Cet article est rédigé par des élèves. Il peut comporter des oublis et imperfections, autant que possible signalés par nos relecteurs dans les notes d'édition.

Cryptographie

Des textes à décoder...

De Haro Inès TeS2
Sirot Cellya 1eS1

Lycée Jean Puy Roanne(42)

Professeurs :
Me Gotte Me Martinelli
Chercheur :
Mr Chardard
(Université de St Etienne)

2013-2014

Introduction

Lors de la première séance, nous avons rencontré Mr Chardard, chercheur en mathématiques à l'université Jean Monnet de St Etienne. Il nous a présenté plusieurs sujets de recherche. Celui sur la cryptographie nous a intéressées :

Sauriez-vous déchiffrez ce texte:

mf efqbsufnfou ef mb mpjsf gbju qbsujf ef mb sfhjpo sipof-bmqft

(César se serait servi de ce type de code secret)
Et celui-ci:

iwp pjtuigap igtup rwp kogigtp diwppwta cgt ngwvq r vtw ijtuvwvq cgtagtw agva pvhhgnjta wa diwcw svjtr pgttw i zwvqw yw cw pgvkowtp rwp ygvqp jtnowtp wa yw fiwvqw wa yw c wt kjop jv kwta cjvkjop svo c wcfgqaw rwnj rwij fjqwoi j ij hwvoiiw cgqaw

Ce dernier, plus difficile, remplace chaque lettre de l'alphabet par une autre lettre de l'alphabet (chiffre de substitution). Bien qu'il semble très difficile à décoder sans connaître cette substitution, c'est en fait possible.

Bien d'autres codes, sous une apparence complexe, ne protège pas vraiment ce qui les utilisent et sans doute pas de membres de Maths En Jeans motivés.

Nous avons fait quelques recherches rapides concernant le vocabulaire lié à ce travail :

D'après Wikipédia : ***Décrypter*** consiste à retrouver le texte original à partir d'un message chiffré sans posséder la clé de (dé)chiffrement ; ***la cryptanalyse*** est la science qui consiste à tenter de déchiffrer un message ayant été chiffré sans posséder la clé de chiffrement et ***la cryptographie*** est une discipline qui s'attache à protéger des messages (assurant confidentialité, authenticité et intégrité) en s'aidant souvent de secrets ou clés, elle rend un message inintelligible à autre que qui-de-droit.

Notre but était donc de décrypter les deux textes donnés plus haut afin de connaître leur signification initiale.

Par la suite, M. Chardard nous a proposé d'autres textes à décrypter.

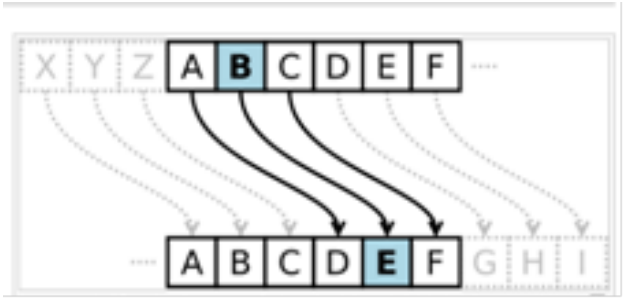
Lors de nos recherches, nous avons rencontré quatre types de cryptage qui sont : le Code César, le codage par substitution, le procédé de Vigenère et le cryptage Diffie-Hellman. Nous avons fait la cryptanalyse des trois premiers codes, ce que nous présentons dans cet article.

I- Le premier texte à décoder et le « code César »

Sauriez-vous déchiffrer ce texte ?
mf efqbsufnfou ef mb mpjsf gbju qbsujf ef mb sfhjpo sipof-bmqft

(César se serait servi de ce type de code secret)

Le premier texte à décrypter nous a amenés à faire des recherches documentaires sur le code



César.

César (-100 ; -44), général romain cryptait ses messages afin de les faire parvenir à ses troupes. Il utilisait un principe assez simple: le décalage de lettres. Chaque lettre du texte initial est remplacée par une autre que l'on trouve dans l'alphabet en se décalant vers la droite.
Il utilisait le plus souvent un décalage de 4 lettres, ainsi le A était codé par un E et le B par un F.

Pour coder le texte, on décale les lettres vers la droite ; pour le décoder on les décale vers la gauche.
Pour décrypter notre texte, on a tout d'abord essayé de décaler les lettres d'un rang vers la gauche. On a alors obtenu une phrase qui paraît cohérente et que nous avons acceptée comme solution :

1ER TEXTE décodé : Le département de la Loire fait partie de la région Rhône-Alpes.

II- Le 2ième texte à décoder et le « codage par substitution»

Et celui-ci ?

IWP PJTUIGAP IGTUP RWP KOGIGTP DIWPPWTA CGT NGWVQ R VTW IJTUVWVQ CGTGAGTW AGVA PVHHGNJTA WA DIWCW SVJTR I ZWVQW YW CW PGVKOWTP RWP YGVQP JTNOWTP WA YW FIWVQW WA YW C WT KJOP JV KWTA CJVKJOP SVO C WCFGQAW RWNJ RWIJ FJQWOI J IJ HWVOIIW CJQAW.

Ce dernier, plus difficile, remplace chaque lettre de l'alphabet par une autre lettre de l'alphabet (chiffrement de substitution).

Pour décrypter ce nouveau texte, nous avons utilisé la piste de recherche donnée.
Nous avons commencé par des recherches documentaires concernant le codage par substitution.

En utilisant le codage par substitution une lettre de l'alphabet est toujours codée dans le texte par une même autre lettre mais le codage ne suit pas un décalage fixe dans l'alphabet contrairement au code César.

Nous avons fait l'hypothèse que ce texte initial était à l'origine écrit en français. Nos recherches documentaires nous ont permis de trouver une stratégie pour décrypter le texte :

Nous avons calculé les fréquences d'apparition des lettres dans le texte codé puis on les a comparées aux fréquences d'apparition des lettres dans un texte quelconque de la langue française.

On a ensuite associé les lettres dont les fréquences étaient proches : (1)

$$F_i = \frac{\textit{nombre de lettres } i}{\textit{nombre de lettres total}}$$ (dans le texte donné)

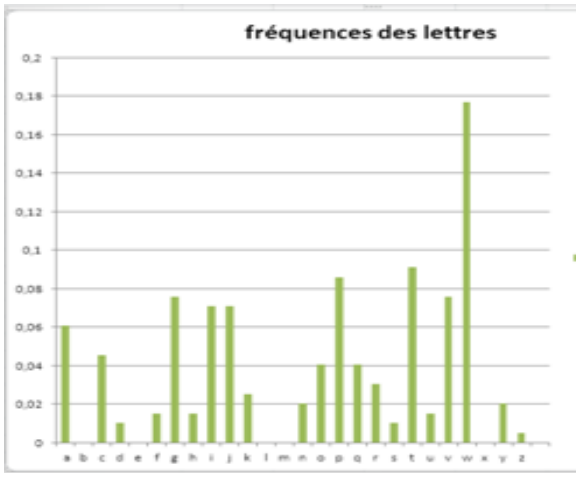
Dans notre texte :

Lettre codée	Fréquence en %
J	7,1
D	1,01
N	2,02
R	3,03
W	17,67
H	1,51
U	1,51
Z	0,5
O	4,04
Y	2,02
/	/
I	7,1
C	4,54
T	9,1
G	7,6
F	1,51
S	1,01
Q	4
P	8,6
A	6,06
V	7,57
K	2,5
/	/
/	/
/	/
/	/

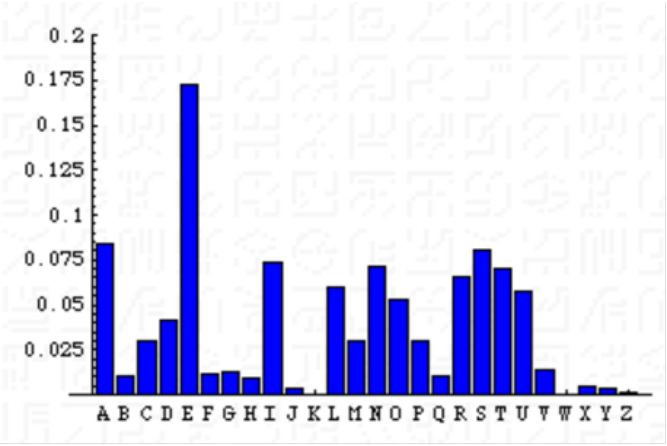
Dans la langue française :

Alphabet Clair	Fréquence en %
A	8,40
B	1,06
C	3,03
D	4,18
E	17,26
F	1,12
G	1,27
H	0,92
I	7,34
J	0,31
K	0,05
L	6,01
M	2,96
N	7,13
O	5,26
P	3,01
Q	0,99
R	6,55
S	8,08
T	7,07
U	5,74
V	1,32
W	0,04
X	0,45
Y	0,30
Z	0,12

Dans notre texte :



Dans la langue française :



On a associé les fréquences les plus proches entre les deux tableaux et graphiques :

Celle de la lettre E dans la langue française est très proche de celle de la lettre W dans le texte codé, celle de la fréquence du T dans la langue française est assez proche de celle du A dans le texte codé …

Des écarts sont remarqués mais peuvent s'expliquer car le texte est court ce qui conduit à des fréquences qui ne sont pas toujours significatives. De plus, la fluctuation d'échantillonnage explique que dans un texte quelconque écrit en français, les fréquences des différentes lettres peuvent fluctuer autour des fréquences données. (2)

Nous avons décodé une partie du texte :

-ES SA-G---S ---GS DES VI----S --ESSE--MO- C-E-R D --E -A-G-E-R M-----E ---- S-------- E- --E-E --A-- S---E - -E--E -E -E S------ -ES ----S ----E-S E- -E --E--E E- -E - -- ---S -- -E-- ------S --- E-----E -E-- -E-- ---E-- - ---- -E----E ----E.

Nous avons ensuite travaillé par tâtonnements :

On a émis comme conjecture que la première lettre du texte était un « L », « D », « M » ou « T ».
Car les mots les plus fréquents, de trois lettres, se terminant par ES et qui se rencontrent en début de phrase dans la langue française semblent être : LES, DES, MES ou TES.

Le troisième mot nous a aussi mis sur la voie :
Si la première lettre du texte est un « L » alors le troisième mot s'écrirait « L - - GS ».
On a supposé qu'il pouvait alors s'agir du mot « LONGS » .

Si cette hypothèse est vraie alors la lettre I se « décode » par un L, la lettre G par un O et la lettre T par un N.

Le texte devient alors un peu plus clair. Nous obtenons :

LES SANGLO-S LONGS DES VIOLONS -LESSEN-MON COE-R D -NE LANG-E-R MONO-ONE -O-- S-------- E- -LE-E --AN- SONNE L -E--E -E -E SO--EN- -ES -O--S -N--ENS E- -E -LE--E E- -E -EN -A-- A- -EN- -A--A-S --- - E--O--E -E-A -E-A -A-E-L A LA -E--LLE -O--E.

Nous avons continué de travailler par tâtonnements :
On a conjecturé que le deuxième mot était "SANGLO T S" ce qui nous a permis de décoder la lettre A par un T. On a ensuite complété le texte avec les « T ».

En continuant nos conjectures, nous avons « décodé » les dernières lettres.

Voilà le texte décodé final :
Les sanglots longs des violons blessent mon cœur d'une langueur monotone tout suffocant et blême quand sonne l'heure je me souviens des jours anciens et je pleure et je m'en vais au vent mauvais qui m'emporte deçà delà pareil à la feuille morte.

La cohérence du texte ainsi que son identification (Il s'agit d'un texte de *Paul Verlaine*) permettent de valider les conjectures que l'on a émises au cours de notre raisonnement.

III- Le troisième texte à décoder et le « procédé Vigenère»

Comme nous avions fini de décrypter les deux textes donnés en début d'année, Mr Chardard nous a confié un fichier contenant le texte à décoder ci-dessous :

De nouveau, M. Chardard nous a donné une piste pour débuter nos recherches. Il nous a invitées à nous renseigner sur le « Chiffre de Vigenère »

```
kw iue xtèd hkqjknmnp hg oégqoit ni opwulkg
xekd expg wy tgf hg aitdézétlrep sp j
etcmxp tcc gqyxtp mn qewe yp eizei cdwgk
pqyk elv utrqy sp yi rpyv aeu wi hlmtp. ip
pjhpx n'pjhpgvtj fp gjluwp hqffnzr gdx vcsr
aivtx fp qêop pgd jtébygyggd hgd pgextpw pp
wqyx rlw utkptjknevtzgd.
```

Le chiffre de Vigenère repose sur une clé composée de plusieurs lettres.

- La première lettre du message à coder est alors décalée de la position dans l'alphabet de la première lettre de la clé.
- La seconde lettre du message est ensuite décalée de la position dans l'alphabet de la seconde lettre de la clé.

- Et ainsi de suite… Lorsque l'on a épuisé toutes les lettres de la clé, on reprend la première lettre, puis la deuxième lettre…

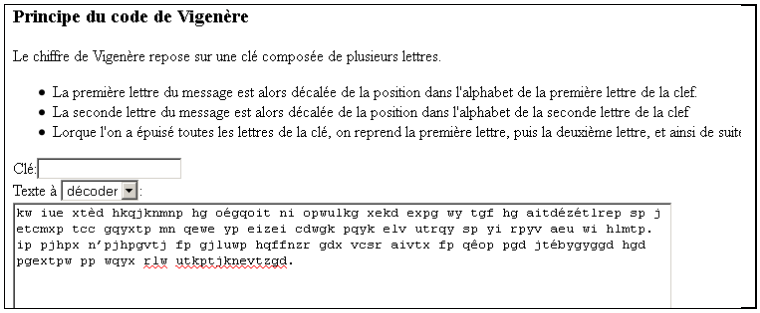
Le procédé de cryptage est nommé ainsi au XIX^{e} siècle en référence au diplomate du XVI^{e} siècle Blaise de Vigenère, qui le décrit dans son traité des chiffres paru en 1586.

Dans un premier temps nous avons essayé de nous familiariser avec ce principe en cryptant et décryptant de petits textes donnés par nos professeurs à l'aide de clés connues.

Ensuite, la difficulté qui s'est présentée pour décrypter notre troisième texte, c'est que nous n'avions aucune idée de la clé qui avait servi à coder le texte !

Nous avons demandé de l'aide à M.Chardard qui nous a conseillé de consulter la page internet suivante :
http://perso.univ-st-etienne.fr/cf39911h/vigenere.html

Nous avons entré le texte à décoder dans une fenêtre :



Et l'ordinateur nous a donné les « occurrences » des différentes paires de lettres dans ce texte :

Occurences des différentes paires de lettres

gd: 157 175 184 187 217
pg: 41 137 143 174 189
qy: 73 97 106 199
pp: 127 173 196
vt: 139 166 214
pj: 61 128 134
tp: 76 124 193
pw: 29 194 197
hg: 18 48 186
xt: 6 75 192
ex: 5 39 191

En effet, lorsque l'on n'a pas la clé qui a servi au codage, il faut la trouver.

Pour cela il faut d'abord déterminer **sa longueur** à l'aide du logiciel évoqué plus haut.

Le logiciel repère les groupes de lettres qui se répètent dans le texte et leur placement dans le texte.

Par exemple, la paire « gd » apparait en positions 157 ; 175 ;184 ;187 et 217 dans le texte.

On étudie ensuite les différences entre les positions données, pour chaque paire. M. Chardard nous a expliqué que les valeurs trouvées sont dans la plupart des cas des multiples de la longueur de la clé. (3)

Nous avons donc calculé toutes les différences de positions pour trouver un diviseur commun à toutes ces valeurs. Voici une partie de ces recherches :

Pour la paire gd :
175-157= 18=2 ×3 ×3 184-175= 9=3 ×3
187-184=3 217-187=30=2 ×3 ×5

Pour la paire pg :
137-41=96=2^5×3 143-137=6=2×3
174-143=31 189-174=15=5×3

Pour la paire pp :
173-127=2×23 196-173=23
…

On a cherché la décomposition en produit de facteurs premiers (à l'aide de nos professeurs) de chacune des valeurs pour déterminer un diviseur commun à presque toutes ces valeurs.

On a conjecturé qu'un diviseur commun à presque toutes ces valeurs est 3 et ainsi on a conjecturé que la **longueur de la clé est 3.**

Si la longueur de la clé est 3 cela signifie qu'une lettre sur 3 est codée par la même lettre de la clé.
On peut ensuite s'intéresser aux fréquences d'apparition des lettres dans le texte codé mais au lieu de prendre toutes les lettres du texte on prend les lettres espacées de la longueur de la clé.

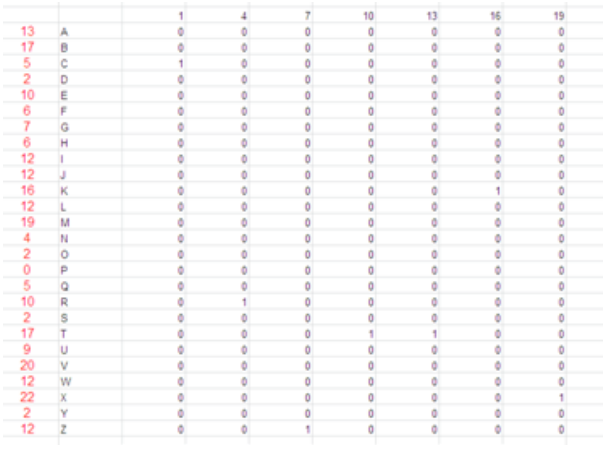
Comme on suppose que la longueur est 3, on calcule la fréquence de chaque lettre dans un texte composé uniquement de la première, la quatrième, la septième et ainsi de suite jusqu'à la dernière lettre du texte :

k w i u e x t è d h k q j k n m n p h g o é g q o i t n i o p w u l k g x e k d e x p g w y t g f h g a i t d é z é t l r e p s p j e t c m x p t c c g q y x t p m n q e w e y p e i z e i c d w g k p q y k e l v u t r q y s p y i r p y v a e u w i h l m t p. i p p j h p x n' p j h p g v t j f p g j l u w p h q f f n z r g d x v c s r a i v t x f p q ê o p p g d j t é b y g y g g d h g d p g e x t p w p p w q y x r l w u t k p t j k n e v t z g d.

On procèdera de même pour un texte composé de la deuxième, la cinquième, la huitième lettre etc...

k w iu e xt è d h k q j k nm n p h g oé g q o i t n i op w ul k g x e kd e xp g wy t gf h g a i td é zé t lr e p s p j e t c m xp t cc g qy x t p m n q e we y p e i ze i cd w gk p qy k el v ut r qy s p y i rp y v a e u w i hl m t p. i p pj h px n' pj h pg v tj f p g j lu w p h q ff n zr g dx v cs r a i v t x f p q ê op p gd j té b yg y gg d hg d pg e xt p w p p wq y x r l w u t kp t j k n ev t z g d.

Et encore de même pour un texte composé de la troisième, la sixième, la neuvième lettre etc...
Après avoir obtenu ces fréquences, on les compare aux fréquences des lettres dans la langue française.
Pour calculer ces fréquences, on s'est servi d'une feuille de calcul excel. (4)

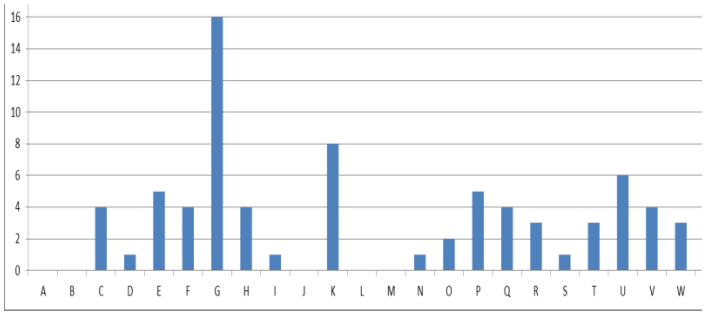


Sur la première ligne on lit le rang de la lettre dans le texte codé,
Dans la colonne associée on lit « 1 » lorsque la lettre de l'alphabet associée est la bonne

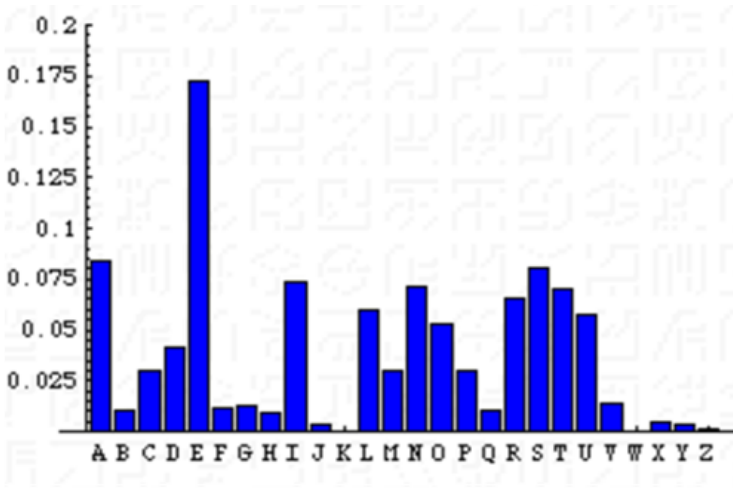
Exemples : La lettre de rang 1 est un « C » et celle de rang 4 est un « R »

Dans la première colonne on lit l'effectif de chaque lettre. On peut ainsi calculer les fréquences des lettres dans le texte composé de la première, la quatrième, la septième et ainsi de suite jusqu'à la dernière lettre du texte.

Voici le diagramme en bâton donnant les fréquences observées pour les 1e 4e 7e … lettres



et le diagramme en bâtons
des lettres dans la langue française



En comparant ces deux diagrammes en bâtons, on peut faire par déduction une association entre les lettres comme pour le codage par substitution.

Par exemple, dans le texte, on s'intéresse à la suite des lettres entourées :

k w i u e x t è d h k q j k n m n p h g o é g q o i t n i o p w u l k g x e k d e x p g w y t g f h g a i t d é z é t l r e p s p j e t c m x p t c c g q y x t p m n q e w e y p e i z e i c d w g k p q y k e l v u t r q y s p y i r p y v a e u w i h l m t p. i p p j h p x n' p j h p g v t j f p g j l u w p h q f f n z r g d x v c s r a i v t x f p q ê o p p g d

j té b yg y gg d hg d pg e xt p w p p wq y x r l w u t kp t jk n ev t zg d .

En comparant les diagrammes en bâtons des fréquences, on peut conjecturer que la lettre G se « décode » par la lettre E et ainsi conjecturer qu'il y a eu un décalage de 3 lettres donc que la première lettre de la clé est la lettre « C » .
Ce qui nous permet ensuite de conjecturer un décryptage de toutes les lettres entourées.

On travaille ensuite sur une autre suite de lettres entourées :

k w iu e xt è d h k qj k nm n p h g oé g qo i t n i op w ul k g x e kd e xp g wy t gf h g a i td é zé t lr e p s p j e tc m xp t cc g qy x tp m n q e we y p e i ze i cd w gk p qy k el v ut r qy s p y i rp y v a e u w i hl m tp. i p pj h px n'pj h pg v tj f p g j lu w p h q ff n zr g dx v cs r ai v tx f p q ê op p gd j té b yg y gg d hg d pg e xt p w p p wq y x r l w u t kp t jk n ev t zg d .

On peut conjecturer que la lettre T se « décode » par la lettre E et ainsi conjecturer au vu du décalage (12 lettres) que la deuxième lettre de la clé est la lettre « L ». Ce qui nous permet ensuite de conjecturer un décryptage de toutes les lettres entourées.

On travaille finalement sur le dernier texte constitué des lettres n°3 ; 6 ;9…
On peut conjecturer que la lettre A se « décode » par la lettre E et ainsi conjecturer que la dernière lettre de la clé est la lettre « E ». Ce qui nous permet ensuite de conjecturer un décryptage des dernières lettres du texte.

La clé que l'on a trouvée est CLE !

La première lettre K a donc été décalée de 3 lettres (car C est la 3e lettre de l'alphabet donc c'est un I dans le texte d'origine. (on « remonte » de 3 lettres en partant de K incluse)
De même la lettre W a été décalée de 12 lettres (car L est la 12e lettre de l'alphabet donc c'est un L dans le texte d'origine (on « remonte » de 12 lettres en partant de W incluse) .

On décrypte donc le texte proposé en :

Il est très difficile de décoder ce message mais avec un peu de persévérance on y arrive par contre il faut un texte assez long car sinon on ne peut pas le faire. En effet l'effectif de chaque doublon est trop petit de même les fréquences des lettres ne sont pas significatives.

La méthode de cryptanalyse utilisée est celle trouvée par Kasiski .

Le procédé de Vigenère n'est **pas une technique de cryptage très fiable** puisqu'il faut peu de temps pour trouver la clé et décoder le message (à condition d'avoir des ordinateurs à disposition…)

IV- Le cryptage Diffie-Hellman :

Nous avons poursuivi en travaillant sur un système de codage plus récent avec clés privées et publiques : le cryptage de Diffie-Hellman. M. Chardard nous a demandé de faire des recherches documentaires.

Le cryptage de Diffie-Hellman s'utilise avec deux clés de décodage. L'une est publique et visible par tous alors que l'autre n'est connue que des interlocuteurs, elle est dite secrète.
Cellya et moi (Ines) avons testé ce dispositif.
Contrairement aux autres codes, le protocole de Diffie-Hellman est encore utilisable, car on ne sait pas en faire la cryptanalyse.

D'abord, nous avons choisi ensemble un très grand nombre premier **p=4051** (un nombre dont les seuls diviseurs sont 1 et lui-même) et un entier a tel que **1≤ a ≤p-1** ;
nous avons pris **a=12**.
Ensuite, nous avons choisi secrètement deux nombres entiers $\mathbf{X_1}$et $\mathbf{X_2}$.
Cellya a choisi $X_1=4$ et j'ai pris $X_2=6$.
On a chacune calculé :
$a^{X1}=20736$ et $a^{X2}=2985984$.
Puis nous avons échangé nos résultats modulo p pour calculer $(a^{X1})^{X2}$ modulo p. (5)

	Cellya	Ines	EXPLICATIONS
Etape 1	P=4051 a=12 (public)	P=4051 a=12 (public)	On choisit p un très grand nombre premier et a un entier tel que 1≤a≤p-1
Etape 2	Cellya choisit X_1 secrètement $X_1 = 4$	Ines choisit X_2 secrètement $X_2 = 6$	
Etape 3	Elle calcule $Y_1 \equiv a^{X1}$ [mod p] $Y_1 \equiv$ 481[mod p]	Elle calcule $Y_2 \equiv a^{X2}$ [mod p] $Y_2 \equiv$397 [mod p]	

Etape 4			Elles échangent Y1 et Y2
Etape 5	Elle calcule $Y_2^{X1} \equiv (a) [mod\ p]$ $Y_2^{X1} \equiv 2615 [mod\ p]$	Elle calcule $Y1^{X2} \equiv (a) [modp]$ $Y_1^{X2} = 2615 [mod\ p]$	
Etape 6			Elles obtiennent un nombre B qui permet de coder et décoder des textes secrètement

Ainsi pour crypter un texte il faut utiliser les étapes suivantes : (6)

Etape 1 : On associe un chiffre à chacune des lettres du texte (ex : A --> 1) et on traduit le texte par une série de chiffres.

Etape 2 : On détermine une clé grâce au procédé de Diffie-Hellman

Etape 3 : On fait la somme des deux séries de chiffres afin de trouver une nouvelle série.

Etape 4 : On associe au nouveau texte chiffré des lettres de l'alphabet, le nombre donnant la position de celle-ci.

Dans ce cryptage, il est important que seuls les interlocuteurs soient actifs. Si une troisième personne a la possibilité de connaître la clé privée, elle pourra coder et décoder le message mais aussi le modifier à sa guise.

Le procédé de Diffie-Hellman porte le nom de ses deux auteurs :
Bailey Whitfield Diffie (né le 5 juin 1944) est un cryptologue américain. Il est l'un des pionniers de la cryptographie asymétrique (utilisation d'une paire de clés publique et privée) en collaboration avec Martin Hellman (2 octobre, 1945) lui aussi un cryptologue américain reconnu.

Nous n'avons pas eu le temps d'étudier le système RSA fréquemment utilisé aujourd'hui. Néanmoins nous sommes satisfaites d'avoir pu mener des recherches dans le cadre de cet atelier cette année. Nous remercions nos professeurs et M. Chardard de nous avoir accompagnées et encouragées.

Notes d'édition

(1) F_i correspond à la fréquence d'une lettre quelconque, qu'on notera la i-ème lettre (i étant un entier entre 1 et 26), et pas uniquement à la fréquence de la lettre «i».
Ainsi F_1 correspond à la fréquence de la lettre «A»

(2) C'est essentiellement le même phénomène : plus le texte est court, plus la fluctuation d'échantillonnage est importante. De plus certains textes peuvent être extrêmement atypiques : imaginez travailler par exemple avec un extrait du roman La Disparition (de Georges Perec), écrit sans la lettre «E».

(3) Certaines paires de lettres sont plus fréquentes dans la langue française que d'autres (c'est le cas de la paire « ES » par exemple). Si deux occurrences d'une paire dans le texte codé, par exemple gd, correspondent à la même paire dans le texte en clair, cela signifie qu'elles ont été codées de la même manière, donc qu'elles sont séparées par un nombre entier de fois la clé. Si le texte est assez long, cela se produira assez souvent pour les paires fréquentes en français.
Bien évidemment, une paire gd peut aussi arriver d'une autre manière : cela explique que, plus loin dans l'article, les différences des positions des paires pg ne sont pas toujours multiples de 3.

(4) Il semble que ce tableau, malheureusement peu lisible, ne correspond pas à l'exemple du texte codé donne dans l'article, mais à un texte codé qui commencerait par
C_ _R_ _Z_ _T_ _T_ _K_ _X

(5) a^x modulo p désigne le reste de la division euclidienne de a^x par p. Comme p est premier et a<p, le reste n'est jamais 0.

(6) L'édition regrette l'absence d'exemple pour crypter un texte avec ce dernier procédé ainsi que les raisons pour lesquelles ce cryptage fonctionne.

Sources :

fr.wikipedia.org
http://www.bibmath.net/crypto
http://exo7.emath.fr/cours/ch_crypto.pdf
http://www.dcode.fr/
Le livre de Simon Singh :
Histoire des codes secrets