Chiffrement par substitution.

Classiquement, les cryptosystèmes étaient des algorithmes fondés les lettres de l'alphabet. Les cryptosystèmes substituaient les caractères, permutaient (ou transposaient) les caractères, ou faisaient une combinaison de ces opérations.

Notation

Nous noterons l'alphabet du message par $\mathscr{A}$ et l'alphabet de texte chiffré par $\mathscr{A}_0$. Nous écrivons E_K pour l'application de chiffrement et D'_K pour l'application de déchiffrement, où K et K' sont des clés pour chiffrer et déchiffrer.

Chiffrement par substitution

Nous identifions quatre types différents de chiffrement par substitution.

La substitution simple

Dans ce cryptosystème, **l'algorithme** est une substitution de caractères, la **clé** étant la liste de substitutions de l'alphabet. En d'autres termes, un chiffrement simple par substitution est défini par une application $\mathscr{A} \mapsto \mathscr{A}_0$.

Supposons que nous codions d'abord un message en éliminant tous les caractères non alphabétique (par exemple nombres, espaces, et ponctuation) et en changeant tous les caractères en majuscule. Alors la taille de la clé, qui borne la sécurité du système, est 26 lettres. Par conséquent le nombre de clés est

$$26! = 403291461126605635584000000 \simeq 4.10^{26} \simeq 2^{88}$$

qui est un nombre énorme.

Néanmoins, nous verrons que la substitution simple est très susceptible d'attaques cryptanalytique.

Exemple. Considérez ce paragraphe, codé de cette façon, nous obtenons le message:

SUPPOSONSQ. . .

Utilisons la clé de chiffrage ULOIDTGKXYCRHBPMZJQVWNFSAE,

nous obtenons le cryptogramme

QWMMPQPBQZWDBPWQOPIXPBQIULPJIWBHDQQUGDDBDRXHXBUBVV

PWQRDQOUJUOVDJDQBPBURMKULDVXZWDMUJDSDHMRDBPHLJDQD

QMUODQDVMPBOVWUVXPBDVDBOKUBGDUBVVPWQRDQOUJUOVDJD

QDBHUYWQOWRD

Des chiffrement par substitution simples peuvent être facilement cassés parce que le chiffrement ne change pas les fréquences des symboles du message.

Chiffrement affines

Un cas spécial des chiffrement par substitution simples sont les chiffrements affines. Si nous codons numériquement l'alphabet
comme éléments 0, 1. . . , 25 de $\mathbf{Z}/26\mathbf{Z}$ alors nous pouvons opérer sur les lettres par des transformations de la forme $X \mapsto ax + b$, pour tout a pour lequel $PGCD(a, 26) = 1$.

Un chiffrement affine lequel $a = 1$ est dit **chiffrement par translation**. Le chiffrement par translation est réalisé par les opérations de décalage circulaire b fois ($A \mapsto B, B \mapsto C$, etc...) sur l'alphabet.

Chiffrement de César

Classiquement un chiffrement par translation est dit chiffrement de César, d'après Jules César, qui avait l'habitude de cette méthode pour communiquer avec ses généraux. Par exemple,si nous employons b = 3 nous obtenons le chiffrement $A \mapsto D, B \mapsto E, \ldots, Z \mapsto C$.

Le chiffre de César est un chiffre monoalphabétique (c'est-à-dire que chaque lettre est remplacée par une autre lettre de l'alphabet), consistant simplement à décaler l'alphabet clair. Le décalage est la clé du chiffrement.

Exemple

On veut chiffrer le mot CRYPTOGRAPHIE avec un décalage de 3. Pour cela on écrit les alphabets clair et chiffré comme suit:

ABCDEFGHIJKLMNOPQRSTUVWXYZ

DEFGHIJKLMNOPQRSTUVWXYZABC

Il suffit alors de remplacer les lettres du message clair à l'aide le l'alphabet chiffré:

CRYPTOGRAPHIE

FUBSWRJUDSKLH

Chiffrement homophone

Dans ce cryptosystème le déchiffrement est une fonction d'un plus grand alphabet $\mathscr{A}_0$ vers l'alphabet $\mathscr{A}$, mais pour chiffrer un message on peut prendre n'importe quel caractère dans l'image.

Un moyen pour réaliser un chiffrement homophe doit commencer par différentes clés de substitution de m, et par chaque substitution, font un choix aléatoire de la clé à employer. Par exemple, supposez que nous prenons $\mathscr{A}$ l'alphabet ordinaire à 26 éléments, et soit l'alphabet $\mathscr{A}_0$ de caractère chiffrés l'ensemble de paires de caractère. Supposez maintenant que nous ayons la paire de clés de substitution dans l'alphabet de texte chiffré :

```
LV MJ CW XP QO IG EZ NB YH UA DS RK TF MJ XO SL PE NU FV TC QD RK YH GW AB ZI

UD PY KG JN SH MC FT LX BQ EI VR ZA OW XP HO DJ CY RN ZV WT LA SF BM GU QK IE
```

comme clé homophone.

Pour coder le message: “Chiffrement homophone”

on le transforme en CHIFFREMENTHOMOPHONE

Alors chacun des messages chiffrés suivant est valable:

CWLXBQIGIGRNQOTFQOMJTCLXXOOWXODJNBHOXPQO

KGLXBQMCMCRNQOTFQOXPTCNBHOTFHOSLLXHOXPQO

KGLXBQMCIGNUQOOWSHXPWTNBHOTFXOSLLXHOXPSH

De plus, chacun se déchiffre en le message original.

Chiffrement par substitution polyalphabetique

Un chiffrement par substitution polyalphabetique, comme le chiffre homophone, emploie des clés multiples, mais le choix de la clé n'est pas choisi aléatoirement, il est déterminé plutôt d'après la position dans le message.

La plupart des chiffrements polyalphabetiques sont des chiffrement de substitution périodiques, qui substituent le $mj + i^{\text{ième}}$ caractère du message en utilisant la $i^{\text{ième}}$ clé. Le nombre m s'appelle la période.

Chiffrement de Vigenère.

Le chiffrement de Vigenère est un chiffrement par translation polyalphabetique, c'est-à-dire que chacune des clés m indique une translation affine. Supposons que nous prenions l' alphabet ordinaire {A, B,. . . , Z} en bijection avec $\mathbf{Z}/26\mathbf{Z}$ = 0, 1. . . , 25.

L'addition doit être effectuée dans $\mathbf{Z}/26\mathbf{Z}$, avec la bijection définie par la table suivante :

A	B	C	D	E	F	G	H	I	J	K	L	M
0	1	2	3	4	5	6	7	8	9	10	11	12

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

Chiffrement

		Lettre en clair																									
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
Clé Utilisée	**A**	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
	C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
	D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
	E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
	F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
	G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
	H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
	I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
	J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
	K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
	L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
	M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
	N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
	O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
	X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
	Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Lettre Codée

- Pour chaque lettre en clair, on sélectionne la colonne correspondante

- pour une lettre de la clé on sélectionne la ligne adéquate,

- puis au croisement de la ligne et de la colonne on trouve la lettre codée.

La lettre de la clé est à prendre dans l'ordre dans laquelle elle se présente et on répète la clé en boucle autant que nécessaire.

Exemple

clé : MUSIQUE

texte : j'adore ecouter la radio toute la journee

Texte en clair : `j'adore ecouter la radio toute la journee`
Clé répétée : `M USIQU EMUSIQU EM USIQU EMUSI QU EMUSIQU`

Colonne O, ligne I: on obtient la lettre W.
Colonne D, ligne S: on obtient la lettre V.
Colonne A, ligne U: on obtient la lettre U.
Colonne J, ligne M: on obtient la lettre V.

Le texte chiffré est alors :

V'UVWHY IOIMBUL PM LSLYI XAOLM BU NAOJVUY.

Pour déchiffrer ce texte,

- on regarde pour chaque lettre de la clé répétée la ligne correspondante,

- on y cherche la lettre codée.

- La première lettre de la colonne que l'on trouve ainsi est la lettre décodée.

Texte codé : `V'UVWHY IOIMBUL PM LSLYI XAOLM BU NAOJVUY`
Clé répétée : `M USIQU EMUSIQU EM USIQU EMUSI QU EMUSIQU`

Ligne I, on cherche W: on trouve la colonne O.
Ligne S, on cherche V: on trouve la colonne D.
Ligne U, on cherche U: on trouve la colonne A.
Ligne M, on cherche V: on trouve la colonne J.

Principe mathématique

Mathématiquement, on considère que les lettres de l'alphabet sont numérotées de 0 à 25 (A=0, B=1 ...). La transformation lettre par lettre se formalise simplement par :

Codé = (Texte + Clé) modulo 26

(Texte + Clé) modulo 26 correspond au "reste de la division entière de (Texte + Clé) par 26",

En fait il suffit d'effectuer l'addition des deux caractères puis de trouver le numéro correspondant à la lettre codée, notre alphabet étant circulaire (après Z on a A), la congruence nous assure que notre résultat sera compris entre 0 et 25.

Chiffrement par substitutions polygramme

Le chiffrement par substitutions polygramme est un cryptosystème dans lequel la substitution opère sur les blocs de caractères.

Par exemple (pour une clé particulière) le AA pourrait devenir NO, AB devenir IR, JU devenir AQ, etc...

Ces cryptosystèmes rendent la cryptanalyse plus dure en détruisant les fréquences de caractère, préservées sous des chiffrements par substitution simples.

Chiffrements affines généraux.

Les chiffrements affines peuvent être généralisés aux chiffrement polygrammes. Plutôt qu'une application $m \mapsto c = ma + b$, nous pouvons appliquer une transformation linéaire des vecteurs

$$u = (m_1, \ldots, m_n) \mapsto (c_1, \ldots, c_n) = uA + v,$$

par une certaine matrice inversible $A = (a_{ij})$ et un vecteur $v = (b_1, \ldots, b_n)$. Comme avant, nous codons numériquement un alphabet {A, B,. . . , Z} par les éléments $\{0, 1 \ldots, 25\}$ de $\mathbf{Z}/26\mathbf{Z}$. Puis chaque mots de caractères $m_1 m_2 \ldots m_n$ est identifié avec le vecteur $u = (m_1, m_2, \ldots, m_n)$. La multiplication de matrice est définie comme d'habitude, de sorte que

$$c_j = \left(\sum_{i=1}^{n} m_i a_{ij} \right) + b_j,$$

avec le résultat interprèté modulo 26 comme élément de $\mathbf{Z}/26\mathbf{Z}$.

Comme cas spécial, considerons des polygrammes à 2 caractères, de telle sorte que AA = (0, 0), . . . , ZY = (25, 24), ZZ = (25, 25). La matrice A donnée par

$$\begin{pmatrix} 1 & 8 \\ 21 & 3 \end{pmatrix}$$

et le vecteur v = (13, 14) definissent une application $AA = (0,0) \mapsto (13,14) = NO, \ldots, ZY = (25,24) \mapsto (18,23) = WA,\ ZZ = (25,25) \mapsto (18,23) = RD$ ce qui est une substitution simple sur les polygrammes à 2 caractères.

Le nombre de chiffrement affines est de beaucoup inférieur à celui de toutes les substitutions possibles, mais augmente exponentiellement quand le nombre n de caractères augmente.

Chiffrement par transposition

Un chiffrement par substitution permute les caractères de l'alphabet. Dans un chiffrement par **transposition**, les symboles du message demeurent inchangés, mais leur ordre est permuté par une permutation des positions d'indice. A la différence des chiffrement par substitution, les chiffrement par transposition sont des chiffrement **par blocs**.

Groupes de permutation

Le groupe symétrique S_n est l'ensemble de toutes les applications bijectives de l'ensemble $\{1, \ldots, n\}$ vers lui-même, et les éléments de S_n sont des permutations.

La clé est la donnée d'une permutation d'un bloc de n éléments.

On note $[i_1, \ldots, i_n]$ la permutation $\sigma(1) = i_1, \ldots, \sigma(n) = i_n$.

Un chiffrement par transposition de longueur de bloc n induit $n!$ différentes permuations.

Pour $n = 7$, ceci donne 5040 permutations, mais pour la longueur de bloc égale à 36, ceci donne

$$371993326789901217467999448150835200000000 \simeq 4.10^{41} \simeq 2^{138}$$

possibilités.

En dépit du grand nombre de permutations possibles, la **structure du message** permet à un adversaire de déchiffrer le texte chiffré par transposition.

Cryptanalyse

Analyse fréquentielle

L'analyse fréquentielle, ou analyse de fréquences, est une méthode de cryptanalyse qui consiste à examiner la fréquence des lettres employées dans un message chiffré. Cette méthode est fréquemment utilisée pour décoder des messages chiffrés par **substitution** (comme par exemple le Chiffre de Vigenère ou le Chiffre de César).

L'analyse fréquentielle est basée sur le fait que, dans chaque langue, certaines lettres ou combinaisons de lettres apparaissent avec une certaine fréquence. Par exemple, en français, le **e** est la lettre la plus utilisée, suivie du **s** et du **a**. Inversement, le **w** est peu usité.

Ces informations permettent aux cryptanalystes de faire des hypothèses sur le texte clair, à condition que l'algorithme de chiffrement conserve la répartition des fréquences, ce qui est le cas pour des substitutions mono-alphabétiques et poly-alphabétiques.

Une deuxième condition est nécessaire pour appliquer cette technique : c'est la longueur du message à décrypter.

En effet, un texte trop court ne reflète pas obligatoirement la répartition générale des fréquences des lettres.

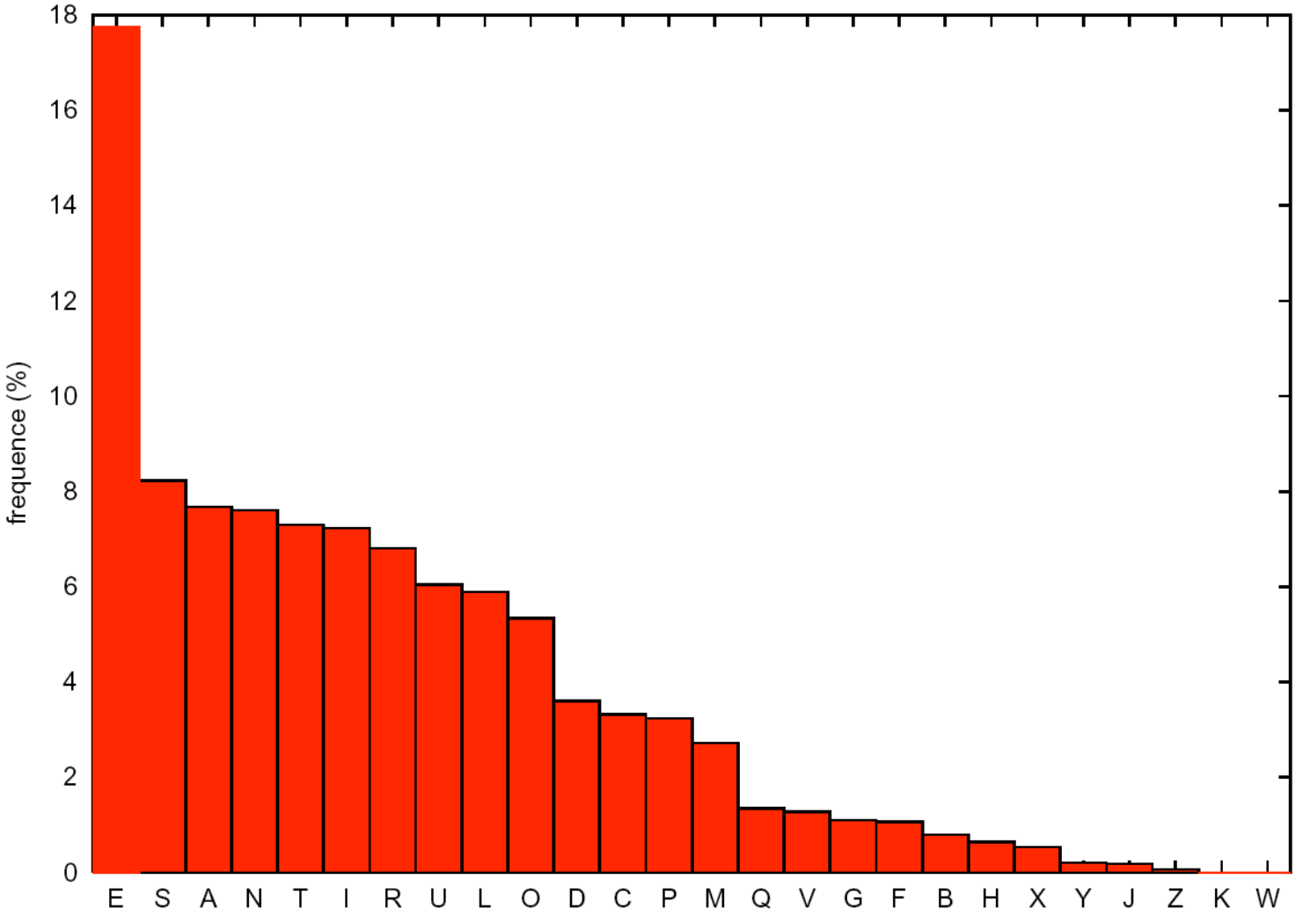
De plus, si la clé est de la même longueur que le message, il ne pourra y avoir des répétitions de lettres et l'analyse fréquentielle sera impossible.

Un exemple

Un texte chiffré

nvxlbgi avxw n ctxnbw ubn dvttbn r bhxqacyb awbggbgi
rbn cueciwvn lcnibn vqnbcxz rbn tbwn hxq nxqlbgi
qgrvubgin mvtacygvgn rb lvscyb ub gclqwb yuqnncgi
nxw ubn yvxoowbn ctbwn c abqgb ubn vgi qun rbavnbn
nxw ubn aucgmdbn hxb mbn wvqn rb u ckxw tcucrwvqin
bi dvgibxz ucqnnbgi aqibxnbtbgi ubxwn ywcgrbn cqubn
eucgmdbn mvttb rbn clqwvgn iwcqgbw c mvib r bxz
mb lvscybxw cqub mvttb qu bni ycxmdb bi lbxub uxq
gcyxbwb nq ebcx hx qu bni mvtqhxb bi ucqr u xg
cycmb nvg ebm clbm xg ewxubyxbxub u cxiwb tqtb bg
evqicgi u qgoqwtb hxq lvucqi ub avbib bni nbteuceub
cx awqgmb rbn gxbbn hxq dcgib uc ibtabib bi nb wqi
rb u cwmdbw bzqub nxw ub nvu cx tquqbx rbn dxbbn
nbn cqubn rb ybcgi u btabmdbgi rb tcwmdbw

Fréquence des lettres en Français



Analyse de la fréquence des lettres

Dans le **chiffré** :

B	N	C	U	X	Q	G	I	W	V
18,7	9,91	7,78	6,90	6,72	6,37	5,84	5,84	5,30	4,60

En **français** :

E	S	A	N	T	I	R	U	L	O
17,8	8,23	7,68	7,61	7,30	7,23	6,81	6,05	5,89	5,34

$$B \longrightarrow E$$
$$N \longrightarrow S$$
$$C \longrightarrow A$$

svxlegi avxw s atxsew ues dvttes r ehxqaaye
aweggegi res aueaiwvs lasies vqseaxz res tews
hxq sxqlegi qgrvuegis mvtaaygvgs re lvsaye
ue galqwe yuqssagi sxw ues yvxoowes atews a
aeqge ues vgi qus reavses sxw ues auagmdes
hxe mes wvqs re u akxw tauarwvqis ei dvgiexz
uaqssegi aqiexsetegi uexws ywagres aques euagmdes
mvtte res alqwvgs iwaqgew a mvie r exz me
lvsayexw aque mvtte qu esi yaxmde ei lexue
uxq gayxewe sq eeax hx qu esi mvtqhxe ei
uaqr u xg ayame svg eem alem xg ewxueyxexue u
axiwe tqte eg evqiagi u qgoqwte hxq lvuaqi
ue aveie esi seteuaeue ax awqgme res gxees
hxq dagie ua ietaeie ei se wqi re u awmdew
ezque sxw ue svu ax tquqex res dxees ses aques
re yeagi u etaemdegi re tawmdew

Fréquence des bigrammes

Bigrammes les plus fréquents dans le chiffré :

ES	U**E**	GI	R**E**	**E**G	**E**X	I**E**	**SE**	QU	T**E**	U**A**	**E**W	**A**G	**A**Q	H
25	17	13	12	9	8	8	8	8	8	8	7	7	7	7

Bigrammes les plus fréquents en français :

ES L**E** **E**N D**E** R**E** NT ON **E**R T**E** **SE** **E**T **E**L QU **A**N N**E** OU **A**I

$$
\begin{aligned}
U &\longrightarrow L \\
R &\longrightarrow D \\
G &\longrightarrow N \\
Q &\longrightarrow I
\end{aligned}
$$

Fréquence des bigrammes

Bigrammes les plus fréquents dans le chiffré :

ES	**LE**	**N**I	**DE**	**EN**	**E**X	I**E**	**SE**	**IL**	T**E**	**LA**	**E**W	**AN**	**AI**	HX
25	17	13	12	9	8	8	8	8	8	8	7	7	7	7

Bigrammes les plus fréquents en Français :

ES LE EN DE R**E N**T O**N E**R T**E SE E**T **EL** QU **AN NE** OU **AI**

$$I \longrightarrow T$$

svxlent avxw s atxsew les dvttes d ehxiaaye awennent des aleatwvs lastes viseaxz des tews hxi sxilent indvlents mvtaaynvns de lvsaye le naliwe ylissant sxw les yvxoowes atews a aeine les vnt ils deavses sxw les alanmdes hxe mes wvis de l akxw taladwvits et dvntexz laissent aitexsetent lexws ywandes ailes elanmdes mvtte des aliwvns twainew a mvte d exz me lvsayexw aile mvtte il est yaxmde et lexle lxi nayxewe si eeax hx il est mvtihxe et laid l xn ayame svn eem alem xn ewxleyxexle l axtwe tite en evitant l inoiwte hxi lvlait le avete est setelaele ax awinme des nxees hxi dante la tetaete et se wit de l awmdew ezile sxw le svl ax tiliex des dxees ses ailes de yeant l etaemdent de tawmdew

Quelques mots du chiffré :

indvlent, vnt	$V \longrightarrow O$
oiseaxz	$X \longrightarrow U$
	$Z \longrightarrow X$
a aeine,	$A \longrightarrow P$
leuws	$W \longrightarrow R$
taladroits	$T \longrightarrow M$
yrandes	$Y \longrightarrow G$

soulent pour s amuser les dommes d ehuipage
prennent des aleatros lastes oiseaux des mers hui
suilent indolents mompagnons de losage le nalire
glissant sur les gouoores amers a peine les ont ils
deposes sur les planmdes hue mes rois de l akur
maladroits et donteux laissent piteusement leurs grandes
ailes elanmdes momme des alirons trainer a mote
d eux me losageur aile momme il est gaumde et
leule lui naguere si eeau hu il est momihue et laid
l un agame son eem alem un erulegueule l autre
mime en eoitant l inoirme hui lolait le poete est
semelaele au prinme des nuees hui dante la tempete
et se rit de l armder exile sur le sol au milieu des
duees ses ailes de geant l empemdent de marmder

Chiffrement par transpositions

La cryptanalyse de ce système se fonde sur le fait qu'il est **linéaire**:

il existe une matrice (ici, une matrice de permutation) A_σ telle que $c = m \times A_\sigma$. Le terme $(A_\sigma)_{i,j}$ de cette matrice qui est situé à la $i^{\text{ème}}$ ligne et à la $j^{\text{ème}}$ colonne vaut 1 si $i = \sigma(j)$ et 0 sinon.

Voici une attaque possible sur les cryptosystèmes linéaires.

Soit un cryptosystème tel que $c = m \times A$ (où la matrice A dépend de la clé secrète, nous considérerons qu'elle est la clé secrète).

On suppose que l'attaquant dispose de n paires de clairs-chiffrés et on va montrer qu'il peut avec une probabilité non négligeable retrouver la clé, c'est dire retrouver A (on dit qu'il fait une attaque à clair connu).

On met les n clairs en une matrice M carrée $n \times n$ et les n chiffrés correspondants en une matrice C. On a $C = M \times A$.

Si la matrice M est inversible il en déduit $A = M^{-1} \times C$ en n^3 opérations élémentaires (par la méthode classique du pivot de Gauss). La proportion de matrices inversibles parmi les matrices carrées étant non négigeable, la probabilité que l'attaque réussisse est non négligeable.

Chiffrement de Vigenère

La première étape doit déterminer la longueur de la clé qui est le nombre entier m.

Une méthode est celle de **Kasiski** et une autre emploie **l'indice de coïncidence**.

La **méthode de Kasiski** a été décrit par Friedrich Kasiski en 1863 (et par Babbage avant lui). On se fonde sur l'observation que deux segments identiques du message seront chiffrés de la même manière toutes les fois que leur occurrence dans le message sont séparée de d positions, où $d \equiv 0 \pmod{m}$.

Réciproquement, si nous observons deux segments identiques de texte chiffré, chacune de la longueur au moins trois par exemple alors il y a de fortes chances qu'ils correspondent aux segments identiques du message.

La méthode de Kasiski fonctionne comme suit. Nous recherchons dans le texte chiffré des paires de segments identiques de longueur au moins trois, et enregistrons la distance entre les positions de départ des deux segments. Si nous obtenons plusieurs distances d_i, nous conjecturerons que m divise le pgcd des d_i.

KQOWEFVJPUJUUNUKGLMEKJINMWUXFQMKJBGWRLFNFGHUDWUUMBSVLPS

NCMUEKQCTESWREEKOYSSIWCTUAXYOTAPXPLWPNTCGOJBGFQHTDWXIZA

YGFFNSXCSEYNCTSSPNTUJNYTGGWZGRWUUNEJUUQEAPYMEKQHUIDUXFP

GUYTSMTFFSHNUOCZGMRUWEYTRGKMEEDCTVRECFBDJQCUSWVBPNLGOYL

SKMTEFVJJTWWMFMWPNMEMTMHRSPXFSSKFFSTNUOCZGMDOEOYEEKCPJR

GPMURSKHFRSEIUEVGOYCWXIZAYGOSAANYDOEOYJLWUNHAMEBFELXYVL

KQOWEFVJPUJUUNUKGLMEKJINMWUXFQMKJBGWRLFNFGHUD**WUU**MBSVLPS

NCMUEKQCTESWR**EEK**OYSSIWCTUAXYOTAPXPLWPNTCGOJBGFQHTD**WXIZA**

YGFFNSXCSEYNCTSSPNTUJNYTGGWZGR**WUU**NEJUUQEAPYMEKQHUIDUXFP

GUYTSMTFFSH**NUOCZGM**RUWEYTRGKMEEDCTVRECFBDJQCUSWVBPNLGOYL

SKMTEFVJJTWWMFMWPNMEMTMHRSPXFSSKFFST**NUOCZGMDOEOYEEK**CPJR

GPMURSKHFRSEIUEVGOYC**WXIZAY**GOSAANY**DOEOY**JLWUNHAMEBFELXYVL

On regarde la distance entre les répétitions.

On cherche les facteurs pour chaque paire :

Séquence répétée	Distance entre les répetitions	Longueurs de clef possibles (diviseurs de la distance)			
		2	3	5	19
WUU	95			x	x
EEK	200	x		x	
WXIZAYG	190	x		x	x
NUOCZGM	80	x		x	
DOEOY	45		x	x	

Les facteurs premiers du nombre de caractères entre deux débuts de séquences figurent dans le tableau (ex. 95 = 5 x 19). Il apparaît dans le tableau que toutes les périodes sont divisibles par 5. Tout se cale parfaitement sur un mot-clef de 5 lettres.

Une meilleure approximation de la valeur de m peut être obtenue par **l'indice de coïncidence.** Ce concept a été défini par Wolfe Friedman en 1920, comme suit.

Définition 1 *Supposons que $x = x_1 \dots x_n$ soit une suite de n lettres. L'indice de coïncidence de x, noté $I_c(x)$, est défini comme étant la probabilité que deux éléments aléatoires de x d'indice distincts soient identiques.*

Notons le nombre d'occurences de A ($\dots$ Z) dans x par f_0 ($\dots f_{25}$) respectivement. Nous pouvons choisir deux éléments distincts de x de $\binom{n}{2}$ manières différentes. Pour chaque i, il y a $\binom{f_i}{2}$ manières différentes de choisir les deux éléments distincts qui soient la lettre correspondant à l'indice i. Par conséquent nous avons la formule

$$I_c(x) = \sum_{i=0}^{25} \frac{f_i(f_i-1)}{n(n-1)} \simeq \sum_{i=0}^{25} \left(\frac{f_i}{n}\right)^2$$

Supposons maintenant que x est une suite d'un **texte en français**. Notons les probabilités prévues de l'occurrence des lettres A...Z par $p_0,\ldots,p_{25}$ respectivement. Alors nous nous attendrions à ce que

$$I_c \simeq \sum_{i=0}^{25} p_i^2 = 0,074$$

puisque la probabilité que deux éléments aléatoires tous les deux soient A est de p_0^2, etc...

Si c'est une **suite aléatoire** on a $p_i = 1/26$, donc

$$I'_c = \sum_{i=0}^{25} (1/26)^2 = 26(1/26)^2 = 1/26 \simeq 0,038$$

Le même raisonnement s'applique si x est un texte chiffré obtenu au moyen de n'importe quel chiffrement monosyllabique (car les probabilités sont décalées).

Supposons que nous ayons par un texte chiffré $y = y_1 \dots y_n$ qui a été construit en employant un **chiffrement de Vigenère.** Définissons m sous-chaînes $Y_1 \dots Y_m$ de y en prenant pour Y_i les lettres y_{am+i}.

Par exemple avec m = 5:

$$\begin{aligned} Y1 &= y1y6\dots \\ Y2 &= y2y7\dots \\ &\dots \\ Y5 &= y5y10\dots \end{aligned}$$

Si m est la longueur de la clé chaque $I_c(Y_i)$ devrait être à peu près égal à $0,074$.

Si m n'est pas la longueur de la clé, alors les sous-chaînes Y_i seront beaucoup plus aléatoires, puisqu'ils auront été obtenus par chiffrage avec des clés différentes, on devrait trouver $I'_c \simeq 0,038$.

Les deux valeurs sont suffisamment éloignées pour que nous puissions déterminer correctement la longueur principale.

Une fois que la longueur de la clé est obtenue, comment obtenir la clé ? Il est utile de considérer **l'indice mutuel de coïncidence de deux suites.**

Définition 2 *Supposons que* $y = y_1 \dots y_n$ *et* $y' = y'_1 \dots y'_n$ *sont des suites de* n *et* n' *lettres respectivement. L'indice mutuel de coïncidence de* y *et de* y', $MI_c(y, y')$ *est défini comme la probabilité qu'un élément aléatoire de* y *est identique à un élément aléatoire de* y'.

Supposons que nous dénotons le nombre d'occurences de A (... Z) dans y par f_0 (... f_{25}) respectivement et le nombre d'occurences de A (... Z) dans y' par f'_0 (... f'_{25}) respectivement. Alors

$$MI_c(y, y') = \sum_{i=0}^{25} \frac{f_i f'_i}{nn'}$$

Supposons que $K = (k_1 \dots k_m)$ soit la clé.

La probabilité que les deux caractères soient A dans Y_i, Y_j est $p_{-k_i} p_{-k_j}$, la probabilité que tous les deux soient B est $p_{1-k_i} p_{1-k_j}$ etc...

Par conséquent nous estimons que

$$MI_c(Y_i, Y_j) \simeq \sum_{h=0}^{25} p_{h-k_i} p_{h-k_j} = \sum_{h=0}^{25} p_h p_{h+k_i-k_j}$$

Observons que la valeur de cette évaluation dépend seulement de la différence $k_i - k_j$.

Pour $k_i = k_j$, cette valeur est 0,078 tandis que pour un différence non nulle cette valeur est moins de 0,045.

L'idée est maintenant de fixer un Y_i (avec fréquences (f_h), de le décaler par n'importe quelles des 26 lettres g (appelons Y_i^g le résultat), et de calculer pour chaque Y_j (avec fréquences (f'_h))

$$MI_c(Y_j, Y_i^g) = \sum_{i=0}^{25} \frac{f_h f'_{h+g}}{nn'} = \sum_{h=0}^{25} p_h p_{h+k_i-k_j+g}$$

Quand $g = k_i - k_j$ la valeur de MI_c devrait être environ de 0,078. Nous obtenons de cette façon (en changeant également Y_i) un ensemble de différences et nous pouvons exprimer tout les k_i en termes de l'un d'entre eux. Il faut faire 26 tentatives.

Remarque. Les Y_i ont été obtenus par un chiffrement par décalage qui est un cas particulier de substitution. Un autre moyen serait d'appliquer les méthodes statistiques habituelles.