A DIPLOID GENETIC ALGORITHM WITH EXOGENOUS RECOMBINATION FOR BREAKING MERKLE-HELLMAN KNAPSACK

Sarab M. Hammed, *Mayafda F. AbdulHalim, Bara’a A. Attea
Department of Computer Science, College of Science, University of Baghdad. Baghdad- Iraq
* University of Bahrain

Abstract

Cryptanalysis is the science and study of methods of breaking cryptographic techniques. Cryptanalytic attack on Merkle-Hellman knapsack using Genetic Algorithm (GA) was done by Spillman. Then Garg et al improved the Spillman GA. The objective of this paper is to harness the power of GA used by Garg et al. to get more reliable results and in less computation time. This paper utilizes a GA**,** known as diploid GA with *exogenous* recombination scheme for translating each number in ciphertext into the correct ASCII code for the plaintext characters. Our results are compared with Garg et al. result and they proved that diploid GA with exogenous recombination scheme is more efficient and highly successful in finding the correct bit pattern for the hard knapsack sum.

الخوارزمية الجينية الثنائية مع اعادة تجميع الابعاد لكسر حقيبة ميركل هيلمان

سراب مجيد حميد، *ميادة فيصل عبد الحليم، براء علي عطية
قسم علوم الحاسبات، كلية العلوم، جامعة بغداد. بغداد – العراق.
* جامعة البحرين.

الخلاصة

تحليل الشفرة هو علم ودراسة طرق كسر الشفرة. تحليل حقيبة ميركل هيلمان باستخدام الخوارزمية الجينية طبقت من قبل سبلمان. كارج واخرون حسنوا الخوارزمية الجينية لسبلمان. هذا البحث يساهم في تقوية الخوارزمية الجينية المستخدمة من قبل كارج للحصول على نتائج ذات موثوقية عالية وبوقت قليل. استخدمت الخوارزمية الجينية الثنائية مع اعادة تجميع الابعاد لتحويل كل رقم في النص الشفر الى رمز ASCII للحرف في النص الصريح. قورنت نتائجنا مع نتائج كارج واخرون. اثبت النتائج ان استخدم الخوارزمية الجينية الثنائية مع اعادة تجميع الابعاد ناجحة في ايجاد البت الصحيح لمجموع الحقيبة الصعبة وذات كفاءة عالية.

1. Introduction

With the exponential growth of networked system and application such as eCommerce, the demand for effective internet security is increasing. Cryptology is the science and study of systems for secret communication. It consists of two complementary fields of study cryptography and cryptanalysis.

Cryptography is the science and the study of secret writing and it is practitioner by cryptographer. Cryptanalysts are practitioner of cryptanalysis, the science and study of methods of breaking ciphertext. In other words it can be described as the process of searching for flaws or oversights in the design of ciphers [1] [2].

One of first knapsack cipher was proposed by Merkle and Hellman in 1978 which utilized a NP-complete problem for its security [3]. This cryptosystem belongs to major categories of public/private key cryptosystem. The public/private key aspect of this approach lies in the fact that there are actually two different knapsack problems referred to as the easy Knapsack and hard knapsack. The Markle-Hellman algorithm is based on this property. The hard knapsack becomes the public key and the easy knapsack is the private key. The public key can be used to encrypt the message but can not be used to. Spillman presented a cryptanalytic attack on Merkle-Hellman knapsack using genetic algorithm in [4]

This paper is looking for a new solution that further improves the robustness of cryptanalytic attack with high effectiveness. The idea is based on utilizing diploid GA with *exogenous* recombination scheme to recover the plaintext from ciphertext without knowing the key. The paper objectives are to satisfy: reliability, efficiency, and implementation simplicity.

The remained of this paper is organized in four sections. Section 2 illustrates the basic concepts behind Merkle-Hellman knapsack. Section 3 discusses the core of this paper: how to turn the mechanism of diploid GA with exogenous recombination scheme into profitable account to attack Merkle-Hellman knapsack cipher. Section 4 presents our results and finally, section 5 concludes this work.

2. Merkle-Hellman Knapsacks

The first practical public-key cryptosystem was invented by Merkle and Hellman, soon after the basic principles of public-key cryptography had been stated by Diffie and Hellman. Merkle and Hellman algorithm based on *knapsack problem* [5] [6].

The Merkle-Hellman knapsack cipher attempts to disguise an easily solved instance of the subset sum problem, called a superincreasing subset sub problem, by modular multiplication and a permutation [7]. A superincreasing sequence is a sequence $b_1, b_2, ..., b_n$ of positive integers with the property that $b_i > \sum_{j=1}^{i-1} b_j$ for each i, $2 \le i \le n$.

The integer *n* is a common system parameter. $b_1, b_2, ..., b_n$ is the superincreasing sequence and u is the modulus, selected such that $u > b_1 + b_2 + ... + b_n$. w is a random integer, such that $1 \le w \le u - 1$ and $\gcd(w, u) = 1$. The public key of *A* is $a_1, a_2, ..., a_n$ where $a_i = wb_i \bmod u$ and the private key of A is $(u, w, b_1, b_2, ..., b_n)$.

The steps in the transmission of a message *m* are as follows [7]:

In order to encrypt *m*, the sender should do the following:

- Obtain the receiver authentic public key $a_1, a_2, ..., a_n$
- Represent the message *m* as a binary string of length n, $m = m_1, m_2, ..., m_n$
- Compute the integer $c = m_1 a_1 + m_2 a_2 + ... + m_n a_n$
- Send the ciphertext c to the receiver.
- On the other hand, in decryption process (to recover the plaintext *m* from *c*), the receiver should do the following:
- Compute $d = w^{-1} c \bmod u$
- By solving a superincreasing subset sum problem, find the integers $r_1, r_2, ..., r_n$, $r_i \in \{0,1\}$ such that $d = r_1 b_1 + r_2 b_2 + ... + r_n b_n$
- The message bits are $m_i = r_i, i = 1,2,...n.$

3. The Cryptanalytic Algorithm

This section presented how diploid GA with exogenous recombination scheme is utilized to attack Merkle-Hellman knapsacks. Cryptanalyst has ciphertext only and want to recover plaintext from it. The ciphertext forms as integer number that represents target sum of a hard knapsack problem.

A diploid GA algorithm was proposed by Goldberg and Simith [8] [9]. The three main characteristics of this GA that distinguish it from the simple, so called haploid, GA are:

Each individual is represented as a pair of homologous chromosomes instead of only one single chromosome.

The existence of dominance operator for eliminating any conflict between the homologous chromosomes.

Crossover is accomplished via two stages (gametogenesis and fertilization) instead of single stage in haploid GA.

Then, in [10] Mayada proposed another diploid recombination scheme, so-called exogenous recombination scheme.

The primary idea of diploid GA with exogenous recombination is as follows. A given population consists of diploid parents is created. Then to produce offspring for the next generation, selection is applied to struggle among parent individuals in a way that an individual with more favorable properties produces on average more offspring than a one less adapted to the environment. After that, a process of combing genetic information through exogenous recombination idea is revealed. In this recombination process, two diploid parents are drawn from the mating pool and recombined in their traits in a form of pair against pair and mate against mate in the sense as a synonym for the type of the substance of inheritance. The exogenous recombination is continued to other parents in the mating pool until a full population of new offspring is created. Additional source of variation is added to the offspring through mutation and dominance shift to perverse the diversity required in the population [8].

3.1individual Representation and Initialization

Diploid GA work on a population of individuals. The population represents a set of candidate solutions to the problem at hand. The individual is triallelic and its values are 0, 1, and -1. The number of bits in each individual is equal to the number of elements key. Here, the individual length is equal to eight. Figure 1 depicts an example of individual representation.

1	-1	0	1	1	1	-1	0
1	0	0	0	1	1	-1	-1

Figure 1: Individual Representation

At the beginning of the diploid cycle, a random number of individuals is created, the size of which is equal to pop_{size}.

3.2 Fitness Evaluation

To evaluate the individual, first we apply Hollstien rule to control dominance between homologous genes. It say that both 1 and -1 map to 1, but 1 dominates 0 and 0 dominates -1. Figure 2 depicts this dominance scheme. With this scheme the more effective allele becomes dominant, thereby shielding the recessive [9].

	0	-1	1
0	0	0	1
-1	0	1	1
1	1	1	1

Figure 2: Triallelic Dominance Map

The expressed value of figure 1will be as figure 3. Then, the expressed value is evaluated using fitness function given by equation (5) [4].

1	0	0	1	1	1	1	0

Figure 3: Expressed Value

Let $M = \{m_1, m_2, ..., m_8\} \in \{0,1\}$ be an arbitrary solution and the public key $A = \{a_1, a_2, ..., a_8\}$, then:

$$Sum = \sum_{j=1}^{8} a_j m_j \qquad (2)$$

$$Target = \sum_{j} a_j \qquad (3)$$

$$FullSum = \sum_{j=1}^{8} a_j \qquad (4)$$

$$Maxdiff = \max\{Target, FullSum\}$$

$$Individual\ Fitness = \begin{cases} 1 - \left(\dfrac{|Target - Sum|}{T\arg et} \right)^{1/2} & if\ Sum \le Target \\ 1 - \left(\dfrac{|Target - Sum|}{MaxDiff} \right)^{1/6} & if \quad Sum > Target \end{cases} \qquad (5)$$

3.3 Diploid GA Operators

After defining the representation of individual solutions and computing the fitness function for each individual, the diploid GA operators are applied to get better solutions. These operators are selection, exogenous recombination, and mutation.

Given a population of individuals, each one with a fitness value, the algorithm progresses by randomly selecting two for mating. The binary tournament selection is used [11] [12]. In tournament, a pair of individuals is competed. The individual with a high fitness value is copied into the mating pool. That is, individuals with a high fitness value have a greater chance of being selected to generate children for the next generation. This process is repeated until the mating pool is filled with number of individuals equal to population size (*pop_{size}*).

Then, the exogenous recombination scheme is applied between mating parents to obtain a new generation of population. The crossover operator operates with probability P_c.

After the new generation has been determined, the individual are subjected to a low rate mutation function. The mutation converts each allele type to one of the other types with equal probability. Thus, mutation accounts to changes in value (i.e. from a -1 valued allele to a 0 and vise versa) and changes in dominance (i.e. from dominant 1 to recessive -1) [9, 10]. The mutation operates with probability P_m.

The process of applying diploid GA with exogenous recombination is repeated until stopping criteria is met. Here, the stopping condition is satisfied when the fitness function equal to 1 which means that we reach to exact solution that represent the ASCII code of *one* character in the plaintext. The overall diploid GA can be continued for each cipher character in the ciphertext and so on until the end of the ciphertext.

4. Experimental Results

The performance of the diploid GA with exogenous recombination is measured and compared with the traditional GA used by Garge et al. [13]. The Merkle-Hellman algorithm was used with the following parameter settings:

1. *Private Key* (easy knapsack): (1 3 7 13 26 65 119 267 504 1007 2013 4027 8053 16107 32213).
2. *u*= 65423.
3. *w* = 21031 and compute w^{-1} = 5363.
4. *Public key* (hard knapsack): (21031 63093 16371 11711 23422 58555 16615 54322).

After determining the parameters of the algorithm, the encryption process is applied on plaintext" MACRO" and the following ciphertext is obtained 65728 37646 100739 103130 128821 each number represents a sum of the difficult knapsack. The ciphertext is attacked using diploid GA with exogenous recombination with the following parameters setting:

1. pop_{size} =5.
2. P_c = 0.9.
3. P_m = 0.1.

The cryptanalyst obtains the result in ASCII code. That is, it correctly identified the enciphered characters which represent the plaintext. The result is: 10110010 10000010 11000010 01001010 11110010 which represent the ASCII code for MARCO. Table 1 illustrates Garg et al. result. Figure 4 demonstrates the best fitness function in the population across generation number for obtaining plaintext MACRO.

Table 1: Garg et al Results

Character	M	A	C	R	O	**Average**
pop$_{size}$	17	112	271	89	87	**115**

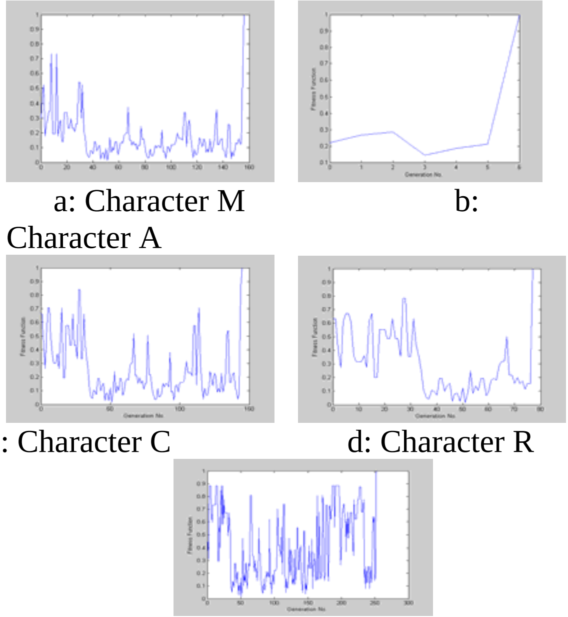
Diploid GA embodies a form of temporal memory that is distributed across the population, and this fact can give the GA more power in finding the correct bit pattern for the hard knapsack sum.

Comparing the approach with that of Garg et al. that uses simple GA, the following encouraging remarks are drawn:

- Doploid GA with exogenous recombination approach is more efficient than Garg et al. The diploid GA works with average *pop*$_{\text{size}}$ equal to 5, while Garg et al GA has to work with, on average, 115 individuals. Additionally, When the adopted diploid GA converges to the accurate solution for the first number in ciphertext which is normally occurred when the fitness function equal to 1, it continues to re-evolve this final population according to the next number in the ciphertext and so on without requiring to re-initialize the population randomly from the beginning as this is the case of the Garg et al GA approach. This re-evolving ability comes from the fact that diploid GA has a diverge number of solutions saved besides those converged solutions that expressed when population continues to evolve toward a given solution. However, the GA of Garg et al. evolves and thus converges its population towards only one solution, and to re-change solution requires re-initializing population instead of re-evolving it.
- Code implementation of the diploid GA with exogenous recombination idea is still as easy as the traditional GA.

5. Conclusion

In this paper, a diploid GA with exogenous recombination scheme is used to cryptanalysis the Merkle-Hellman knapsack cipher algorithm. The approach is simple, efficient, and offers a powerful tool to cryptanalysis knapsack cipher that successfully finds the correct bit pattern for the hard knapsack sum. This approach is found to be more efficient than Garg et al.



a: Character M b: Character A

c: Character C d: Character R

e: Character O

Figure 4: Best Fitness Function in the Population across Generation Number

References

1. Denning, D. **1982**. *Cryptography and Data Security*. Addison Wesley.
2. Schneier, B. **1996**. *Applied Cryptography Algorithms, Protocols and Source Code in C. Second edition*, John Wiley & Sons.
3. Merkle, R. C. and Martin, E. H. **1978**. Hiding Information And Signatures In Trapdoor Knapsacks, *IEEE transaction on Information Theory*, **IT-24:** pp. 525-530.
4. Spillman, R. **1993**. Cryptanalysis of Knapsack Ciphers Using Genetic Algorithms, Cryptologies, **17**(4):367-377.
5. Stallings, W. **1999**. *Cryptography and Network Security: Principles and Practice.* Second edition, Prentice-Hall Inc.
6. Pheeger, C. P. and Pfleger, S. L. , **1997**. *Security in Computing*. Third edition, Prentice-Hall Inc.
7. Menezes, A.; van Orschot, P. and Vanstone S. **1997**. *Handbook of Applied Cryptography Boca Raton*, CRC Press.
8. Simth, R. E. **1988**. *An Investigation of Diploid Genetic Algorithms for Adaptive Search of Non stationary Functions*. M. Sc. Thesis, Alabama, Tuscaloosa.
9. Simth, R. E. and Goldberg, D.E. **1992**. Diploidy and Dominance in Artificial Genetic Search. *Complex Systems*, **6**:251-285.
10. Abdul Haleen, M.F. **2002** . *An Alternative Recombination Scheme for Diploid Genetic Algorithm*. Ph. D. Thesis, University of technology. Baghdad. Iraq.
11. Goldberg, D. E. **1989**. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley.
12. Mitchell, M. **1996**. *An Introduction to Genetic Algorithm*, MIT press.
13. Garg, P.; Shastri A. and Agarwal, D.C. **2006**. An Enhanced Cryptanalytic Attack on Knapsack Cipher Using Genetic Algorithm. *Transaction on Engineering, Computing and Technology*, **12.**