Cryptographie et analyse fréquentielle

Laurent Signac – CC-BY-SA – 15-09-20 2151 e0ff103987790d2b0e4c

1 Substitutions et transpositions

Historiquement, on distingue trois types de méthodes de chiffrement basées sur les lettres :

- Substitutions monoalphabétiques
- Substitutions polyalphabétiques
- Transpositions

Dans une **substitution monoalphabétique**, chaque lettre est remplacée par une autre, toujours la même. C'est le cas du chiffre de César (décalage : 3) :

```
clair   : BIENVENUE
chiffré : ELHQYHQXH
```

Dans une **substitution polyalphabétique**, chaque lettre est remplacée par une autre, mais **pas** toujours la même. C'est le cas du chiffre de Vigénère.

Par exemple, avec la clé `bj` et sachant que `a` décale de 0, le `b` de 1,… le `j` de 9,… le `z` de 25) :

```
clair   : BIENVENUE
clé     : bjbjbjbjb
décal.  : 191919191
chiffré : CRFWWNODF
```

Dans le cas d'une **transposition**, les lettres restent identiques mais on change leur position dans le message.

Exemple : prendre une lettre sur 4 (en cyclant correctement)

```
clair   : BIENVENUE
chiffré : BVENUENIE
```

2 Analyse fréquentielle

L'analyse fréquentielle date du IXe siècle. Elle consiste à tirer parti du fait que les différentes lettres apparaissent avec des fréquences différentes dans un texte intelligible.

Si on compare les fréquences des lettres en français avec celles du texte suivant, elles sont très similaires (figure 1) :

```
PUISQUEVOUSETESRESOLUSDEMOURIRJETONSLESORTPOURVOIRQUISERACELUIQUIDEVRAETRETUELEPREMIER
PARCELUIQUILESUIVRAETCONTINUONSTOUJOURSDENUSERDELAMEMESORTEAINSILESORTFUTJETEETCELUISU
RQUIILTOMBAITTENDITLAGORGEACELUIQUILEDEVAITTUERCEQUICONTINUAJUSQUACEQUILNERESTAPLUSQUE
JOSEPHETUNAUTRESOITQUECELAARRIVATPARHASARDOUPARUNECONDUITEPARTICULIEREDEDIEUALORSJOSEP
HVOYANTQUESILEUTENCOREJETELESORTOUILLUIAURAITFALLUTREMPERSESMAINSDANSLESANGDUNDESESAMI
SILLUIPERSUADADEVIVREAPRESLUIAVOIRDONNEPAROLEDELESAUVER
```

Dans le cas d'un chiffrement pas transposition, les lettres ne changent pas, et donc l'histogramme reste inchangé entre le texte clair et chiffré.

Dans le cas d'une substitution monoalphabétique, l'histogramme (figure 2) est modifié, mais les valeurs des fréquences sont simplement mélangées (on a toujours un pic pour le chiffrement du `E` et un creux pour le chiffrement du `K`). Dans le cas d'un décalage (César), c'est encore plus simple puisque l'histogramme est simplement décalé.

Une substitution rend le texte chiffré «plus aléatoire», d'autant plus qu'il y a beaucoup d'alphabets de chiffrement (d'autant plus que la clé est longue dans le cas de la méthode de Vigénère). L'histogramme est lissé (figure 3).

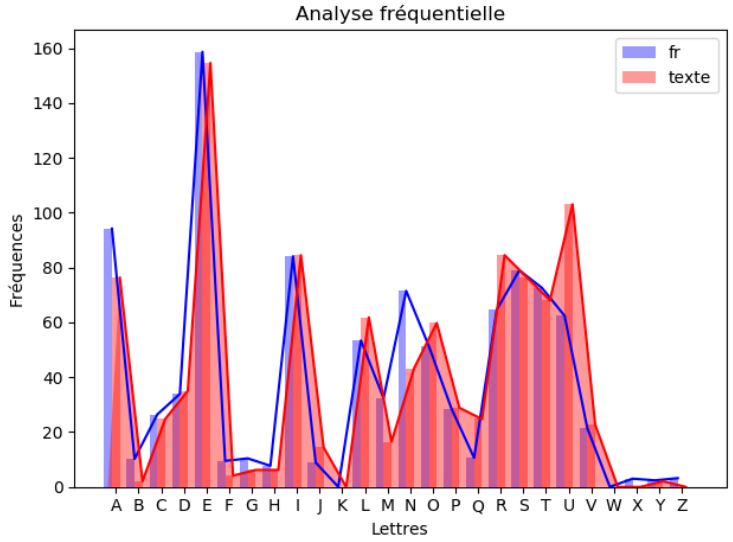

Figure 1: Fréquence d'apparition de chaque lettre. Pour un texte raisonnablement long, en français, l'histogramme est remarquablement constant

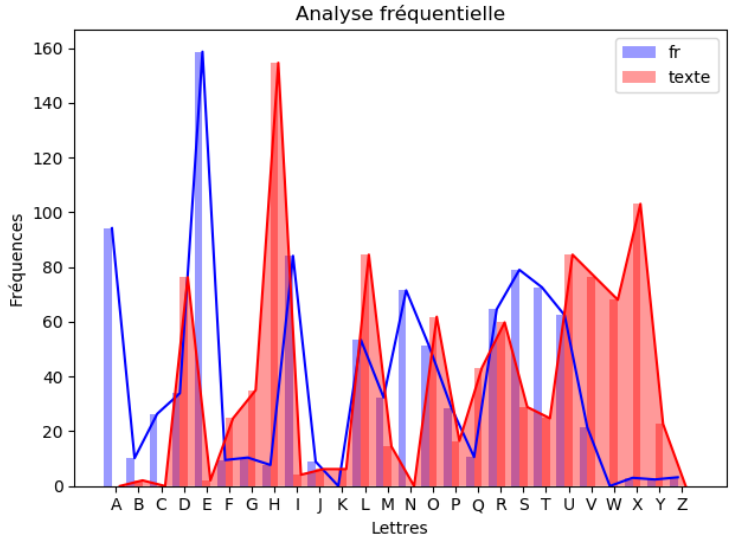

Figure 2: Dans le cas du chiffre de César, on déduit le décalage, ici 3.

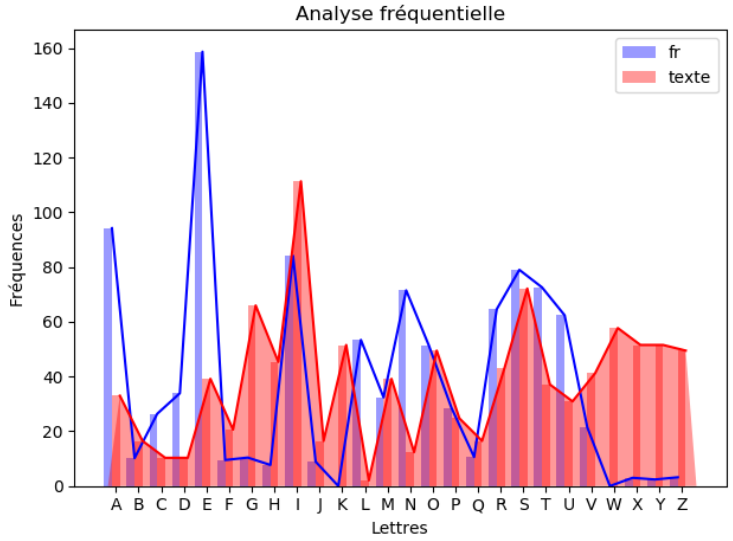

Figure 3: Effet lissant d'une substitution polyalphabétique

En conclusion :

- On peut savoir si on a affaire à une transposition (mais pas forcément la décrypter)
- **On peut savoir si on a affaire au chiffre de César (et le décrypter)**
- On peut savoir si on a affaire à une subst. poly (et c'est tout)

3 Indice de coïncidence

L'indice de coïncidence a été inventé par Friedman au début du XXe siècle. Sa valeur indique, pour un texte donné, la probabilité qu'en tirant deux lettres au hasard, ce soit la même :

$$I_C = \sum_{i=A}^{i=Z} \frac{n_i(n_i-1)}{n(n-1)}$$

Sa valeur est *caractéristique* d'une langue donnée :

Langue	allemand	anglais	espagnol	français
I_c	0.072	0.065	0.074	0.074

Dans le cas de lettres choisies au hasard (texte aléatoire), on obtiendrait : $I_c \approx 0.0384$

Voyons quel est l'effet des manipulations de texte sur l'indice de coïncidence :

- Texte clair en Français : ≈ 0.074
- Transposition : ≈ 0.074
- Substitution mono. : ≈ 0.074
- Substitution poly. : $<< 0.074$ (≈ 0.05)
- Un **sous-texte** d'un texte français : ≈ 0.074 👍

Ces constatations permettent de déterminer la longueur de la clé de Vigénère utilisée !

```
BIENVENUE  => B.E.V.N.E
bjbjbjbjb     b.b.b.b.b
191919191     1.1.1.1.1
CRFWWNODF     C.F.W.O.F
```

On remarque qu'en prenant une lettre sur 2, à partir de la première, on retombe sur une substitution monoalphabétique (décalage 1), et en prenant une lettre sur 2, à partir de la seconde, on retombe sur une substitution monoalphabétique (décalage 9).

Sur le texte utilisé plus haut, chiffré avec la clé `BJ`, si on prend toutes les lettres, on obtient l'histogramme de la figure 4 et un indice de coïncidence très inférieur à $0,07$. Mais si on prend une lettre sur deux, on obtient l'histogramme de la figure 5 (si on commence sur la lettre 1) et un indice proche de celui du français.

4 Exemple de déchiffrement de la méthode de Vigénère

On reprend le texte donné plus haut, chiffré avec la clé `EGO` (qu'on ignore).

Le message chiffré commence par `TAWWWIIBCYYSXKGVKGSRIWJSQUIVOFN...` (longueur réelle : 485 caractères)

On calcule les indices de coïncidence en prenant toutes les lettres, une sur deux, une sur trois ou une sur quatre, et en commençant par la lettre 1, ou 2 ou 3…

	1	2	3	4
1/1	0.05			
1/2	0.052	0.049		
1/3	**0.084**	**0.074**	**0.078**	

	1	2	3	4
1/4	0.047	0.056	0.056	0.045

L'indice de coïcidence correspond au Français pour la ligne 1/3. La longueur de la clé est donc 3.

On considère maintenant le sous-texte 1 lettre sur 3 en commençant par la première lettre :

```
TAWWWIIBCYYSXKGVKGSRIWJSQUIVOFN...
T  W  I  Y  X  V  S  W  Q  V  N
```

L'histogramme (figure 6) permet de retrouver le décalage utilisé. On constate que le décalage est 4, et donc la lettre de la clé de Vigénère est un `E`.

On recommence la même opération, en commençant par extraire un sous texte à partir de le seconde lettre:

```
TAWWWIIBCYYSXKGVKGSRIWJSQUIVOFN...
 A  W  B  Y  K  K  R  J  U  O
```

L'histogramme (figure 7) nous permet de trouver le décalage de 6 et la lettre de la clé `G`.

On termine en cherchant la troisième lettre de la clé :

```
TAWWWIIBCYYSXKGVKGSRIWJSQUIVOFN...
  W  I  C  S  G  G  I  S  I  F
```

L'histogramme (figure 8) nous permet de trouver le décalage de 14 et la lettre de la clé `O`.

On a reconstitué la clé : `EGO` et il ne reste plus qu'à déchiffrer.

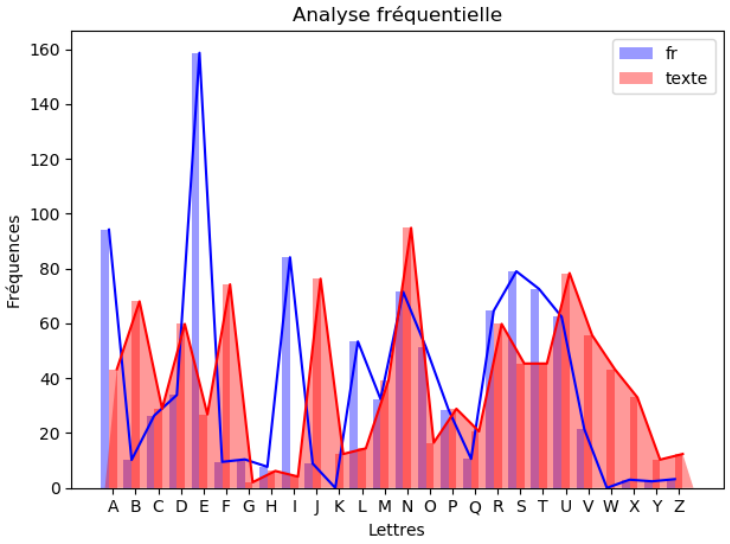

Figure 4: $I_c = 0.0536$

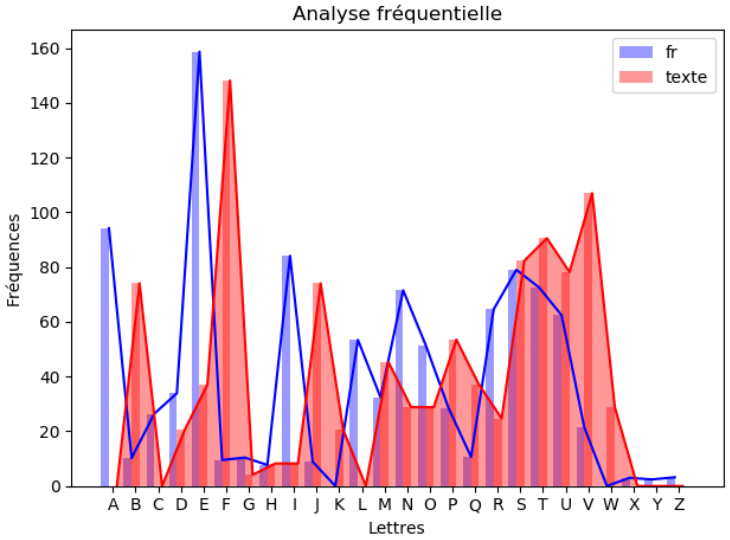

Figure 5: $I_c = 0.0734$

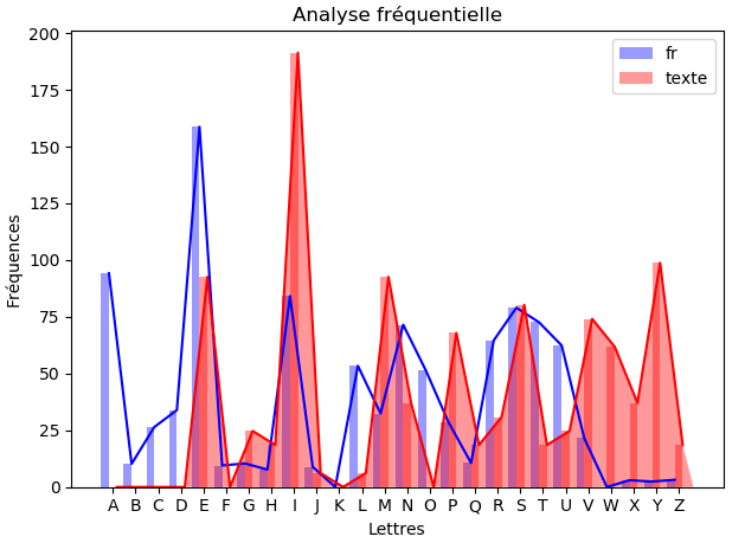

Figure 6: Le décalage est de 4

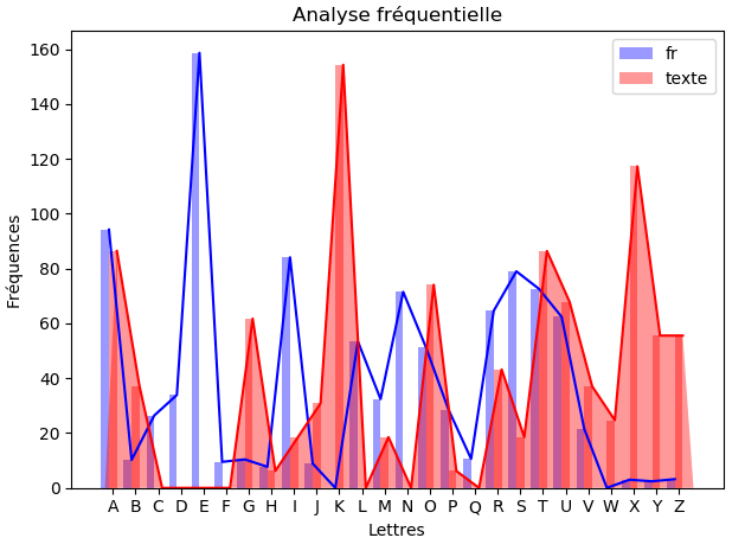

Figure 7: Le décalage est de 6

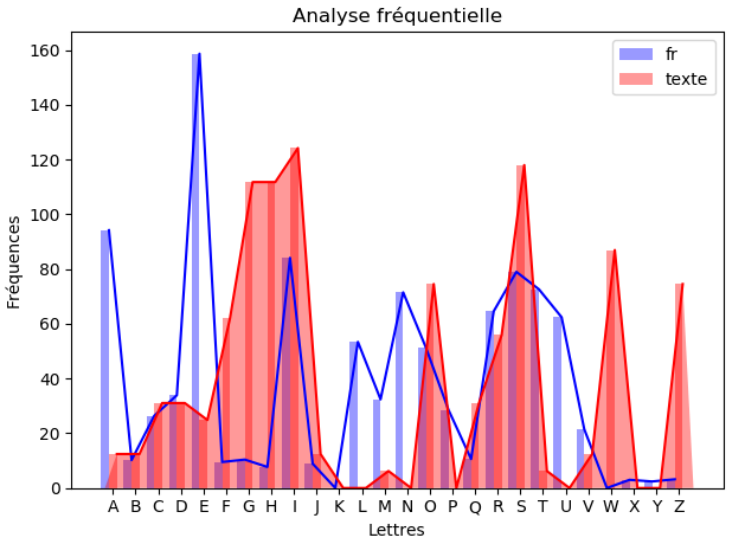

Figure 8: Le décalage est de 14