Relatório 1 de ACH2076 Segurança da Informação

Daniel Augusto de Melo Moreira, Felipe Brigatto, Georges Basile Stavracas Neto, Marcelo Gaioso Werneck

Escola de Artes, Ciências e Humanidades (EACH) – Universidade de São Paulo (USP)

Introdução

Este relatório consiste em três experimentos abordando a temática inicial da matéria Segurança da Informação. A primeira experiência requer demonstrar a frequência de ocorrência de cada letra utilizando um texto grande para as quatro cifras: Monoalfabética, Vigènere, Hille Vigènere incrementado. Na segunda experiência verifica-se a dificuldade de encontrar a chave utilizando força bruta das seguintes cifras: Hill, Vigènere Monoalfabética. Na terceira experiência precisa-se descobrir o texto criptografado a partir de dois textos.

Para realizar estes experimentos, utilizamos a linguagem Octave. A seguir encontra-se nossos resultados e considerações.

Experimento 1

No primeiro experimento foi implementado os algoritmos de criptografia e decriptografia das cifras de Monoalfabética, Vigènere, Vigènere Incrementado e Hill, na linguagem do Matlab/Octave. Esses algoritmos são funções que recebem como parâmetro uma chave e um texto e retorna o texto criptografado ou decriptografado seguindo a respectiva cifra.

A ideia do experimento é, além de entender o funcionamento de cada uma, analisar a frequência de ocorrência das letras no texto cifrado de cada uma das cifras, pois estas são de grande valor para adversários que tentam quebrar as cifras, já que quanto maior a disparidade de ocorrência entre a frequência das letras mais fácil é de um adversário quebrar a cifra aplicando análise estatística sobre a mesma.

Como resultados, obtemos que a cifra Monoalfabética (a mais simples de todas) tem uma grande disparidade entre as frequências de letras do texto criptografado em relação às outras cifras. Isso era o esperado, visto que o algoritmo simplesmente troca uma letra do texto original por outra, de acordo com a chave, mantendo assim a frequência de letras na cifra igual a frequência do texto original.

A cifra de Vigènere é uma exceção, pois a frequência das letras é influenciada pelo tamanho da chave: quanto maior for a quantidade de caracteres da chave (se bem escolhidos), menor será essa variância. A razão disso está no fato de que a chave será repetida várias vezes até chegar no mesmo tamanho do que o texto original, portanto com chave pequena haverá uma maior disparidade entre as frequências das letras.

Agora a cifra de Vigènere Incrementado já se difere das demais cifras nesse sentido pois o crescimento da porcentagem de ocorrência das letras é muito menor em relação à todas as

outras, parecendo até constante no gráfico com as demais cifras. Ela funciona como Vigènere, só que usando o próprio texto original em conjunção com a chave para criptografar ou decritografar o texto. Apesar desse algoritmo conseguir manter uma frequência entre todas as letras da cifra quase constante, o que é desejável, esse é o algoritmo mais fácil de ser quebrado utilizando força bruta, pois apenas o primeiro caractere da chave é escondido e os demais caracteres são a própria cifra.

Por fim, a cifra de Hill é a que possui a menor variância entre as porcentagens de ocorrência das letras em relação as demais cifras. No máximo ela possui uma porcentagem maior que a de Vigènere Incrementado mas não em quantidade relativamente significativa. Válido lembrar que a cifra de Hill é a única cifra em blocos de todas as trabalhadas, ou seja, enquanto as outras cifras fazem a criptografia e decriptografia caractere por caractere, a cifra de Hill faz isso por blocos de caracteres.

Tabela 1. Quadro comparativo entre algoritmos de Hill, Vigenere, Monoalfabetica e Vigenere Incrementado

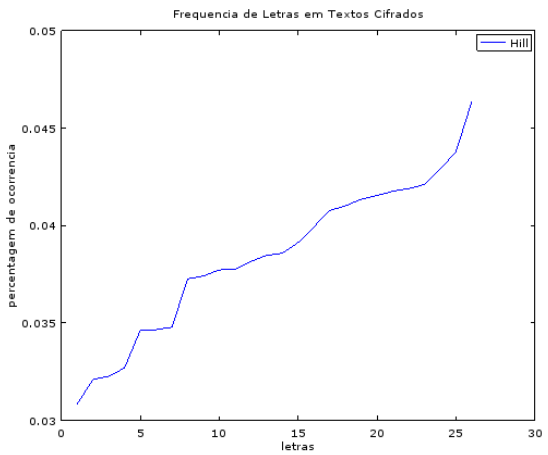


Figura 1. Algoritmo Hill para chave de comprimento 3x3.

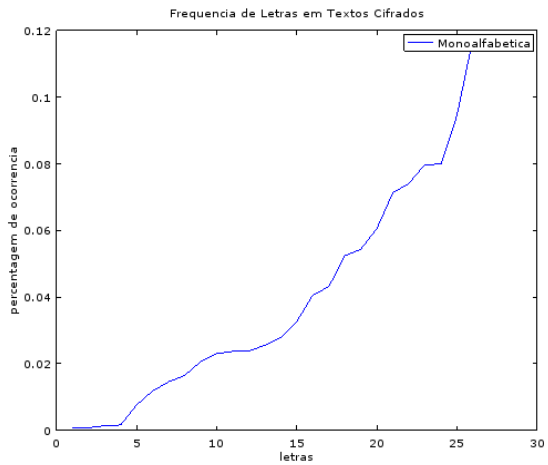
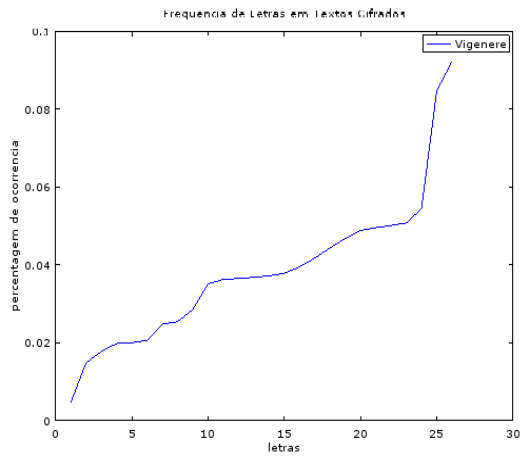


Figura 2. Monoálfabetica para chave de comprimento 26.



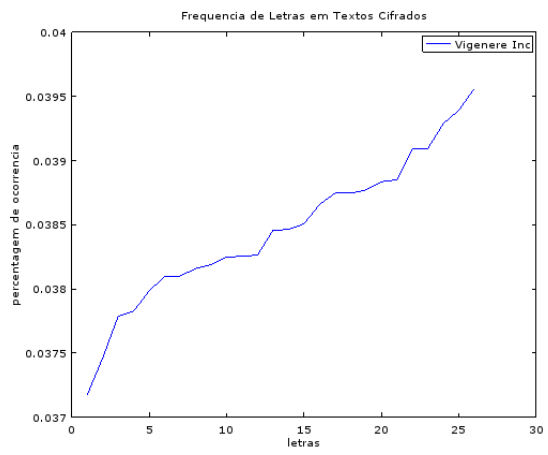


Figura 3. Vigènere para chave de comprimento 3.	Figure 4. Vigènere Incrementado com chave inicial de comprimento 1.

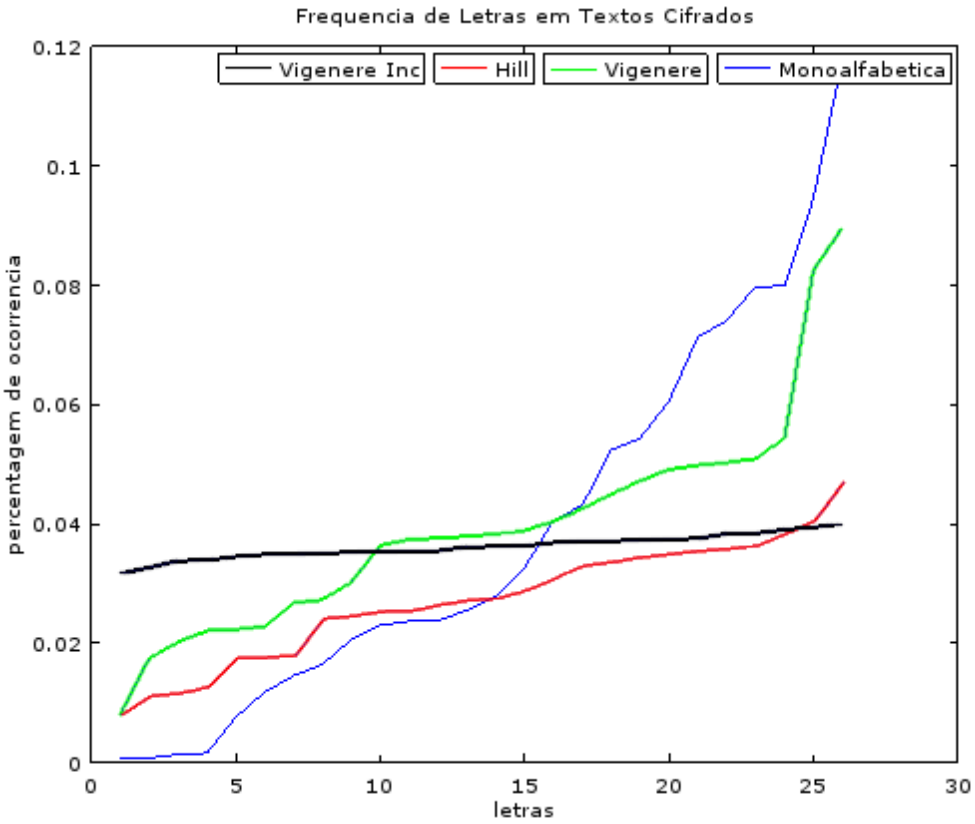


Figura 5. Sobreposição dos gráficos dos algoritmos Hill, Vigènere, Monoálfabetica e Vigènere Incrementado

Experimento 2

Nesse experimento foram testados os seguintes três algoritmos clássicos de criptografia:

- Algoritmo de Vigènere com chaves de tamanho três e valores variando de 0 e 16 (inclusive), resultando em um espaço de chaves K = 4913;
- Algoritmo Monoálfabetico com chaves permutando os 7 primeiros caracteres entre "etaoins", resultando em um espaço de chaves K = 5040, e;
- Algoritmo de Hill com chaves de tamanho 2x2 e valores variando de 0 e 10 (inclusive), resultando em um espaço de chaves K = 4524.

Para cada algoritmo foi analisado o tempo de avaliar todas as chaves $k \in K$ e a verossimilhança da chave k ser a chave correta, considerando monogramas e digramas.

Em cada teste realizado com textos de tamanho entre 10 e 100 (inclusive) com incrementos de 10 e 50 textos analisados por tamanho, o algoritmo de Vigènere e Hill demoram cerca de 1:30h cada e o Monoálfabetico demorou cerca de 50 minutos para avaliarem todo o espaço de chaves K.

Como podemos observar nos gráficos abaixo as cifras de Vigènere e de Hill convergem de forma semelhante quando a chave k é analisada considerando digramas, ambas precisando de textos de aproximadamente 50 e 30 caracteres respectivamente para obtermos certeza da chave k ser a chave verdadeira utilizada na criptografia da mensagem.

Porém, as cifras de Vigènere e Hill convergem diferentemente quando a chave k é analisada considerando monogramas. Na cifra de Vigènere a probabilidade da chave correta em relação ao tamanho do texto analisado converge linearmente e com textos de aproximadamente 100 caracteres obtemos acertos da chave k com probabilidade próximas de 1. Já na cifra de Hill a probabilidade da chave correta em relação ao tamanho do texto analisado converge em escala logarítmica e a probabilidade máxima para qualquer texto suficientemente grande será de aproximadamente 0.5.

Finalmente, a cifra Monoálfabetica converge linearmente quando a chave k é analisada considerando tanto monogramas quanto digramas, porém a taxa de crescimento são bem distintas. Quando analisamos digramas obtemos certeza da chave correta com textos de tamanho próximos de 100 caracteres, porém com monogramas mesmo analisando textos 10 vezes maior do que com digramas a probabilidade máxima encontrada foi de apenas 0.06.

Tabela 2. Monogramas

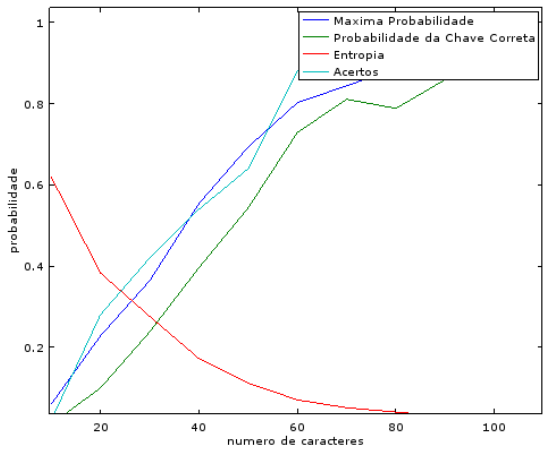


Figura 6. Vigènere.

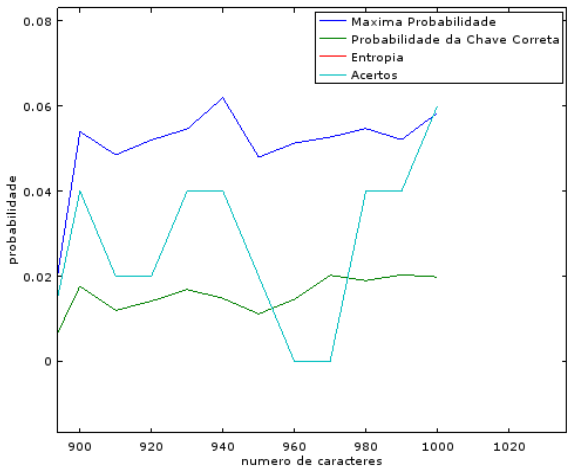


Figura 7. Monoalfabética.

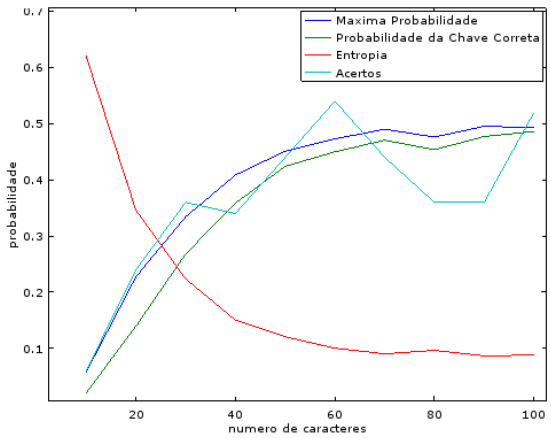


Figura 8. Hill.

Tabela 3. Diagramas

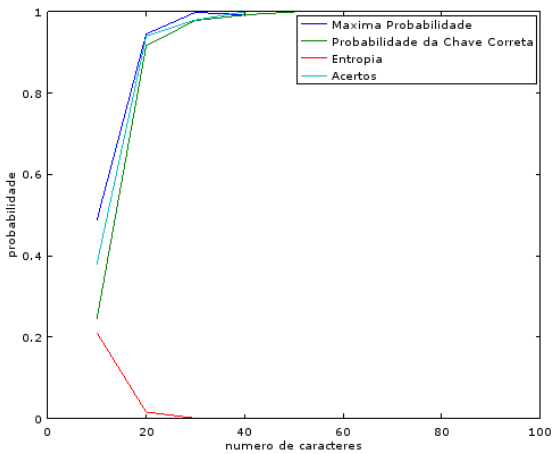


Figura 9. Vigènere.

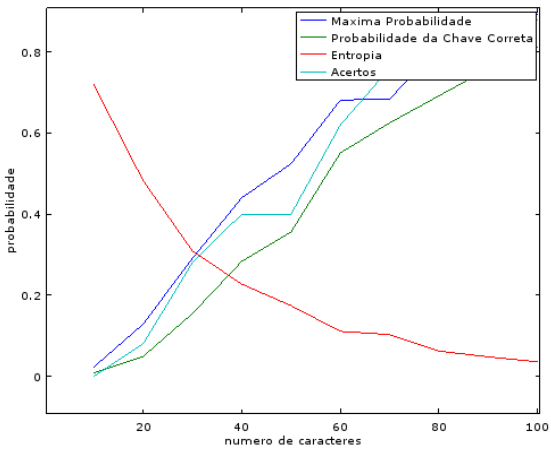


Figura 10. Monoalfabética.

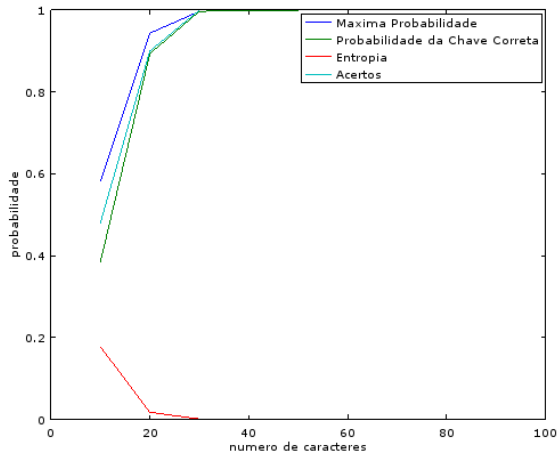


Figura 11. Hill.

Experimento 3

O primeiro passo foi ler as cifras dos respectivos arquivos dados. Os arquivos estão organizados em *[prefixo]{A|B}.txt*, em que há em cada par de arquivo com o mesmo prefixo, a mesma chave foi utilizada para criptografar suas mensagens e possuem cem caracteres cada. Com as cifras lidas, faz-se a subtração de seus caracteres. O resultado é a distância entre cada caracter do texto, já sem a interferência da chave *k*.

$$diff = c1 - c2$$

Agora com o vetor de diferenças (*diff*), há duas abordagens comuns para a dedução do texto original: a análise estatística e a detecção de padrões (MASON, WATKINS, *et al*, 2006). Para este experimento, entretanto, foi decidido utilizar uma terceira forma de análise, onde o espaço *M* de mensagens é conhecido (no caso, "NotesUnderground.txt"), com as seguintes considerações:

- Todos os caracteres estão em caixa baixa
- Ambas as mensagens estão presentes no espaço de mensagens

Assim para encontrarmos as mensagens originais avaliamos iterativamente todas as mensagens $m1 \in M$, e para cada mensagem m1 avaliada encontramos uma mensagem *m2* subtraindo, caractere por caractere, m1 da diferença entre as cifras c1 e c2.

$$m2 = (m1 - diff) \; mod \; 26$$

Realizamos essa operação para todo *m1* até que encontremos uma mensagem $m2 \in M$, quando teremos encontrado os textos referentes a *c1* e *c2*.

O método acima pode ser resumido em encontrar um *m2* que satisfaça a seguinte formula: $\exists (m2 \in M) \forall (m1 \in M) \; m2 = (m1 - diff) \; mod \; 26$.

O nosso grupo ficou encarregado de encontrar o texto referente ao terceiro par de cifras e com o método acima obtivemos os seguintes textos:

> A - rut and had a whole hearted terror of any kind of eccentricity in myself but how could i live up to it i was morbidly sensiti
>
> B - rhaps never will have an object for your spite that it is as leight of hand a bit of juggling a cardsharpers trick that it is

Referências

MASON, J.; WATKINS, K,; *et al*. **A Natural Language Approach to Automanted Cryptanalysis of Two-time Pads**, 2006. Disponível em <http://www.cs.jhu.edu/~jason/papers/mason+al.ccs06.pdf>