

Lois uniformes et normales de chaînes discrètes

Uniform and normal laws of discrete strings

S. Rebecchi

J.-M. Jolion

Université de Lyon, F-69361 Lyon
INSA de Lyon, F-69621 Villeurbanne
CNRS, LIRIS, UMR 5205

{sebastien.rebecchi,jean-michel.jolion}@liris.cnrs.fr

Résumé

Dans cet article, nous proposons l'extension du concept de distribution statistique à l'espace des chaînes discrètes, ceci plus particulièrement pour les lois uniformes et normales. Ces définitions permettent la manipulation statistique d'ensembles de chaînes discrètes, via des algorithmes génératifs et prédictifs. Nous proposons également une méthode d'estimation de paramètres de distributions de chaînes discrètes, appliquée dans cet article à l'estimation d'écart-type de la loi normale, et donnant des résultats encourageants. Ce travail est une première avancée dans l'objectif de reconnaissance statistique de formes structurées.

Mots clef

Chaînes discrètes, loi uniforme, loi normale, moyenne, écart-type, estimation.

Abstract

We propose in this paper the generalization of concept of statistical distribution to the space of discrete strings, this more particularly for uniform and normal laws. These definitions allow statistical handling of sets of discrete strings, via generative and predictive algorithms. We also propose a method of estimation of parameters of distributions of discrete strings, applied in this paper to the estimation of standard deviation of the normal law, and giving encouraging results. This work is a first trend toward a general statistical framework for recognition of structured patterns.

Keywords

Discrete strings, uniform law, normal law, mean, standard deviation, estimation.

1 Introduction

1.1 Contexte et motivations

Depuis quelques années, plusieurs équipes de recherche tentent de fusionner les deux approches principales de la

reconnaissance de formes, à savoir l'approche statistique et l'approche structurelle. Ce choix est motivé par l'envie de pouvoir à la fois profiter des avantages des deux méthodes tout en se détachant de leurs inconvénients.

La reconnaissance de formes statistique se base sur un codage des données sous forme de vecteurs numériques, incapable bien souvent de reproduire fidèlement la complexité des données brutes. Cependant ce choix est justifié par la large palette d'outils et algorithmes statistiques parus dans la littérature et reconnus comme performant pour la manipulation de données numériques.

Le problème est inverse pour la reconnaissance de formes structurelle, où la partie codage est riche car s'appuyant sur des structures de données de grande expressivité (graphes, chaînes, arbres...), permettant notamment de représenter de manière adéquate toute sorte de relations intra/inter formes (réflexivité, séquentialité, hiérarchie...). Par contre, les outils liés à la manipulation de structures sont souvent bien trop restrictifs (isomorphisme, distance d'édition...) et pas assez robustes pour les applications spécifiques à la reconnaissance de formes. Une autre limitation vient de l'absence de formalisme structurel pour le traitement d'ensembles de données, en ce sens où les outils classiquement utilisés sont le plus souvent basés sur des opérateurs seulement unaires ou binaires. Enfin, l'association entre d'une part la taille des structures de données utilisées, et d'autre part la complexité des algorithmes mis en œuvre, tend à rejeter cette approche pour le traitement de grands volumes de données.

Pour pouvoir réconcilier ces deux approches, il est au préalable nécessaire de définir une caractérisation statistique des espaces de structures discrètes, ce qui est l'objectif de notre travail. Nous nous focalisons dans cet article sur les lois uniformes et normales de structures de type chaîne. Après avoir posé les notations utilisées tout au long de cet article, nous faisons un bref rappel des descripteurs statistiques, ou apparentés comme tels, proposés dans la littérature pour caractériser des ensembles de chaînes discrètes

(section 2). Puis nous introduisons nos définitions des lois uniformes et normales de chaînes discrètes (section 3), munies toutes deux d'algorithmes permettant de générer des chaînes sous contrôle statistique, et de prédire la probabilité de toute chaîne. Enfin nous proposons l'application d'une méthode de recherche d'estimation d'écart-type de la loi normale, basée sur la minimisation de la distance entre la loi proposée comme estimation et un échantillon généré sous contrôle statistique de la loi à estimer (section 4).

1.2 Notations et définitions préliminaires

Soit A un alphabet fini, *i.e.* un ensemble non vide fini, dont les éléments sont appelés lettres¹. Une chaîne (définie) sur A est une séquence de lettres de A . Notons $|X|$ la longueur de la chaîne X , X_i la i -ème lettre de X , $|A|$ le nombre de lettres de A , A^n l'ensemble des chaînes de longueur n sur A , A^* l'ensemble des chaînes de longueur finie sur A , et λ la lettre ou la chaîne vide ($|\lambda| = 0$). Pour remarque, une chaîne définie sur A est définie sur tout sur-alphabet (alphabet sur-ensemble) de A , et λ est défini sur tout alphabet.

Une sous-séquence X' d'une chaîne X est une chaîne obtenue en supprimant des lettres à X , tout en préservant l'ordre relatif dans X des lettres restantes : $X' = X_{i_1} \dots X_{i_m}$, avec $i_1 < \dots < i_m$ et $m \leq |X|$. Si $i_{j+1} = i_j + 1$ pour tout $j \in \{1, \dots, m-1\}$, alors X' est une sous-chaîne de X . Si en plus $i_1 = X_1$ (resp. $i_m = X_{|X|}$), alors X' est le préfixe (resp. suffixe) de longueur m de X . Supposons l'existence de l'opérateur de concaténation entre chaînes ou lettres (chaînes de longueur 1), noté par un point, souvent omis par souci d'allègement des notations, et défini comme suit : pour tout couple (X, Y) de chaînes, $X.Y =_{\text{notation}} XY = Z$ ssi $Z = X_1 \dots X_{|X|} Y_1 \dots Y_{|Y|}$. Dès lors Z est défini sur l'alphabet union des alphabets sur lesquels sont respectivement définis X et Y , et la longueur de Z est égale à la somme des longueurs respectives de X et Y . Pour remarque, $\lambda X = X\lambda = X$ pour toute chaîne X .

Une opération d'édition sur A est un couple (a, b) de lettres de $A \cup \{\lambda\}$, noté $a \rightarrow b$. Nous disons qu'une telle opération transforme a en b , ou consomme a pour produire b . Si $a = \lambda$ (resp. $b = \lambda$), il s'agit de l'insertion de b (resp. la suppression de a), et si $a \neq \lambda$ et $b \neq \lambda$, il s'agit de la substitution de a par b .

Soit $c()$ une fonction de coût d'édition sur $A : \forall (a, b) \in (A \cup \{\lambda\})^2$, $a \neq b$, $c(a \rightarrow a) = 0$ et $c(a \rightarrow b) > 0$.

Une chaîne d'édition (ou chemin d'édition) sur A est une chaîne d'opérations d'édition sur A . Soit $Z = X_1 \rightarrow Y_1 \dots X_n \rightarrow Y_n$ une telle chaîne. Nous disons que Z consomme (resp. produit) la chaîne formée de la concaténation des lettres consommées (resp. produites) par ses opérations, *i.e.* $X = X_1 \dots X_n$ (resp. $Y = Y_1 \dots Y_n$). Le coût de Z par (extension de) $c()$ est défini comme la somme des coûts de ses opérations d'édition : $c(Z) = \sum_{i=1}^{|Z|} c(Z_i)$. La distance d'édition (de Levenshtein) associée à

¹La notion de lettre est à comprendre au sens large, et nous l'utilisons aussi bien pour désigner un élément de type symbolique ou numérique

$c()$, $d(X, Y)$, entre deux chaînes X et Y sur A , est définie comme le coût minimal d'une chaîne d'édition transformant X en Y [9] : $d(X, Y) = \min\{c(Z) \mid Z \text{ chaîne d'édition transformant } X \text{ en } Y\}$.

Une chaîne d'édition est dite optimale ssi son coût est égal à la distance d'édition entre sa chaîne consommée et sa chaîne produite. L'algorithme classiquement utilisé pour calculer la distance d'édition entre deux chaînes X et Y , et le chemin d'édition optimal associé, repose sur une méthode de programmation dynamique de complexité temporelle et spatiale $O(|X| \times |Y|)$, proposée initialement dans [15]. Il est à noter qu'il existe généralement plus d'un chemin optimal entre X et Y .

2 Descripteurs statistiques d'ensembles de chaînes discrètes

2.1 Moyenne

Le terme *moyenne* est souvent ambigu de part son association à plusieurs concepts. Il peut être utilisé pour désigner, entre autres, celui de moyenne arithmétique d'un ensemble de données, ou celui d'espérance mathématique (moyenne théorique) d'une loi de probabilité.

Le seul concept clairement traduit de manière à pouvoir être utilisé dans le domaine des chaînes est celui de moyenne pondérée de deux chaînes, proposé dans [1], et défini comme suit : une moyenne pondérée de deux chaînes X et Y est une chaîne Z vérifiant cette équation : $d(X, Z) + d(Z, Y) = d(X, Y)$. Dans ce cas, le poids de X (resp. Y) dans Z est égal à $d(Z, Y)/d(X, Y)$ (resp. $d(X, Z)/d(X, Y)$). La méthode de construction de Z est également détaillée dans [1].

Cependant, le concept qui nous intéresse ici est celui d'espérance mathématique (ou simplement espérance) d'une distribution. Il n'existe pour le moment aucune procédure permettant de calculer cette moyenne dans le domaine des chaînes, car ceci nécessite au préalable de préciser ce que sont les probabilités et donc la notion de distribution dans ce domaine, ce qui est l'objet de cette étude.

2.2 Médiane

Sous acceptation d'une analogie géométrique, la médiane d'un ensemble fini S de chaînes est définie dans [7] comme la chaîne de A^* minimisant la somme de ses distances d'édition aux chaînes de S : $\text{médiane}(S) = \arg \min_{X \in A^*} \sum_{Y \in S} d(X, Y)$.

Cette approche se substitue le plus souvent à l'approche probabiliste en l'absence de notion de distribution. Pour remarque, toujours selon l'analogie géométrique, on obtient une définition équivalente de la chaîne moyenne comme étant celle minimisant la somme des carrés de ses distances aux chaînes de S . Cependant certains préfèrent toujours employer le terme médiane, avec simplement un changement de distance, passant de la distance d'édition traditionnelle à une distance correspondant au carré de cette dernière, et semblant donner de légèrement meilleurs résultats

en terme de classification [11].

Dans tous les cas cette définition ne garantit pas la propriété d'unicité, à savoir qu'il peut exister plusieurs chaînes validant la contrainte de minimisation, et en outre le calcul d'une telle médiane se heurte à de plus ou moins solides obstacles. D'une part, il a été prouvé comme étant un problème NP-difficile [2, 13], et d'autre part, il nécessite la mise en place de la fonction de coût $c()$ sur A .

Pour réduire la complexité des calculs, de nombreuses méthodes de résolution s'appuyant sur une réduction de l'espace de recherche ont été proposées : recherche gloutonne [8], locale [10], dynamique [4], génétique [5].

Pour ce qui est de la mise en place de $c()$, il existe trois solutions. La première consiste à fixer le coût de transformation d'une lettre en une autre (différente) à une constante, cette méthode étant utilisée dès qu'il n'y a aucune connaissance *a priori* quant à la dissimilarité entre les lettres. La deuxième solution consiste à s'appuyer cette fois-ci sur une telle connaissance, pouvant être fournie par un expert du domaine, ou déduite de manière *ad hoc* (distance "géographique" sur un clavier de caractères, alphabet numérique...). Enfin, la dernière solution est celle de l'apprentissage automatique de $c()$, comme effectué dans [12].

2.3 Écart-type

L'écart-type d'une loi de probabilité L est défini comme l'espérance des écarts à l'espérance de L . Pour le cas spécifique des chaînes, un écart à l'espérance est traduit par une chaîne d'édition consommant l'espérance (chaîne définie sur l'alphabet $(A \cup \{\lambda\})^2$). Comme nous l'avons vu, le concept de chaîne espérance n'est pas encore utilisable, et il est par conséquent d'usage d'approximer l'écart-type d'un ensemble S de chaînes par la médiane des écarts à sa médiane, toujours sous acceptation d'une analogie géométrique, et tel qu'initialement proposé dans [6]. Cependant, même dans les situations où l'analogie géométrique pourrait être supposée valable, l'approximation de l'espérance par la médiane d'un ensemble n'est pertinente que dans le cas où l'ensemble est supposé être un échantillon généré sous contrôle d'une distribution symétrique et unimodale. Finalement, cette approximation hérite doublement de tous les inconvénients de la médiane (cf. 2.2).

3 Lois statistiques de chaînes discrètes

3.1 Loi uniforme

La loi uniforme sur un ensemble fini S d'éléments est celle qui associe la même probabilité $|S|^{-1}$ à tout élément de S . Nous proposons ici une définition de la loi uniforme sur l'ensemble A^n , pour tout n donné.

Proposition 1 Une chaîne X est dite réalisation d'une loi uniforme sur A^n si chacune de ses lettres est indépendamment réalisation d'une loi uniforme sur A .

Preuve La probabilité de toute lettre de X étant indépendamment fixée selon une loi uniforme sur A , nous avons : $\forall X_i, p(X_i) = |A|^{-1}$. Dès lors, nous avons : $p(X) = \prod_{i=1}^n p(X_i) = \prod_{i=1}^n |A|^{-1} = |A|^{-n}$. Or $|A^n|^{-1} = (|A|^n)^{-1} = |A|^{-n}$.

Cette définition donne aisément lieu à la mise en place de processus génératifs et prédictifs, respectivement détaillés par les algorithmes 1 et 2. L'algorithme génératif applique le processus itératif sous-entendu par la proposition 1, de complexité $O(n)$.

Données : Un entier positif n Résultat : Une chaîne X générée selon une loi uniforme sur A^n début $X \leftarrow \lambda;$ pour $i \leftarrow 1$ à n faire $l \leftarrow$ tirage aléatoire uniforme d'une lettre de A; $X \leftarrow X.l;$ fin retourner X; fin
--

Algorithme 1 : Génération d'une chaîne sous contrôle d'une loi uniforme

Données : Une chaîne X sur A, un entier positif n Résultat : La probabilité de X suivant une loi uniforme sur A^n début si $X \neq n$ alors retourner 0.0; retourner $A ^{-n}$; fin
--

Algorithme 2 : Prédiction de la probabilité d'une chaîne selon une loi uniforme

3.2 Loi normale

Il n'est pas inapproprié de rapprocher les deux concepts mathématiques de chaîne et de vecteur. En effet ceux-ci partagent un grand nombre de points communs, comme par exemple leur considération vis à vis des notions importantes d'ordre et de duplication. Si l'on échange la position de deux lettres dans une chaîne, on obtient une chaîne différente, et il est possible d'avoir deux lettres égales à deux positions différentes d'une chaîne. Or ces deux faits se retrouvent dans le cas des coordonnées de vecteurs. La seule réelle différence entre chaîne et vecteur réside dans l'utilisation que l'on en fait. En effet, le concept de vecteur s'adresse plus classiquement aux espaces numériques, discrets ou continus, avec lesquels il est muni de la définition d'un grand nombre d'opérateurs particuliers (addition, soustraction, produit...). D'un autre côté, en insérant autant de lettres λ que voulu à n'importe quelles positions où il est possible de le faire dans une chaîne X (impossible

aux positions supérieures à $|X| + 1$), on obtient X . Suivant ce raisonnement, nous proposons de définir une loi normale de chaînes sur $A \cup \{\lambda\}$, et donc sur A , par une loi normale multivariée de vecteurs constitués de lettres de $A \cup \{\lambda\}$.

Or la loi normale multivariée est paramétrée par son espérance et sa matrice de covariance, et le concept de covariance nous est pour le moment inconnu dans l'espace des chaînes. Cependant, nous connaissons la traduction du concept d'écart-type dans ce domaine (cf. 2.3 : chaîne d'opérations d'édition consommant l'espérance). Dans le cas précis où est supposé connue l'unique donnée de l'écart-type (et par là même de l'espérance), et en supposant par conséquent l'indépendance deux-à-deux de chaque lettre composant une chaîne, nous proposons cette traduction du concept de loi normale au domaine des chaînes :

Proposition 2 *Une chaîne X est dite réalisation d'une loi normale sur A^* , d'écart-type $\sigma = M_1 \rightarrow D_1 \dots M_n \rightarrow D_n$ (d'espérance $M = M_1 \dots M_n$), si X est égale à la concaténation de n lettres X_i respectivement et indépendamment réalisation d'une loi normale sur $A \cup \{\lambda\}$, d'écart-type $M_i \rightarrow D_i$ (d'espérance M_i).*

Preuve Soient $x_1, \dots, x_n$, n variables aléatoires normales, respectivement d'espérance μ_i et écart-type σ_i , et indépendantes deux-à-deux. D'après la théorie statistique multivariée, nous savons que le vecteur $[x_1, \dots, x_n]$ suit une loi normale (multivariée) d'espérance $[\mu_1, \dots, \mu_n]$, et de matrice de covariance diagonale (i.e. une matrice d'indépendance) avec vecteur de variance $[\sigma_1^2, \dots, \sigma_n^2]$, et donc vecteur d'écart-type $[\sigma_1, \dots, \sigma_n]$.

Néanmoins nous ne pouvons pas utiliser cette définition telle quelle. En effet, une fois donnés l'espérance et l'écart-type d'une loi normale (univariée), sa densité de probabilité en tout point x dépend de la valeur de l'écart de x à l'espérance μ , écart étant traduit par exemple dans le domaine classique numérique par $|x - \mu|$. C'est ici que nous avons recours à la fonction de coût $c()$ sur A , $c(a \rightarrow b)$ traduisant effectivement une notion d'écart de b à a , en tant que coût de la transformation de a en b ($(a, b) \in (A \cup \{\lambda\})^2$). Ce raisonnement nous permet d'affiner notre définition de la manière suivante :

Proposition 3 *Une chaîne X est dite réalisation d'une loi normale sur A^* , d'écart-type $\sigma = M_1 \rightarrow D_1 \dots M_n \rightarrow D_n$ (d'espérance $M = M_1 \dots M_n$), si X est égale à la concaténation de n lettres X_i respectivement et indépendamment réalisation d'une distribution L_i , telle que la probabilité de X_i selon L_i soit égale à la probabilité de $c(M_i \rightarrow X_i)$ selon une loi normale sur l'ensemble des coûts d'édition de M_i par une lettre de $A \cup \{\lambda\}$, d'espérance $c(M_i \rightarrow M_i) = 0$ et d'écart-type $c(M_i \rightarrow D_i)$.*

Par l'expression "loi normale sur l'ensemble des coûts d'édition de M_i par une lettre de $A \cup \{\lambda\}$ ", nous sous-entendons le recours à une distribution discrète instanciée de manière à respecter un processus gaussien sur l'ensemble considéré. Dans notre cas cette discrétisation, car il

s'agit de cela, est aisément productible. En effet, A (et donc $A \cup \{\lambda\}$) étant fini, L_i correspond simplement à la distribution associant une probabilité à toute lettre $l \in A \cup \{\lambda\}$ comme étant égale à la densité de probabilité de $c(M_i \rightarrow l)$ selon la loi normale associée (d'espérance 0 et d'écart-type $c(M_i \rightarrow D_i)$), normalisée par la somme de ces densités sur l'ensemble des $c(M_i \rightarrow b)$, avec $b \in A \cup \{\lambda\}$.

Preuve La preuve de la proposition 3 découle de celle de la proposition 2, si l'on accepte, d'une part, la possibilité de traduction de processus gaussien dans le domaine discret, voire fini, et d'autre part, la justesse de la méthode utilisée pour instancier cette discrétisation dans le cas fini (méthode dont un exemple est donné ci-dessus par la construction de L_i).

Les processus génératifs et prédictifs sont respectivement détaillés par les algorithmes 3 et 4. L'algorithme génératif applique le processus itératif sous-entendu par la proposition 3, de complexité $O(|\sigma|)$. Pour ce qui est de l'algorithme prédictif, la difficulté est autrement plus élevée. En effet, une chaîne X étant inchangée avec l'insertion d'autant de λ que voulu à n'importe quelles positions où il est possible de le faire, le nombre de possibilités de production de X est égal au nombre de combinaisons de $|X|$ objets parmi $|\sigma|$. Pour résoudre ce problème, nous avons recours à une méthode de programmation dynamique de complexité temporelle $O(|X| \times |\sigma|^3)$ et spatiale $O(|\sigma|)$. À chaque itération j de la boucle centrale principale de l'algorithme 4, le tableau $PP[i]$ contient la probabilité de production du préfixe de longueur j de X , selon une loi normale d'écart-type le préfixe de longueur i de σ . L'initialisation est effectuée de manière à prendre en considération le fait qu'il est improbable que la première lettre de X soit produite par la loi normale d'écart-type σ_i avec $i > |\sigma| - |X| + 1$. Il est en effet dans ce cas improbable de produire les $|X| - 1$ lettres restantes de $|X|$, différentes de λ , avec la succession d'exécutions unitaires de $|\sigma| - i < |X| - 1$ lois normales de lettres. La finalisation est effectuée de manière à tenir compte de toutes les possibilités de fin de production de X , i.e. les probabilités de fin de production de X pour chaque position de σ , éventuellement complétées par des successions de production de λ . Enfin la boucle i contenue dans la boucle j principale met à jour la probabilité de production de X_j en position i de σ , de manière à tenir compte de l'ensemble des possibilités de production de X_{j-1} en position k inférieure à i , éventuellement complétées par des successions de production de λ entre les positions supérieures à k et inférieures à i . Le cas spécial où $X = \lambda$ ne rentre pas dans ce schéma, et est donc traité à part.

4 Expérimentations

4.1 Alphabet et fonction de coût

L'alphabet A que nous utilisons pour les expérimentations est un sous-ensemble de l'alphabet latin basique (ou ASCII, intervalle [U+0000 ; U+007F] dans la norme Unicode), dans lequel ont été retirés les caractères de contrôle

Données : Une chaîne X sur A , une chaîne d'édition $\sigma = M_1 \rightarrow D_1 \dots M_n \rightarrow D_n$ sur A

Résultat : La probabilité de X suivant une loi normale sur A^* , d'écart-type σ

début

/* Cas spécial $X = \lambda$ */

si $X = \lambda$ **alors retourner** $p(\lambda)$; // algorithme 5

/* Initialisation : production de la première lettre de X */

pour $i \leftarrow 1$ **à** $|\sigma| - |X| + 1$ **faire**

 // Production de X_1 en position i de σ

$PP[i] \leftarrow p(X_1 \text{ à la position } i)$; // algorithme 6

 // Complétion à gauche avec la production de λ

pour $l \leftarrow 1$ **à** $i - 1$ **faire**

$PP[i] \leftarrow PP[i] \times p(\lambda \text{ à la position } l)$; // algorithme 6

fin

fin

pour $i \leftarrow |\sigma| - |X| + 2$ **à** $|\sigma|$ **faire**

 // Impossible de commencer la production à cette position i de σ

$PP[i] \leftarrow 0.0$;

fin

/* Itérations : production des autres lettres de X */

pour $j \leftarrow 2$ **à** $|X|$ **faire**

 // Possibilités de production de X_j en position i de σ

pour $i \leftarrow |\sigma|$ **à** 1 **faire**

$PP[i] \leftarrow 0$; // initialisation

 // Possibilités de production de X_{j-1} en position $k < i$ de σ

pour $k \leftarrow 1$ **à** $i - 1$ **faire**

 // Production de X_{j-1} en position k de σ

$PLUS \leftarrow PP[k]$;

 // Complétion avec la production de λ en positions $(k + 1) \leq l \leq (i - 1)$

pour $l \leftarrow k + 1$ **à** $i - 1$ **faire**

$PLUS \leftarrow PLUS \times p(\lambda \text{ à la position } l)$; // algorithme 6

fin

 // Sommation des probabilités de production de X_{j-1}

$PP[i] \leftarrow PP[i] + PLUS$;

fin

 // Production de X_j en position i de σ

$PP[i] \leftarrow PP[i] \times p(X_j \text{ à la position } i)$; // algorithme 6

fin

fin

/* Finalisation des possibilités de production de X */

$P \leftarrow 0.0$;

pour $i \leftarrow 1$ **à** $|\sigma|$ **faire**

 // La production de X s'est stoppée à cette position i de σ

$PLUS \leftarrow PP[i]$;

 // Complétion à droite avec la production de λ

pour $l \leftarrow i + 1$ **à** $|\sigma|$ **faire**

$PLUS \leftarrow PLUS \times p(\lambda \text{ à la position } l)$; // algorithme 6

fin

 // Sommation des probabilités de production de X

$P \leftarrow P + PLUS$;

fin

retourner P ;

fin

Algorithme 4 – Prédiction de la probabilité d'une chaîne selon une loi normale

Données : Une chaîne d'édition
 $\sigma = M_1 \rightarrow D_1 \dots M_n \rightarrow D_n$ sur A
Résultat : Une chaîne X générée selon une loi normale sur A^* , d'écart-type σ

début
 $X \leftarrow \lambda$;
pour $i \leftarrow 1$ **à** n **faire**
 $g \leftarrow$ loi normale d'espérance 0 et d'écart-type $c(M_i \rightarrow D_i)$;
 $l \leftarrow$ tirage aléatoire d'une lettre de $A \cup \{\lambda\}$ selon la distribution suivante :

$$\forall a \in A \cup \{\lambda\}, p(a) = \frac{g(c(M_i \rightarrow a))}{\sum_{b \in A \cup \{\lambda\}} g(c(M_i \rightarrow b))}$$

 $X \leftarrow X.l$;
fin
retourner X ;
fin

Algorithme 3 : Génération d'une chaîne sous contrôle d'une loi normale

Données : Une chaîne d'édition
 $\sigma = M_1 \rightarrow D_1 \dots M_n \rightarrow D_n$ sur A
Résultat : La probabilité de λ suivant une loi normale sur A^* , d'écart-type σ

début
 $P \leftarrow 1.0$;
pour $i \leftarrow 1$ **à** n **faire**
 $P \leftarrow P \times p(\lambda \text{ à la position } i)$; // algorithme 6
fin
retourner P ;
fin

Algorithme 5 : Prédiction de la probabilité de λ selon une loi normale

Données : Une lettre $l \in A \cup \{\lambda\}$, une opération d'édition
 $o = m \rightarrow d$ sur A
Résultat : La probabilité de l suivant une loi normale sur $A \cup \{\lambda\}$, d'écart-type o

début
 $g \leftarrow$ loi normale d'espérance 0 et d'écart-type $c(o)$;
 $p \leftarrow \frac{g(c(m \rightarrow l))}{\sum_{b \in A \cup \{\lambda\}} g(c(m \rightarrow b))}$;
retourner p ;
fin

Algorithme 6 : Prédiction de la probabilité d'une lettre selon une loi normale

			•	•	•		
	•					•	
	•					•	
	•					•	
	•					•	
		•	•	•			

TAB. 1 – Codage visuel du caractère O par une matrice binaire de dimensions 7×7

C0 (intervalle [U+0000 ; U+001F], ainsi que U+007F). Les 95 lettres de type caractère de A sont donc accessibles par n'importe quel clavier utilisant l'alphabet latin.

Quant à la fonction de coût $c()$, elle est purement visuelle : $\forall (a, b) \in A^2, c(a \rightarrow b) = c(b \rightarrow a)$, et il s'agit de la distance de Hamming [3] entre les codages visuels respectifs de a et b selon une matrice binaire de dimensions 7×7 . Un exemple de tel codage est donné dans le tableau 1. Cette méthode de codage peut être appliquée à tout alphabet de lettres traduites sous forme de matrices binaires, comme c'est le cas par exemple pour des symboles codés par des matrices 3×3 , proposé dans [14] pour le codage d'images à base de contraste, et donnant de bons résultats en indexation et recherche dans des bases d'images naturelles. Les coûts des insertions et suppressions sont quant à eux tous fixés à $+\infty$, résultant en le fait que la probabilité de toute chaîne X selon une loi normale d'écart-type σ (cf. 3.2) est nulle dans le cas où $|X| \neq |\sigma|$. Ce choix est motivé par le désir de forcer notre processus de génération de chaînes gaussiennes (algorithme 3) à ne produire que des chaînes de même longueur ($|\sigma|$), ce qui est toujours le cas pour la loi uniforme (cf. 3.1). De ce fait, les mesures d'adéquations entre échantillons et lois (cf. 4.2) ne s'effectuent que sur des échantillons de chaînes de même longueur. Le contraire aurait été très pénalisant pour ce qui est de l'adéquation de la loi uniforme à l'échantillon gaussien, et les mesures n'auraient donc pas été fiables. Ce choix implique également un aspect positif en terme de performance en prédiction de probabilité d'une chaîne selon la loi normale. Le processus général (algorithme 4) est dans ce cas simplifié à un processus itératif, détaillé dans l'algorithme 7, aligné sur celui de l'algorithme génératif (algorithme 3), et de complexité $O(|\sigma|)$. En effet, il est dans ce cas inutile de conserver la méthode dynamique proposée par l'algorithme général, car il n'y a qu'une seule combinaison possible de $|X| = |\sigma|$ objets parmi $|\sigma|$.

4.2 Validation

Nous souhaitons mesurer la qualité de nos modèles de lois présentés, en terme de contrôle statistique de génération d'ensembles de chaînes discrètes. Pour ce faire, nous mettons en place le protocole expérimental suivant : des échantillons de chaînes sont créés *via* l'utilisation des algorithmes génératifs, et nous mesurons la qualité des algorithmes prédictifs quant à la modélisation de la distribution

Données : Une chaîne X sur A , une chaîne de substitutions $\sigma = M_1 \rightarrow D_1 \dots M_n \rightarrow D_n$ sur A

Résultat : La probabilité de X suivant une loi normale sur A^* , d'écart-type σ

début

 si $|X| \neq n$ alors retourner 0.0;

$P \leftarrow 1.0$;

pour $i \leftarrow 1$ à n **faire**

$g \leftarrow$ loi normale d'espérance 0 et d'écart-type $c(M_i \rightarrow D_i)$;

$p \leftarrow \frac{g(c(M_i \rightarrow X_i))}{\sum_{b \in A} g(c(M_i \rightarrow b))}$;

$P \leftarrow P \times p$;

fin

retourner P ;

fin

Algorithme 7 : Prédiction de la probabilité d'une chaîne selon une loi normale, dans le cas spécifique où la fonction de coût interdit les insertions et suppressions

sous-jacente à chaque échantillon.

Ces mesures de qualité de modélisation sont effectuées via des calculs de *perplexité*. La perplexité permet d'estimer la distance statistique entre un échantillon S de données et une distribution $\hat{L}$ proposée comme hypothèse de distribution génératrice de S . Elle correspond à l'inverse de la moyenne géométrique de la probabilité selon $\hat{L}$ d'un élément de S : $\text{Perp}(S|\hat{L}) = \left(\prod_{x \in S} p(x|\hat{L}) \right)^{-1/|S|}$, avec $|S|$ la taille de S . Plus cette valeur est faible, et plus $\hat{L}$ est supposée statistiquement proche de la distribution L réellement sous-jacente à S . Il est en pratique d'usage de calculer cette valeur par un passage au logarithme, de manière à éviter un dépassement de capacité de représentation des réels, pouvant survenir dans le cas d'un grand nombre de données à faibles probabilités : $\text{Perp}(S|\hat{L}) = \exp \left((-1/|S|) \sum_{x \in S} \ln p(x|\hat{L}) \right)$.

Pour des raisons de visibilité et d'interprétabilité des résultats, les mesures présentées correspondent en effet au rapport $\text{Perp}(S|\bar{L})/\text{Perp}(S|L)$, avec $\bar{L}$ le modèle non sous-jacent à la génération des données, et L celui effectivement responsable de celle-ci. De fait, une telle valeur R supérieure (resp. inférieure) à 1 signifie que l'adéquation entre L et S est R fois supérieure (resp. $1/R$ fois inférieure) à celle entre $\bar{L}$ et S , en ce sens où il est en moyenne R fois plus (resp. moins) vraisemblable qu'une chaîne de S ait été générée selon L plutôt que selon $\bar{L}$.

Les figures 1 et 2 présentent les évolutions de ces mesures en fonction du nombre de chaînes générées, dans les cas respectifs où L est une loi uniforme (U) et $\bar{L}$ une loi normale (G), et l'inverse. Les chaînes générées sont de longueur 100, et les écarts-types des lois normales sont initialisés aléatoirement.

Nous pouvons constater les résultats attendus, à savoir, d'une part, la bien meilleure adéquation à S du modèle

l'ayant généré, en comparaison avec la distribution qui en est indépendante, et, d'autre part, la variabilité de cette adéquation pour des échantillon de faible taille (en comparaison avec le nombre 95^{100} de chaînes à probabilité non nulle), corrigée et contrôlée dans le cas d'échantillons plus représentatifs. Les valeurs affichées ne sont en elles mêmes pas très importantes car ces courbes ne représentent bien entendu que quelques expérimentations parmi tant d'autres possibles, dépendantes des paramètres utilisés.

4.3 Estimation d'écart-type d'une loi normale

La loi uniforme étant non paramétrée, cette partie ne concerne effectivement que la loi normale.

Nous estimons l'écart-type d'une loi normale G par une chaîne d'édition $\widehat{D}_{\min}$ impliquant la minimisation de la perplexité, selon la loi normale $\widehat{G}_{\min}$ d'écart-type $\widehat{D}_{\min}$, d'un échantillon S de chaînes généré selon G : $\widehat{D}_{\min} = \arg \min_{\widehat{D} \in ((A \cup \{\lambda\})^*)} \text{Perp}(S|\widehat{G})$, avec $\widehat{G}$ la loi normale d'écart-type $\widehat{D}$. Nous intuitions malheureusement que ce problème d'optimisation est NP-difficile. Cependant, dans notre cas spécifique où la fonction de coût rend improbable toute insertion ou suppression (cf. 4.1), et où donc nous savons que toutes les lettres à la même position dans les chaînes de S ont été générées selon la même loi normale (de lettres), nous avons : $\widehat{D}_{\min} = \widehat{d}_{1 \min} \dots \widehat{d}_{|\sigma| \min}$, avec, $\forall 1 \leq i \leq \sigma$, $\widehat{d}_{i \min} = \arg \min_{\widehat{d}_i \in A^2} \text{Perp}(S_i|\widehat{G}_i)$, S_i l'ensemble (le multi-ensemble) formé de toutes les lettres à la position i dans une chaîne de S , et $\widehat{G}_i$ la loi normale d'écart-type $\widehat{d}_i$. Il nous est donc effectivement possible de trouver $\widehat{D}_{\min}$ en $|A|^2 \times |\sigma|$ calculs de perplexité d'un échantillon S_i de taille $O(|S|)$, soit un processus global de complexité $O(|A|^2 \times |\sigma| \times |S|)$.

Deux types d'expérimentations sont menées, correspondant à deux niveaux d'estimation d'écart-type : avec espérance connue, et avec espérance inconnue. Dans le premier cas seule la chaîne produite par l'écart-type est à estimer, la chaîne consommée étant bien entendue fixée à l'espérance. Par contre dans le deuxième cas la chaîne consommée par l'écart-type est également à estimer, constituant par là même une estimation de l'espérance.

La figure 3 présente les évolutions des qualités des modèles estimés (G estimé) en fonction du nombre de chaînes générées. Pour les mêmes raisons que précédemment (cf. 4.2), celles-ci sont effectivement normalisées par celles du modèle à estimer (G réel). Ces évolutions sont également comparées avec celle, présentée comme repère en figure 4 pour des raisons de visibilité, de la qualité d'un modèle de loi normale dont l'écart-type est initialisé aléatoirement. Les chaînes générées sont de longueur 100, et l'écart-type de la loi normale à estimer est initialisé aléatoirement.

Les résultats sont positifs, à savoir qu'espérance connue ou inconnue, la loi estimée est au moins aussi proche de l'échantillon S que celle l'ayant généré, parfois même plus proche, ceci étant simplement du au fait que S , lorsque

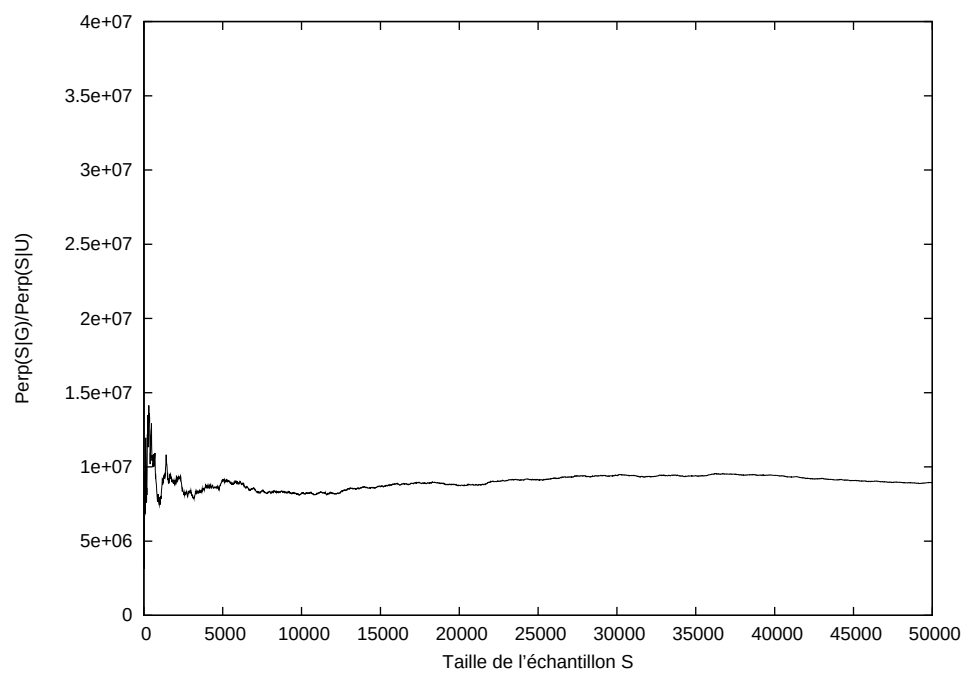


FIG. 1 – Évolution du rapport de perplexité en fonction de la taille de l'échantillon généré selon une loi uniforme (algorithme 1)

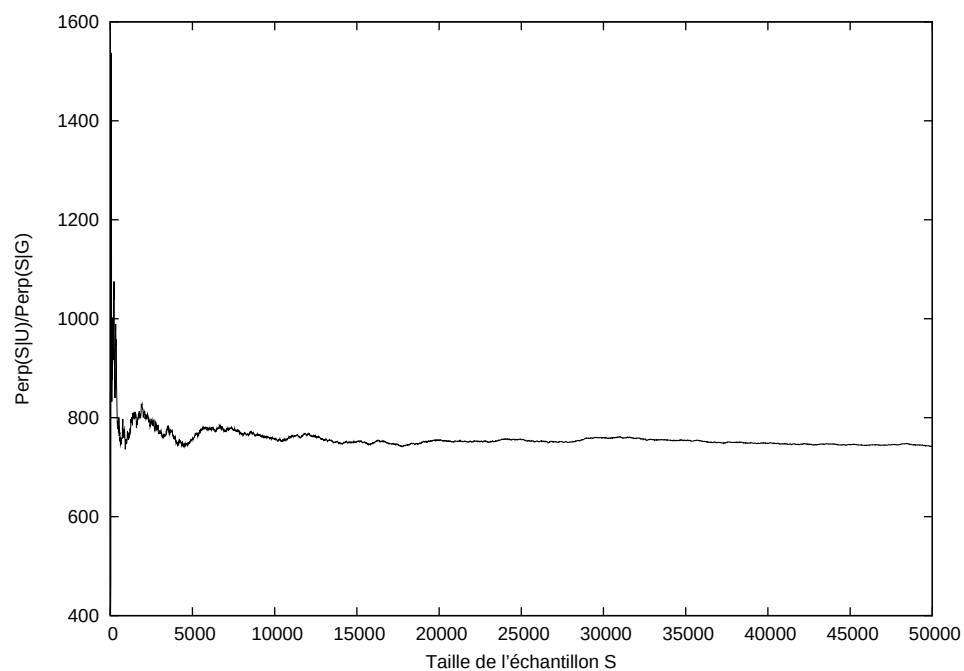


FIG. 2 – Évolution du rapport de perplexité en fonction de la taille de l'échantillon généré selon une loi normale (algorithme 3)

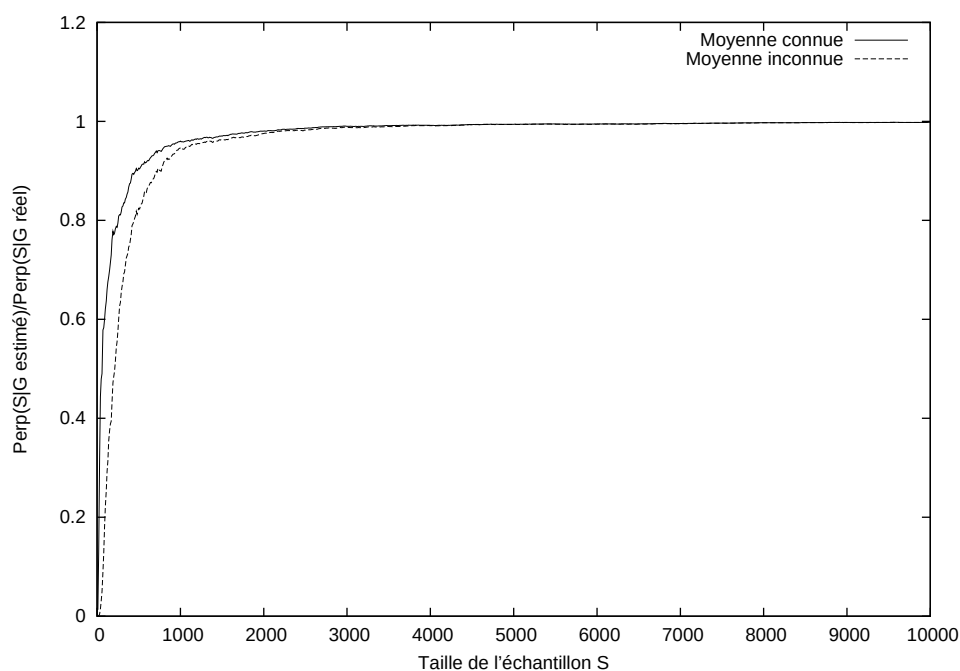


FIG. 3 – Évolution des rapports de perplexité des modèles estimés en fonction de la taille de l'échantillon généré selon une loi normale

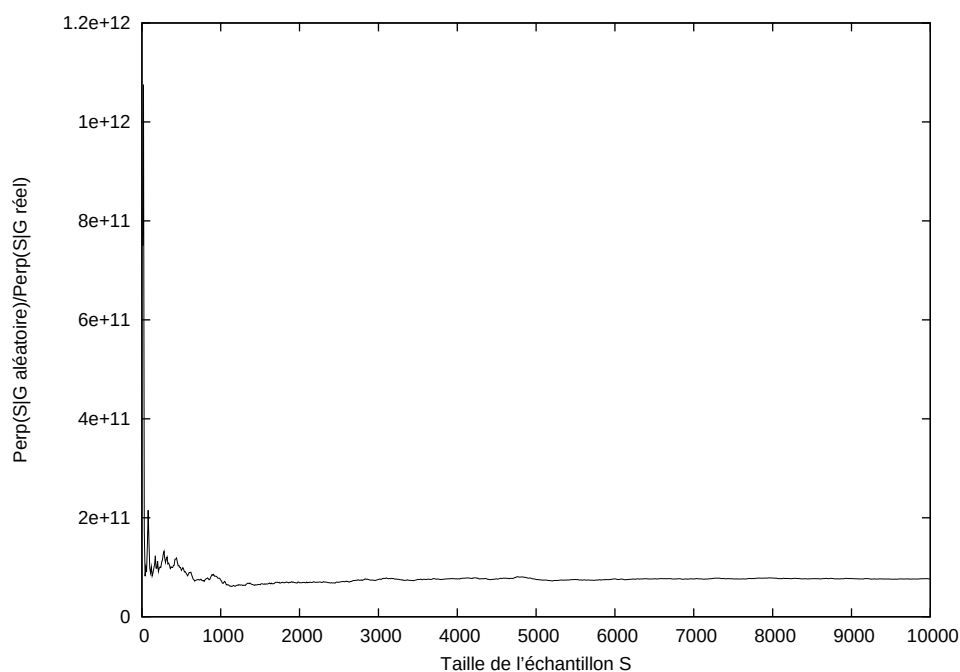


FIG. 4 – Évolution des rapports de perplexité du modèle normal d'écart-type aléatoire en fonction de la taille de l'échantillon généré selon une loi normale

sa taille est faible, ne peut assurer une bonne représentativité du processus aléatoire à partir duquel il a été généré (cf. 4.2). En tous cas nous pouvons dire que, sous couvert de représentativité de S , les estimations proposées sont proches du paramètre réel, et ceci sans besoin de mesurer cette proximité par une quelconque distance d'édition entre chaînes d'édition, qui aurait nécessité la justification d'une fonction de coût d'édition sur $(A \cup \{\lambda\})^2$. De plus cette méthode est totalement déterministe pour un échantillon donné. Enfin, la piètre qualité de l'estimation aléatoire témoigne de la difficulté de la tâche et donc également de la qualité des estimations fournies.

Une autre observation, s'il en est encore besoin, est celle du manque de fiabilité de toute estimation à base de maximisation de la vraisemblance d'un échantillon S de faible taille². Il serait alors intéressant d'appliquer une autre mesure de distance statistique, plus adaptée à de tels cas. Nous pensons à la distance du χ^2 , obtenue par comparaison des effectifs observés et théoriques (*i.e.* qui devraient être observés dans le cas d'une estimation parfaite) de S regroupé au sein k classes, dont l'union doit constituer une partition du domaine d'existence de la distribution estimée. Ce processus nécessite de définir ce que constituent ces classes dans l'espace des chaînes, et la classification engendrée doit tenir compte de l'aspect structurel des chaînes pour que les résultats soient supposés fiables. Il serait alors intéressant de pouvoir adapter k à $|S|$, de manière à établir un bon compromis entre validité du processus (augmentant avec le nombre minimal d'observations dans une classe) et fiabilité des résultats (augmentant avec k). Ceci fait actuellement l'objet d'une attention particulière de notre part.

5 Conclusion

Nous avons proposé dans cet article le moyen de contrôler statistiquement la génération d'ensembles de chaînes discrètes par des lois normales ou uniformes, ainsi qu'une méthode d'estimation de paramètres de distributions de chaînes s'appuyant sur une recherche optimisant la distance d'un échantillon à la distribution proposée. Cette méthode n'est pour le moment qu'appliquée à l'estimation de l'écart-type d'une loi normale, avec des résultats encourageants, et pourrait aisément être étendue à d'autres distributions paramétrées, nécessitant la formalisation préalable de l'extension de ces lois à l'espace des chaînes discrètes. Ce travail ne constitue donc qu'une première avancée et de nombreuses perspectives en découlent, parmi lesquelles : l'étude des propriétés de la loi normale, la formalisation d'autres lois statistiques dans le domaine des chaînes discrètes, l'utilisation d'une distance statistique plus appropriée pour estimer les paramètres d'une distribution à partir d'un échantillon de faible taille, l'application à la classification de données réelles codées par des chaînes discrètes, la généralisation aux arbres et graphes discrets... L'objectif étant la formalisation d'un cadre statistique général permettant de réconcilier les paradigmes numériques

et symboliques, et appliqué plus particulièrement à la reconnaissance de formes structurelle.

Références

- [1] H. Bunke, X. Jiang, K. Abegglen, and A. Kandel. On the weighted mean of a pair of strings. *Pattern Analysis and Applications*, 5(1) :23–30, 2002.
- [2] C. de la Higuera and F. Casacuberta. Topology of strings : median string is NP-complete. *Theoretical Computer Science*, 230(1-2) :39–48, 2000.
- [3] R. W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2) :147–160, 1950.
- [4] X. Jiang, K. Abegglen, H. Bunke, and J. Csirik. Dynamic computation of generalised median strings. *Pattern Analysis and Applications*, 6(3) :185–193, 2003.
- [5] X. Jiang, H. Bunke, and J. Csirik. Median strings : a review. In *Data Mining in Time Series Databases*, pages 173–192. World Scientific, 2004.
- [6] J.-M. Jolion. The deviation of a set of strings. *Pattern Analysis and Applications*, 6(3) :224–231, 2003.
- [7] T. Kohonen. Median strings. *Pattern Recognition Letters*, 3(5) :309–313, 1985.
- [8] F. Kruzsliz. Improved greedy algorithm for computing approximate median strings. *Acta Cybernetica*, 14(2) :331–340, 1999.
- [9] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8) :707–710, 1966.
- [10] C. D. Martínez-Hinarejos, A. Juan, and F. Casacuberta. Use of median string for classification. In *ICPR*, pages 2903–2906, Barcelona, Spain, September 2000.
- [11] C. D. Martínez, A. Juan, and F. Casacuberta. Improving classification using median string and NN rules. In *Spanish Symposium on Pattern Recognition and Image Analysis*, volume 2, pages 391–395, Benicàssim, Spain, May 2001.
- [12] J. Oncina and M. Sebban. Learning stochastic edit distance : application in handwritten character recognition. *Pattern Recognition*, 39(9) :1575–1587, 2006.
- [13] J. S. Sim and K. Park. The consensus string problem for a metric is NP-complete. *Journal of Discrete Algorithms*, 1(1) :111–117, 2003.
- [14] I. Simand and J.-M. Jolion. Représentation d'images par chaînes de symboles : application à la recherche par le contenu. In *GRETSI : Traitement du signal et des images*, volume 2, pages 925–928, 2005.
- [15] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1) :168–173, 1974.

²Minimiser la perplexité revient à maximiser la vraisemblance