RECONHECIMENTO DE PADRÕES EM CRIPTOGRAMAS

*Nilson Mori Lazarin*Jóse Antônio Moreira Xexéo*
Instituto Militar de Engenharia, Seção de Engenharia da Computação–Praça General Tibúrcio, 80, 22290-270, Praia Vermelha, Rio de Janeiro, RJ, Brasil.
**nlazarin@cefet-rj.br*

RESUMO

Ao tentar obter o conteúdo de uma mensagem criptografada, partindo apenas do texto cifrado capturado, um invasor pode utilizar a técnica da busca exaustiva, sendo esse o pior caso de ataque, pois é necessário testar todas as chaves de todos os algoritmos que possivelmente foram utilizados na cifração. Para tentar diminuir o custo computacional desse tipo de ataque, uma das armas utilizadas pelo invasor é o reconhecimento de padrões, que pode servir para identificar o algoritmo gerador de um criptograma, utilizando apenas textos cifrados. Essa técnica, além de possibilitar uma diminuição do custo computacional, pode levar à descoberta de vulnerabilidades que possibilitem a obtenção do conteúdo de um criptograma, de forma desautorizada. Este artigo contribui para a evolução das técnicas de Recuperação da Informação aplicadas ao reconhecimento de padrões em criptogramas, do Grupo de Segurança da Informação do IME.

Palavras-chave: criptologia, reconhecimento de padrões, classificação.

ABSTRACT

When trying to get the contents of an encrypted message, leaving only the captured ciphertext, an attacker can use the brute force. This is the worst case of attack, because is necessary to test all the keys of all the algorithms that were possibly used in encryption. To reduce the computational cost of such attack, one of the weapons used by the attacker is the pattern recognition, which may serve to identify the algorithm generator of a cryptogram, using ciphertext only. This technique, besides enabling a reduction in the computational cost, can lead to the discovery of vulnerabilities that allow obtaining the contents of a cryptogram from unauthorized manner. This paper contributes to the evolution of information retrieval techniques applied to cryptology.

Keywords: Cryptology, pattern recognition, classification.

INTRODUÇÃO

A criptografia é o ramo da Criptologia que reúne as técnicas para a proteção do conteúdo de uma mensagem, entre outras funções. Compartilhando algoritmo chave, remetente e destinatário podem trocar mensagens de forma segura. De forma que a segurança de uma mensagem criptografada reside na robustez do algoritmo e no sigilo da chave.

Contemporaneamente, têm-se buscado modelos matemáticos cada vez mais sofisticados para a produção de algoritmos criptográficos. E um dos requisitos para garantir a confiabilidade de um algoritmo é a inexistência de assinaturas nos criptogramas gerados, para isso um algoritmo criptográfico deve dissipar qualquer padrão que possa existir na mensagem, sendo submetido a testes estatísticos para garantir sua eficiência, credibilidade e segurança (SCHNEIER, 1996) (LAMBERT, 2004) (OLIVEIRA, 2011).

Os testes estatísticos são comumente utilizados para avaliação de algoritmos criptográficos. Em (SOTO; BASSHAM, 2000), por exemplo, o conjunto de testes estatísticos proposto por (RUKHIN *et al.*, 2010) foi utilizado como parâmetro de avaliação dos candidatos do concurso AES (Advanced Encryption Standard), no qual o algoritmo Rijndael (DAEMEN; RIJMEN, 1999) foi o vencedor, podendo ser adotado como algoritmo criptográfico por qualquer organização.

Entretanto, a adequabilidade dos testes estatísticos propostos pelo NIST (National Instituteof Standards and Technology) no concurso AES, foi contestada por (MURPHY, 2000), que aconselhou a realização de testes complementares para análise de aleatoriedade, já que nenhum teste estatístico pode efetivamente certificar um algoritmo como seguro.

Essa contestação foi reforçada pela suposição de (CARVALHO, 2006) de que padrões linguísticos da mensagem pudessem ser propagados ao criptograma. Tal hipótese culminou no início das pesquisas de ataque de distinção baseados em Recuperação da Informação (RI) no IME, suscitando um método de identificação do período da chave utilizada na encriptação com a cifra de Viginère.

Nessa mesma linha de pesquisa, (SOUZA, 2007) utilizou técnicas de linguística computacional para subsidiar a busca de padrões em criptogramas, no qual se pôde identificar fragilidades em alguns algoritmos, quando utilizados em modo ECB. Possibilitando o agrupamento de criptogramas maiores que 1KB, gerados pelos algoritmos DES, AES e RSA, com base na chave de cifração utilizada.

Posteriormente, (TORRES *et al.*, 2010) aplicou técnicas de agrupamento, baseadas em teoria dos grafos, possibilitando o agrupamento de criptogramas maiores que 4KB, cifrados em modo ECB, oriundos dos cinco finalistas do AES: Rijndael, MARS, Serpent, Twofish e RC6.

O último trabalho apresentado pela linha de pesquisa do IME é o de (OLIVEIRA, 2011) que, através do uso de algoritmos genéticos modelados, possibilitou a identificação da quantidade de chaves distintas, utilizadas na cifração de conjuntos de criptogramas de tamanho maior que 6KB, gerados em modo ECB.

Várias outras pesquisas relacionadas à classificação de criptogramas têm sido apresentadas desde 2001, tais como: (MAHESHWARI, 2001), (CHANDRA, 2002), (RAO, 2003), (DILEEP; SEKHAR, 2006), (NAGIREDDY, 2008), (SAXENA,

2008), (SHARIF *et al.*, 2010) e (MANJULA; ANITHA, 2011).Todas tentam resolver o problema da identificação do algoritmo gerador, partindo apenas do texto cifrado. Entretanto nenhuma dessas pesquisas utiliza técnicas de RI aplicadas à Criptologia, foco principal deste trabalho.

Uma vez que remetente e destinatário utilizam sistemas criptográficos, para a troca de mensagens através de meios inseguros de comunicação, um invasor (terceiro, não autorizado) pode facilmente capturar as mensagens criptografadas. Após capturar a informação, o criptoanalista pode tentar identificar o algoritmo utilizado, para que se possa explorar uma vulnerabilidade conhecida ou aplicar técnicas específicas na tentativa de obtenção do conteúdo da mensagem.

Segundo (LIMA *et al.*, 2009), a identificação de padrões em criptogramas se caracteriza como uma ferramenta importante para as técnicas de criptoanálise. Além disso, o conhecimento desses padrões pode levar à quebra de algoritmos criptográficos, porém esta ferramenta também pode ser utilizada pela criptografia para avaliar novos algoritmos, tornando-os mais seguros.

Sendo assim, é importante destacar que a motivação deste trabalho é apresentar um método não supervisionado de reconhecimento de padrões para auxiliar na identificação do algoritmo criptográfico utilizado para cifrar uma mensagem. Contribuindo assim, para a evolução da linha de pesquisa em criptologia do IME, expandindo a classificação de criptogramas para o modo de operação CBC.

MODOS DE OPERAÇÃO

Um algoritmo simétrico de bloco é projetado para criptografar texto em claro de tamanho , esta limitação é dada pelo tamanho do bloco do algoritmo. Para que se possa criptografar texto de tamanho , onde , faz-se necessário dividir o texto em blocos de tamanho . Cada é submetido ao algoritmo. A forma de submissão cada bloco pode ser distinta, por isso o NIST convencionou alguns modos de operação através da FIPS SP 800-38A.

O *Eletronic Code Book* (ECB) é o modo mais simples de operação, já que os blocos são cifrados independentemente. Dessa forma, blocos de texto em claro idênticos, resultam em blocos cifrados igualmente idênticos, quando se utiliza a mesma chave (RIBEIRO; ROIHA, 2005)com baixo custo computacional, eles conseguem mascarar similaridades do texto legível e poupar o usuário das limitações do tamanho de bloco. Neste trabalho, será apresentado um overview dos principais modos de operação de confidencialidade, destacando suas vantagens e desvantagens.", . A Figura 1 representa o modo ECB.

No modo de operação *Cipher-Block Chaining*(CBC), um vetor de inicialização é operado, através de ou-exclusivo, com o primeiro bloco da mensagem. Posteriormente, todos os blocos são operados, em ou-exclusivo, com o bloco cifrado imediatamente anterior(RIBEIRO; ROIHA, 2005). Nesse modo de operação, textos em claro idênticos resultam em blocos cifrados distintos, ainda que se utilize a mesma chave. A Figura 2 representa o modo de operação CBC.

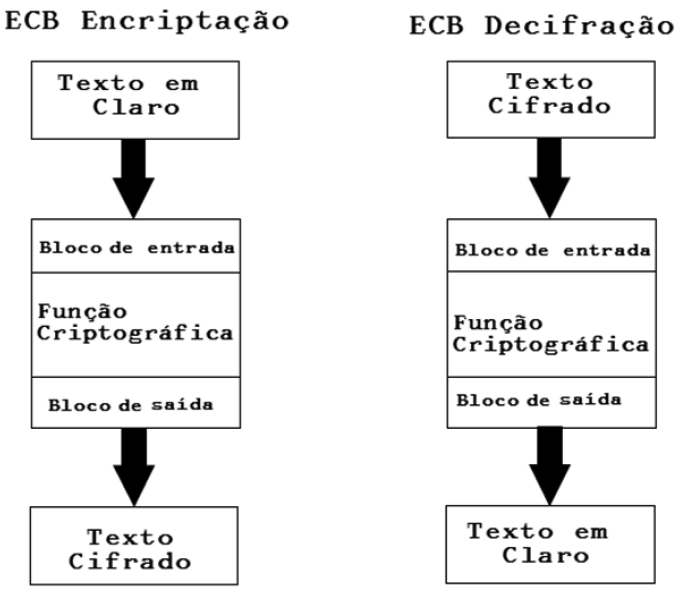


Figura 1 Modo de operação ECB, adaptado de (DWORKIN, 2001).

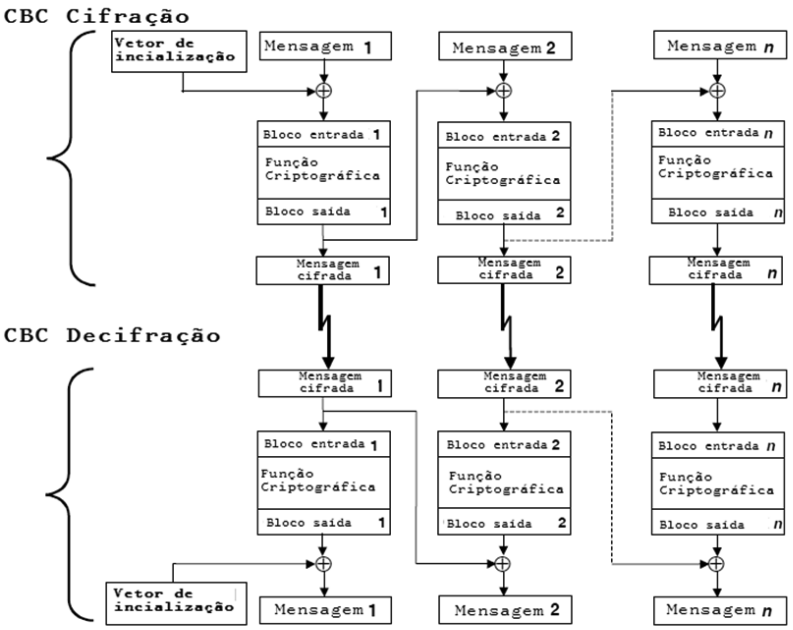


Figura 2 Modo de operação CBC, adaptado de (DWORKIN, 2001).

CRIPTOTERMO

Um texto qualquer é constituído de semântica, sintaxe e léxico. Um criptograma, por sua vez, não é oriundo de uma linguagem convencional e não possui características semânticas ou sintáticas. Em (SOUZA, 2007, p. 65), criptotermo é definido como uma palavra binária de tamanho *n*, onde *n* é o tamanho do bloco cifrado pelo algoritmo gerador do criptograma.

DISTÂNCIA DE HAMMING

A Distância de Hamming é uma métrica que calcula a proximidade léxica entre

dois vetores, retornando um valor inteiro que representa a quantidade de coordenadas em que os dois vetores diferem (MILIES, 2009). Fórmula

Neste trabalho, a distância de Hamming será utilizada para retornar a quantidade de bits diferentes entre dois criptotermos. Esta distância será usada como base de comparação entre os criptotermos de um conjunto de amostras, com os criptotermos dos criptogramas analisados.

COLORAÇÃO EM GRAFOS

A coloração de um grafo é a atribuição de uma cor para cada vértice do grafo, atendendo a restrição de que vértices adjacentes possuam cores distintas (SZWARCFITER, 1986). Essa técnica será utilizada para formar grupos amostrais de criptotermos, através do algoritmo abaixo.

ColoreGrafo(C,E,cor)

Entrada: (C, E, cor), onde: $C = \{c_1, c_2, \ldots, c_n\}$, $E = \{\overrightarrow{e_1}, \overrightarrow{e_2} \ldots \overrightarrow{e_n}\}$ e $cor \equiv$ *número atual da cor*

1: **se** $cor \equiv null$ **então**
2: $cor = 1$
3: **fim se**
4: **para** $c_i = 1$ até c_n **faça**
5: **se** $c_i.cor \equiv null$ **então**
6: $c_i.cor = cor$, $cor = cor + 1$
7: $\bar{A} = \{a_1, a_2, \ldots, a_n\}$, $\bar{A} \subseteq C$, tal que: $\forall a_j \neg\exists \overrightarrow{e_k}\{c_i, a_j\}$ {*subconjunto de vértices não adjacentes ao vértice* v_i}
8: *ColoreGrafo*$(\bar{A}, E, cor)$ {*chamada recursiva ao algoritmo, enviando o subconjunto não adjacente*}
9: **fim se**
10: $i = i + 1$
11: **fim para**

MÉTODO PROPOSTO

O propósito de um método de RP (Reconhecimento de Padrões) é classificar um dado objeto desconhecido com base em uma coleção de objetos previamente catalogada. A característica da coleção usada para tomada de decisão distingue o método em *supervisionado* ou *não supervisionado*. O método *supervisionado* é aquele no qual se conhece a catalogação de cada objeto da coleção. E o *não supervisionado,* por sua vez, ocorre quando não se tem informações sobre a coleção dos objetos.

Um sistema de RP *não supervisionado*, embora obtenha resultados inferiores ao de um sistema *supervisionado*, é a única solução para casos onde não há informação sobre as classes dos dados (MARQUES, 2005). A característica de um método *não supervisionado* é interessante em casos onde há apenas textos cifrados e não existem informações sobre o algoritmo ou chave utilizados na cifração.

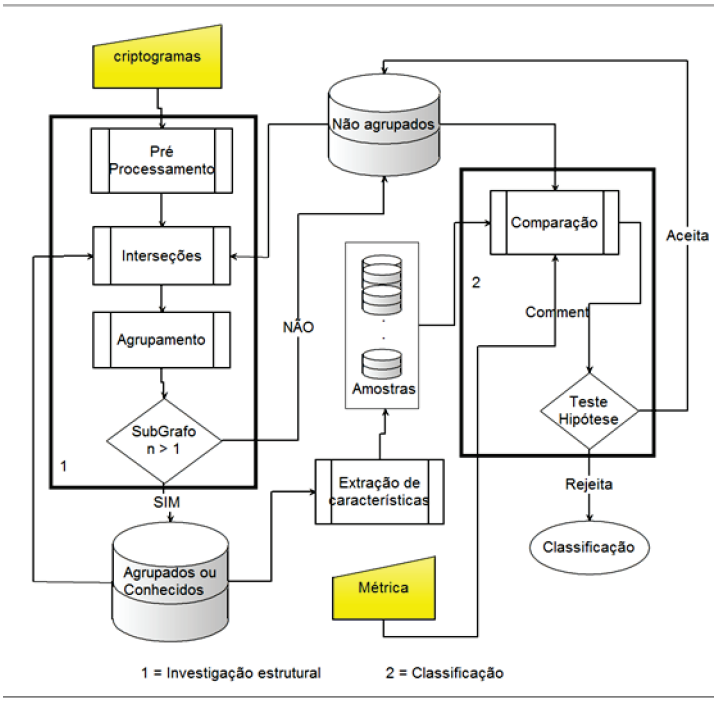


Figura 3 Visão geral do método proposto.

Deste modo, o método proposto é formado de três fases: *investigação estrutural*, *extração de características* e *classificação*, descritas abaixo. Pode-se observar na Figura 3 uma visão geral do método proposto.

Investigação estrutural

Investigação Estrutural é a fase inicial do método de classificação, onde o conjunto de criptogramas apresentado é processado para formar grupos de criptogramas unidos por interseção de blocos binários. Essa fase é divida em três etapas, sendo elas:

1. ***Pré-processamento*** é uma etapa de indexação, semelhante ao de técnicas de RI, utilizado para tratar a coleção de criptogramas submetidos ao classificador, gerando uma estrutura que armazenará a coleção de criptogramas. Esta etapa está baseada nos trabalhos de (CARVALHO, 2006) e (SOUZA, 2007).
2. ***Interseções*** é a etapa do método que busca por interseções, ou seja, criptotermos idênticos, entre os integrantes do conjunto de criptogramas. Esta etapa está fundamenta-se no trabalho de (OLIVEIRA, 2011).
3. ***Agrupamento*** é a etapa que produz um grafo das relações entre os criptogramas analisados. Esta etapa se apoiará no trabalho de (TORRES *et al.*, 2010).

Pré-processamento

Segundo (SOUZA, *et al.*, 2009), um criptograma pode ser considerado como um texto legível escrito em idioma desconhecido, utilizando alfabeto binário. Assim, cada bloco de 64bits de um criptograma será considerado como uma palavra binária desse idioma desconhecido, doravante denominada criptotermo. O Pré-processamento pode ser definido como uma quádrupla (C, TO, p) , onde:

- C é o conjunto de criptogramas $\{c_1, c_2 \ldots c_n\}$, submetidos ao classificador.
- T é conjunto de criptotermos $\{c_1, c_2 \ldots c_n\}$.
- O é conjunto de ocorrências $\{\overrightarrow{o_1}, \overrightarrow{o_2} \cdots \overrightarrow{o_n}\}$, onde é $\overrightarrow{o_i} = \{o_i.cript, o_i.term\}$ um vetor que relaciona um criptograma (c_i) a um criptotermo (t_k) , atendendo a condição de : $t_k \in c_i$.
- $.p$ é um procedimento de indexação que recebe um criptograma como entrada e armazena seu conteúdo na estrutura de dados definida, conforme: (((p(entrada) → C) → O).

Interseções

Define-se E como um conjunto de interseções $\{\overrightarrow{e_1}, \overrightarrow{e_2}, \cdots \overrightarrow{e_n}\}$ dos integrantes de , onde $\overrightarrow{e_i}$ é um vetor que armazena a identificação dos criptogramas (c_j, c_k), se e somente se atender a seguinte condição:

- $\forall \overrightarrow{e_i} \in E$, existe ao menos uma dupla em O $(\overrightarrow{o_j}, \overrightarrow{o_k})$ que satisfaça:
  - $o_j.cript \neq o_k.cript$
  - $o_j.term = o_k.term$

Agrupamento

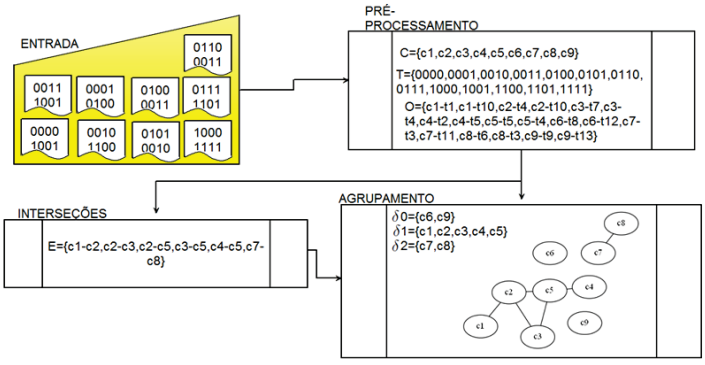
A etapa de agrupamento produz um grafo formado por um conjunto de vértices (C) e um conjunto de arestas (E) onde:

1- $C = \{c_1, c_2, \cdots c_n\}$, ou seja, conjunto dos criptogramas do *Pré-processamento*.

2- $E = \{\overrightarrow{e_1}, \overrightarrow{e_2}, \cdots \overrightarrow{e_n}\}$, isto é, conjunto das interseções identificadas.

A estrutura desta etapa, projetada para teoria dos grafos, é primordial para a etapa seguinte, onde serão formados grupos amostrais, através da coloração em grafos. O processo de agrupamento de criptogramas, por conseguinte, consiste em dividir um conjunto C em n grupos, mantendo uma relação de proximidade e considerando as seguintes características:

- δ_0 é um subconjunto de C que atende a seguinte condição: $\forall c_i \in \delta_0 | grau(c_i) = 0$, ou seja, todo criptograma que não possui interseção com outro criptograma da coleção pertencerá ao grupo δ_0, doravante denominado *não-agrupados*.
- $\delta_i | \{1 \leq i \leq n\}$, n é quantidade de subgrafos conexos com vértices de $grau \geq 1$ é um conjunto de vértices de um subgrafo conexo de C que atende a seguinte condição: $\forall c_k \in \delta_i \mid grau(c_k) \geq 1$.



A Figura 4 representa a fase de Investigação Estrutural.

Extração de características

A fase de *Extração de Características*, representada na Figura 5, processa os grupos identificados pela fase de investigação estrutural, gerando diversas bases amostrais, que serão utilizadas para comparação na próxima fase. Acrescenta-se que o processo se dará através da técnica de coloração em grafos. As bases formadas serão utilizadas para comparação de proximidade na próxima fase.

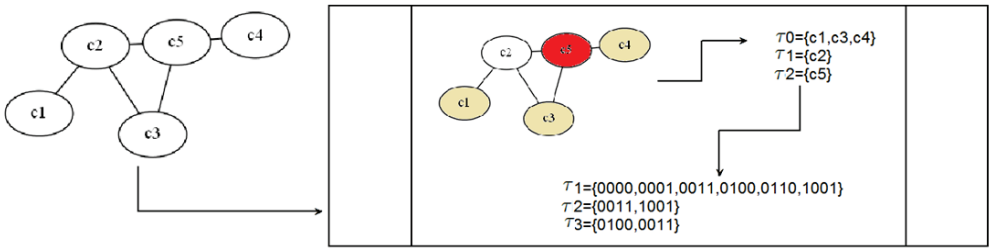


Figura 5. Extração de características.

Dado um grupo (subgrafo conexo) δ_i, esta etapa retorna as amostras $(\tau_0, \tau_1, \cdots \tau_n)$ do grupo, onde τ_i é um conjunto de criptotermos ocorrentes nos criptogramas que pertencem a um mesmo grupo, mas não possuem criptotermos idênticos entre si, do grupoδ_i. Um criptotermo t_k pertencerá ao conjunto τ_i se, e somente se, atender as seguintes condições:

- $\tau_i \subset T$.
- $(t_l \in c_j) \wedge (c_j \in \tau_k) \wedge (\tau_k \in \delta_i)$.
- τ_k é um subgrupo de δ_i, que agrega os criptogramas de cor k.

Classificação

Classificação, formada pelas etapas de *Comparação* e *Teste de hipótese*, é a fase onde os criptogramas que não foram agrupados, são mensurados por uma determinada métrica e submetidos a um teste de hipótese para definir a classificação do criptograma.

Comparação

Na etapa de comparação, o criptograma desconhecido é mensurado através de uma métrica qualquer e comparado com as bases amostrais geradas pela fase anterior. Para isso, desenvolveu-se uma métrica para representar um índice de proximidade entre o criptograma e uma base conhecida.

Seja D um criptograma desconhecido e δ_i um conjunto de amostras. O Índice de confiabilidade por paridade de criptotermos calcula o coeficiente de variação da menor distância de hamming entre D, δ_i . O resultado [0-1] representa o nível de proximidade de um criptograma D para um grupo δ_i, baseado na

dispersão da *distância de Hamming* de seus criptotermos, representado por:

$$Índice_{confiabilidade} = 1 - \frac{desviopadrão(D, \delta_i)}{média(D, \delta_i)}$$

onde:

- $D = \{d_1, d_2, \cdots d_n\}$é um criptograma desconhecido o qual se deseja classificar.
- $\delta_i = \{\tau_1, \tau_2, \cdots \tau_n\}$é um conjunto de amostras de criptotermos conhecidos.
- $desviopadrão(D, \delta_i) = \sqrt{\frac{\sum_{i=1}^{n}(peso(D,\tau_i) - média(D,\delta_i))^2}{n-1}}$
- né a quantidade de amostras de δ_i, com a restrição de $n \geq 2$.
- $\tau_i = \{t_1, t_2 \cdots t_n\}$é uma amostra de criptotermos gerada pela fase *Extração de características*.
- $média(D, \delta_i) = \frac{\sum_{i=1}^{n} peso(D,\tau_i)}{n}$
- $peso(D, \tau_i)$é uma função que retorna o peso da aresta de um dado criptograma com um $\tau_i$qualquer, levando em consideração a *distância de Hamming* de cada criptotermo de D em relação a todos os criptotermos de T, exceto δ_0 . Esta função está descrita no algoritmo abaixo.

$peso(D, \tau_i)$

Entrada: D, τ_i (criptograma desconhecido e grupo criptogramas conhecidos)
Saída: n (peso da aresta entre D, τ_i)

```
1:  peso = 0
2:  para d_i = 1 até d_n faça
3:      distT = 64,  distD = 64,  aresta = 0
4:      para t_i = 1 até t_n, onde t_i ∈ τ_i faça
5:          h = Dist_H(d_i, t_i)
6:          se h < distD então
7:              aresta = 1,  distD = h
8:          senão
9:              se h = distD então
10:                 aresta = aresta + 1
11:             fim se
12:         fim se
13:     fim para
14:     para t_i = 1 até t_n, onde t_i ∈ T \ (δ_0 ∪ τ_i) faça
15:         h = DistH(d_i, t_i)
16:         se h < distT então
17:             distT = h
18:         fim se
19:     fim para
20:     se distD ≤ distT então
21:         peso = peso + aresta
22:     fim se
23: fim para
24: retorna peso
```

- $Dist_H(d_i, t_i)$é a *distância de Hamming* entre dois criptotermos.
- d_ié um criptotermo ocorrente no criptograma desconhecido.

Teste de hipótese

A decisão sobre a classificação de um criptograma analisado é realizada através de um grafo de decisão. Ao rejeitar a *hipótese de nulidade* (H_O), adiciona-se 1 ao peso da *aresta* (D, δ_i) .O teste de hipótese utilizado na decisão é o seguinte:

(H_O) O criptograma analisado foi gerado por um algoritmo desconhecido.

(H_1) O criptograma analisado foi gerado pelo algoritmo gerador de .

A regra de decisão é rejeitar se o valor de referência é menor que o nível de significância escolhido, através de:

$$valor_{referência} = 1 - Índice_{confiabilidade}(D, \delta_i)$$

Caso a quantidade de amostras na fase de *Extração de Características* seja > 2, devem ser submetidas à fase de *Comparação* todos os pares de combinações possíveis de amostras, acrescendo os resultados no grafo de decisão.

O criptograma analisado será classificado como oriundo do algoritmo gerador do grupo δ_i que possuir aresta de maior peso. Caso contrário, o criptograma não deverá ser classificado, pois não existem evidências suficientes através da métrica utilizada para classificar o criptograma.

EXPERIMENTOS

Os experimentos realizados visam analisar a influência de parâmetros variáveis do modelo no resultado da classificação de criptogramas. Analisou-se a influência da quantidade de $\tau \in \delta i$; a influência da quantidade de $d \in D_i$; e a eficiência de classificação com base homogênea e heterogênea.

Para isso, foram utilizados nove algoritmos com características de tamanho de chave e bloco distintos e as chaves criptográficas descritas na Tabela 1. Os criptogramas são oriundos de texto em claro, extraídos da bíblia em português, cifrados em modo CBC.

Tabela 1. Algoritmos Utilizados

Algoritmo	Tamanho da chave	Tamanho do bloco	Tipo	Chave utilizada
3DES	168	64	bloco	vKnb8KlqmjX9Cd81CUN6i
AES	256	128	bloco	vKnb8KlqmjX9Cd81CUN6i62aJHeyhetp
BF	256	64	bloco	vKnb8KlqmjX9Cd81CUN6i62aJHeyhetp
Camellia	256	64	bloco	vKnb8KlqmjX9Cd81CUN6i62aJHeyhetp
Cast5	128	64	bloco	vKnb8KlqmjX9Cd81
DES	56	64	bloco	vKnb8Kl
RC2	128	64	bloco	vKnb8KlqmjX9Cd81
RC4	128	-	fluxo	vKnb8KlqmjX9Cd81
Seed	128	64	bloco	vKnb8KlqmjX9Cd81

A regra de decisão adotada para classificação em todos os experimentos foi rejeitar a hipótese de nulidade, caso o valor de referência fosse menor que um nível de significância escolhido arbitrariamente. O teste baseia-se nas seguintes hipóteses:

- (H_O) O criptograma analisado foi gerado pelo algoritmo DES.
- (H_1) O criptograma analisado não foi gerado pelo algoritmo DES.

Os experimentos realizados possuem a característica de que, ao se aumentar o nível de significância, tende-se a não rejeitar H_O, fazendo com que os criptogramas sejam considerados como oriundos do DES. Ao se diminuir o nível de significância, tende-se a rejeitar H_O, considerando-os como não oriundos do DES. O ideal é encontrar um nível de significância que possa identificar corretamente mais que 50% dos criptogramas gerados pelo DES e dos não gerados pelo DES.

Experimento 1: influência do tamanho de .

Este experimento, dividido em duas fases, tem como objetivo avaliar a influência da quantidade de amostras (τ) do conjunto , utilizadas para comparação na base de criptotermos conhecidos. Para isso, foram construídas duas bases de comparação, conforme Figura 6, com um milhão de criptotermos, diferindo apenas na quantidade de grupos amostrais.

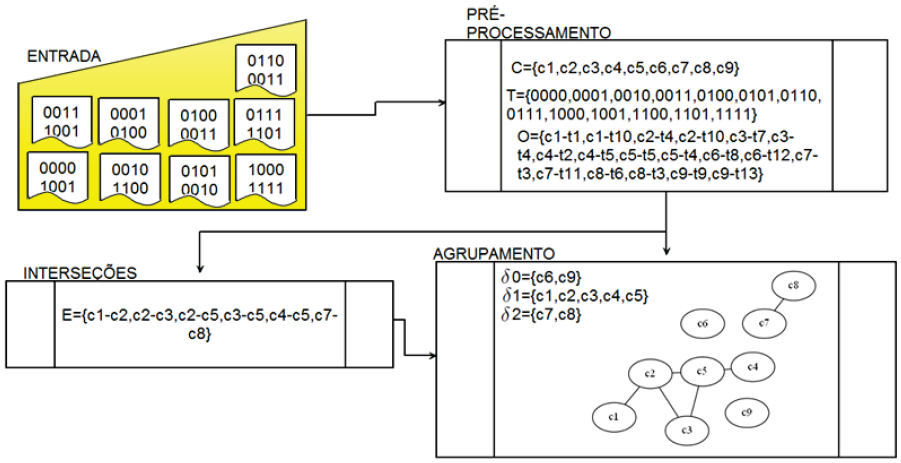


Figura 6. Construção de bases a partir da Bíblia em português.

Os criptotermos são oriundos de 782KB (100 mil palavras binárias de 64bits), extraídas da bíblia em português, cifrados em modo CBC pelo algoritmo DES, utilizando dez chaves distintas, conforme Tabela 2.

Na primeira fase, a base foi dividida em duas amostras ($\tau = 2$) de 500 mil criptotermos, cada amostra cifrada com cinco chaves. Na segunda fase, a base foi dividida em cinco amostras ($\tau = 5$) de 200 mil criptotermos, cada amostra cifrada com duas chaves distintas.

Tabela 2. Chaves de cifração da base de comparação.

k Chave (56bits)	k Chave (56bits)
k1 cGg5Y94	8a2dl7V
k2 Pq8N54C	PsQob15
k3 S19U1D9	v2bdZ22
k4 hk9pDcY	km1RnuI
k5 iqd6iea	W312xq8

Foram submetidos à classificação 900 criptogramas de 8KB, oriundos de texto em claro extraídos da bíblia (cifrado pelos nove algoritmos da Tabela 1), 100 criptogramas por algoritmo. Os resultados obtidos nas duas fases estão descritos na Tabela 3.

Tabela 3. Resultado do experimento 1

	1ª Fase				2ª Fase			
Nível de Significância	0,021	0,023	0,024	0,025	0,047	0,0475	0,048	0,0485
Algoritmo	**ACERTO***							
DES	50%	52%	54%	55%	50%	52%	54%	55%
Não DES	55%	50%	48%	46%	50%	49%	48%	46%
Desvio Padrão (Não DES)	0,035721				0,012657			

Para comparar o desempenho de cada fase, optou-se por analisar a variação do resultado de identificação correta dos criptogramas não oriundos do DES em quatro níveis de significância distintos. A comparação ocorreu da seguinte maneira: fixou-se o valor de acerto no reconhecimento de criptogramas gerados pelo DES e verificou-se o desvio-padrão dos resultados de acerto na identificação dos criptogramas não oriundos do DES.

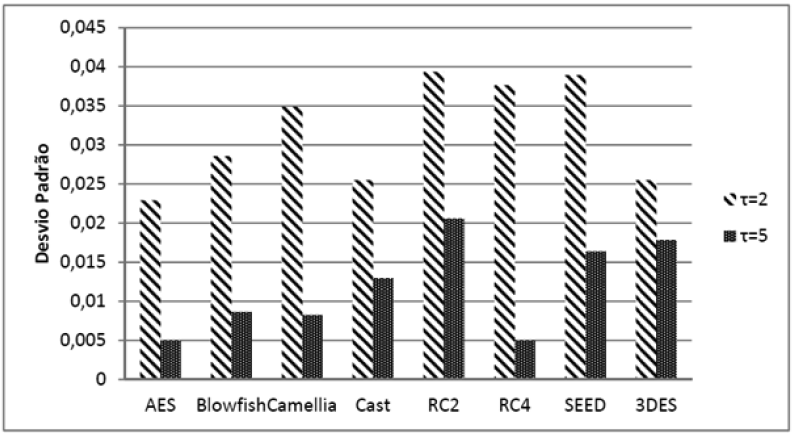


Figura 7. Desvio padrão do resultado de classificação (por algoritmo).

A Figura 7 representa o gráfico dos resultados do desvio-padrão, (por algoritmo), entre os resultados da fase1 ($\tau = 2$) e da fase 2 ($\tau = 5$).

Considera-se como ideal a amostra que gerar resultados menos variáveis, uma vez que os valores de significância utilizados são muito próximos e se espera que os resultados de acerto não apresentem variação muito grande. Os resultados obtidos com $\tau = 5$ são claramente menos variáveis que os resultados obtidos com uma menor quantidade de amostras.

Assim, pode-se concluir que a quantidade de amostras conhecidas interfere no resultado. Neste caso, deve-se encontrar um $\tau = n$ apropriado para cada tipo de algoritmo, pois acredita-se que, quanto maior o número de amostras, mais confiáveis serão os resultados.

Experimento 2: influência do tamanho de D_i.

O objetivo deste experimento é avaliar a influência do tamanho do criptograma analisado (quantidade de criptotermos) no resultado da classificação. Para isso, foram cifrados 800KB de texto em claro, extraídos da bíblia em português, em modo CBC, utilizando os algoritmos e senhas descritos na Tabela 1.

O experimento foi realizado em duas fases, conforme Figura 8. Na primeira fase, os textos 800KB, cifrados por cada algoritmo, foram divididos em criptogra-

mas de 1024 criptotermos (8KB), totalizando 900 criptogramas. Na segunda fase, os 800KB de texto cifrado, gerado por cada algoritmo, foi dividido em criptogramas de 32768 criptotermos (32KB), totalizando 225 criptogramas.

A base de comparação usada continha 1 milhão de criptotermos cifrados pelo algoritmo DES, utilizando 10 chaves distintas, descritas na Tabela 2. Destaca-se que base foi dividida em cinco amostras ($\tau = 5$) com 200 mil criptotermos por amostra, cifrados por duas chaves (100 mil por chave).

Na primeira fase, utilizando criptogramas de 8KB, foi possível identificar apenas os criptogramas cifrados pelo algoritmo SEED, como não oriundos do DES, com . A Tabela 4, destaca o resultado dessa fase. Não foi possível distinguir os criptogramas cifrados pelos outros sete algoritmos (acerto inferior a 50%). Os níveis de significância testados foram: .

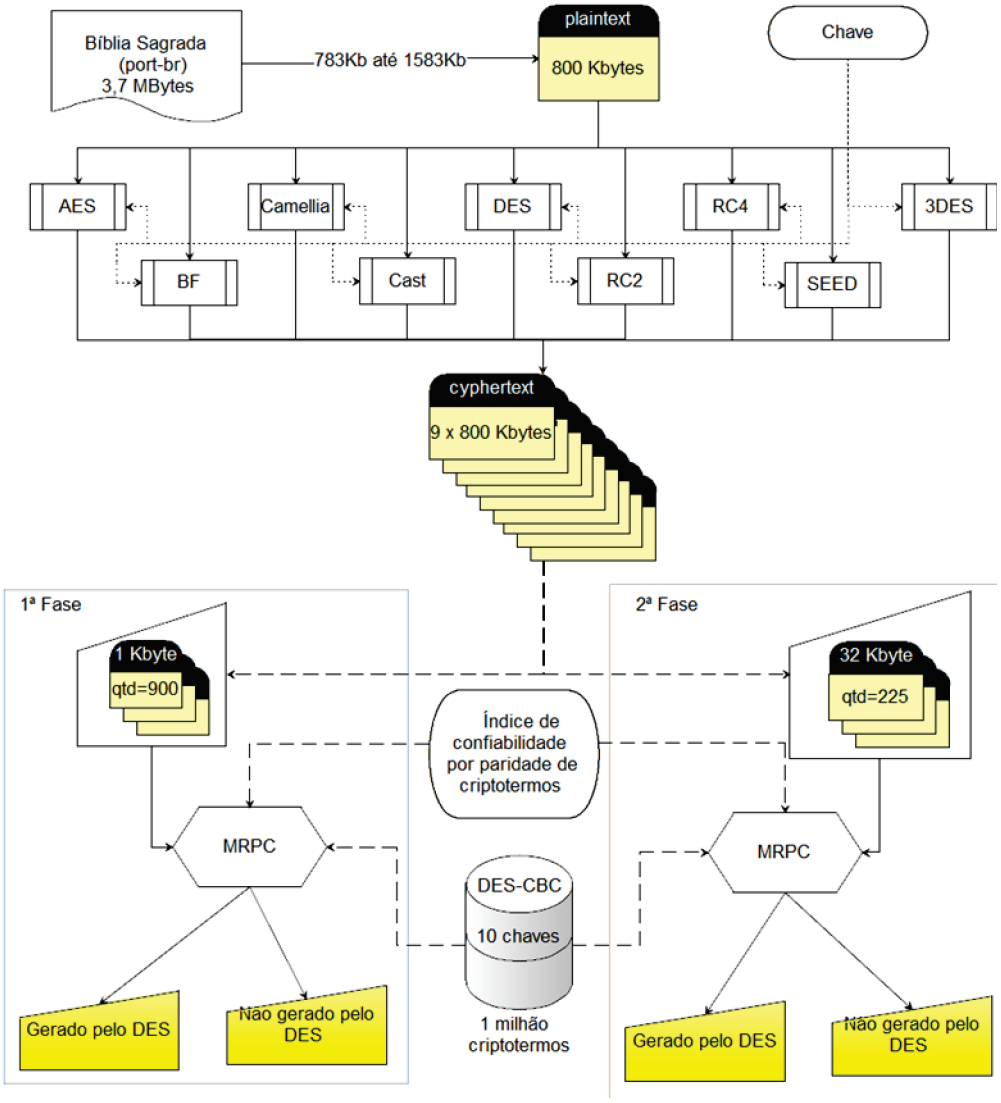


Figura 8. Metodologia do experimento 2.

Tabela 4. Classificação de criptogramas de 8KB.

	Submetidos à classificação	Classificados corretamente	Nível de acerto na classificação
DES	100	52	44%
BF	100	44	
Camellia	100	42	
AES	100	42	
SEED	100	51	
RC4	100	37	
CAST	100	47	
RC2	100	47	
3DES	100	35	
Total	900	397	

Na segunda fase, utilizando criptogramas de 32KB, foi possível identificar os criptogramas cifrados pelos algoritmos Blowfish, Camellia, Cast, RC2, RC4, SEED e 3DES como não oriundos do DES, com $\alpha = 0{,}021$. Na . A Tabela 5, destaca o resultado dessa fase. Não foi possível, no entanto, distinguir os criptogramas cifrados pelo algoritmo AES, (acerto inferior a 50%). Os níveis de significância testados foram: $\alpha = \{ 0{,}02; ...; 0{,}025\}$.

Tabela 5. Classificação de criptogramas de 32KB.

	Submetidos à classificação	Classificados corretamente	Nível de acerto na classificação
DES	25	15	**54%**
BF	25	13	
Camellia	25	13	
AES	25	12	
SEED	25	13	
RC4	25	13	
CAST	25	16	
RC2	25	15	
3DES	25	13	
Totais	225	123	

Os resultados de classificação, obtidos com criptogramas de 32KB, identificaram corretamente a grande maioria dos algoritmos testados, como oriundos ou não do DES. Assim, pode-se concluir que o tamanho do criptograma analisado interfere no resultado. Neste caso, deve-se encontrar um tamanho apropriado para cada tipo de algoritmo. Acredita-se que quanto maior o número de amostras mais confiáveis serão os resultados.

Experimento 3: eficiência de classificação com base homogênea.

O objetivo deste experimento é verificar se o método, através de uma base especializada (homogênea), pode identificar se dado criptograma, de tamanho igual a 128KB, foi ou não gerado pelo mesmo algoritmo da base conhecida.

A base é considerada homogênea por possuir texto de apenas uma obra (Bí-

blia Sagrada). A base de comparação usada continha 1 milhão de criptotermos cifrados pelo algoritmo DES, utilizando 10 chaves distintas, descritas na Tabela 1. A base foi dividida em cinco amostras ($\tau = 5$) com 200 mil criptotermos por amostra, cifrados por duas chaves (100 mil por chave).

Foram submetidos 132 criptogramas de seis algoritmos distintos (DES, BF, AES, RC4, RC2 e 3DES) e extraídos 2.8MB de texto em claro da bíblia em português, dividido em 22 arquivos de 128KB, cifrados em modo CBC.

Tabela 6. Resultado de classificação do experimento 3.

	Submetidos a Classificação	Classificados corretamente	Nível de acerto na classificação
Oriundos DES	22	12	**60%**
Não Oriundos do DES	110	68	
Totais	132	80	

Em outra análise, do mesmo experimento, agora considerando os resultados de classificação de cada algoritmo, temos:

- 55% dos criptogramas gerados pelo DES foram classificados corretamente como oriundos do DES;
- 59% dos criptogramas gerados pelo Blowfish, AES e 3DES, foram classificados corretamente como não oriundos do DES;
- 64% dos criptogramas gerados pelo RC4 foram classificados corretamente como não oriundos do DES; e
- 68% dos criptogramas gerados pelo RC2 foram classificados corretamente como não oriundos do DES.

Experimento 4: eficiência de classificação com base heterogênea.

O objetivo deste experimento é verificar se o método, através de uma base especializada (heterogênea), pode identificar se dado criptograma, de tamanho igual a 32KB, foi ou não gerado pelo mesmo algoritmo da base conhecida.

A base é considerada heterogênea por possuir texto de diversas áreas, todos cifrados em modo CBC utilizando o algoritmo DES. Composta por:

- E-mail: 50 mil palavras binárias de 64bits, oriundas de listas de e-mail da Sociedade Brasileira de Computação.
- Enciclopédia: 50mil palavras binárias de 64bits, oriundas de páginas aleatórias da Wikipédia em português.
- Legislação: 50 mil palavras binárias de 64bits, de legislação aleatória do site da Casa Civil.
- Literários: 50 mil palavras binárias de 64bits, de obras literárias em português disponíveis no Projeto Gutenberg.
- Técnicos: 50 mil palavras binárias de 64bits, oriundos do Guia Foca Linux (Avançado).

Todos os textos em claro foram cifrados por dez chaves distintas (conforme Tabela 2), totalizando 2.5 milhões de criptotermos, conforme:

$$Base_{heterogênea} = \left(\overbrace{5 * 10^4}^{\substack{palavras \\ binárias}} * \underbrace{10}_{\substack{chaves \\ criptográficas}} \right)$$

A base de comparação foi dividida em cinco amostras (τ = 5) com 500 mil criptotermos por amostra, cifrados por duas chaves (250 mil por chave). Além disso, foram submetidos 175 criptogramas de sete algoritmos distintos (DES, BF, Camellia, Cast, RC2, SEED e 3DES), conforme Tabela 1, bem como extraídos 800KB de texto em claro da bíblia em português, dividido em 25 arquivos de 32KB e cifrado em modo CBC.

Tabela 7. Resultado de classificação do experimento 4.

	Submetidos a Classificação	Classificados corretamente	Nível de acerto na Classificação
Oriundos DES	25	15	**63%**
Não Oriundos do DES	150	96	
Totais	175	111	

Em outra análise, do mesmo experimento, agora considerando os resultados de classificação de cada algoritmo, temos:

- 60% dos criptogramas gerados pelo DES foram classificados corretamente como oriundos do DES;
- 44% dos criptogramas gerados pelo Blowfish foram classificados corretamente como não oriundos do DES;
- 52% dos criptogramas gerados pelo CAST e 3DES, foram classificados corretamente como não oriundos do DES;
- 68% dos criptogramas gerados pelo RC2 foram classificados corretamente como não oriundos do DES;
- 80% dos criptogramas gerados pelo Camellia foram classificados corretamente como não oriundos do DES; e
- 88% dos criptogramas gerados pelo SEED foram classificados corretamente como não oriundos do DES.

ANÁLISE DOS RESULTADOS DOS EXPERIMENTOS

Embora o método proposto seja projetado para classificar criptogramas em modo ECB, optou-se pela utilização de criptogramas cifrados em modo CBC, pois esse modo de operação parecia imune à classificação, através dos diversos trabalhos até então apresentados.

Através dos experimentos realizados, pode-se concluir que a quantidade de amostras produzidas na fase de extração de características afeta o desempenho. Observou-se, inclusive, que é um valor ideal.

O tamanho do criptograma analisado é outro fator de influência nos resultados de classificação. O tamanho de melhor custo-benefício (custo computacional X resultado de acerto) encontrado foi , ou seja, 32KB.

Por fim, outro fator importante na classificação é a formação da base de comparação. A base heterogênea apresentou, ademais, ótimos resultados de classificação, mesmo com criptogramas cifrados em modo CBC.

CONCLUSÃO

Classificar criptogramas cifrados em modo CBC, ainda é um problema em aberto. Este trabalho contribuiu para as técnicas de Recuperação da Informação (RI) aplicadas ao reconhecimento de padrões em criptogramas, apresentando um método não supervisionado baseado na distância de Hamming. Esse método, além de ser utilizado como ataque de distinção sobre criptogramas, ele pode servir como parâmetro de avaliação de algoritmos criptográficos.

A comparação dos resultados obtidos através desse método com outras técnicas não foi realizada, até porque, boa parte das pesquisas disponíveis sobre classificação e agrupamento de criptogramas não informam o modo de operação utilizado e os únicos trabalhos, baseados em RI, que tentaram classificar criptogramas em modo CBC ((CARVALHO, 2006) e (SOUZA, 2007)) , não obtiveram sucesso.

Acrescente-se ainda que os experimentos realizados demonstraram a influência da quantidade de amostras, do tamanho do criptograma analisado e da formação da base no desempenho da classificação. O quarto, inclusive, experimento apresentou resultados satisfatórios, com criptogramas de 32KB, classificando os criptogramas gerados pelos algoritmos Camellia e SEED como não oriundos do DES, com uma taxa de acerto acima de 80%.

Deste modo, como trabalhos futuros, um ajuste no calculo do valor de referência do teste de hipótese pode ser realizado, levando-se em consideração o tamanho e a quantidade de amostras e o tamanho dos criptogramas. De fato, percebeu-se que os níveis de significância variam quando esses parâmetros são alterados. Novas métricas poderão ser propostas, então, para serem utilizadas nesse método, pois o Índice de confiabilidade por paridade de criptotermos tem complexidade , onde *n* é a quantidade de criptotermos do criptograma analisado e m é a quantidade de criptotermos da base de comparação.

REFERÊNCIAS BIBLIOGRAFIAS

- *CARVALHO. O uso de técnicas de recuperação de informações em criptoanálise. Rio de Janeiro: Instituto Militar de Engenharia,* ***2006****.*
- *CHANDRA, G. Classification of Modern Ciphers. Kanpur, Indian: Indian Institute of Technology Kanpur,* ***2002****.*
- *DAEMEN, J.; RIJMEN, V. AES Proposal: Rijndael. [S.l: s.n.],* ***1999****.*
- *DILEEP, A. D.; SEKHAR, C. C. Identification of Block Ciphers using Support Vector Machines. IJCNN. Anais... [S.l: s.n.]. ,* ***2006***
- *DWORKIN, M. Recommendation for block cipher modes of operation. 2001 ed. ed. [S.l.]: U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology : For sale by the Supt. of Docs., U.S. G.P.O., Gaithersburg, MD :,* ***2001****.*
- *LAMBERT, J. A. Cifrador simétrico de blocos: projeto e avaliação. Rio de Janeiro: Instituto Militar de Engenharia,* ***2004****.*
- *LIMA, A. P.; XEXÉO, J A M; GOMES, P. R. Novas formas de ataque de distinção a cifradores: caminho para o futuro. Revista Militar de Ciência e Tecnologia, v. XXVI, p. 12-17,* **2009***.*
- *MAHESHWARI, P. Classification of Ciphers. Kanpur, Indian: Indian Institute of Technology Kanpur,*

2001.

- *MANJULA, R.; ANITHA, R. Identification of Encryption Algorithm Using Decision Tree. In: MEGHANATHAN, N.; KAUSHIK, B. K.; NAGAMALAI, D. (Eds.). Advanced Computing. Communications in Computer and Information Science. [S.l.]: Springer Berlin Heidelberg, 2011. v. 133p. 237-246.*
- *MARQUES, J. S. Reconhecimento de Padrões: métodos estatísticos e neuronais. Portugal: IST Press, **2005**.*
- *MILIES, F. C. P. Introdução a Teoria dos Códigos Corretores de Erros. . Campo Grande, MS: [s.n.]. Disponível em: <http://www.coloquiodematematica.ufms.br/conteudo/material/mc09.pdf>. Acesso em: 21 jun. 2011. , **2009***
- *MURPHY, S. The power of NIST's Statistical Testing of AES Candidates. , Information Security Group. U.K.: Royal Holloway, University of London. Disponível em: <http://csrc.nist.gov/archive/aes/round2/conf3/papers/09-smurphy.pdf>. Acesso em: 20 jul. 2011. , **2000***
- *NAGIREDDY, S. A pattern recognition approach to block cipher identification. Chennai, Indian: Indian Institute of Technology Madras, **2008**.*
- *OLIVEIRA, G. A. A aplicação de Algoritmos Genéticos no Reconhecimento de Padrões Criptográficos. Rio de Janeiro: Instituto Militar de Engenharia, **2011**.*
- *RAO, M. B. Classification of RSA and IDEA Ciphers. Kanpur, Indian: Indian Institute of Technology Kanpur, **2003**.*
- *RIBEIRO, C.; ROIHA, L. Estudo Comparativo dos Modos de Operação de Confidencialidade: um Overview para Iniciantes. Revista Ciência e Tecnologia, v. 8, n. 13, **2005**.*
- *RUKHIN, A.; SOTO, J.; NECHVATAL, J. et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications. . [S.l.]: NIST - Special Publication 800-22. Disponível em: <http://goo.gl/rnv5v>. Acesso em: 12 ago. 2011. , abr **2010***
- *SAXENA, G. Classification of Ciphers using Machine Learning. Kanpur, Indian: Indian Institute of Technology Kanpur, **2008**.*
- *SCHNEIER, B. Applied Cryptography. 2. ed. New York, NY, USA: John Wiley & Sons, **1996**.*
- *SHARIF, S. O.; KUNCHEVA, L. I.; MANSOOR, S. P. Classifying encryption algorithms using pattern recognition techniques. . Beijing: IEEE. Disponível em: <http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5689769>. , 17 dez **2010***

*SOTO, J.; BASSHAM, L. Randomness Testing of the Advanced Encryption Standard Finalist Candidates. NIST IR - 6483. Anais... [S.l.]: National Institute of Standards and Technology. Disponível em: <http://www.nist.gov/customcf/get_pdf.cfm?pub_id=151216>. Acesso em: 21 jul. 2011. , **2000***

- *SOUZA, WILLIAM AUGUSTO RODRIGUES. Identificação de padrões em criptogramas usando técnicas de classificação de textos. Rio de Janeiro: Instituto Militar de Engenharia, **2007**.*
- *SOUZA, WILLIAM AUGUSTO RODRIGUES; CARVALHO, L. A. V.; XEXÉO, JOSÉ A. M. Agrupar Textos Cifrados é Equivalente a Agrupar Textos Legíveis. [STIL] - Brazilian Symposium in Information and Human Language Technology. Anais... [S.l: s.n.]. Disponível em: <http://goo.gl/h6Dr5>. Acesso em: 19 jul. 2011. , **2009***
- *SZWARCFITER, J. L. Grafos e algoritmos computacionais. Rio de Janeiro: Campus, 1986.*
- *TORRES, R. H.; OLIVEIRA, GLÁUCIO ALVES DE; XEXÉO, JOSÉ A. M.; SOUZA, WILLIAM A. R.; LIDEN, R. Detecção de padrões em criptogramas como suporte às atividades de criptoanálise. CADERNOS DO IME - Série Informática, v. 29, p. 37-44, jun **2010**.*