Using Genetic Algorithm to Break a Classical Cryptosystem (Transposition Cipher)

Maha Ali Hussain

Abstract

The genetic algorithm (GA) is an adaptive search method that has the ability for a smart search to find the best solution and to reduce the number of trials and time required for obtaining the optimal solution .the practicality of using the GA is to solve complex problems compared with traditional search techniques.

Many of the GA-based attacks lacked information required for comparison to the traditional attacks. Dependence on parameters unique to one GA-based attack does not allow for effective comparison among the studies approaches.

GA's are a class of optimization algorithms .GA's attempts to solve problems through modeling a simplified version of genetic process. There are many problems for which a GA approach is useful .it is however, untraditional if cryptanalysis is such a problem.

The aim of this work was to implement cryptanalysis attack algorithms, called GA- cryptanalysis system, in classical cryptographic systems (simple transposition) in the performance of genetic algorithm. We have also studied the effects of changing the parameters and variables (cipher text length, mutation rate, population size and number of generation) for controls the algorithm.

The GA –cryptanalysis system suggested new fitness for simple transposition. The suggested fitness depends on the most frequent diagram and trigram in the test text.

References

[1]. **Tim Kientzle** "A Programmer's Guide to Sound" **Addison – Wesely Developer Press.1998**

[2]. **Ernest L. Hall** "Computer Image Processing and Recognition" **Academic Press.1979**

[3]. John Watkinson **"Compression in Video and Audio"** Hartnolls Limited.1997

[4]. Kruti Dangarwala and Jigar Shah **"Implementation and Comparison of Companding and Silence audio Compression Techniques"**.2010

[5]. M. A. Shaalan **"New High Synthetic Coding Methods for Compressing digital Speech Signals"** M. Sc. Thesis, University of Baghdad.2000

[6]. Yon Q. Shi and Huifang Sun **"Image and Video Compression for Multimedia Engineering"** CRC Press.2000

[7]. Daniel Gent **"Speech and Music Compression"** 2008

The experimental results reported will shed more light into how parameters affect the GA's search power in the context of cryptographic problems.

1- Related Works

The first paper published in 1993 [1] by R.A.J Matthews ,uses an order-based GA to attack a simple transposition cipher .it covers background material and problem considerations well .Additional parameters are used to expand the range of parameters value.

A PH.D thesis "Use of GA in cryptanalysis of a class of stream cipher system "which introduced by Dr. Alageelee S. A in 1998 ,this work used Genetic Algorithm in cryptanalysis of class of stream cipher system depending on finding correlation between cipher text and output of some of LFSR[2].

Another paper published in 2007 [3]by R.Toemeh,S. Arumugam "Breaking Transposition Cipher with Genetic Algorithm". The aim of the research is to investigate the use of genetic algorithm in the cryptanalysis of transposition cipher. The applicability of genetic algorithms for searching the key space of encryption scheme is studied. The frequency of bigram and trigram is used as an essential factor in objective function.

2- The Difference of GA's from Traditional Methods

Genetic algorithms are different from more normal optimization and search procedures in four ways:

1. GA's work a coding of the parameter set, not the parameters themselves.
2. GA's search from a population of points, not a single point.

3. GA's use payoff (objective function) information, not derivatives or other auxiliary knowledge.
4. GA's use probabilistic transition rules ,not deterministic rules.

3- The Basic GA's Cycle :

the basic GA's cycle based on the three processes (selection, mating, and mutation) as shown in figure (1)

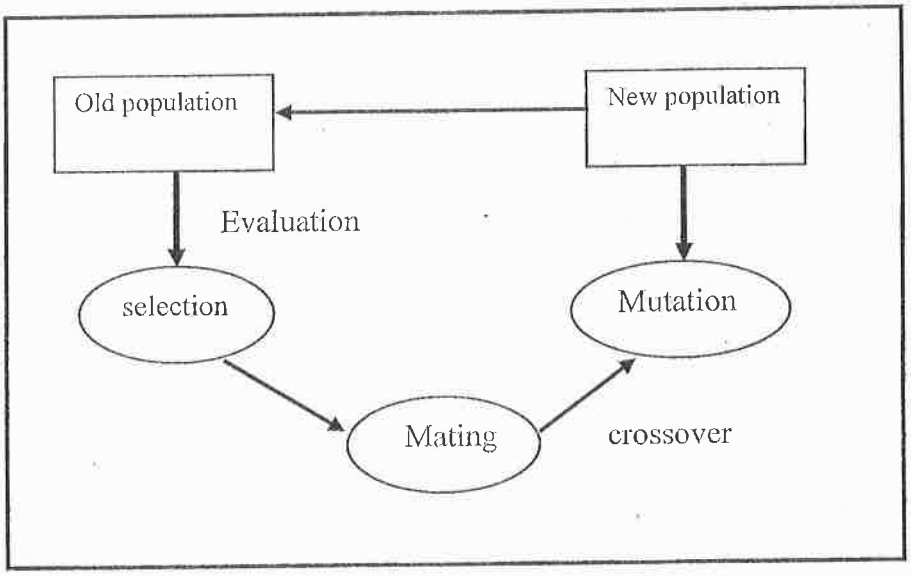

Figure (1) the basic GA cycle

Genetic algorithms (GA's) are search algorithms based on the mechanics of natural selection and natural genetics .they combine survival of fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search.

GA's attempt to identify optimal solution by applying the techniques of natural selection to a population of solutions ,the solutions are evaluated ,the led solutions are killed ,the

remaining solutions are recombined (mate)to form a new generation of solution.

Thus, a GA is an iterative procedure, which maintains a constant size population of candidate solution. During each iteration step (Generation) the structures in the current population are evaluated.

An abstract view of the GA is:

```
Generation =0;
Initialize G(P);  {G=Generation ;P=Population }
Evaluate G(P);
While (GA has not converged or terminated)
    Generation =Generation +1 ;
    Select G(P) from  G(P-1);
    Crossover G(P);
    Mutate G(P);
    Evaluate G(P);
End (While)
Terminate the GA.
```

The GA main steps are as follows:

1. The selection process, which determines which string in the current generation, will be used to create the next generation .by using the best point to determine the next population.
2. the mating process that determines the actual form of the string in the next generation .two of the selected parents are paired .if the length of each string is r, then a random number between 1 and r is selected, says, the mating process is one swapping bits s+1 to r of the first parents with bits s+1 to r of the second parent. In this way two new strings are created as shown in figure (1-2).

3. The mutation process. A fixed mutation probability is set at the start of the algorithm. Bits in all the new string are then subject to change based on this mutation probability.

These three steps are repeated to create each new generation it continues in this fashion until some stopping condition is reached such as a maximum number of generation or a specified fitness value (threshold) .

4- Classical Cryptosystems (Simple Transpositions):

A transposition is an encryption in which the letters of the message are rearranged .with a transposition the goal is diffusion, spreading the information from the message or the key out widely across the cipher text. The columnar transposition is a rearrangement of the characters of the plaintext into columns.

The columnar transposition is a rearrangement of the characters of the plaintext into columns.
The following example is a five-column transposition .the plaintext characters are separated into blocks of five and arranged one block after another, as shown here.

c1 c2 c3 c4 c5
c6 c7 c8 c9 c10
c11 c12 etc.

The resulting cipher text is formed by transferring the columns.
c1 c6 c11 .....c2 c7 c12.... c3 c8 etc..

4-1 Cryptanalysis of Transpositions (Traditional Method):

Before we talk about the transposition cryptanalysis we have to know something about Diagram and Trigram.

Just as there are characteristic letter frequencies ,there are also characteristic patterns of pairs of adjacent letters ,called diagrams ,letter pairs such as "-RE-", "-TH-" ,"-EN-",and "-ED-" appear very frequently .Table (1) lists the 32 most common diagrams and trigrams (groups of three letters) in English (they are shown with the most frequent ones first).

The first step in analysis of transposition is to compute the letter frequencies .the fact that all letters will appear with their normal frequencies implies that a transposition has been performed .given a string of text ,the trick is to break it into columns [5].

Table (1) most common diagram and Trigram

Diagram	Trigram
AN	THI
AR	NTH
AS	THE
AT	ENT
EA	AND
ED	HER
EN	ETH
ER	DTH
ES	WAS
ET	ONE
HA	ION
HE	THA
HI	OUR
IN	FOR
IS	IVE

IT	TIO
ND	ING
NG	
NT	
OF	
ON	
OR	
OU	
RE	
SE	
ST	
TE	
TF	
TH	
TI	
TO	

The process involves exhaustive comparison of strings of cipher text .the process will compare a block of cipher text characters against characters successively farther away in the cipher text .imagine a moving window that locates a block of characters for checking .assume the block being compared is seven characters .the first comparison is c1 to c8 ,c2, to c9…, c7 to c14 .then the window of comparison shifts and c1 is compared to c9, c2 to c10 ,and so forth .the window shifts again to c1 against c10.

5- Design and implementation of GA cryptanalysis system

5-1 Use GA to break transposition cipher

5-1-1 Problem Definition

The idea behind a simple transposition cipher (STC) is to keep the plaintext characters unchanged ,but alter this position by rearrangement using a transposition .although simple

transposition ciphers change the dependencies between consecutive characters ,they fairly not very hard to recognized since the frequency distribution of the character is preserved.

The type of transposition ciphers, which want be attacked, encrypts text according to the following classical two-stage algorithm:

a- A key of length N takes the form of a permutation of the integers 1 to N .the plaintext ,of L characters ,is written beneath the key to form a matrix N characters wide and at least L mod N characters deep.

b- The text is then enciphered by reading it off in columns in the order dictated by the integers making up the key.

The breaking of transposition cipher involves two stage processes first, the length of transposition sequence (i.e. N) must be found, and then the permutation of the N integers determined. If the length of a keyword is allowed to be anything up to B integers long ,then the total number of permutations possible for the transposition system is P where :

$$P(N) = \sum^{B} N!$$

This number rapidly increases with N, P(6) =873 ,while P(12) $\approx 5.23 * 10^{8}$, making brute- force methods quickly impractical.

5-1-2 Key Representation

First ,we should address an important question connected with chromosome representation ,should we leave a chromosome to be an integer vector ,or permutation of 1..N integers .when using the suitable representation allowed us to use the suitable crossover and mutation

operators ,applying these operators we got legal offspring ,i.e. offspring within the search space .As this representation indicates, the size of the key space is N! Which suggests that a poorly random search is not acceptable?

5-1-3 Initial Population

The cryptanalysis process begins with a set of chromosomes consisting of permutations of the integers 1 to N ,e.g. "86513247" for N=8 .it then applies each such guessed key to the cipher text ,and assesses the "fitness "of each by determining the extent to which the attempted decryption matches certain characteristics of plain English.

5-1-4 Fitness Function

The fitness rating helps the transposition cryptanalysis algorithm achieve breaking by awarding scores according to the number of times two and three letter combinations (bi-and tri-grams) commonly found in English actually occur in the decrypted text .the more columns correctly put next to one another by this algorithm ,the higher the fitness rating ascribed to that trail permutation .whenever these combinations were found in decrypted text ,points were awarded ,the highest being given for the appearance of trigrams ,as their appearance suggests that three columns have been correctly aligned .

A study has been made in English text consists of more than 4500 letters ,taken the 29 high frequency diagram letters and 12 high frequency trigram letters in

Table (2) gives 41 combinations letters and their percentage frequency in standard English text .were the percent calculated by :

Let frd [D_i] denotes the frequency of the diagram letters ,where D_i={TH,HE,....,HA} ,where $1 \leq i \leq 29$,and frq[Tj] denotes the frequency of the trigram letters ,and $T_j \in$ {ARE,THE ,....,DTH} ,where $1 \leq j \leq 12$,then percent (diagram)= frq [D_i] / (L-1) and percent (trigram)= frq[T_j] / (L-2) ,(we use (L-2) because the diagram and trigram used with overlap),and

L:is the text length then :

Total percent (diagram)= $\sum_{i=1}^{29}$ frq[D_i] / (L-1) and ,

Total percent(trigram)= $\sum_{j=1}^{12}$ frq [T_j] / (L-2)

So the total percent will represent the fitness value of this algorithm:

fitness= $\sum_{i=1}^{29}$ frq [D_i] / (L-1) + $\sum_{j=1}^{12}$ frq [T_j] /(L-2)

Table (2) 29 diagram and 12 trigram letters combinations

29 Diagram Letters				12 trigram Letters	
Diagram	Percent	Diagram	Percent	Trigram	Percent
TH	0.0250791	EA	0.0079078	ARE	0.0006778
HE	0.0237235	NG	0.0103931	THE	0.0153638
IN	0.0171713	AS	0.0112969	ING	0.0085856
AN	0.0153638	OR	0.0117488	AND	0.0085856
RE	0.0115228	TI	0.0092634	ENT	0.0042928
ED	0.0092634	IS	0.0092634	THA	0.0033891
ON	0.0108450	ET	0.0067781	NTH	0.0027113
ES	0.0131044	IT	0.0040669	WAS	0.0045188
ST	0.0151378	AR	0.0056484	HER	0.0011297
EN	0.0108450	TE	0.0070041	ETH	0.0038409
AT	0.0131044	SE	0.0079078	FOR	0.0020334
TO	0.0101672	HI	0.0063263	DTH	0.0009038
NT	0.0099413	OF	0.0074559	-------	---------

ND	0.0122006	HA	0.0079078	--------	---------
OU	0.0112969	---------	----------	--------	---------
Di-Total	Percent = 0.3246724			Tri-Total	0.0564844
Total	Percent =0.3811568 =fitness of plaintext				

It's clear that the other non-plaintext fitness values <= 0.3811568. Its important to mentioned that , the text fitness increased as the text length increased ,and vice versa.

5-1-5 Genetic Operators

1. **Selection Operator**

   The transposition cryptanalysis algorithm applies "roulette wheel" selection, in which the probability of a chromosome being selected is proportional to its normalized fitness. This constitutes the crucial "survival of the fitness" fitter. Clearly ,fitter chromosomes stand a better chance of passing through it, but it is important to note that even chromosome with a very low initial fitness have some chance of getting through .this ensures that each has some chance of developing into better chromosome by subsequent breeding.

   As there is a risk of fittest chromosome not surviving the fitter, however ,the algorithm incorporates "elitism" ,which ensures the fittest ("elite") chromosome found up to that point is always allowed into the breeding pool.

2. **Crossover Operator :**

   The breeding process is achieved using the position-based crossover technique proposed by swswerda [5] ,whose extensive numerical work has shown to be particularly effective for scheduling GA's ,the family of order-based GA's to which this

algorithm belongs. two chromosomes are taken for breeding ,and a random bit-string of the same length is generated .for example ,for N=9 we may have:

Chromosome A : 4 5 1 7 8 3 9 2 6
Chromosome B : 4 3 1 9 2 8 7 6 5
Bit-string : 1 0 0 1 1 0 1 0 1

To form one child of A and B ,take those elements of A corresponding to a 1:

4 * * 7 8 * 9 * 6

and enter the elements of A omitted 5,1,3 and 2 in the order they appear in chromosome B,i.e 3, 1 , 2 , 5 giving

child A : 4 5 1 7 8 3 9 2 6

to form the other child of A and B ,take those elements of B corresponding to a 0 in the bit-string:

* 3 1 * * 8 * 6 *

and enter the elements of B omitted ,i.e 4,5,7,9,2 giving :

child B : 4 3 1 5 7 8 9 6 2

3. Mutation Operator

Following crossover ,the algorithm applies the mutation operators to introduce genetic diversity into the evolving population of permutation .The used mutation operator is a simple two-point mutation, which randomly selects two elements in the chromosome and swap them . Thus 4 3 1 5 7 8 9 6 2 become 4 3 6 5 7 8 9 1 2 .Numerical studies of scheduling GA's by

syswerda [5] found this to be better than other mutation operators.

4. Genetic Parameters

The following parameters are be used :population size pop size =20 ,probability of crossover Pc= 0.8 ,probability of mutation Pm= 0.1 ,and 100 generations=100.

5. Experimental Results

The breaking of a transposition cipher can be divided into two elements: finding the correct key length N, and then finding the correct permutation of that key, i.e. the permutation of integer 1 to N.

To investigate the performance of the cryptanalysis algorithm in the first task, we encrypted the standard text given above using a target transposition key, K ,then we search for the best possible decryption of the text using (11) keys of length ranging from 6 to 16 ,one of which was the same as K .Table (3) shows the results of cryptanalysis GA algorithm for 100 generation using 178 letters cipher text length.

Table (3) Results of 100 generations for STC

Gen.	Fitness	Chr.no	Best Chromosome
0	0.2303371	3	578192463
9	0.2415730	7	817492635
17	0.2640449	10	871492635
29	0.2752809	3	486359217
37	0.2977528	14	836592417
41	0.3202247	5	681735924
63	0.3258427	8	861735924
69	0.3370787	13	592417863
80	0.3595506	17	359241786
86	0.3651685	8	863592417

The best chromosome after (100) generations was:

8 6 3 5 9 2 4 1 7

which is equal to the real transposition key with fitness=0.3651685

6- Role of GA –Parameters in Cryptanalysis

The parameters of GA are: cipher text length (L) ,number of generation (NG),population size (p) and mutation probability (p_m) .when we change one parameter of the mentioned parameters ,we let the others fixed ,the fixed values of these parameters are :L=250 ,NG=100,NP=20 and P_m=0.1.

6-1 Role of Cipher text Length

In this study we take three samples of cipher text length (L=100 ,150,250) then compare the obtained fitness and the key with origin . Table (4) shows these results.

Table (4) the fitness and Key results for L=100,150,250

L	STC	
	Fit.	Key
100	0.282	4 1 7 2 6 3 5
150	0.304	1 7 6 3 5 2 4
250	0.433	6 3 5 2 4 1 7
Origin	0.433	6 3 5 2 4 1 7

6-2 Role of Number of Generation

In this study we take three numbers samples of number of generations (NG=50,100,150) then compare the obtained fitness and the key with origin. Table (5) shows these results.

Table (5) the fitness and Key results for L=100,150,250

NG	STC	
	Fit.	Key
50	0.345	2 4 1 7 6 3 5
100	0.433	6 3 5 2 4 1 7
150	0.433	6 3 5 2 4 1 7
Origin	0.433	6 3 5 2 4 1 7

6-3 Role of Population Size

In this study we take three samples of number of population size (p=20,25,30), then compare the obtained fitness and the key with origin table (6) shows these results.

Table (6) the fitness and Key results for L=100,150,250

P	STC	
	Fit.	Key
20	0.433	6 3 5 2 4 1 7
25	0.345	7 6 3 5 2 4 1
30	0.433	6 3 5 2 4 1 7
Origin	0.433	6 3 5 2 4 1 7

4-2-4 Role of Mutation Probabilities:

In this study we take three samples of mutation probability (P_m=0.1, 0.3, 0..5) ,then compare the obtained fitness and the key with origin. Table (7) shows these results.

Table (7) fitness and Key results for P_m=0.2, 0.3, 0.5

P_m	STC	
	Fit.	Key
0.1	0.433	6 3 5 2 4 1 7
0.3	0.433	6 3 5 2 4 1 7
0.5	0.345	2 4 1 7 6 3 5
Origin	0.433	6 3 5 2 4 1 7

7- Conclusions

1. the cipher text length has serious effects on GA cryptanalysis results for the simple transposition ciphers .the obtained results are more precise whenever the length of cipher text as high as possible ,as noticed in Table (4).

2. there is no clear role for the number of population size when it applied in the two ciphers ,Table (5)describes this situation.
3. As be recommended, the mutation probabilities must be as low as possible in order to approach the correct or heuristic key of the two cipher and that what the results obtained from table (6) be certain on.
4. We can improve the fitness value by adding more diagram strings ,or change the fitness value by using positive or negative weights for any diagram and trigram.
5. Visual demonstration of the plaintext of the two ciphers is required most of the time to determine the exact key.

8- Future works

1. the GA cryptanalysis system can be extended to break more classical ciphers like substitution , ploy alphabet, vigenere, vernam...etc.

2. The GA cryptanalysis system can be extended to break more modern ciphers systems like public key and stream cipher.
3. It is able to make other kinds of studies on the GA parameters, like crossover types.

9- References :

[1] Matthews ,R. A. J. ,"The Use of Genetic Algorithm in Cryptanalysis", crypto logia ,17(4) ,187-201 ,April 1993.

[2] Al-Ageelee S. A. "Use of GA in Cryptanalysis of a class of Stream Cipher System "PH.D thesis, Technology University, 1998.

[3] R. Toemeh, S. Arumugam." Breaking Transposition Cipher with Genetic Algorithm" , Electronics and Electrical Engineering. -- Kaunas: Technologija, 2007.

[4] Sinkov A. "Elementary Cryptanalysis ", mathematical Association of America,1989.

[5] Syswerda ,G "Uniform Crossover in Genetic Algorithms", proceedings of the 3th International Conference on Genetic Algorithms, San Mateo pp.2-9,1989.

[6] Goldberg, D. E ., "Genetic Algorithm in Search ,Optimization ,and Machine Learning ",Boston :Addison –Wesley ,1989.