

# IFT-4001 (Hiver 2016) Projet d'exploration

Alexandre Cormier (111 101 150)

Alexandre Picard-Lemieux (111 103 625)

Patrick Côté (111 103 743)

Vincent Beaudoin (111 103 778)

24 avril 2016

# 1 Introduction

L'optimisation combinatoire est une branche de l'informatique et des mathématiques appliquées. C'est la recherche d'une solution au coût minimal d'un problème dont l'espace des solutions est discret.

Pour le projet d'exploration, nous devons apprendre quelque chose sur les solveurs ou l'optimisation combinatoire qui n'a pas été mentionné dans le cours et en faire mention dans notre rapport. Nous ferons une étude de la viabilité de la programmation par contraintes pour attaquer la cryptographie classique par substitution.

Ce rapport sera séparé en plusieurs parties. Nous commencerons par faire une description du problème. Ensuite, nous proposerons des approches. Par la suite, le rapport sera finalisé avec le protocole d'expérimentation, les résultats ainsi qu'une discussion sur ces résultats.

## 2 Description du problème

Le chiffrement par substitution est une façon de cacher un message en remplaçant chaque lettre par une autre. On parle de substitution monoalphabétique lorsque qu'une lettre est toujours remplacée par la même autre lettre. Autrement, on parle de substitution polyalphabétique.

Un exemple de chiffrement par substitution monoalphabétique est le chiffre de César. Pour ce chiffre, on associe à chaque lettre sa position dans l'alphabet, commençant par 0. La clé est une lettre et chaque lettre du message est décalée d'un nombre de positions correspondant à la clé. Par exemple, le message ABC chiffré avec la clé B devient BCD. Les calculs de décalage se font en modulo 26, tel que XYZ chiffré avec la même clé devient YZA.

Le chiffre de vigenère est un exemple de chiffrement par substitution polyalphabétique. La clé est composée d'une ou plusieurs lettres et le message est placé divisé en partie de la même longueur que la clé. Si les différentes parties du message sont placées l'une en-dessous de l'autre, il suffit d'appliquer le chiffre de César à chaque colonne  $i$  avec la  $i^{eme}$  lettre de la clé. Par

exemple, le message ABCD chiffré avec la clé BC devient BDDF.

La substitution est toujours une composante à la base de la cryptographie moderne, mais ces méthodes de chiffrement simples ne sont plus utilisées comme systèmes cryptographiques à part entière puisqu'elles sont vulnérables à différentes attaques. Il est possible, notamment, d'analyser la fréquence relative des lettres dans le message chiffré et comparer avec la fréquence relative des lettres dans la langue du message d'origine pour trouver les clés les plus probables.

Dans le cadre de ce projet, nous étudierons une méthode alternative, utilisant la programmation par contrainte, pour retrouver la clé de tels chiffrements à partir du message chiffré. Il est évidemment impossible pour un programme automatisé de trouver la clé de façon certaine, car il n'y a pas nécessairement moyen de reconnaître le message d'origine lorsque la bonne clé est trouvée. Par contre, nous pourrions ordonner les clés les plus probables selon, notamment, la fréquence relative de chaque lettre dans le message déchiffré.

L'objectif ici est d'étudier la viabilité d'une telle approche, par contrainte, pour l'attaque du chiffrement par substitution. Nous étudierons l'efficacité de l'approche selon la longueur du message et, dans le cas du chiffre de Vigenère, de la longueur de la clé.

Par exemple, en prenant le message encrypté LXFOPVEFRNHR, le solveur doit trouver que la clé la plus probable est LEMON et que le texte original est ATTACKATDAWN. L'implémentation utilisée utilisera les fréquences des lettres trouvées d'une langue donnée.<sup>1</sup> Cette implémentation sera donc plus efficace si la chaîne est longue et si elle utilise des vrais mots de la langue choisie.

---

1. Wikipedia, [En ligne]. [https://en.wikipedia.org/wiki/Letter\\_frequency](https://en.wikipedia.org/wiki/Letter_frequency) (Page consultée le 3 avril 2016)

### 3 Approche(s) proposée(s)

On prend  $n$  qui est la longueur de la chaîne à déchiffrer,  $c$  qui est la longueur de la clé,  $l$  qui est le nombre de lettres dans l'alphabet, et  $p$  qui est la précision de la fréquence.

On déclare une variable pour chaque lettre de la chaîne cryptée. Ainsi, la variable  $X_i$  représente la valeur de la lettre de la chaîne cryptée à la position  $i$  pour  $1 \leq i \leq n$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $l - 1$ .

On déclare une variable pour chaque complément de clé. Ainsi, la variable  $K_i$  représente la valeur de la lettre de la clé à la position  $i$  pour  $1 \leq i \leq c$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $l - 1$ .

On déclare une variable pour le texte intermédiaire. Ainsi, la variable  $I_i$  représente la valeur de la lettre de la chaîne à la position  $i$  pour  $1 \leq i \leq n$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $2 * (l - 1)$ .

On déclare une variable pour le texte brut. Ainsi, la variable  $B_i$  représente la valeur de la lettre de la chaîne à la position  $i$  pour  $1 \leq i \leq n$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $l - 1$ .

On déclare une variable pour les comptes. Ainsi, la variable  $C_i$  représente le compte de chaque lettre dans la chaîne à la position  $i$  pour  $1 \leq i \leq l$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $n$ .

On déclare une variable pour les comptes multipliés. Ainsi, la variable  $M_i$  représente le compte multiplié par la précision de la fréquence de chaque lettre dans la chaîne à la position  $i$  pour  $1 \leq i \leq l$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $n * p$ .

On déclare une variable pour les fréquences multipliées. Ainsi, la variable  $F_i$  représente la fréquence multipliée de chaque lettre dans la chaîne à la position  $i$  pour  $1 \leq i \leq l$ . Le domaine de chaque variable est l'ensemble des entiers entre 0 et  $p$ .

On déclare une variable pour les fréquences multipliées négatives du lan-

gage. Ainsi, la variable  $N_i$  représente la fréquence multipliée négative de chaque lettre de l'alphabet à la position  $i$  pour  $1 \leq i \leq l$ . Le domaine de chaque variable est l'ensemble des entiers entre  $-p$  et  $0$ .

On déclare une variable pour les différences. Ainsi, la variable  $D_i$  représente la différence entre les fréquences multipliées et les fréquences multipliés du langage à la position  $i$  pour  $1 \leq i \leq l$ . Le domaine de chaque variable est l'ensemble des entiers entre  $-p$  et  $p$ .

On déclare une variable pour les différences absolues. Ainsi, la variable  $A_i$  représente valeur absolue de la différence entre les fréquences multipliées et les fréquences multipliés du langage à la position  $i$  pour  $1 \leq i \leq l$ . Le domaine de chaque variable est l'ensemble des entiers entre  $0$  et  $p$ .

On déclare une variable  $S$  pour la somme des différence des fréquences. Le domaine de la variable est l'ensemble des entiers entre  $0$  et  $l * p$ .

Pour chaque lettre  $i$  de la chaine, nous avons ces contraintes.

$$X_i + K_{i\%c} = I_i \quad (1)$$

$$I_i \% l = B_i \quad (2)$$

Pour chaque lettre  $i$  de l'alphabet, nous avons ces contraintes.

$$Compte(i, B, C_i) \quad (3)$$

$$C_i * p = M_i \quad (4)$$

$$M_i / n = F_i \quad (5)$$

$$F_i + N_i = D_i \quad (6)$$

$$A_i = |D_i| \quad (7)$$

Pour finir, nous avons cette contrainte.

$$A_1 + A_2 + \dots + A_l = S \quad (8)$$

La fonction objective est de minimiser S. L'heuristique par défaut a été utilisé.

Nous avons donc  $n+c+n+n+l+l+l+l+l+l+1 \in \Theta(n+c+l)$  variables.

Nous avons donc

$$\begin{aligned} & nl+cl+n(2l-1)+nl+l(n+1)+l(np+1)+l(p+1)+l(p+1)+l(2p+1)+l(p+1)+(lp+1) \\ = & nl+cl+2nl-n+nl+ln+l+nlp+l+lp+l+lp+l+2lp+l+lp+l+lp+1 \\ = & 5nl+cl-n+6l+nlp+6lp+1 \\ & \in \Theta(nlp+cl+ln+lp+l-n) \end{aligned}$$

valeurs.

Nous avons n contraintes de type (1) et (2). Nous avons l de type (3), (4), (5), (6) et (7). Nous avons une contrainte de type (8). Nous avons donc au total  $2n+5l+1 \in \Theta(n+l)$ .

Ceci est donc notre modèle finale. Lors de nos premières itérations, nous avons mis que la fréquence des lettres était restée dans des variables réelles. Par contre, grâce à la variable de précision, nous convertissons les fréquences, qui sont réelles, en entiers.