Chiffrement par substitution : rappels

Ce procédé de chiffrement repose sur la notion de substitution. De façon générale, une substitution d'un alphabet A est une bijection de A dans lui-même. Ainsi une subsitution transforme chaque lettre de l'alphabet A en une autre lettre du même alphabet. Si π est une substitution de A, alors il existe une unique substitution σ de A pour laquelle on a quel que soit $a \in A$, $\pi(\sigma(a)) = a$ et $\sigma(\pi(a)) = a$. La substitution σ est l'**inverse** de π, généralement notée π^{-1}.

Pour le chiffrement par substitution, les messages clairs et chiffrés sont des lettres d'un alphabet A (par exemple, l'alphabet latin usuel). Les clefs secrètes sont choisies parmi les substitutions de A. Soient alors π une substitution de A et $M \in A$ une lettre. On a alors :

$$E_\pi(M) := \pi(M) \ .$$

Soit alors $C \in A$ le chiffré correspondant, $C = \pi(M)$. Alors on a également

$$D_\pi(C) := \pi^{-1}(C) = M \ .$$

Pour chiffrer une suite de lettres $M_1 M_2 \dots M_n$ prises dans l'alphabet A, on calcule

$$C = E_\pi(M_1)E_\pi(M_2)\dots E_\pi(M_n) \ .$$

Posons $C_i := E_\pi(M_i)$ pour $i = 1, \dots, n$. Pour déchiffrer $C = C_1 C_2 \dots C_n$, on calcule

$$D_\pi(C_1)D_\pi(C_2)\dots D_\pi(C_n) = M_1 M_2 \dots M_n = M \ .$$

Énoncé

Dans cet exercice, on s'intéresse à une technique de cryptanalyse permettant de casser un procédé de chiffrement par substitution. Cette technique est basée sur l'analyse des fréquences d'occurrence des lettres dans un texte écrit dans une langue donnée (par exemple, l'anglais ou le français). Dans le cas présent, on effectue une hypothèse simplificatrice : on suppose que le texte clair est **un message rédigé en anglais sans ponctuations ni espaces**.

Plusieurs personnes ont estimé la probabilité d'apparition des vingt-six lettres de l'alphabet en faisant des statistiques sur de nombreux romans, magazines et journaux quotidiens écrits en anglais. Les estimations suivantes sur la langue anglaise ont été obtenues par Beker et Piper.

Fréquences d'occurrences des lettres dans les textes écrits en anglais (Beker & Piper)

lettre	proba	lettre	proba
a	$0,082$	n	$0,067$
b	$0,015$	o	$0,075$
c	$0,028$	p	$0,019$
d	$0,043$	q	$0,001$
e	$0,127$	r	$0,060$
f	$0,022$	s	$0,063$
g	$0,020$	t	$0,091$
h	$0,061$	u	$0,028$
i	$0,070$	v	$0,010$
j	$0,002$	w	$0,023$
k	$0,008$	x	$0,001$
l	$0,040$	y	$0,020$
m	$0,024$	z	$0,001$

À partir de ces résultats, Beker et Piper ont classé les 26 lettres en cinq groupes :

1. "e", ayant pour probabilité d'environ 0,120 ;
2. "t", "a", "o", "i", "n", "s", "h" et "r", ayant une probabilité entre 0,06 et 0,09 ;
3. "d" et "l" ayant une probabilité d'environ 0,04 ;
4. "c", "u", "m", "w", "f", "g", "y", "p" et "b" ayant une probabilité entre 0,015 et 0,028 ;
5. "v", "k", "j", "x", "q" et "z" ayant une probabilité inférieure à 0,01.

Il peut être utile également d'étudier la probabilité d'occurrence de deux ou trois lettres consécutives, appelés **digrammes** ou **trigrammes**. En anglais, les trente digrammes les plus fréquents sont (par ordre décroissant) "th", "he", "in", "er", "an", "re", "ed", "on", "es", "st", "en", "at", "to", "nt", "ha", "nd", "ou", "ea", "ng", "as", "or", "ti", "is", "et", "it", "ar", "te", "se", "hi" et "of". Les douze trigrammes les plus fréquents sont (par ordre décroissant) "the", "ing", "and", "her", "ere", "ent", "tha", "nth", "was", "eth", "for" et "dth".

Maintenant que l'on sait que les lettres n'apparaissent pas toutes avec la même fréquence, on veut tirer profit de ce biais statistique afin de cryptanalyser le procédé de chiffrement pas substitution.

On considère le texte chiffré suivant obtenu par substitution.

$$yifqfmzrwqfyvecfmdzpcvmrzwnmdzvejbtxcddumj$$
$$ndifefmdzcdmqzkceyfcjmyrncwjcszrexchzunmxz$$
$$nzucdrjxyysmrtmeyifzwdyvzvyfzumrzcrwnzdzjj$$
$$xzwgchsmrnmdhncmfqchzjmxjzwiejyucfwdjnzdir$$

Question : En utilisant la fréquence d'apparition des lettres en anglais, ainsi que les digrammes et trigrammes les plus fréquents, retrouver le message clair (écrit en anglais) ayant produit ce message chiffré.