International Journal of Trend in Scientific Research and Development, Volume 1(4), ISSN: 2456-6470
www.ijtsrd.com

Application of Genetic Algorithm in Cryptanalysis of Mono-alphabetic Substitution Cipher

Dr. Prabha Shreeraj Nair
Dean Research
Tulsiramji Gayakwade Patil College of Engineering and Technology, Nagpur

ABSTRACT

Today, security is a vital concern in computer science, cryptography is used vastly for implementation of the same. Cryptanalysis is a process in which the security is attempted to breach and the complexity of this process is considered as security measurement. As cryptographic algorithm is open to all, the whole strength lies in the complexity of the key i.e. efforts to crack the key. Mostly the strength of the key is shown through its length, eventually the number of communication (Brute- force method). Genetic algorithms are considered to be a tool for meta heuristic applications. In this work an attempt is made to carry out cryptanalysis, through genetic algorithms.

In this, mono-alphabetic substitution cipher technique is considered. The experiment is carried out for four key samples, and attempt to break with variations in genetic operators i.e. selection, crossover and mutation. Regarding variations, for selection- random with elitism, roulette wheel and tournament options are used, for crossover – 1 -point, 2-point and Uniform options are used, with interchanging mutation.

Keywords: Genetic algorithm, Ciphers, Encryption, Roulette wheel selection

1 INTRODUCTION

Today, security is a vital concern in computer science for all type users and when it regards to communication it become more important. In computer science, cryptography or security is used vastly, it is a very common tool to protect the data, for example, protecting web traffic HTTPS is used for wireless traffic 802.11i WPA2 is used for files disk protection and content protection EFS (encryption file system) and CSS (content scrambling system) are used. The encryption algorithm and decryption algorithm are publicly known, but only the thing key is kept secret between sender and receiver. This implies the basic model of the crypto system for a secure data exchange or massage communication and the whole strength of the crypto system lie in the key. In other words, the degree of security can be measured in terms of difficulty or efforts to obtain the key which even can crack the whole crypto system [10].

At the time of literature review, it is found that mostly the security related literature or articles are in cryptography, key exchange algorithm, attempts to breach the security, hacking and cracking, cryptanalysis, protection from various attacks and security threats, etc. very few papers or articles were found to measure the strength of cryptography or on techniques of measurement of security. The cryptanalysis is selected as the area of research here because most of security techniques are based on cryptography.

In this research work the cryptanalysis of mono-alphabetic substitution cipher is done with three different efforts and the by measuring the results conclusion is given about the measurement of security of mono-alphabetic substitution cipher by calculating difficulty or efforts to crack the encryption system.

A. Genetic Algorithm

"A genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. The algorithm repeatedly modifies a population of individual solutions." [27]

For many real-world optimization problems polynomial time algorithm does not exist that making them hard or to solve. Many heuristic algorithm shave been devised to solve such hard problems. These heuristics can give well (might not be best) and

acceptable solutions in a reasonable computational time effort [25].

Soft Computing is an area where some of meta-heuristics solution such as evolutionary algorithms, artificial neural networks, and simulated annealing derived from physical, biological and natural phenomena has also been used to solve these problems shown capabilities to solve hard problems. Going towards the security in computer science it is obvious that it is very vast and a vital field so here in the research the capabilities of such algorithms used in the vital vast and a very typical ancient area of computer science.

In 1975, Holland proposed an idea in his book "Adaptation in natural and artificial systems". He elaborated how to implement the principles of the natural genesis to optimization problems and developed the first Genetic Algorithms. Holland's theory has been further extended and now Genetic Algorithms (GA) has taken the form of a powerful tool to exercise search and optimization problems. Genetic algorithms have bases on the concept of genetics and evolution. Genetic Algorithms represent to a collection of computational models guided by evolution. These genetic algorithms give a potential solution to specific problems on simple chromosome like data structures as preserving critical information by using recombination operators to these structures. Although Genetic algorithms are seen as function optimizers, they can be used and adapted in quite broad area. A Genetic Algorithm starts with an initial population of (typically random) chromosomes. After that, these structures are evaluated and promoted for reproductive opportunities in a manner that the chromosomes showing a better probability to solve the target problem are given extra edge and chances to "reproduce" than others. The "goodness" of a solution is generally described against the current population [28][26][24].

II. RELATED WORK

At the time of literature review, it is found that mostly the security related literature or articles are in cryptography, key exchange algorithm, attempts to breach the security, hacking and cracking, cryptanalysis, protection from various attacks and security threats, etc., very few papers or articles were found to measure the strength of cryptography or on techniques of measurement of security. Few of which are as

- S. S. Omran, A. S. Al-Khalid, and D. M. Al-Saadystudied the basic cryptosystem and Genetic Algorithms and proposed a new Genetic Algorithm to attack on mono-alphabetic cipher and poly-alphabetic substitution cipher (Vigenère cipher) in their different papers [1][4].
- Ali S.Al-Khalid and AlaaO. Al-Khfagi presented the scenario of attacking the Hill cipher using Genetic Algorithms [2].
- Aditi Bhateja and Shailender Kumar presented a method decryption of the Vigenere cipher with Genetic Algorithms using elitism with a novel fitness function [3].
- Ayman M B. Albassal and Dr. Abdel-Moneim A. Wahdan designed algorithm to find the key of a Substitution Permutation Network using Genetic Algorithms [7].
- Joseph Alexander Brown, Sheridan Houghten, and Beatrice Ombuki-Berman explore of the use of Genetic Algorithms (GA) upon a Substitution Permutation Network (SPN) cipher [8].
- Mohammad Faisal Uddin and Amr M. Youssef investigated Particle Swarm Optimization (PSO) applications in classical simple substitution cipher' sautomated cryptanalysis [9].
- Feng-Tse Lin and Cheng-Yan Kao described a method for cryptanalysis based on Genetic Algorithms to break the Vernam cipher. The approach used is ciphered text-only attack in which attacker is unaware of plain-text, the only thing attacker has to know about plaintext is that plain text is the English document [6].
- Jitin Luthra and Saibal K. Pal discussed the integration of the genetic operators used in Genetic Algorithms with the Firefly Algorithm for cryptanalysis of the mono-alphabetic substitution cipher [5].
- Salabat Khan, Waseem Shahzad, and Farrukh Aslam Khan discussed the cryptanalysis of DES (Data Encryption Standard) they provide the description of four rounded DES and proposed an algorithm for the cryptanalysis based on Binary Ant Colony Optimization (BACO) [11].

III. PROBLEM STATEMENT

The problem statement of the work is to find the most efficient genetic algorithm. As in previous sections it is described that a lot of work has done on genetic algorithm in the field of cryptanalysis but being specific to genetic algorithms there are various types of genetic operators that may provide a variation to

the results gained previously. Some of main genetic operators are listed in table I with their types. Here comes to the point the problem statement for this research to confirm the type of the operator and the combination of operators which gives the best result of genetic algorithm for cryptanalysis the mono-alphabetic substitution Cipher. In other words, use different genetic operators which may give the best result over others.

TABLE I

Classification of Genetic Operators

Selection	Crossover	Mutation
Random Selection	1-point Crossover	Flipping Mutation
Roulette Wheel	2-point Crossover	Interchanging Mutation
Selection		
Rank Selection	Multipoint	-
	Crossover	
Tournament	Three-parent	-
Selection	Crossover	
-	Uniform	-
	Crossover	

IV. PROPOSED WORK

According to the above problem statement the proposed work can be described as to design different Genetic Algorithm to crypt analyse the Mono-alphabetic Substitution Cipher using different type of genetic operator and then analyze and compare the results with respect to parameters considered for the best. Figure 1 depicts the scenario of basic model for proposed Work.

V. METHODOLOGY

The Genetic Algorithm starts with creating a random initial generation of individuals or chromosomes. Chromosomes represent the possible solution from a solution space chosen randomly. Members of the current population are selected in pairs and "mated" performing a crossover operation to generate members for the next generation.

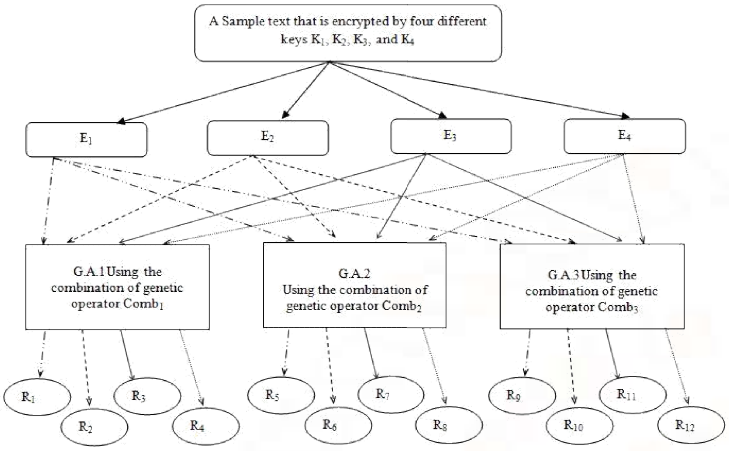

Fig. 1: Base Model for the Proposed Work.

Mutation is also applied to the randomly selected members of the current generation exchanging some random portion. Then a fitness function is evaluated and the fittest members are determined in each population. The crossover and mutation operations done with the base of crossover rate and mutation rate attributes, which is set by algorithm parameters for the individuals of population that has to be operated.

To fulfill the objectives of the research three combinations of genetic operator are used to design three different genetic algorithms for the experiment. Such combinations are as follows.

1 GA1: Random Selection, 1-point Crossover, Interchanging Mutation with Elitism.
2 GA2: Roulette wheel Selection, 2-Point Crossover, Interchanging Mutation
3 GA3: Tournament Selection, Uniform Crossover, Interchanging Mutation.

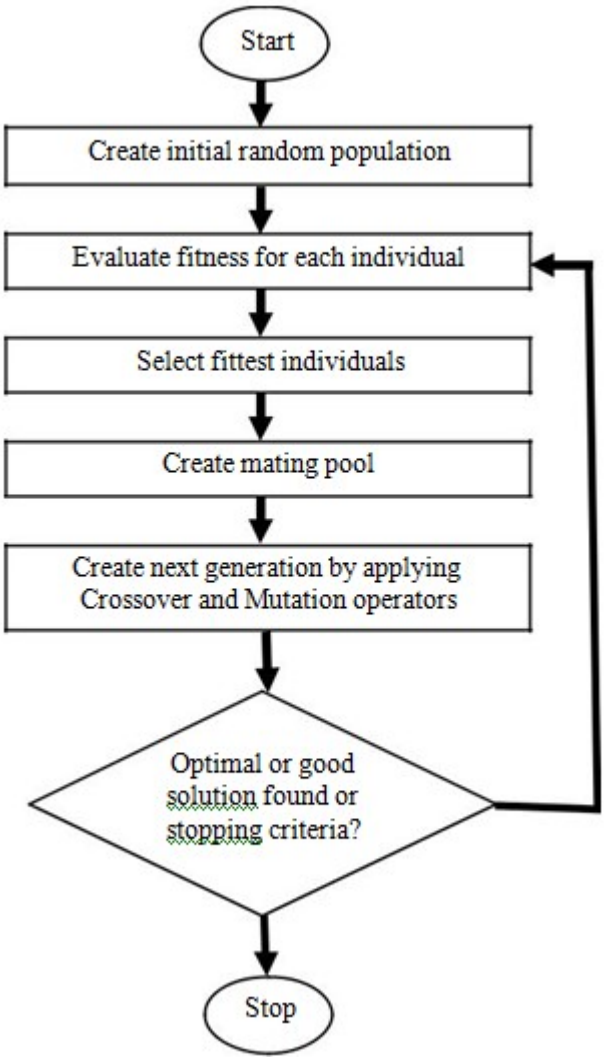

Fig. 2: Flow chart of Genetic Algorithm [15].

A. Pseudo Code

Genetic Algorithm (E, F [])

1. Input the Encrypted Text E.
2. Input the standard character frequencies of English language.
3. Define stopping criteria.
4. Start with a set of random keys of length 26 as initial population (let set of 15/20).
5. For 1 to population size do:
   I. Decrypt the given encrypted text using all keys in population.
   II. Calculate the fitness for all individuals (Keys) using fitness function.
   III. Apply Selection operator
   IV. Arrange selected individuals in pairs in mate pool.
   V. Apply Crossover operator to all individual pairs in mate pool.
   VI. Apply Mutation operator.
   VII. Replace initial population to get new population.
6. Go to step 5.
7. Repeat step 5 for new generation till stopping criteria

B. Initialization

The Algorithm starts with a set of randomly generated keys for decryption called as initial population and keys as individuals or chromosomes. All keys are applied to the encrypted text and the decrypted text is produced. Now the measurement of fitness is done according to the fitness function defined for the algorithm. Selection of best or fittest individuals for reproduction is made and they are kept in a mate pool after using Crossover or mating and Mutation are applied. As result new population is generated. The process is re applied to new generation again and again until stopping criteria is reached [17] [22] [23].

C. Fitness Measurement

The fitness calculation of the individuals for the algorithm is based on the devotion of frequency of characters in the English language. The frequency of a character in the text is used for attacks against cryptographic cipher. For this the deviation of frequency of occurrence of a particular character in the decrypted text corresponding to the frequency of occurrence in Standard English language (U.K) is calculated for each candidate solution or key or

individual. The lowest deviation represents the highest fitness. For example particular set of keys (let K1, K2, K3,…)is applied to decrypt the sample encryption E, the decrypted text is analyzed and the frequency of each character of the English language in decrypted text is recorded. After that the difference of frequencies of standard English text and decrypted text is calculated for each letter and the addition of all differences defines the fitness. If the difference is low that shows high fitness because that make it near to English language text. The frequencies of letters typically occurring in English languages are well known and documented [12] [15] [16]. Figure 3 shows the occurrences of letters in English language text of length 10000 characters. The formula for calculating the fitness is as below [1][3][4][5] [7].

$$\text{Fitness} = 1 - \alpha \times \sum \forall |\text{SDM i} - \text{OFM i}| + \beta \times \sum \forall |\text{SDB i} - \text{OFB i}| \quad \ldots(1)$$

The formula above is used when a set of monograms and bigrams and their frequencies are known, but for research purpose, it is found that the monogram frequency is sufficient so the formula used in research is as below.

$$\text{Fitness} = 1 - \sum \quad | \quad - \quad | \ldots \ldots \ldots\ldots\ldots (2)$$

Where:

SDM = Standard character frequency in English language text of ith character (monogram).

SDB = Standard character frequency in English language text of ith bigram.

OFM= Observed character frequency in decrypted text of ith character (monogram).

OFB = Observed character frequency in decrypted text of ith bigram.

α and β are two constants shows the ratio of monogram and bigram respectively in a standard English language text.

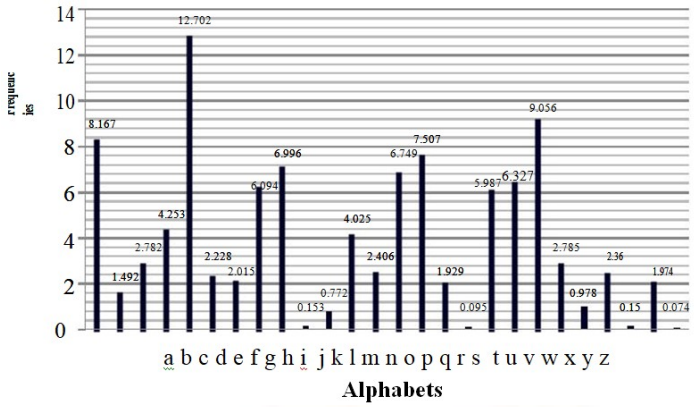

Fig. 3: Character frequencies in English language text [1].

VI. EXPERIMENTAL SETUP

To analyze the various aspects of proposed Genetic Algorithm for cryptanalysis of Mono-alphabetic substitution cipher, DOS Box-0.74 IDE is used as the development environment. In the development environment, the programming language C is used for coding. Turbo C-3.0 is used as compiler. The IDE and the compiler are setup on a machine which is configured with Intel Pentium(R) Dual Core CPU T6500 @ 2.10 GHz processor and 2 GBDDR3 RAM, and installed with Windows 7 operating system. With respect to above setup the input and testing parameters are as follows

A. Experimental input parameters

Following input parameters are used to get the experimental results.

1. Sample text: For implementation of experiment following sample text was used for analysis of the results as

"When, in the course of human events, it becomes necessary for one people w dissolve the political bands which have connected them with another, and to assume among the Powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation. We hold these truths to be self-evident: that all men are created equal, that they are endowed by their Creator with certain unalienable rights, that among these are Life, Liberty, and the pursuit of Happiness."

The space and punctuation marks are eliminated manually at the time of execution [6].

2. Four combinations of original key for encryption: four different combination of keys to encrypt the sample text for analysis of various parameters of sample texts were used which are:

i. Key1 -5mod 27: "ejotychmrwafkpuzdinsxbglqv"

ii. Key2A-N to M-Z: "zyxwvutsrqponmlkjihgfedcba"

iii. Key3 Random: "lprfvdanyzboqchejuikmtswx"

iv. Key4 A-Z/Z-A: "nopqrstuvwxyzabcdefghijklm"

3. Character frequency statistic of English language: For calculating the fitness function the character frequencies are used as input parameters shown in figure 3.

4. Basic genetic operator's definitions.

B. Experimental testing parameters

1. Time Constraints: For a particular key (by which sample text is encrypted) how much time is taken to find that key by the algorithm

2. Initial population size: Number of keys guessed for initialize the process or the initial solution space size.

3. Fitness level: Level of fitness achieved by the algorithm or the maximum fitness achieved by the algorithm

4. Character matched: Number of characters matched with sample text by the final solution

5. Sample text size: the size of the given text on which algorithm is applied after encrypting.

Above five parameters decide that which algorithm performed well on that basis the research conclude about the genetic operator combination.

VII. EXPERIMENTAL RESULT AND ANALYSIS

The graphs show that GA1 performs well as it takes good start and the maximum fitness of the keys increasing as the number of generation increases to threshold extent.

GA1, the combination of Random Selection operator with 1-point Crossover and Interchange Mutation with Elitism, here the elitism guaranties that the fittest key of the previous generation will definitely go to meting pool for reproduction and number of times means the frequency of the fittest key in mate pool is grater so while the crossover and mutation are applied the fitness increase that make it smarter and successful.

While the graph of GA2 is not increasing all the time, sometimes it is decreasing that because that the mate pool selection is using the Roulette wheel selection which is pure random, but due to that is biased towards good fitness so some time its fitness increase, it neither guarantees that the best individual will go to mate pool nor the worst individual will not go to mate pool because of its randomness this algorithm did not perform well as compared to the GA1.

In GA3 the combination of genetic operators used is Tournament Selection with uniform crossover and Interchange Mutation, here the selection strategy is responsible for the performance because the tournament selection guarantees that the worst individuals will never go to mate pool for reproduction. The tournament selection is like the sport tournament some individual fights for points who have the points above cut off will go to the next round and this process repeated in the next round so finally the best and a runner up will enter into final that increase the success rate and reliability more over it starts slow but reaches to a very good statistic. In the graph, initially it starts moderate because as the computation starts all will be at the same place but as the tournament goes on fire the algorithm starts performing well. Graph for showing the impact of population size as the population size increases the speed and the increase and the fitness is also increase the red line shows the population size as 40 where blue line shows the population size as 20. The start of the red line is better and goes constant towards the final fitness value while the start of the blue line is lesser then red line, but at the end it goes near to the red line which shows if the population size is bigger

then the start of the algorithm will be good. The above scenario applies to all four keys with little defalcation as the randomness of key increases the starting trouble for all three algorithms as Key4 is randomly generated with random substitutions the graph shows the start of all three algorithms is little disturbed as a curve and crusts for a while for Key1 there is a particular pattern for substitution that shows the better and a constant graph line.

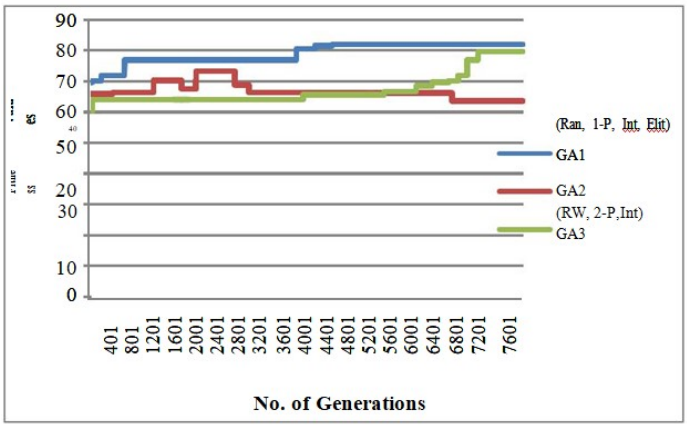

Fig. 4: Graph of GA1, GA2, and GA3 between fitness and number of generations for key1.

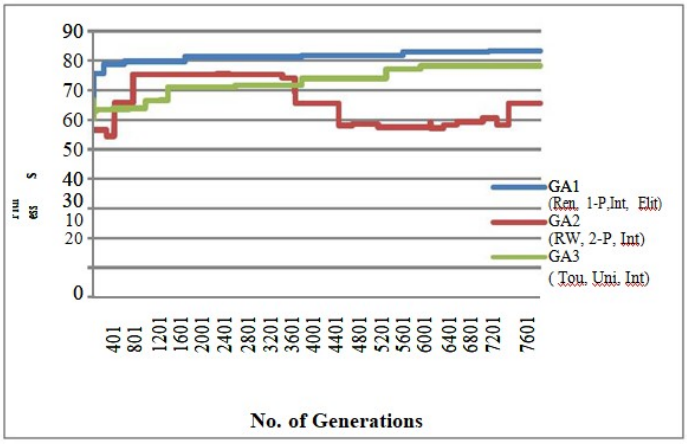

Fig. 5: Graph of GA1, GA2, and GA3 between fitness and number of generations for key2.

VIII. CONCLUSION

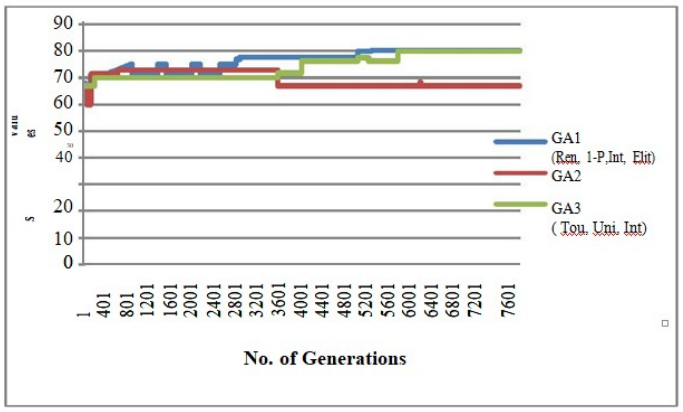

Fig. 6: Graph of GA1, GA2, and GA3 between fitness and number of generations for key3.

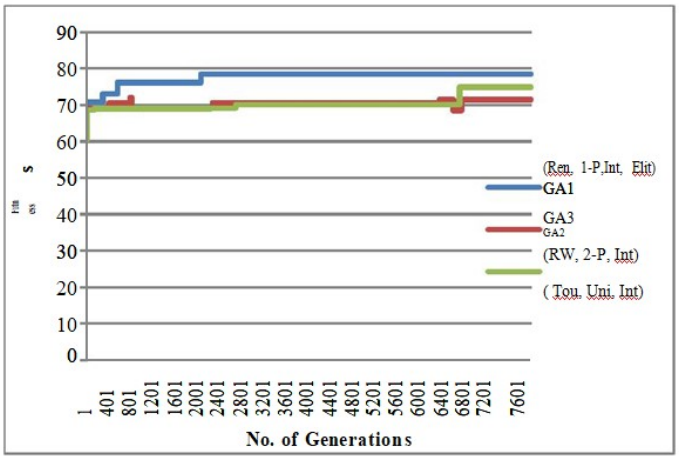

Fig. 7:Graph of GA1, GA2, and GA3 between fitness and number of generations for key4.

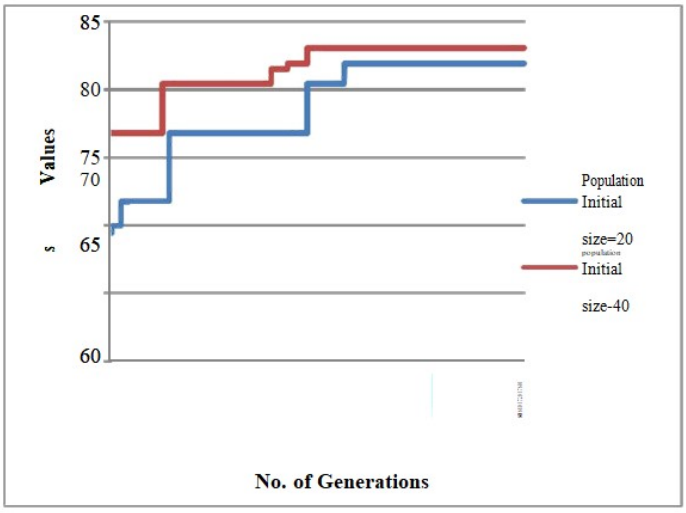

Fig. 8: Graph of GA1 between fitness and number of generations for different Population Size

After studying and analyzing the experimental results the conclusion of the research came to the point that the genetic algorithm GA1 combination of random selection with elitism along with 1-point cross over and interchange mutation performs well as it converges to fittest key faster than the GA3 and GA2 where the GA3 is combination of tournament selection along with uniform crossover and interchange mutation performed good but not as GA1 it starts slow and the fitness level of key achieved by GA2 is lesser than the fitness level of key achieved by GA1.

In case of GA2 which is the combination of roulette wheel selection, along with 2-point crossover and interchange mutation do not perform very stable manner because of the randomness of the roulette wheel selection some time it crosses the fitness level of key achieved by GA3 but finally the overall performance of GA2 does not reach to a satisfactory level.

Finally, it can be concluded that for this research context the GA1 is a winner and GA3 is runner up. Shows the cryptanalysis of Mono-alphabetic substitution cipher technique using GA1 is faster and converge towards correctness while using GA3 it became little slower than GA1 and consistent towards correctness. In case of GA2 the randomness makes it inconsistent and sometimes fast convergence towards correctness sometimes towards the opposite.

REFERENCES

[1] S. S. Omran, A. S. Al-Khalid, and D. M. Al-Saady, “Using Genetic Algorithm To Break A Mono - Alphabetic Substitution Cipher”, IEEE Conference on Open Systems (ICOS -2010) , Kuala Lumpur, Malaysia December 5-7, 2010.

[2] Ali.S.Al-Khalid and Alaa.O.Al-Khfagi, “Cryptanalysis of a Hill Cipher using Genetic algorithm”, IEEE, 2015.

[3] AditiBhateja and Shailender Kumar, “Genetic Algorithm with Elitism for Cryptanalysis of Vigenere Cipher”, IEEE, 2014.

[4] S. S. Omran, A. S. Al-Khalid, and D. M. Al-Saady, “A Cryptanalytic Attack on Vigenère Cipher Using Genetic Algorithm”, IEEE Conference on Open Systems (ICOS -2010), Kuala Lumpur, Malaysia, September 25 - 28, 2011.

[5] JitinLuthra and Saibal K. Pal, “A Hybrid Firefly Algorithm using Genetic Operators for the Cryptanalysis of a Monoalphabetic Substitution Cipher”, IEEE, 2011.

[6] Feng-Tse Lin and Cheng-Yan Kao, “A Genetic Algorithm for Cipher-text-Only Attack in Cryptanalysis”, IEEE, 1995.

[7] Eng. Ayman M B. Albassal, and Prof. Dr. Abdel-Moneim A. Wahdan, “Genetic Algorithm Cryptanalysis of the basic Substitution Permutation Network”, IEEE, 2004.

[8] Joseph Alexander Brown, Sheridan Houghten, and Beatrice Ombuki-Berman, “Genetic Algorithm Cryptanalysis of a Substitution Permutation Network”, IEEE, 2008.

[9] Mohammad Faisal Uddin and Amr M. Youssef, “Cryptanalysis of Simple Substitution Ciphers Using Particle Swarm Optimization”, IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.

[10] Jun Song, Huanguo Zhang, QingshuMeng, and Zhangyi Wang, “Cryptanalysis of Four-Round DES Based on Genetic Algorithm”, IEEE, 2007.

[11] Salabat Khan, WaseemShahzad, and FarrukhAslam Khan, “Cryptanalysis of Four-Rounded DES using Ant Colony Optimization”, IEEE, 2010.

[12] XuDewu and Chen Wei, “A Survey on Cryptanalysis of Block Ciphers”, International Conference on Computer Application and System Modeling (ICCASM 2010), 2010.

[13] Nalini N and RaghavendraRao G, “Cryptanalysis of Block Ciphers via Improved Simulated Annealing Technique”, 9th International Conference on Information Technology (ICIT'06), IEEE, 2006.

[14] Raphael C.-W. Phan and Mohammad Umar Siddiqi, “A Framework for Describing Block Cipher Cryptanalysis”, IEEE Transactions on Computers, Vol. 55, No. 11, November 2006.

[15] Eng. Ayman M. B. Albassal and Dr. Abdel-Moneim A. Wahdan, “Genetic Algorithm Cryptanalysis of a Feistel Type Block Cipher”, IEEE, 2004.

[16] Tao Li1,Jiguo Li, and Jing Zhang, “A Cryptanalysis Method based on Niche Genetic

Algorithm", Natural Sciences Publishing Corporation, 2014.

[17] Song Y Yan, "Computability, Learnability and Breakability in Cryptanalysis", CIMCA 2008, IAWTIC 2008, and ISE 2008.

[18] Christophe De Canniere, Alex Biryukov, and Bart Preneel, "An Introduction to Block Cipher Cryptanalysis", Proceedings of the IEEE, Vol. 94, No. 2, FEBRUARY 2006.

[19] O P Verma, RituAgarwal, DhirajDafouti, and ShobhaTyagi, "Peformance Analysis of Data Encryption Algorithms", IEEE, 2011.

[20] Imad F.T. Yaseen and H.V. Sahasrabuddhe, "A Genetic Algorithm for the cryptanalysis of Chor-Rivest knapsack Public Key Cryptosystem (PKC)", 2000.

[21] TaharMekhaznia and AbdelmadjidZidanien, "Genetic algorithm for attack of image encryption scheme based chaotic map", IEEE, 2016.

[22] Ho Yean Li, AzmanSamsudin, and BahariBelaton, "Heuristic Cryptanalysis of Classical and Modern Ciphers", IEEE, 2005.

[23] Andrew Clark, "Modern Optimisation Algorithms for Cryptanalysis", IEEE, 1994.

[24] RichaGarg and Saurabh Mittal, "Optimization by Genetic Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014.

[25] MehrdadDianati, Insop Song, and Mark Treiber, "An Introduction to Genetic Algorithms and Evolution Strategies", University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada, 2004.

[26] Anita Thengade and RuchaDondal, "Genetic Algorithm – Survey Paper", MPGI National Multi Conference, 2012.

[27] Mitchell and Melanie "An Introduction to Genetic Algorithms", MIT Press, 1996.

[28] D. E. Goldberg. "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison-Wesley, 1989.