
# Automatisation de la cryptanalyse des cryptosystèmes classiques à l'aide d'algorithmes modernes

Helder Brito
O'nel Hounnouvi

# Table des matières

# 1 Substitution monoalphabétique

## 1.1 Introduction

La substitution monoalphabétique est l'une des plus anciennes méthodes de chiffrement. Elle consiste à remplacer dans le message clair une lettre donnée de l'alphabet par une autre lettre. Voici un exemple :

| 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

Le message *SUBSTITUTION* devient *PRYPQFQRQFLK*.

L'alphabet latin comporte 26 lettres. Cela permet donc de construire $26! = 4 \times 10^{26}$ permutations. Soit de l'ordre de $2^{88}$. Sachant qu'environ $2^{58}$ secondes se sont écoulées depuis la création de l'univers, il serait impossible d'explorer toutes les permutations. Ce chiffre donne une impression de sûreté qui est toutefois trompeuse...

## 1.2 Cryptanalyse

La substitution monoalphabétique possède de grosses faiblesses structurelles. Les chiffres utilisant cette méthode sont "facile" à casser par analyse fréquentielle. Notre analyse ici sera basée sur l'analyse des fréquences d'apparitions des n-grammes dans le message chiffré.

Un *n-gramme* est une séquence de $n$ lettres consécutives dans un texte. Par exemple, dans le mot *CRYPTANALYSE*, les bigrammes ($n = 2$) incluent *CR*, *RY*, *YP*, ..., tandis que les trigrammes ($n = 3$) sont *CRY*, *RYP*, *YPT*, .... En utilisant un dictionnaire de référence contenant les fréquences relatives des n-grammes dans un large corpus de textes en français, il est possible d'estimer la probabilité qu'un texte donné soit écrit dans cette langue.

Pour évaluer la qualité des solutions potentielles et identifier celle qui correspond le mieux à du français, nous attribuons un score à chaque solution avec une fitness function.

La fonction de score utilisée dans ce projet est basée sur la somme des probabilités logarithmiques des n-grammes. C'est la fonction log-vraisemblance Elle est définie comme suit :

$$score = -\sum \log(frequence(c_1 \dots c_n))$$

L'objectif ici est de minimiser cette fonction à cause du changement de signe. La solution ayant le plus petit est celle qui sera "le plus" français.

## 1.3 Métaheuristiques

Une *métaheuristique* est un algorithme d'optimisation visant à résoudre des problèmes d'optimisation pour lesquels on ne connaît pas de méthode classique plus efficace. Les métaheuristiques sont généralement des algorithmes stochastiques [1] itératifs, qui progressent vers un optimum global (c'est-à-dire l'extremum global d'une fonction).

1. Un processus stochastique est un processus qui intègre des éléments d'aléatoire, c'est-à-dire dont l'évolution est déterminée par des phénomènes aléatoires.

### 1.3.1 Hill Climbing

L'idée générale pour trouver la clef de déchiffrement est la suivante :

1. Partir d'une clef aléatoire ;
2. Utiliser la clef pour déchiffrer le cryptogramme ;
3. Calculer le score du texte obtenu ;
4. Si ce score est meilleur que le score précédent, prendre cette nouvelle clef comme clef courante, sinon garder la clef précédente ;
5. Modifier légèrement la clef courante ;
6. Retourner en 2 tant qu'on n'a pas trouvé la bonne clef.