2.4 Classical Ciphers Cryptanalysis

Most of classical cipher cryptanalysis methods make use of the statistical properties of the natural languages. Among these properties is the letter frequency distribution, which gives the percentage frequency of the characters in the given text. Another property of natural language is the frequency of pairs and triples of letters in the target language. After encryption, some information will remain in the cipher text, especially for simple cipher systems. Cryptanalysis will rely heavily on such information to analyze the cipher text.

The knowledge of the above mentioned properties is quite sufficient for simple substitution ciphers and monoalphabetic ciphers, because the cipher text alphabet is a rotation of the plaintext alphabet and not an arbitrary permutation. Therefore the statistical information unchanged during the encryption process. Homophonic substitution ciphers do not obscure all of the statistical properties of the plaintext, hence it is slightly harder than simple substitution ciphers to break. In polyalphabetic ciphers, if the key length (period) is equal to one (**d=1**), then polyalphabetic ciphers become monoalphabetic (simple substitution), and hence, it is as easy as its equivalent to break. However, as period increased, it becomes harder and harder.

To solve a periodic substitution cipher, the cryptanalyst must determine the period of the cipher. Two earlier methods of classical ciphers cryptanalysis have been used, **Index of Coincidence (IC)**, and **Kasiski** method, which help to determine the period.

2.4.1 Statistical Cryptanalysis.

All natural languages have statistical characteristics, which mean that each character in the alphabet has its own frequency in any text of 1000 characters or

more. Since these frequencies are so consistent, then an approximate probability can be attached to letter. For example p(e) is much greater than every other probability, we would deduce that the most significant letter in a monoalphabetic cipher is equivalent for (e). However, English characters can be grouped into five sets according to their frequencies:

group	letters	frequency
I	e	9.614
II	t , a ,o , i ,n , s , h, r	6.855 - 4.532
III	d , l	3.219 - 3.047
IV	c, u , m , w , f, g , y , p , b	2.106 - 1.129
V	v, k , j , x , q , z	0.741 - 0.056

Digrams, trigrams, also have consistent frequencies.

Example: in a given ciphertext, after calculating characters frequency, the following results are obtained:

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 .
5	24	19	23	12	7	0	24	21	29	6	21	1	3	0	3	1	11	14	8	0	27	5	17	12	45

2.4.2 Index of Coincidence.

Index of coincidence measures the variation in the frequencies of the letters in ciphertext. If the period of cipher is one (1), then it means that a simple substitution cipher has been used, and there will be a considerable variation in the letter frequencies, and the index of coincidence (IC) will be high. As the period (d) is increased, the variation gradually eliminated due to diffusion, and the IC will be low. IC is defined by the following formula:

$$IC = \frac{\sum_{i=1}^{n} Fi(F-1)}{N(N-1)}$$

Where Fi : is the frequency of the i^{th} letter in the cipher text.

N : is the length of cipher text.

n: is the number of alphabet.

Once IC is calculated, the period **d** can be estimated using the following formula:

exp(IC) = (N-d)/ d(n-1) (0.066) + N(d-1)/d(N-1) (0.038)

IC is developed from the message which ranges from flat distribution (infinite period) to 0.066 for English cipher with period 1.Thus IC varying from 0.038 for an infinite period to 0.066 for a period of 1 .as shown in table (2-4)

Period (d)	IC
1	0.066
2	0.052
3	0.047
4	0.045
5	0.044
10	0.041
Large	0.038

Table (2-4) expected index of coincidence.

To estimate the period of a given cipher, measure the frequencies of letters in the ciphertext, compute IC, and finally compare this with expected values shown in table 2-4.

2.4.3 Kasiski Method.

Kasiski method uses repetition in the ciphertext to give clues to the cryptanalyst of the period. IC method analyzes repetitions in the ciphertext to determine the exact period. Repetitions occur in the ciphertext when a plaintext pattern repeats at a distance equal to a multiple of the key length. IC method is useful when it used with Kasiski method, to confirm the period 1 found by the later method.

Example:

M: T O B E O R N O T T O B E

K: H A M H A M H A M H A M H

Z (M): A O N L O D U O F A O N L

If C ciphertext repetitions are found at intervals Ij (m > j > I), the period is likely to be some number that divides most of the c intervals.

Recently, many researchers in cryptanalysis have introduced several methods and techniques using modern tools such as automated cryptanalysis. Genetic algorithms and genetic programming. All these methods make use of the statistical properties of the natural language. Among these methods, the work of Carroll and Martin which is an automated cryptanalysis procedure to solve simple substitution ciphers in ciphertext only attack. It was a rule based Artificial Intelligence program written in Prolog language.

Carroll and Lynda Robins present another work, which is an extension to the work published by Carroll and Martin. In this work the execution time have been reduced or simple substitution ciphers, with the addition of solving polyalphabetic ciphers.

Jolin C. King and Dennis RJ3ahlei present a framework for representing homophone ciphers graphically using the genetic systems. Spillman, and others present a method of using Genetic Algorithms finding the key for simple substitution ciphers, although it is not believed that this approach will always find the exact key. However, it gives evident that in a short run it will come close enough to the exact key that a visual inspection of the resulting plaintext could be used to determine any misplaced letters. Matliews has showed the possibility of using genetic algorithms as an efficient tool searching large key spaces and breaking classic cryptographic systems. More recently, two methods have been presented, the first one using genetic algorithm, and the second one using the adaptive genetic programming. All the above mentioned methods rely on the properties of the natural language and make use of tlie statistical information gained from the language redundancy remains in the ciphertext. Finally, a black-box attack, using neuro-identifier, has proved rigorously that ANN can build a cryptanalysis model for all classical ciphers, and equivalent transfer function which emulate the cipher system without the knowledge of statistical characteristics.