

Arabic Text Cryptanalysis Using Genetic Algorithm

Rokaia Shalal Habeeb

Computer Engineering Department
Faculty of Engineering
Mustansiriyah University
Baghdad, Iraq
rokaia_shalal@yahoo.com

Abstract In this paper a Genetic Algorithm (GA) is proposed to attack an Arabic encrypted text by Vigenere cipher. The frequency of occurrence of Arabic letters has been calculated by using the text of the holy book of Quran, since it has rich language features compared to many other books. The algorithm is tested to find the key letters for different ciphertext sizes and key lengths. The results shows 100% correct letters retrieved from medium size ciphertext and short key length, while 90% of the ciphertext is retrieved from long ciphertext and medium key length, and 82% of the ciphertext is retrieved from long ciphertext and long key.

Keywords: Arabic, Cryptanalysis, Genetic Algorithm, Vigenere cipher,

I. INTRODUCTION

Intelligent systems show high capability in solving nonlinear multivariable problem. One of these problems is cryptanalysis. In recent years Genetic Algorithms shows efficient results in search space of large, complex and dynamic systems. Many researchers in the field of cryptanalysis are interested in developing automated attacks on ciphertext. Using of GA in cryptanalysis has attracted much attention [1]. Creed F. Jones and Michael Christman use GA to develop solutions to a Vigenere alphabetic code for English language in [2], the authors in this paper suggest a fitness function that based on word look-up dictionary of most known English words. This can help the GA to converge quickly, however, if the original message contains a word which is not included in the dictionary, an incorrect solution case could appear. Ragheb Toemeh and Subbanagounder Arumugam apply GA for cryptanalysis of Vigenere cipher for English language [3], they used GA to guess the size of the key, then find the correct key. Although, using traditional methods such as Kasiski examination to find the key size [4], is less complicated than using GA. S. S. Omran et al. [5] use counting coincidence technique to find the key length, then apply the GA to break the Vigenere cipher for English language. The fitness

function based on monogram frequencies only, however, in the results, almost all the key letters are successfully retrieved, but, the ciphertext size is not mentioned. Aditi Bhateja and Shailender Kumar [6] introduce a method of deciphering encrypted messages of Vigenere cipher cryptosystems by GA using elitism strategy for English language, their results show that for large key sizes elitism increases the performance of GA. Yahya Alqahtani et al. [7] introduce a new approach of Arabic encryption/decryption technique using Vigenere cipher on modulus 39. Shaza D. Rehan and Saif Eldin F.Osma propose a cryptography technique for Arabic language using a genetically tuned neural network [8]. The authors in [9-11] introduce Arabic language encryption technique using symmetric key algorithm. In this paper a GA is used to find the key for an Arabic ciphertext encrypted using Vigenere cipher. The proposed method assumes that the key length is known. The rest of the paper is organised as follow:

Section two describes a theoretical background of the Vigenere Cipher technique. GA is illustrated in section three. Section four explains the frequency analysis of Arabic letters. The proposed algorithm is presented in section five. The experimental results are introduced in section six, while final conclusions are addressed in section seven.

ciphers. Al-Kindi took a text of 3667 letters to find out the frequency of occurrence of Arabic letters and he come up with Table III.

TABLE III
ALKINDI’S WORK ON THE ORDER OF LETTER
FREQUENCY [15]

Letter	Order	Frequency of occurrence	Percentage of Occurrence (%)
أ	1	600	16.36
ل	2	*437	11.91
م	3	320	8.72
ه	4	273	7.44
و	5	262	7.14
ي	6	*252	6.87
ن	7	221	6.02
ر	8	155	4.22
ع	9	131	3.57
فا	10	122	3.32
نا	11	120	3.27
با	12	112	3.05
كا	13	112	3.05
د	14	92	2.5
س	15	91	2.48
ق	16	63	1.71
ح	17	57	1.55
جم	18	46	1.25
ذ	19	35	0.95
ص	20	32	0.87
ش	21	*23	0.63
ظ	22	*20	0.55
نح	23	20	0.55
ثا	24	17	0.46
ز	25	*16	0.44
ط	26	15	0.41
غ	27	15	0.41
نظ	28	8	0.22
		3667	100

* These numbers have been corrected according to quotations made by Ibn Dunaynir and Ibn Adlan of Al-Kindi, who work on the text after Al-Kindi.

V. PROPOSED ALGORITHM

The frequency of occurrence of Arabic letters has been calculated by using the text of the holy book of Quran, since it has rich language features compared to many other books. In this paper the Quran text is downloaded from www.tanzil.net the text size of the Quran is 412184 characters without spacing. Table IV shows the calculated frequencies of Arabic letters.

TABLE IV
ARABIC LETTERS FREQUENCIES

Order	Letter	Frequency of Occurrence	Percentage of Occurrence (%)
1	ء	1954	0.47
2	آ	1871	0.45
3	أ	11293	2.74
4	ؤ	833	0.20
5	إ	6326	1.53
6	ئ	1464	0.36
7	ا	54339	13.18
8	ب	14369	3.49
9	ة	2903	0.70
10	ت	13028	3.16
11	ث	1751	0.42
12	ج	4108	1.00
13	ح	5404	1.31
14	خ	3092	0.75
15	د	7419	1.80
16	ذ	6108	1.48
17	ر	15637	3.79
18	ز	1980	0.48
19	س	7584	1.84
20	ش	2630	0.64
21	ص	2566	0.62
22	ض	2088	0.51
23	ط	1576	0.38
24	ظ	1056	0.26
25	ع	11647	2.83
26	غ	1512	0.37
27	ف	10832	2.63
28	ق	8711	2.11
29	ك	13000	3.15
30	ل	47851	11.61
31	م	33525	8.13
32	ن	33910	8.23
33	ه	18529	4.50
34	و	30728	7.46
35	ى	3210	0.78
36	ي	27350	6.64

Fig. 1 and Fig. 2 shows the calculated frequency of occurrence of Arabic letters that are used in the experimental tests.

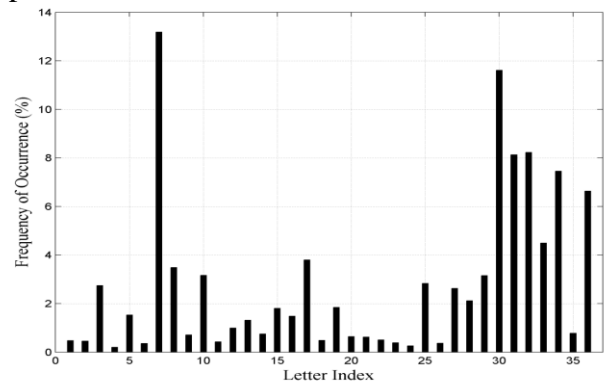


Fig. 1 Frequency of occurrence of Arabic letters.

TABLE VI
PLAINTEXT

Size (no spacing)	Plaintext
400	ملحمة جلجامش اللوح الأول هو الذي رأى كل شيء فغني بذكره يا بلادي وهو الذي خبر جميع الأشياء وافاد من عبرها وهو الحكيم العارف بكل شيء لقد ابصر الأسرار وعرف الخفايا المكتومة وجاء بأنباء الأيام مما قبل الطوفان لقد أوغل في الأسفار البعيدة حتى حلَّ به الضنى والتعب ففقد في نصب من الحجر كل ما عاناه وما خيره بنى اسوار (أوروک) وحرّم (أي – أنا) المقدس والمستودع الطاهر فأنظر إلى سوره الخارجي تجد شرفاته تتألق كالنحاس وأنعم النظر في سوره الداخلي الذي لا يمائله شيء واستلم أسكفته الحجرية الموجودة منذ ال
600	ملحمة جلجامش اللوح الأول هو الذي رأى كل شيء فغني بذكره يا بلادي وهو الذي خبر جميع الأشياء وافاد من عبرها وهو الحكيم العارف بكل شيء لقد ابصر الأسرار وعرف الخفايا المكتومة وجاء بأنباء الأيام مما قبل الطوفان لقد أوغل في الأسفار البعيدة حتى حلَّ به الضنى والتعب ففقد في نصب من الحجر كل ما عاناه وما خيره بنى اسوار (أوروک) وحرّم (أي – أنا) المقدس والمستودع الطاهر فأنظر إلى سوره الخارجي تجد شرفاته تتألق كالنحاس وأنعم النظر في سوره الداخلي الذي لا يمائله شيء واستلم أسكفته الحجرية الموجودة منذ القدم أقرب من (أي أنا) مسكن عشتار الذي لا يمائله صنع ملك من الآتين ولا إنسان أعل فوق أسوار (أوروک) وامش عليها تفحص أسس قواعدها وأجر بناؤها وتيقن أليس بناؤها بالأجر المفخور؟ وهلا وضع الحكماء السبعة أسسها بعد ان خلق جلجامش واحسن الآله العظيم خلقه ح
1000	ملحمة جلجامش اللوح الأول هو الذي رأى كل شيء فغني بذكره يا بلادي وهو الذي خبر جميع الأشياء وافاد من عبرها وهو الحكيم العارف بكل شيء لقد ابصر الأسرار وعرف الخفايا المكتومة وجاء بأنباء الأيام مما قبل الطوفان لقد أوغل في الأسفار البعيدة حتى حلَّ به الضنى والتعب ففقد في نصب من الحجر كل ما عاناه وما خيره بنى اسوار (أوروک) وحرّم (أي – أنا) المقدس والمستودع الطاهر فأنظر إلى سوره الخارجي تجد شرفاته تتألق كالنحاس وأنعم النظر في سوره الداخلي الذي لا يمائله شيء واستلم أسكفته الحجرية الموجودة منذ القدم أقرب من (أي أنا) مسكن عشتار الذي لا يمائله صنع ملك من الآتين ولا إنسان أعل فوق أسوار (أوروک) وامش عليها تفحص أسس قواعدها وأجر بناؤها وتيقن أليس بناؤها بالأجر المفخور؟ وهلا وضع الحكماء السبعة أسسها بعد ان خلق جلجامش واحسن الآله العظيم خلقه حياه شمس السماوي بالحسن وخصه اد بالبطولة جعل الآلهة العظام صورة جلجامش تامة كاملة كان طوله احد عشر ذراعا وعرض صدره تسعة أشبار ثلثان منه إله، وثلثه الآخر بشر وهيئة جسمه لا نظير لها وفتك سلاحه لا يصده شيء وعلى ضربات الطبل تستيقظ رعيته لازم أبطال أوروک حجراتهم متذمرين شاكين لم يترك جلجامش ابنا لأبيه ولم تنقطع مظالمه عن الناس ليل نهار ولكن جلجامش هو راعي أوروک السور والحمى أنه راعينا قوي وجميل وحكيم لم يترك جلجامش عذراء لحبيبها ولا ابنة المقاتل ولا خطيبة البطل وأخيراً سمع الآلهة شكواهم فأ

By applying the proposed algorithm for 10 runs, the following results are obtained.

TABLE VII
RESULTS FOR PROPOSED ALGORITHM

Ciphert ext size	Key length	Min. correct letters	Max. correct letters	Average correct letters	Average correct Letters (%)	Average Time (sec.)
400	5	4	4	4	80	110.6
	10	8	9	8.1	81	110.1
	20	10	14	12.7	63.5	110.7
600	5	5	5	5	100	163.9
	10	9	9	9	90	164.7
	20	12	16	14.1	70.5	163.4
1000	5	5	5	5	100	270.5
	10	9	9	9	90	273.3
	20	14	18	16.4	82	284.4

It is noticed from the results above that better results are obtained when longer ciphertext and shorter key length are considered. The Arabic language as any other language contains some letters that have the same n-gram characteristics as shown in Fig. 1 and Fig. 2. The similarity between the n-gram characteristics of the original plaintext and the n-gram characteristics of the Holy Quran text has a crucial impact on the fitness function calculation. Practically, the plaintext has some difference in its n-grams characteristics due to the short length compared to the Holy Quran text. This may result in a target fitness function, which does not necessarily of the highest value, and it is probable that other infeasible key solutions may produce the highest fitness value.

It is observed that the fitness function in some runs converges to a value which is higher than the one of the correct key. This means that the GA is finding a key combination that produces larger fitness value than the target fitness value. To reduce the effect of this phenomenon, higher order of n-grams can be included in the fitness function model.

Until the preparation of this paper, no research is found in the literature, which consider the same case study in this paper. However, comparative results can be extracted from [6] as shown in Table VIII. The comparison shows that the results of the proposed algorithm is comparative to the results in [6], taking in consideration the difference in the language which refers to different number of letters. The algorithm presented in [6] considers different GA parameters such as the elitism percentage, which is 10%, this is as twice as the percentage used in

this paper. The number of generations is not specified in [6].

TABLE VII
COMPARISON RESULTS

Cipher text size	Key length	Proposed Algorithm		Algorithm presented in [6]	
		Average correct letters	Average correct Letters (%)	Average correct letters	Average correct Letters (%)
400	5	4	80	4.6	92
600	5	5	100	4.7	94
	10	9	90	9.2	92
	20	14.1	70.5	16.8	84

VII. CONCLUSION

In this paper a cryptanalysis method is suggested to attack Arabic ciphertext. The proposed method uses GA to search the key space for the correct encryption key. The ciphertext considered to be encrypted using Vigenere cipher with known key length. Different sizes of ciphertext (400, 600, and 1000 letters) with different key lengths (5, 10, and 20 letters) are investigated. The proposed algorithm can decrypt 100% of the ciphertext when a key length of 5 letters is used to decrypt a ciphertext of 600 and 1000 letters.

REFERENCES

[1] J. Song, H. Zhang, Q. Meng, and Z. Wang, "Cryptanalysis of Four-Round DES Based on Genetic Algorithm," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 2007, pp. 2326-2329.

[2] C. F. Jones and M. Christman, "Genetic algorithm solution of Vigenere alphabetic codes," in *Soft Computing in Industrial Applications, 2001. SMCia/01. Proceedings of the 2001 IEEE Mountain Workshop on*, 2001, pp. 59-63.

[3] R. Toemeh and S. Arumugam, "Applying Genetic Algorithms for Searching Key-Space of Polyalphabetic Substitution Ciphers," *International Arab Journal of Information Technology*, vol. 5, pp. 87-91, 2008.

[4] D. R. Stinson, *Cryptography: Theory and Practice, Third Edition*: Taylor & Francis, 2005.

[5] S. S. Omran, A. S. Al-Khalid, and D. M. Al-Saady, "A cryptanalytic attack on Vigenere cipher using genetic algorithm," in *Open Systems (ICOS), 2011 IEEE Conference on*, 2011, pp. 59-64.

[6] A. Bhateja and S. Kumar, "Genetic Algorithm with elitism for cryptanalysis of Vigenere cipher," in *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on*, 2014, pp. 373-377.

[7] Y. Alqahtani, P. Kuppaswamy, and S. Shah, "New Approach of Arabic Encryption-Decryption Technique Using Vigenere Cipher on Mod 39," *International Journal of Advanced Research in IT and Engineering*, vol. 2, 2013.

[8] S. D. Rihan and S. E. F.Osma, "ARABIC Cryptography Technique Using Neural Network and Genetic Algorithm," *International Research Journal of Computer Science*, vol. 3, pp. 35-42, 2016.

[9] P. Kuppaswamy and Y. Alqahtani, "New Innovation of Arabic Language Encryption Technique Using New Symmetric Key Algorithm," *International Journal of Advances in Engineering and Technology*, vol. 7, pp. 30-37, 2014.

[10] M. A. M. Aysan and P. Kuppaswamy, "Hybrid Combination Of Message Encryption Techniques On Arabic Text Using New Symmetric Key And Simple Logarithm Function," *International Journal of Scientific Knowledge (Computing and Information Technology)*, vol. 5, pp. 37-41, 2014.

[11] H. A.-Z. Atee, "Development of A New Way To Encrypt The Arabic Language Letters Using The Symmetric Encryption System," *AL-TAQANI*, vol. 24, pp. 101-111, 2011.

[12] W. Stallings, *Cryptography and Network Security: Principles and Practice*: Pearson/Prentice Hall, 2006.

[13] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*: Addison-Wesley Publishing Company, 1989.

[14] S. N. Sivanandam and S. N. Deepa, *Introduction to Genetic Algorithms*: Springer Berlin Heidelberg, 2007.

[15] M. Mrayati, Y. M. Alam, and M. H. At-Tayyan, *Series on Arabic Origins of Cryptology: al-Kindi's treatise on cryptanalysis*: KFCRIS & KACST, 2003.

[16] T. Baqir, *The Epic of Gilgamesh 1962*, ملحمة كلكامش.