Application of Polyalphabetic Substitution cipher Using Genetic Algorithm

Ghusoon Salim Basheer
College of Computer sciences and Mathematics
University of Mosul

Received on: 16/8/2006 **Accepted on: 23/1/2007**

الملخص

هنالك العديد من البحوث التي تناولت استخدام الخوارزمية الجينية مع علم التشفير و فك الشفرة (Cryptography)، وتشترك جميعها بكونها تستخدم الخوارزمية الجينية لفك شفرة البيانات المشفرة(Cryptanalysis) والحصول على النص الاصلي. في هذا البحث تم تقديم استخدام جديد للخوارزمية الجينية مع الـــ (Cryptography)، حيث تم الحصول على افضل مفتاح سري(Secret Key) لتشفير البيانات بطريقة الـــ(Polyalphabetic Substitution Cipher) ومن ثم استخدامه في التشفير وفك الشفرة وبما يضمن مستوى عالي من السرية. تمت كتابة البرنامج بلغة Matlab (6.5) .

Abstract

Several Genetic Algorithms have been developed for applications of cryptography problem; the primary distinction among all of them being the G.A. used for decryption problem and obtains the plain text. In this paper a new approach is proposed using Genetic Algorithm with cryptography. G.A. is used to obtain a best secret key in polyalphabetic substitution cipher. This key will be used then for encryption and decryption with a high level of security. The program is written in Matlab language (6.5).

1. Introduction:

Many problems that computer scientists encounter are very hard to solve. Some of these problems commonly called NP-hard problems, have no known efficient solution process (i.e., no algorithm that returns a solution in a time that is polynomial with respect to the size of the input). While no efficient process is known for generating an optimal solution all of the time to an NP-hard problem, many NP-hard problems can be solved efficiently much of the time to near optimality using a heuristic. A heuristic is a solution-generating rule that gives near-optimal solutions a high percentage of the time. However, there is no guarantee that a heuristic will ever give a near-optimal solution at all**.**(11)

Several Genetic Algorithms have been developed for decryption for many types of encryption methods. For example, Spillman R.(12) has

shown on his paper that genetic algorithm could be used to easily compromise even high density knapsack ciphers. Gester J. (6) used a simple genetic algorithm to search the key space of cryptograms in an attempt to create a general solver for such problems. In their paper, Spillman R., Janssen M., Nelson B., and Kepner M.,(13) consider a new approach to cryptanalysts based on the application of a genetic algorithm. They showed that such an algorithm could be used to discover the key for a simple substitution cipher. But Dimovski A., Gligoroski D.(4) in their paper presented three optimization heuristics which can be utilized in attacks on the transposition cipher, These heuristics are simulated annealing, genetic algorithm and tabu search.

In this paper our goal of using G.A. with encryption is to generate the best secret key for Monoalphabetic Substitution Cipher which satisfied a high level of security.

2. Cryptography:

Is the science of writing in secret code. In data and telecommunications, cryptography is necessary when communicating over any UN trusted medium, which includes just about any network, particularly the Internet.

There are, in general, three types of cryptographic schemes: secret key (or symmetric) cryptography, public-key (or asymmetric) cryptography, and hash functions. With ***secret key cryptography***, (which is proposed in this paper) a single key is used for both encryption and decryption. The sender uses the key (or some set of rules) to encrypt the plaintext and sends

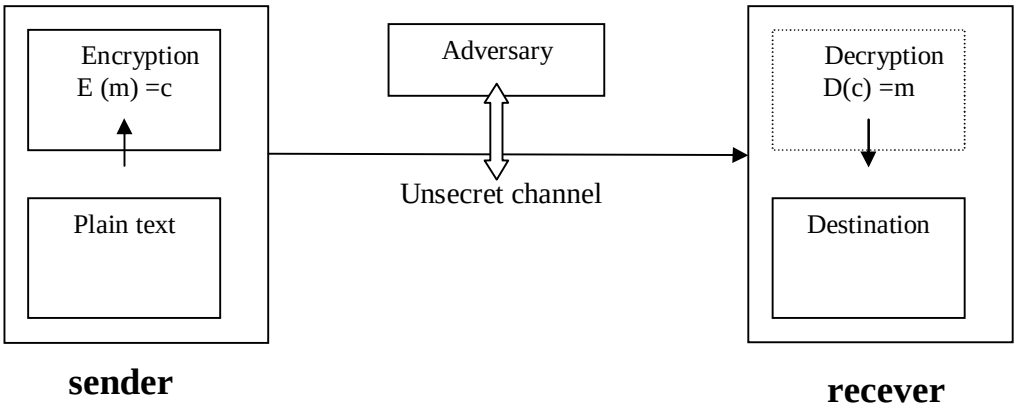

Figure 1: The block diagram for cipher system.

the cipher text to the receiver. The receiver applies the same key (or ruleset) to decrypt the message and recover the plaintext. Because a single key is used for both functions, secret key cryptography is also called ***symmetric encryption***. The initial unencrypted data is referred to as ***plaintext***. It is

encrypted into ***cipher text***, which will in turn (usually) be decrypted into usable plaintext. (5)
figure (1) shows the block diagram for cipher system.

3. Polyalphabetic Substitution cipher:

The Polyalphabetic substitution cipher is a simple extension of the monoalphabetic one. The message is broken into blocks of equal length, say *B*, and then each position in the block (1,…, *B*) is encrypted (or decrypted) using a different simple substitution cipher key. The block size *B* is often referred to as the period of the cipher (3).

An example of a polyalphabetic substitution cipher is shown on figure 1. The block size (i.e., B) is chosen to be three; figure (2) gives an example key and shows the corresponding encryption.

The plain text letters	**p**	**o**	**l**	**y**	**a**	**l**	**p**	**h**	**a**	**b**	**e**	**t**	**i**	**c**
Positions of the letters in the plain text	**16**	**15**	**12**	**25**	**1**	**12**	**16**	**8**	**1**	**2**	**5**	**20**	**9**	**3**
key	**8**	**2**	**20**	8	2	20	**8**	**2**	**20**	8	2	20	**8**	**2**
Positions of the letters in the encrypted text	**24**	**17**	**6**	**7**	**3**	**6**	**24**	**10**	**21**	**10**	**7**	**14**	**17**	**5**
Encrypted letters	**x**	**q**	**f**	**g**	**c**	**f**	**x**	**J**	**u**	**x**	**g**	**n**	**q**	**e**

Figure (2): An example explaining polyalphabetic substitution cipher.

4. Genetic Algorithm:

Genetic algorithms (GA) are general, domain-independent search and optimization techniques developed in the 1970s. These algorithms borrow from nature the concept of natural selection, according to which, the stronger survives to a competition while the weaker will die, so that the genes contained within the chromosomes of dominant individuals will spread within the next generation. As the number of generations' increases, an individual able to withstand the external environmental pressure is likely to be contained in the population. Similarly, GAs is based on an initial population of individuals, each of which represents a possible solution to the problem at hand (9).

Over time, the genetic algorithm will produce few exceptionally fit individual solutions within the population. These individual solutions would be highly, but not completely, optimized. (11)

A G.A. is composed of three main components, which are problem dependent: the **encoding problem** *(***Chromosome Representation***)*, **evaluation function** (**Fitness Function**) and the **operators**.

A. Encoding Problem (Chromosome Representation):
The chromosome representation selected for the problem concerned is a simple one in which each gene is an integer value indicating the key of length 26 integer values. The positions of genes in a chromosome specify the shifting value which each letter in the plain text file should be changed to construct encrypted text file. For example, the following chromosome representation in figure (3) shows that char 1 in the plain text file should be shifted by 4 and char 2 shifted by 6 and so on.

4 6 23 7 9 11 12 16 19 18 25 22 23 24 21 14 20 17 13 10 8 5 2 15 3 1
Figure (3): Chromosome representation, the value of each gene in the chromosome represented the shifting value for each letter in the plain text file.

B. Evaluation Function (Fitness Function):
The evaluation of chromosome is a critical portion of a genetic algorithm. For encryption problem, we used a simple evaluation function in which the fitness value of a chromosome is determined by determining the maximum differentce between letters frequencies in the plain text file and the encrypted file.

C. Genetic Operators:
Crossover: The main genetic operator randomly selects two chromosomes from the population and swaps second part of each gene after a randomly selected point. This is equivalent to assigning a subset of key to other**.**
Mutation: An operator which produces random changes in various chromosomes. Mutation serves the critical role of either replacing the chromosomes lost from the population during the selection process or introducing new chromosomes that were not present in the initial population. The mutation rate controls the rate at which new chromosomes are introduced into the population. (7)

Selection: Is the process of keeping and eliminating chromosomes in the population based on their relative quality or fitness. In most practices, a roulette wheel approach is adopted as the selection procedure. A value-based selection scheme assigns roulette wheel sectors proportional to the fitness value of the chromosomes. (7)
Figure (4) illustrates the basic steps of standard genetic algorithm.

*** Create an initial population randomly**

***While not (termination condition) do**

*** Evaluate each member's fitness**

*** Kill the bottom x% elements of the population**

*** Let the fittest reproduce themselves**

*** Randomly select two members/parents (many other selection methods are also used)**

*** Perform crossover on the selected elements to generate two children (any variations of crossover exist)**

*** Perform mutation**

*** End While**

Figure (4): A basic steps of a typical genetic algorithm.

5. Proposed Algorithm:

In this paper, new algorithm is proposed using G.A. for the encryption and decryption text file.

A. Encryption:

1.Input the plain text file.
2.Measure the letters frequencies of the plain text file.
3.Divide the plain text file into fixed size of blocks (each block consists of 26 chars).
4.Creat initial generation of 10 individuals(each individual representing a secret key),each of them consists of 26 random integer value.
5.Encrypt the plain text file as follows:
- By using each generated key we shift each letter in each block of the plain text file by corresponding value in this key. For each key, we obtain new encrypted file with different frequencies of letters.
6.Measure the frequencies of letters in each encrypted file (10 encrypted files).

7.Determine the fitness function for each individual(each key) as a maximum difference between frequencies of the letters in the plain text file and encrypted files.
8.**Repeat**
9.Creat new generation by selection, crossover and mutation.
10.Compute fitness for new generation.
11.**Until** no fitness improvement is achieved.
12.Determine the best key as the key which corresponds to maximum fitness value.
13.Converting the integer form of the best key to character form and adding it to the ending of the encrypted file.

B. Decryption:

In the second step of the algorithm, a decryption of the encrypted file is performed. After the receptor has the encrypted file do the following:
1. Spread the key from the encrypted file and convert it from character form to integer form.
2. Subtracting each value in the key from corresponding letters values in the encrypted file.

6. Practical Representation:

A. Initializing Population (pop): (by using **rand** order in Matlab language) we build function **pop** which generates array of 10 keys each key consists of 26 columns (randomly values between 1 and 26).
B. Frequency Function (FRE): We built function **FRE** for calculating the frequencies of each letter in the plain text file and in the encrypted file.
C. Dividing the plain text file: After determining the frequencies of the letters in the plain text file we divide it into blocks, each block consists of 26 characters.
D. Encrypting the plain text file: For encrypting the plain text file we used each key in **pop** array (10 row * 26 column) to shift each letter in each block of the plain text file (length of key = length of each block in the plain text file = 26 character). We used key for shifting letters in first block, second block… est., until we reached the ending of the file, as shown bellow:

While not end of file
Encrypted block=position of each letter in the plain text file+ coorespon -ding value in the key.
End

After encryption we obtain 10 encryption files.

E. Fitness Function (Evaluation Function): We measure the differences between frequencies for each letter in the plain text file and in the encrypted files. Assumption of their frequencies of letter represents the fitness function; the largest fitness value represents the better solution for the problem.

F. Selection: We use **roulette-wheel** for selection which is considered the simplest selection scheme which involves the following technique:

1.The individuals are mapped to contiguous segments of a line, each individual's segment is equal in size to its fitness.
2.A random number is generated and the individual whose segment spans the random number is selected.
3. The process is repeated until the desired number of individuals is obtained (called mating population).

Table (1) shows the selection probability for 11 individuals. Individual 1 is the most fit individual and occupies the largest interval, whereas individual 10 as the second least fit individual has the smallest interval on the line (see figure). Individual 11, the least fit interval, has a fitness value of 0 and get no chance for reproduction. (12)

Number of individual	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
fitness value	**2.0**	**1.8**	**1.6**	**1.4**	**1.2**	**1.0**	**0.8**	**0.6**	**0.4**	**0.2**	**0.0**
Selection probability	**0.18**	**0.16**	**0.15**	**0.13**	**0.11**	**0.09**	**0.07**	**0.06**	**0.03**	**0.02**	**0.0**

Table 1: Selection probability and fitness value

For selecting the mating population the appropriate number of uniformly distributed random numbers (uniform distributed between 0.0 and 1.0) is independently generated.
{Sample of 6 random numbers}

0.81, 0.32, 0.96, 0.01, 0.65, 0.42.

Figure (5) shows the selection process of the individuals for the example in table together with the above sample trials.

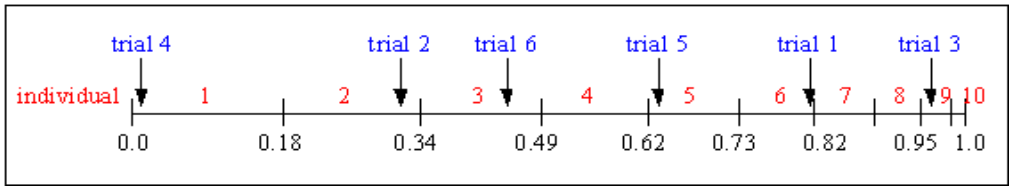

Figure (5): Roulette-wheel selection

After selection the mating population consists of the individuals:

1, 2, 3, 5, 6, 9.

The bad thing about this selection method is that if there is not much difference in the fitness of the chromosomes compared to the absolute fitness value, all chromosomes have a roughly equal chance of being selected to survive.

G. Crossover: We use a simple crossover for generating next generation; we don't neglect the original parents, since we generate a new population from original parents. This causes duplication of size population. (new generation consists of 20 keys in each iteration).
As an example, if the crossover point =10, the result of crossover shown below in figure (6).

H. Mutation: we choose mutation rate equal to 0.01.
I. Stop (Terminating conditions): After executing the above mentioned steps a new generation of keys is created and the steps are repeated until the stop condition is reached. Our algorithm will be stops when the fitness function reaches high value or when algorithm generates 15 populations.

10 5 22 9 18 11 1 13 17 6 2 3 20 26 12 4 8 14 7 25 24 16 15 23 24 21

Parent 1

5 26 4 19 6 22 3 14 7 23 2 13 8 20 1 18 12 9 21 6 11 10 24 7 8 25

Parent 2

10 5 22 9 18 11 1 13 17 6 2 3 20 26 12 4 8 14 7 25 24 16 15 23 24 10

Child 1(old parent 1)

5 26 4 19 6 22 3 14 7 23 2 13 8 20 1 18 12 9 21 6 11 10 24 7 8 25

Child 2(old parent 2)

10 5 22 9 18 11 1 13 17 6 2 13 8 20 1 18 12 9 21 6 11 10 24 7 8 25

Child 3

5 26 4 19 6 22 3 14 7 23 2 3 20 26 12 4 8 14 7 25 24 16 15 23 24 10

Child 4

Figure (6): illustrates the child construction from two parents. Note the first two child the parent themselves, and second two parent result from crossover the parents in position 10 (10 selected randomly).

7. Results and Discussion:

The program has been tested on the plain text from paper(5) which consists of (319) characters. An experimental result for the genetic algorithms was generated with 15 runs using 'Matlab' language. Roulette wheel selection is used to exploit past results to direct the search for efficient secret key. Probabilities of crossover and mutation are fixed. Mutation acts as a safety net to recover good genetic material that may be lost through selection and crossover. The genetic algorithm will stop when the fitness function is unchanged after some predefined number of generation.

-The plain text file:

[does increased security provide comfort to paranoid people or does security provide some very basic protections that we are naive to believe that we don't need during this time when the Internet provides essential communication between tens of millions of people and is being increasingly used as a tool for commerce security becomes a tremendously important issue to deal with.]

(Note: we eliminate all white space from text file before encrypt it)

A. Encryption the plain text file:

-The encrypted file using best key in first iteration of G.A.:

yjxctfidyclsqqzdwdpczpokpayjxctfidycuswbndwdpczpokpayjxctfidy
cusqkldwdpczpoofredxctfidycusqqzdwdpczpokpayjxctjhdycusqjtrwdpozpok
payjxctfidyceswqzdwdpcdpokpayoxctfidycusqseuwdpczpokpayjxcmgtdycu
sqqzdwdpczpokpayjxctfidycuspmndwdpczpokpayjxctfidycusqnaxwdgfzpok
payjxctfidycusqlzdwdpchpokpayjxctfidyculxqzdwdpczpokpa

-The encrypted file using best key in fifth iteration of G.A.:

twpkuppazjiflslrisewfblrgvowpkuspajgiflslrisewfblbgvowpkuppazjiflslrisew
fblbgvowpeupxazoifltlrisewfalbgvowtkupawzjiflslrisewfblbgvowpkupdazjif
lslrisewfblbgvowphupeazqiflslrisewfblbgvowpkuqiazjiflslrisewfhlbgvowpk
upqazjiflslrisewfblbgvohskupqazjiflslrisewfblbgvowpkupdazjiflslrisewfblbg
vofpkuhpazjiflslrisewfblbgv

-The encrypted file using best key in fifteen iteration of G.A.:

rjflyxlxkeymdxochrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcnbimoxiynnlbgzml
vyhhrgrwgcnbimoxiybnlbgzmlvylhrgrwgcnbimoxiibnlbgzmlvywhrgrwgcnb
imoxiygnlbgzmlvehhrgrpgcnbimoxiybnlbgzmlvyhjrgrwgcnbimtxiybnlbgzm
lvyhhrgrqmcnbimwoxiybnlbozmlvyhhrgrwgcnbimoliybnlbgzmlvyhhrgrwgc
nbimoxiybnlbgzmlvchhrgrwgcnbimoxiybnlbgzmlvyehrgrwgcnbi

Figure (7): shows an overview result of encrypted file for a variety of differing keys in several iterations.

Iteration of G.A.	Best key	Sums of all difference-s for all letters
1	24 8 21 25 15 26 2 22 16 19 10 7 4 3 11 14 6 12 23 9 5 20 18 13 17 1	391
5	5 16 24 25 1 4 13 3 23 18 10 22 7 9 20 12 2 14 17 26 21 6 19 15 11 8	405
15	14 21 13 7 16 10 9 19 17 26 24 20 25 20 22 23 19 24 13 1 7 15 5 2 8 18	411

Figure (7): Results for a best generated key and maximum differing frequencies in several iterations." best key" indicates the key which satisfies the largest differences of frequencies value and is to be inspected in the current step.

From figure (7), and depending on the fitness function value which is represented by **Sums of all differences for all letters** in the encrypted file and in the plain text file, we can note that the best key is generated after fifteen iteration, so this key will be used for encrypting the plain text file and add it to the end of the encrypted file before sending it to the receiver.

***Integer form of the best key:**

14 21 13 7 16 10 9 19 17 26 24 20 25 20 22 23 19 24 13 1 7 15 5 2 8 18

***Character form of the best key:**

numgpjisqzxtytvwsxmagoebhr

***Encrypted file after adding the key :**

rjflyxlxkeymdxochrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcnbimoxiybnlbgzml
vyhhrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcn
bimoxiybnlbgzmlvyhhrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcnbimoxiybnlbg
zmlvyhhrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcnbimoxiybnlbgzmlvyhhrgrw
gcnbimoxiybnlbgzmlvyhhrgrwgcnbimoxiybnlbgzmlvyhhrgrwgcnbi**numgpj**
isqzxtytvwsxmagoebhr

B. Decryption the encrypted file:

When receptor has the encrypted file he separates the key from the ending of the file(separate 26 letters),and converts it from character form to integer form , then subtracting each key value from encrypted file to obtain the plain text file.

8. Conclusion:

In this work, we have developed a novel method for using Genetic Algorithm with Cryptography. The method utilizes a genetic algorithm for evaluating the consequences of secret keys for encryption. The experimental results show that this kind of approach is very suitable for the encryption problem.

9. Future Work:

Several future directions are identified for further investigation. One of them includes applying Genetic Algorithm in other types of ciphers, like Monoalphabetic cipher.

REFERENCES

[1] Curtin, M. 1998."Snake Oil Warning Signs: Encryption Software to Avoid ",Nikos Drakos, Computer Based Learning Unit, University of Leeds.

[2] Delman ,B., 2004, "Genetic Algorithms in Cryptography", A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Rochester, New York, July 2004.

[3] Dimovski,I.A., Gligoroski2 D., 2003, "Attack On The Polyalphabetic ubstitution Cipher Using A Parallel Genetic Algorithm", Technical report, Swiss-Macedonian scientific cooperation trought SCOPES project,March 2003, Ohrid, Macedonia

[4] Dimovski, A., and Gligoroski D., 2003, "Attacks on the Transposition ciphers Using Optimization Heuristics", In Proceedings of the Second International Conference on Tools for AI, Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodist University Archimedean b.b., PO Box 162, 1000 Skopje, Macedonia .

[5] Gary C. Kessler, 1998, "An Overview of Cryptography ",Handbook on Local Area Networks, published by Auerbach in September 1998, Champlain College in Burlington.

[6] Gester J., 2003, " Solving Substitution Ciphers with Genetics Algorithm", In Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science.

[7] Jack M. West and John K. Antonio, 1999, "A Genetic Algorithm Approach to Scheduling Communications for a Class of Parallel Space-Time Adaptive Processing Algorithms ",School of Computer Science, University of Oklahoma 200 Felgar Street Norman, OK 73019 .

[8] Ludovic M_e, 2000, "GasSATA, a Genetic Algorithm as an Alternative Tool for Security Audit Trails Analysis ", SUP_ELEC B.P. 28 35511 Cesson _evign_e Cedex ,France.

[9] Mahmood A., 2001,"A Hybrid Genetic Algorithm for Task Scheduling in Multiprocessor Real-Time Systems", Department of Computer Science, University of Bahrain, Bahrain.

[10] McGovern A., 2004, "A Random Number Generation Technique with Encryption and Genetic Algorithm Applications ", Cambridge, Boston: MIT Press, McGraw-Hill.

[11] Phillips, J. , 2004, " Application of Genetic Algorithms to recovering corrupted file streams", A Thesis Presented to the Faculty of Bucknell University In Partial Fulfillment of the Requirements for the Degree of Bachelor of Science with Honors in Computer Science April 2004.

[12] Pohlheim, H., 2005," Genetic and Evolutionary Algorithm", part of version 3.7 of the GEATbx: Genetic and Evolutionary Algorithm Toolbox for use with Matlab.

[13] Song S., Kwok Y.,and Hwang K., 2005, "Security - Driven Heuristics and A Fast Genetic Algorithm for Trusted Grid Job Scheduling",19th IEEE international parallel and distributed processing symposium (IPDPS,05)-papers P.65a, University of Southern California, Los Angeles.

[14] Spillman, R., Janssen, M.; Nelson B., and Kepner, M., 1993,"Use of A Genetic Algorithm in the analysis of simple of simple Substitution Ciphers" , Department of Computer Science Pactfic Lutheran, University Tacoma WA 98447 USA. Email: SPILLMANR@PLU.EDU .