages de Huffman [5] (le plus utilisé en compression informatique) et RLE [6] (« run length encoding », très utilisé de nos jours en imagerie). Cependant, les théories probabilistes utilisées n'abordent pas le lien avec la théorie des programmes, et il est difficile d'établir la fiabilité des procédures de codage optimal pour des probabilités d'occurrence des chaînes à transmettre inconnues. En d'autres termes, cela revient à affirmer que l'entropie du message doit être connue pour déterminer si l'algorithme est optimal et comprime donc le message avec le maximum d'efficacité. En ce sens, la taille du message, ou plus précisément le taux de compression (taille du message codé divisée par la taille du message original), ne peut être reliée à l'entropie que si l'algorithme est optimal. Pour que la taille du message codé puisse devenir une mesure de l'entropie du système, l'algorithme de codage ne doit faire aucune hypothèse a priori sur l'entropie informationnelle du message lui-même. La seule règle que nous devons nous fixer est que parmi tous les algorithmes existants, l'algorithme optimal est celui qui comprimera au maximum le message initial. Comment pouvoir vérifier une telle assertion? La réponse (ou comme nous verrons l'indécidabilité à cette réponse) doit être recherchée dans la théorie de la programmation. La thermodynamique statistique considère un système physique comme un système aléatoire. Le plus probable parmi tous les états accessibles d'un système thermodynamique est celui qui rend son entropie maximum. Statistiquement et physiquement, comme le système évolue continûment d'un état à l'autre, la configuration la plus probable est celle qui correspond au plus grand désordre (qui contient le maximum de complexions) et rend le système purement aléatoire. La notion d'aléatoire, ou plus précisément d'un point de vue de la théorie de l'information ce qu'est une suite aléatoire, a fait l'objet d'un grand nombre de publications. Toutes les théories basées sur la notion de tests statistiques (Martin Hoff, Church) sont vouées à un échec conceptuel, puisque aucun test statistique ne pourra affirmer qu'une hypothèse est vraie à un seuil de confiance fixé [7]. Par contre, il est possible d'affirmer que celle-ci est fausse à une probabilité d'erreur donnée. Une deuxième approche est la *Théorie algorithmique de l'information* [8] introduite par Kolmogorov et développée intensément par Chaitin qui définit l'aléatoire en termes de taille d'un programme informatique : une suite N sera dite aléatoire si la taille du programme permettant de la générer ne peut être inférieure à N.

Cette théorie algorithmique de l'information peut être reliée à la théorie de Shannon ainsi qu'à la thermodynamique statistique. Le message initial représente le système physique comme une succession d'états d'un spin codé en binaire (**00000111111**), séquence qui peut être considérée comme un programme informatique décrivant les états de la source. Mais un programme de taille réduite pourrait être (**5 Rep 0, 6 Rep 1**), où **Rep** signifie l'instruction **Répéter**.

Il est primordial de différencier trois catégories de programmes : le programme permettant le codage (encodeur), celui permettant la lecture du message codé (décodeur) et finalement le programme représentant le système lui-même (descripteur). Les programmes encodeur et décodeur peuvent être très complexes afin d'obtenir un descripteur de taille réduite, mais en aucun cas encodeur et décodeur ne caractérisent le système physique lui-même. Dans notre cas, ces programmes ne sont considérés que comme des outils d'observation dont les degrés de complexité ne représentent qu'une mesure de la compréhension et de notre aptitude à modéliser le système physique qui nous intéresse. La construction d'un encodeur (et de son décodeur associé) permettent d'obtenir un descripteur de taille minimale et induit notre parfaite maîtrise de la physique du système encodé, c'est-à-dire comprimé. Le taux de compression devient une mesure au sens mathématique de notre description du système physique modélisé par

un programme informatique. Csiszar et Körner [9] ont montré que quelle que soit la nature de la distribution P du message à encoder, il existe un encodeur universel donnant un descripteur d'entropie E, tel que $H(P) \leq E$. Ce théorème prouve l'existence d'un protocole universel d'encodage mais pas sa convergence uniforme vers l'entropie du système ni l'existence d'un protocole unique de caractérisation de l'entropie d'un système (particulièrement pour l'étude des processus thermodynamiques irréversibles, c'est-à-dire à production d'entropie). De plus, l'approche de Shannon suppose que l'information est unidimensionnelle et ergodique. Certes, dans le cas d'un système à l'équilibre, cette dernière hypothèse est satisfaite, mais en quittant l'état d'équilibre (soit par fluctuations statistiques, soit par des actions extérieures) aucun théorème ne prouve l'adéquation de l'encodeur. La non-ergodicité est particulièrement vérifiée pour des systèmes multifractaux pour lesquels la dimension d'information caractérise la variation non linéaire de l'entropie avec l'échelle de mesure où cette entropie est mesurée [10]. Pour caractériser un système physique, il paraît donc judicieux d'utiliser la théorie algorithmique de l'information suivant laquelle un système de taille n sera dit à l'équilibre si la taille du programme descripteur varie linéairement avec n.

Dans cette note, nous proposons d'analyser l'entropie de mélange d'un système physique à l'équilibre par la théorie algorithmique de l'information. Ce système est simulé par la méthode de Monte Carlo, puis comprimé par l'algorithme RLE, celui d'Huffman ou par la composition de ces deux algorithmes. En faisant varier les probabilités de présence de différentes complexions ainsi que leurs nombres, nous étudierons le lien entre l'entropie statistique et la taille du programme compressé. Le plan de l'article se compose de la manière suivante : après avoir décrit les méthodes de compression RLE et Huffman, nous introduisons une présentation sommaire du formalisme mathématique développé pour étudier l'entropie par cette technique et décrire les propriétés que doivent suivre les encodeurs. Nous présentons alors les résultats de la simulation Monte Carlo d'un système binaire à l'équilibre pour les différentes classes d'encodeurs étudiés et analysons par des techniques statistiques les isomorphismes entre la taille des programmes codés et l'entropie statistique. Finalement, l'étude est étendue aux systèmes ternaires, et nous présentons une réflexion sur le lien entre l'énergie de simulation et l'énergie libre du système physique simulé.

2. Les algorithmes de compression

2.1. L'algorithme RLE

Le principe de base de la compression RLE est le suivant [6] : Si une des données encodées notée d apparaît n fois, alors les n occurrences sont remplacées par le couple (n, d). Les nombreuses variantes de codage du couple (n, d) dépendent des protocoles utilisés, et sont plus ou moins efficaces en fonction du type de système à comprimer. Cet algorithme semble être efficace pour comprimer les systèmes thermodynamiques, car les écarts par rapport aux états les plus probables auront tendance à allonger certaines séquences et donc à augmenter le taux de compression. De même le changement des probabilités des états aura pour conséquence de changer la longueur moyenne des séquences successives et pourra être détecté. Cependant, un des principaux défauts de cet algorithme demeure dans le codage informatique du nombre de répétitions, puisque la représentation numérique de n, souvent bornée, peut introduire des artefacts. Dans les algorithmes « commerciaux », cette borne est souvent égale à 256. Pour un système à entropie nulle (message uniforme) la taille du message compressé varie en $\log_2(n)$ si le nombre d'états est inférieur à cette borne. Par contre, si le nombre de répétitions d'états est supérieure à $256k$ (avec $k > 1$), alors l'encodeur répète k fois la séquence $(256, d)$, et la taille du descripteur sera de l'ordre de $k \log_2(256)$, ce qui implique une taille du *descripteur* proportionnelle à la taille du message initial, puisque k est proportionnel à la longueur du message à encoder. D'après la théorie algorithmique de l'information, cela laisserait supposer à tort que le message est aléatoire, et le fait d'augmenter la taille du codage sur 16 ou 32 bits ne changerait rien au problème. De plus, l'adaptation du codage en fonction de la longueur de la plus grande répétition aurait pour conséquence d'introduire un artefact dans la méthode d'analyse que nous voulons développer, puisque le taux de compression

risquerait d'être assez éloigné du taux de compression optimum. De ce fait l'analyse entre l'entropie de deux systèmes différents (taille, équilibre...) par le taux de compression risquerait de n'être pas pertinente.

2.2. L'algorithme d'Huffman

Cet algorithme [5], proche de celui de Shannon–Fano fait l'objet de nombreuses recherches dans le domaine de la compression de données pour les ordinateurs personnels. Dans une première étape, l'algorithme analyse la liste des symboles du signal à encoder et crée une table par probabilité décroissante. Un arbre doté d'un symbole à chaque ramification et affecté d'un code binaire est ensuite construit. De par son aspect probabiliste, cette méthode d'analyse est susceptible d'inclure l'aspect statistique de l'entropie physique et semble pertinente pour compresser d'une manière homogène des systèmes de la thermodynamique statistique. En adaptant la construction de l'arbre des probabilités, ce type d'algorithme peut être facilement modifiable pour encoder d'une manière plus pertinente l'information d'un système thermodynamique particulier. Cependant, la décorrélation spatiale ou temporelle des données nuit à l'analyse des systèmes thermodynamiques non ergodiques. Ce défaut peut être partiellement contourné par un pré-codage (ou même une pré-compression) de l'information du système physique incluant les effets de corrélation spatiale (comme l'algorithme RLE).

3. Formalisme mathématique

Pour obtenir un outil mathématique d'étude des différents systèmes physiques X considérés comme des ensembles ordonnés, nous devons développer un formalisme suffisamment universel pour être appliqué aux différentes branches des mathématiques que sont la topologie, les fractales, l'analyse non standard, la statistique, le calcul différentiel, la structure algébrique... Pour créer des isomorphismes entre cette théorie et la mathématique des systèmes physiques, nous avons volontairement abandonné le formalisme de la théorie de l'information trop délicat à utiliser. Nous allons brièvement introduire ce formalisme et quelques théorèmes dont les démonstrations seront publiées ultérieurement.

Soit X le système physique à étudier. Nous demandons que la structure algébrique de X satisfasse les propriétés suivantes :

- Les propriétés physiques (comme par exemple des équations aux dérivées partielles, lois statistiques, aspect fractal ...) applicables sur le système initial (continu ou discret) sont facilement applicables sur le système X lui-même.
- La quantification des erreurs de discrétisation est possible.
- Les règles et des morphismes nécessaires à la construction de l'encodeur considéré maintenant comme une algèbre facile.

De ce fait, X sera considéré comme un espace vectoriel pour traiter un grand nombre de systèmes physiques dans un formalisme standard.

Soit X un ensemble ordonné (système discrétisé), T l'algèbre bijective non linéaire qui transforme le système X en un système Y noté $Y = T(X)$, et $\dim(X)$ la dimension de l'espace X, alors $T\ (X)$ sera dite contractile si :

$$\dim(T(X)) \leq \dim(X) \tag{1}$$

Par conséquent les dimensions de X et de $T(X)$ sont considérées comme celles de deux espaces différents. L'algèbre T peut être vue comme un opérateur de compression sans perte d'information puisque T est bijective et donc : $Y = T(X)$, $X = T^{-1}(Y)$. T est considérée comme une projection

d'un espace vectoriel sur un autre espace pour former un sous-espace, diminuant ainsi la dimension de l'espace original. La réduction de dimension est caractérisée par :

$$R(X) = \frac{\dim(T(X))}{\dim(X)} \tag{2}$$

En réalité, pour pouvoir être appliquées à un grand nombre de systèmes physiques, T et T^{-1} ne peuvent être résumées par une formule analytique mais doivent être considérées comme une succession d'opérateurs complexes. Cette algèbre sera décrite sur une machine de Turing [11], et seul le λ-calcul permet une formulation « analytique » de T ou T^{-1}.

Soit E un espace vectoriel de dimension N dans S, Ψ une famille de sous-espaces de E et G une relation de E dans E, T sera dite monotone dans (E_S, Ψ, G) si :

$$\begin{aligned} &\forall (M, N) \in \Psi^2, \qquad \forall x \in M_G, \qquad \forall y \in N_G \\ &Dim(M) \leq Dim(N) \Rightarrow Dim\,(T\,(x)) \leq Dim\,(T\,(y)) \end{aligned} \tag{3}$$

où M_G est l'ensemble des éléments de M donnés par la relation$G$.**Auteur : quelle est la signification de *Dim*? Est-ce que c'est différent de dim dans l'équation 4?**

T est une projection d'un espace vectoriel E de dimension N dans S sur un espace vectoriel E' de dimension N' dans S' $(N' < N)$.

Soit $\{A\}$ un ensemble d'algèbres, $T_{\min}$ sera dite $\{A\}$-maxi-contractile dans (E_S, Ψ, G) s'il existe une algèbre notée $T_{\min} \in \{A\}$ telle que :

$$\begin{aligned} &\forall X \subset \Psi, \qquad \forall T \in \{A\}, \qquad \exists T_{\min} \in \{A\} \\ &\dim(T_{\min}(X)) \leq \dim(T(X)) \end{aligned} \tag{4}$$

T_{opt} est dite $\{\infty\}$-maxi-contractile dans (E_S, Ψ, G) si T_{opt} est $\{\Omega\}$-maxi-contractile dans (E_S, Ψ, G), où Ω représente toutes les algèbres possibles définies par l'arithmétique.

Trois théorèmes seront utilisés pour la description des systèmes thermodynamiques présentés dans cette note :

Théorème 1 : Il existe au moins une algèbre maxi-contractile sur Ω.

Théorème 2 : Il est impossible par l'arithmétique de proposer une méthode pour construire $T_{\min} \in \{\infty\}$.

Théorème 3 : Il est toujours possible de construire une algèbre $T \in \{\infty\}$ telle que pour tout sous-ensemble x de X :

$$\begin{aligned} &\alpha\,(T)\ card(F(x, T_{\min})) \leq card(F(x, T) \leq \beta\,(T)\, card(F(x, T_{\min})) \\ &\alpha\,(T) \geq 1, \qquad \beta\,(T) > 1, \qquad \alpha\,(T) < \beta\,(T) \end{aligned} \tag{5}$$

Auteur : quelle est la signification de *card*? Est-ce un mot ou 4 variables?

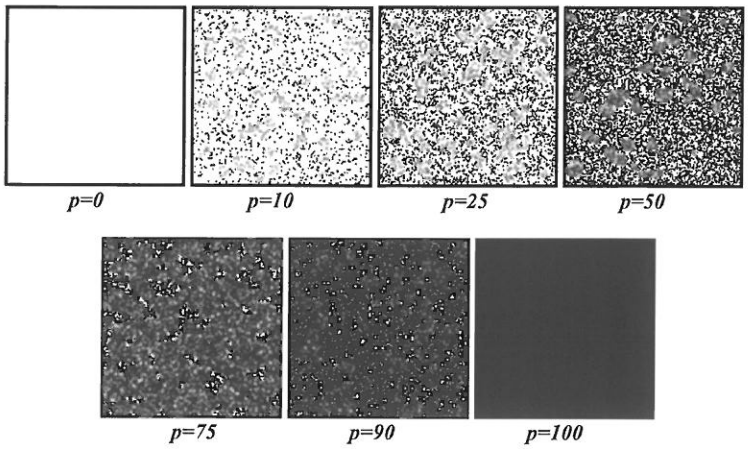
Le théorème 2 provient d'une extension du théorème d'incomplétude de Gödel [7], la proposition $<<$ T_{opt} est dite $\{\infty\}$-maxi-contractile in (E_S, Ψ, G) $>>$ est indécidable. Cependant il est toujours possible de construire un ensemble d'algèbres $\{A\}$ telle que la proposition $<<$ $\{A\}$-maxi-contractile dans (E_S, Ψ, G) $>>$ devienne décidable. Les propriétés définies par l'ensemble $\{A\}$ dépendent principalement du système physique G lui-même.

L'espace projeté peut être décomposé en trois sous-espaces :

$$T(X) = T_{\text{ALGEBRA}} \oplus T_{\text{SYSTEM}}(X) \oplus T_{\text{PHYSICAL}}(X) \tag{6}$$

tels que :

Fig. 1. Résultats de la simulation Monte Carlo d'un système binaire à l'équilibre de taille $r = 512$ (blanc état 0 et noir état 1) pour différentes probabilités p (en %) qu'une cellule soit à l'état 1; avec $p = \{0\ 10\ 25\ 50\ 75\ 90\ 100\}$.



T_{ALGEBRA} est indépendante de X et est créée par la T algèbre pour permettre l'interprétation du système X et construire l'application réciproque T^{-1}. La taille de ce sous-espace correspond à la taille minimale permettant d'affirmer que le système X est vide.

$T_{\text{SYSTEM}}(X)$ dépend de X et contient sa topologie fonction de sa taille $\dim(X)$. Il constitue le noyau $Ker(T(X))$ de $T(X)$ où $\dim Ker(T(X))$ est la dimension du système quand X contient le même élément.

$T_{\text{PHYSICAL}}(X)$ est le sous-espace donné par la relation G.

Si $T_{\text{SYSTEM}}(X) \cap T_{\text{PHYSICAL}}(X) = \phi$ alors :

$$\dim(T(X)) = \dim(T_{\text{ALGEBRA}}) + \dim(Ker(T(X))) + \dim(T_{\text{PHYSICAL}}(X)) \tag{7}$$

4. Simulation Monte Carlo du système physique

Nous proposons d'étudier un système thermodynamique binaire divisé en cellules de tailles égales, chaque cellule occupant l'état 0 avec la probabilité p ou l'état 1 avec la probabilité $1 - p$. Cette schématisation correspond à de nombreux systèmes thermodynamiques comme la répartition d'atomes en interstitiel, de lacunes dans un réseau cristallin, la répartition des éléments A et B d'un alliage binaire à miscibilité totale.

Le système X est simulé de la manière suivante. Nous choisissons une résolution r (le nombre de cellules est donc $n = r^2$), puis affectons à chaque cellule un nombre aléatoire uniforme compris entre 0 et 1. Si ce nombre est supérieur à p, alors l'état de la cellule est 0 et 1 dans le cas contraire. La figure 1 représente sept systèmes de résolution $r = 128$ pour différentes probabilités p (blanc = 0, noir = 1).

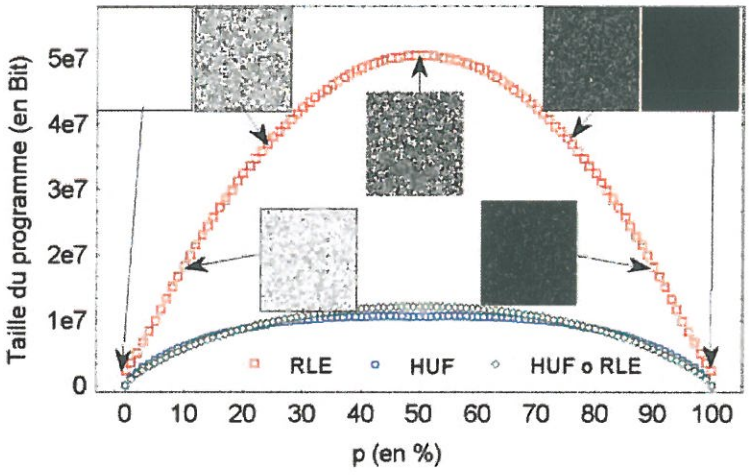
5. Relation entre l'entropie et la dimension de l'espace compressé

5.1. Relation entropie thermodynamique, entropie informationnelle

Pour ce système thermodynamique, l'entropie se limite à l'entropie de mélange :

$$S = -R \sum_{i=1}^{s} p_i \log p_i \tag{8}$$

Fig. 2. Évolution de la taille des programmes (en bit) en fonction de la probabilité p (en %) pour un système binaire de résolution $r = 8192$ en utilisant les trois algèbres, HUF, RLE et HUF $\circ$ RLE.**Auteur : quelle est la signification de HUF? Que représente le cercle entre HUF et RLE?**



où p_i représente la probabilité qu'une cellule soit à l'état i pris parmi s états et R la constante des gaz parfait. Dans notre simulation, $s = 2$ et la valeur maximale de l'entropie qui correspond à $p_1 = p_2 = 0{,}5$ est bornée par $R \log 2$. De même sous l'hypothèse d'équirépartition des états s, le maximum d'entropie aura pour coordonnée $(1/s,\ 1/s, \cdots,\ 1/s\ )$.

L'entropie informationnelle, définie par Shannon [4], est donnée par :

$$H = -K \sum_{i=1}^{s} p_i \log p_i \tag{9}$$

où K est une constante positive qui dépend des unités de mesure du codage (elle sera souvent un multiple de bits/s). En comparant les équations (8) et (9) on déduit que $S = \alpha H$. Le coefficient $\alpha > 0$ constitue le lien entre l'entropie de mélange et l'entropie informationnelle.

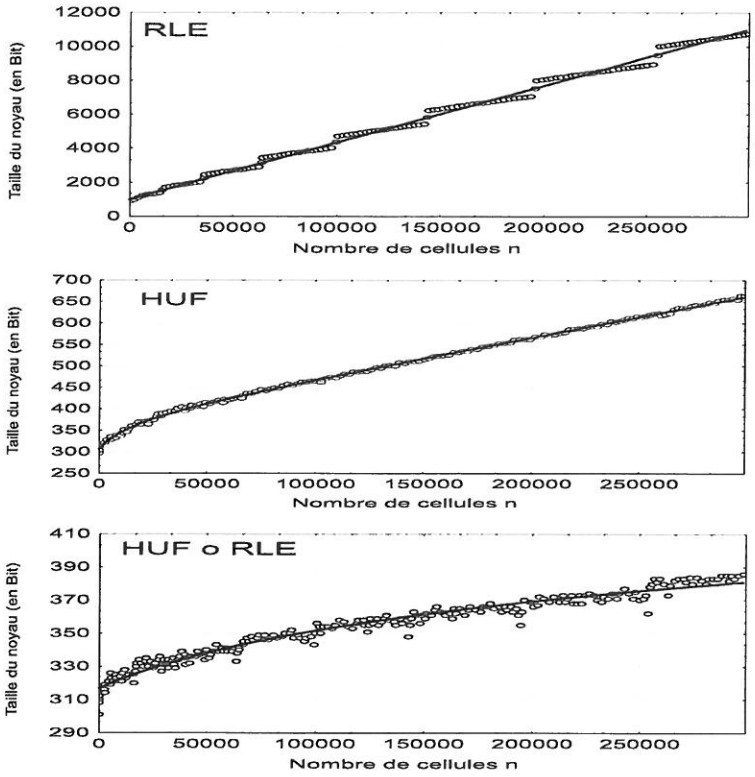
5.2. Étude de la dimension de l'espace compressé

La taille de l'espace compressé est analysée par deux algèbres RLE ($T = \text{RLE}$), Huffman ($T = \text{Huf}$) ainsi que par la composée de ces deux algèbres $T = \text{Huf} \circ \text{RLE}$ **auteur : Est-ce que HUF est différent de Huf?** (le premier espace est formé par le résultat de la compression RLE, puis la compression d'Huffman est appliquée sur celui-ci). Nous noterons $X_{p,n}$ la simulation Monte Carlo sur le système binaire de taille n (avec $n = r^2$) et avec une probabilité p pour que chaque cellule soit à l'état 1. La figure 2 représente la taille en bit des descripteurs $\dim(T(X_{p,n}))$ en fonction de la probabilité p (en %) pour un système de résolution $r = 8192$ en utilisant les algèbres RLE, Huf et Huf $\circ$ RLE. Plusieurs simulations sont effectuées pour diminuer les fluctuations d'ordre statistique en faisant la moyenne de la taille de l'espace compressé (le nombre de simulations diminue avec la résolution r du système). La discussion des résultats obtenus est reportée au paragraphe 5.2.2.

5.2.1. Étude du noyau

Nous étudions d'abord le noyau de l'application de l'algèbre T en imposant un système vide qui correspond à la probabilité $p = 0$ (ou $p = 1$). Comme toutes les cellules du système physique occupent

Fig. 3. Évolution du noyau dim $\left(T(X_{0,n})\right)$ en fonction du nombre de cellules n du système X pour les trois algèbres, HUF, RLE et HUF o RLE. Les courbes de régression données par les équations (11) à (13) sont représentées sur chaque graphique.



le même état, alors dim $\left(T_{\text{PHYSICAL}}(X_{0,n})\right) = 0$ et d'après (7) :

$$\dim\left(T(X_{0,n})\right) = \dim\left(T_{\text{ALGEBRA}}\right) + \dim\left(Ker\left(T(X_{0,n})\right)\right) \tag{10}$$

La figure 3 représente l'évolution de dim $\left(T(X_{0,n})\right)$ en fonction de n ($n = \dim\left(X_{0,n}\right)$. La formulation analytique du noyau est déterminée par régression des moindres carrés pour trouver le modèle qui permet le meilleur ajustement avec les points expérimentaux. Il est montré statistiquement que :

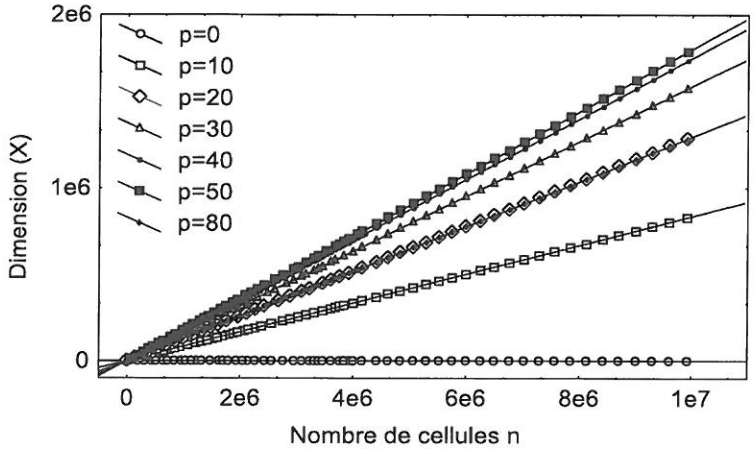
$$\dim\left(\text{RLE}(X_{0,n})\right) = 992 + 0{,}0335n \tag{11}$$

$$\dim\left(\text{HUF}(X_{0,n})\right) = 383 + 0{,}000\,094n - \frac{1907}{24 + 0{,}0015n} \tag{12}$$

$$\dim\left(\text{HUF} \circ \text{RLE}(X_{0,n})\right) = -91 + 36{,}9\ln(64\,007 + n) \tag{13}$$

En posant $n = 0$ dans (11) à (13), dim $(\text{RLE}_{\text{ALGEBRA}}) = 992$, dim $(\text{HUF}_{\text{ALGEBRA}}) = 304$ et dim $(\text{HUF} \circ \text{RLE}_{\text{ALGEBRA}}) = 317$. D'après (11), l'algèbre RLE varie linéairement avec le nombre de cellules en contradiction avec la théorie de Kolmogorov (en effet, les systèmes n'étant pas aléatoires, la complexité du programme ne peut varier proportionnellement à sa taille). Comme nous l'avions évoqué, le « découpage » du codage fait que l'algorithme varie macroscopiquement avec n à partir d'un seuil. Cette particularité se retrouve également pour l'algèbre HUF. Par contre, la composition des ces

Fig. 4. Évolution de la taille du programme compressé pour l'algèbre HUF ∘ RLE en fonction du nombre de cellules n du système binaire pour différentes probabilités (en %).



deux algèbres conduit à une relation induisant une variation du noyau en $\ln(n)$, conforme à la théorie de Kolmogorov. Pour cette raison, et sans perte de généralités, nous nous limiterons principalement dans la suite de l'article à cette composition d'algèbres. D'autres algorithmes de compression ont été testés (« prediction by partial matching » [12], LZ77 et ZZ78 [13], Lempel–Ziv–Welch [14], ...) ont donné des résultats, certes honorables, mais discriminaient statistiquement moins les effets de taille. Seule la combinaison du RLE avec la méthode de Burrows–Wheeler [15] aboutit à des résultats de mêmes qualités que ceux présentés dans cet article.

5.2.2. Étude multi-échelle de l'entropie

La figure 4 représente l'évolution de la taille du programme compressé par l'algèbre HUF ∘ RLE en fonction du nombre de cellules n du système affectées à différentes probabilités d'états. Après régressions linéaires, pour $p \neq \{0,\ 1\}$ une très bonne adéquation des données est obtenue par le modèle linéaire suivant :

$$\dim\left(\mathrm{HUF} \circ \mathrm{RLE}(X_{p,n})\right) = \beta_{\mathrm{HUF}\circ\mathrm{RLE}}\,(p)\ n \tag{14}$$

Si HUF ∘ RLE est $\{\infty\}$-maxi-contractile, alors (14) confirme la définition de l'aléatoire par la théorie de Kolmogorov. Il est montré statistiquement que $\beta_{\mathrm{HUF}\bullet\mathrm{RLE}}\,(p) = \beta_{\mathrm{HUF}\bullet\mathrm{RLE}}\,(1-p)$ **Auteur : pourquoi un point ici et pas un cercle?** puisque le taux de compression demeure inchangé par permutation des différents états. La pente de la droite de régression croit avec p pour atteindre une valeur maximale pour $p = 0{,}5$, valeur qui correspond au maximum d'entropie du système. Plus l'entropie augmente, et plus le taux de compression décroît.

Nous allons maintenant chercher si un isomorphisme existe entre la pente $\beta_T\,(p)$ de la droite donnée par l'équation :

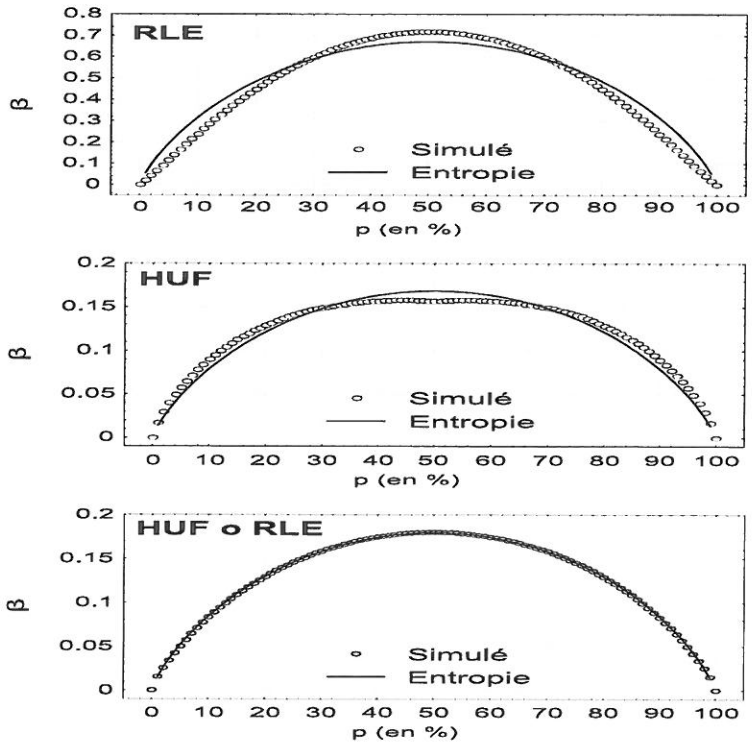
$$\dim\left(T_{\mathrm{SYSTEM}}(X_{p,n})\right) = \beta_T\,(p)\ n \tag{15}$$

et l'entropie physique donnée par l'équation (8) en posant $s = 2$:

$$\beta_T\,(p) = -\alpha_T\left[p\log p + (1-p)\log(1-p)\right] \tag{16}$$

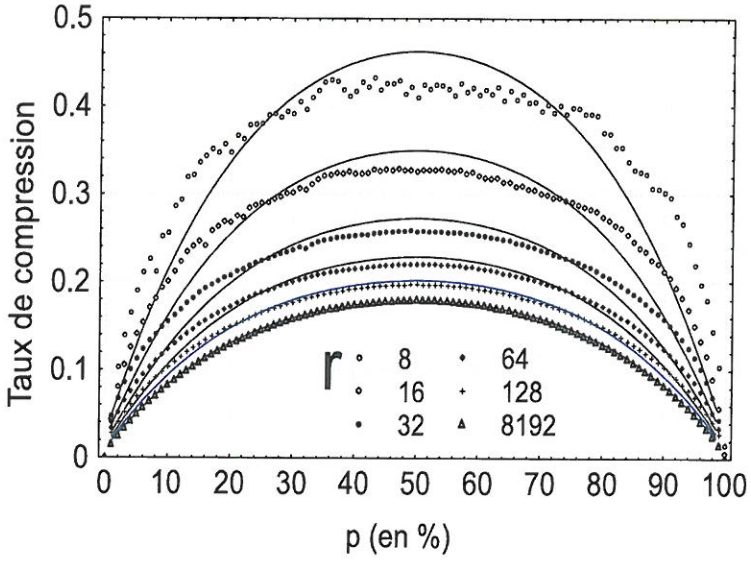
La constante α_T est déterminée par les moindres carrées pour les trois algèbres RLE, HUF et HUF∘RLE.

Fig. 5. Valeurs des pentes $\beta_{\mathrm{RLE}}(p)$, $\beta_{\mathrm{HUF}}(p)$ et $\beta_{\mathrm{HUF}\circ\mathrm{RLE}}(p)$ des droites de régression représentant l'évolution de la taille du programme compressé par l'algèbre HUF o RLE en fonction du nombre de cellules n du système binaire pour différentes probabilités (en %). Les courbes représentent la taille du programme modélisée par l'équation de l'entropie de mélange (16).



Les valeurs calculées et les valeurs modélisées par (16) sont représentées à la figure 5. Les algèbres RLE et HUF ne permettent pas une bonne adéquation avec l'entropie physique (coefficient de corrélation : $r_{\mathrm{RLE}} = 0{,}992$ et $r_{\mathrm{HUF}} = 0{,}996$). Par contre l'algèbre composée HUF o RLE fournit une adéquation quasi parfaite avec un coefficient de corrélation $r_{\mathrm{HUF}\circ\mathrm{RLE}} = 0{,}999\,98$ en prenant la valeur $\alpha_{\mathrm{HUF}\circ\mathrm{RLE}} = 0{,}258\,7 \pm 0{,}000\,1$. La composition des deux algèbres permet donc de construire un isomorphisme entre la taille du programme compressé et l'entropie physique à une constante multiplicative près. Pour comprendre la performance de cette composée, il est nécessaire d'analyser les algorithmes de compression. Dans un premier temps, l'algèbre RLE comptabilise les séquences identiques sous la forme ***n* Rep 0** (ou ***n* Rep 1**), mais ce nombre n possède un codage fini induisant la possibilité d'obtenir une succession identique de la séquence ***n* Rep 0** (ou ***n* Rep 1**) aux faibles entropies, résultats éloignés de la définition d'une algèbre maxi-contractile. D'une manière analogue, le codage Huffman crée un dictionnaire **fini** de différentes séquences pour lequel les séquences les moins probables ne sont pas encodées, ce qui induit également à une algèbre non maxi-contractile. Le fait d'associer l'algèbre RLE puis l'algèbre Huffmann permet de sauvegarder les séquences de répétitions dans le dictionnaire, et par-là de réduire uniformément le taux de compression. En un sens, bien que cette algèbre composée ne puisse être maxi-contractile, elle tend à satisfaire la propriété (5) et donc à approcher l'entropie du système.

Fig. 6. Évolution du taux de compression C (ARJ ∘ RLE, p, n) en fonction de la probabilité p (en %) pour différentes résolutions $r = n^2$ du système binaire.**Auteur : que signifie ARJ?**



5.2.3. Étude de l'entropie d'un système de taille fixée

Dans la partie précédente, nous avons présenté une méthode de calcul de l'entropie qui analyse la variation de la taille du programme compressé en fonction de la taille du système initial. Cependant, est-il possible de calculer l'entropie sans analyser l'évolution de la taille du programme et, si oui, quelle doit être la taille du système pour que la mesure de l'entropie soit pertinente? Pour répondre à cette question, les systèmes $X_{p,n}$ sont simulés pour $n \in \{2, 4, 8, 16, ..., 8192\}$ avec un taux de compression défini par :

$$C(T, p, n) = \frac{\dim\left(T(X_{p,n})\right)}{\dim\left(X_{p,n}\right)} = \frac{\dim\left(T(X_{p,n})\right)}{n} \tag{17}$$

La figure 6 représente les valeurs du taux de compression en fonction de la taille du système. Visuellement, les courbes convergent uniformément vers une courbe limite lorsque la résolution du système augmente. Afin de quantifier cette observation, et vérifier si le taux de compression converge par isomorphisme vers l'entropie physique, nous calculons les valeurs de $\alpha_{T,n}$ par la méthode des moindres carrés pour chaque taille du système telles que :

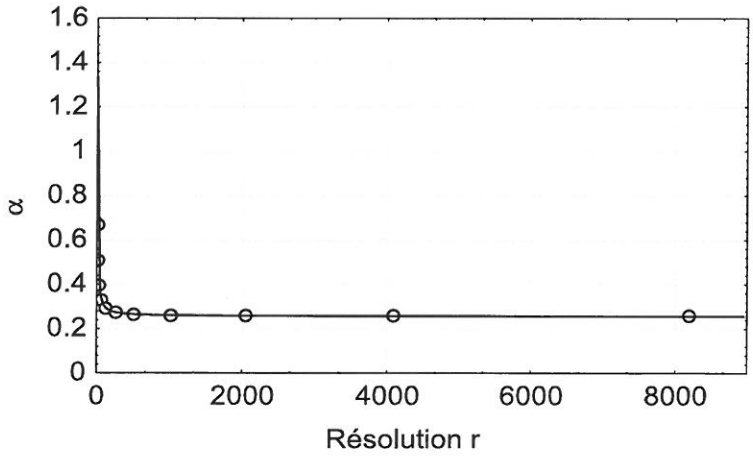
$$C(T, p, n) = -\alpha_{T,n}\left[p \log p + (1 - p) \log (1 - p)\right] \tag{18}$$

Sur la figure 7, qui représente les valeurs de $\alpha_{T,n}$ en fonction de la résolution du système $r = n^2$, on constate que les valeurs de $\alpha_{T,n}$ convergent vers une asymptote d'équation :

$$\alpha_{\text{HUF}\circ\text{RLE},n} = 0{,}258 + \frac{6{,}599}{\left[(r + 5{,}45)^{1{,}07}\right]} \tag{19}$$

L'équation (19) permet d'obtenir la valeur asymptotique $\alpha_{\text{HUF}\circ\text{RLE},\infty} = 0{,}258$. Une étude sur le résidu de la régression montre que pour $r > 512$, les erreurs deviennent négligeables (coefficient de corrélation

Fig. 7. Valeurs de $\alpha_{\mathrm{HUF\circ RLE},n}$ en fonction de la résolution du système $r = n^2$ du système binaire. La courbe représente l'équation de régression donnée par l'équation (19).



$> 0{,}999\,98$) prouvant ainsi l'efficacité de la composition des algèbres pour calculer l'entropie d'un système de taille fixée. La valeur de $\alpha_{T,\infty}$ est proche de la valeur calculée au chapitre précédent ($\alpha_{\mathrm{HUF\circ RLE}} = 0{,}258\,7 \pm 0{,}000\,1$) prouvant ainsi la cohérence et le bien-fondé des deux approches. L'écart pour des tailles de système faibles est expliqué par le fait que le codage informatique est effectué sur la base de l'octet (8 bits), ce qui nécessite de saturer ces octets et donc de travailler avec une taille de système supérieure à $2^8 = 256$ pour obtenir un « bon remplissage » de l'espace X.

5.3. Étude tridimensionnelle de l'espace compressé

Il serait intéressant d'analyser la performance de la composée d'algèbres sur un système ternaire. Pour cela, un système X à trois états de résolution de $r = 1024$ est simulé : seules deux probabilités p_1 et p_2 sont imposées avec $p_1 + p_1 \leq 1$, la troisième p_3 se déduisant par normalisation. La figure 8 représente l'évolution de $\dim\left(\mathrm{HUF} \circ \mathrm{RLE}(X_{p_1,p_2})\right)$ en fonction de p_1 et p_2. Le maximum de la dimension est obtenu pour les valeurs $p_1 = 0{,}33$, $p_2 = 0{,}33$ et $p_3 = 0{,}33$, ce qui correspond au maximum de l'entropie physique (8). La constante de proportionnalité α telle que :

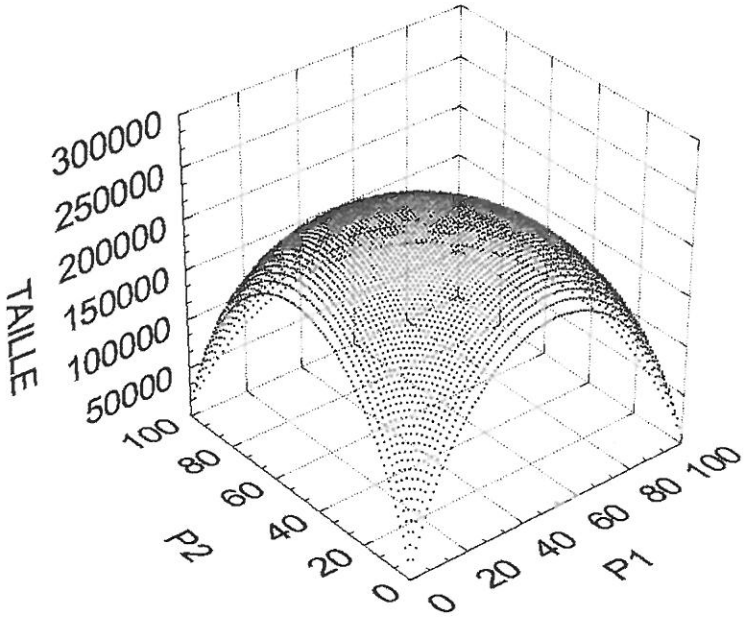
$$C\,(p_1, p_2) = -\alpha\left[p_1 \log p_1 + p_2 \log p_2 + (1 - p_1 - p_2)\log(1 - p_1 - p_2)\right] \tag{20}$$

calculée par les moindres carrés est égale à $\alpha = 0{,}245$, valeur proche de celle obtenue pour le système binaire. Nous en déduisons que cette composition d'algèbre permet de traiter les systèmes ternaires avec la même pertinence.

6. Relation entre l'entropie, l'énergie libre et l'énergie de simulation

Nous avons montré l'existence d'un isomorphisme entre la taille d'un programme qui simule la répartition de particules, d'entités, ou plus généralement des domaines de l'espace des phases et l'entropie statistique du système lui-même, pourvu que ce programme soit le plus petit parmi tous les programmes possibles. **Auteur : que signifie Pr ici? Prandtl number?** D'après l'équation (16), la taille Pr_n du programme minimal décrivant le système de taille n est proportionnelle à l'entropie ($\Delta S_n = \alpha\,\mathrm{Pr}_n$ avec $\Delta S_n = -R\sum_{i=1}^{s} p_i \log p_i$ l'entropie du système de taille n). L'énergie libre est alors $\Delta G_n = \Delta H_n - \alpha T\,\mathrm{Pr}_n$ et, en supposant l'absence de variation d'énergie interne du système qui n'est pas

Fig. 8. Évolution de la dimension dim $\left(\mathrm{HUF} \circ \mathrm{RLE}(X_{p_1,p_2})\right)$ d'un système ternaire en fonction des probabilités p_1 et p_2.



modélisée par notre programme, $\Delta G_n = -\alpha T \Pr_n$. Comme le programme s'exécute sur une machine idéale de Turing [16], l'exécution d'une instruction élémentaire (taille unitaire de programme, p.ex., 1 octet) nécessite une énergie ΔE, et par conséquent l'énergie totale pour simuler le programme informatique est donnée par $\Delta E \Pr_n$ et finalement $\Delta G_n = -\alpha T \Delta E \Pr_n$. L'énergie informatique consommée pour simuler le système physique est alors proportionnelle à l'énergie du système lui-même.

7. Conclusion

Nous avons montré qu'il existe un isomorphisme entre l'entropie de mélange et la taille informatique d'un programme simulant le système physique. Cet isomorphisme est construit par la composition de deux algorithmes : l'algorithme RLE puis celui de Huffman. Par simulation Monte Carlo, nous avons quantifié à l'aide de cet isomorphisme l'entropie de systèmes binaires et ternaires à l'état d'équilibre. L'énergie utilisée pour décrire algorithmiquement le système physique devient alors proportionnelle à l'énergie libre du système. Cette approche offre des perspectives d'analyse intéressantes pour les systèmes physiques. Par exemple, nous avons montré que la mesure de la dimension de l'espace compressé permet de déduire les lois cinétiques pour qu'un système physique atteigne l'équilibre [17]. Cette dimension offre l'avantage de donner une mesure non morphologique et globale du système physique lui-même. Les applications que nous développons sont nombreuses : mesure de la comparaison de systèmes, calcul de la dimension fractale, recherche d'invariants, mesure de gradient ...

Bibliographie

1. M. Horodecki. Fortschr. Phys. **49**, 667 (2001).
2. W.H. Zurek. Complexity, entropy and the physics of information. Addison-Wesley. 1990.
3. L. Brillouin. Science and information theory. Academic Press, New York. 1956.
4. C.E. Shannon. Bell System Tech. J. **27**, 379, 623 (1948).

5. D. Huffman. Proc. IRE, **40**, 1098 (1952).
6. D. Salomon. Data compression. Springer, New York. 1998.
7. J.P Delahaye. Information, complexité, hasard. Hermes, Paris. 1994.
8. G.J. Chaitin. Algorithmic information theory. Cambridge University Press. 1987.
9. I. Csizar et J. Körner. Information theory: Coding theorems for discrete memoryless systems. Academic Press, New York. 1981.
10. H.O. Peitgen, H. Jürgens et D. Saupe. Chaos and fractals new frontiers of science. Springer-Verlag. 1992.
11. E. Nagel, J.R. Newman, K. Gödel, et J.Y. Girard. Le théorème de Gödel. Seuil, Paris. 1989.
12. T. Bell, J. Cleary et I. Witten. IEEE Trans. Commun. **32**, 396 (1984).
13. J. Ziv et A. Lempe. IEEE Trans. Info. Theor. **23**, 337 (1977).
14. T.A. Welch. Computer, **17**, 8 (1984).
15. M. Burrows et D. Wheeler. The Burrows–Wheeler transform. A block sorting lossless data compression algorithm. Tech. Rep. 124. Digital Equipment Corporation. 1994.
16. A.M. Turing. Mind, **59**, 433 (1950).
17. M. Bigerelle et A. Iost. J. Comp. Mater. Sci. **24**, 133 (2002).