Cryptanalysis Using Genetic Algorithm

Number Theory and Cryptology
Sadip Giri (sadipgiri@bennington.edu)

Introduction: Cryptanalysis is the science and study of a method to recover the plaintext and /or key from a ciphertext. Many researchers in the field of cryptanalysis are interested in developing automated attacks on encryption algorithms (ciphers). In the brute force attack, every possible key is tried on a piece of ciphertext until it translates to meaningful plaintext; however, it has the disadvantage of high computational complexity. In the study, the optimization heuristic technique: Genetic Algorithm (GA) is used to crack Vigenere cipher (polyalphabetic cipher).

Genetic Algorithm: GA is a metaheuristic process of natural selection that belongs to the larger class of evolutionary algorithms (EA) to generate high-quality solutions to optimization and search problems based on bio-inspired (Fig.1) operators such as selection, crossover, and mutation.

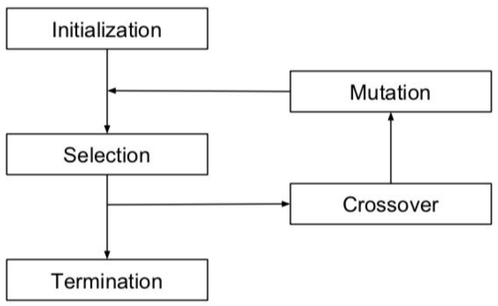

Fig. 1 GA Cycle

Selection: The idea is to give preference to the individuals with good fitness scores and allow them to pass their genes to the successive generations. The process determines which keyword in the current group of keywords will be used to create the next iteration.

Crossover: Single-point crossover (Fig.2) represents mating between individuals where a point on both parents' chromosomes is picked randomly, and designated a 'crossover point'. Bits to the right of that point are swapped between the two parent keywords thus creating a completely new keyword (offspring) each carrying some (genetic) information from both parent keywords.

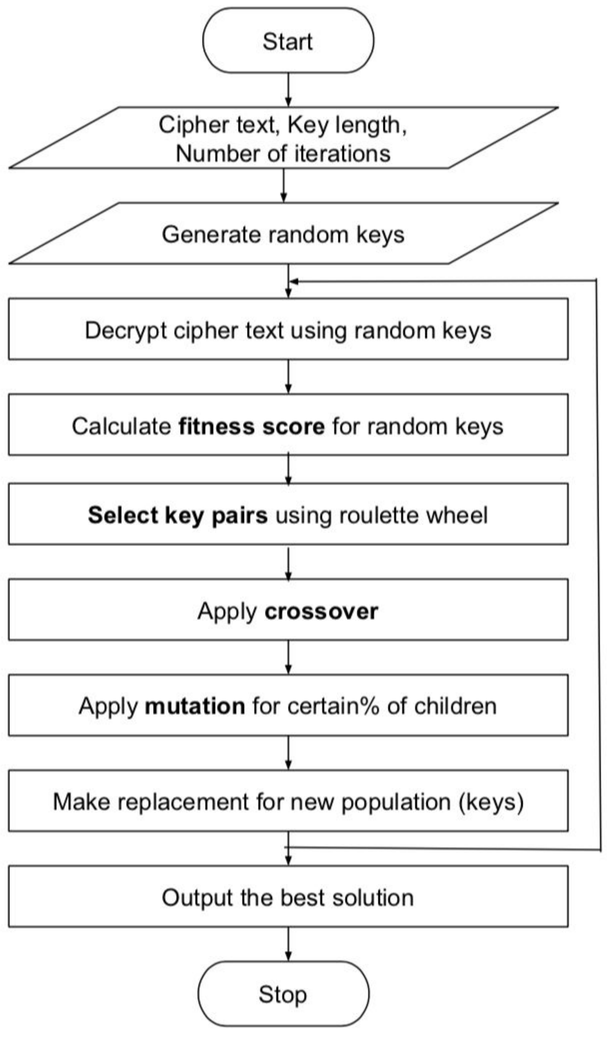

Fig.5 GA Implementation Flowchart

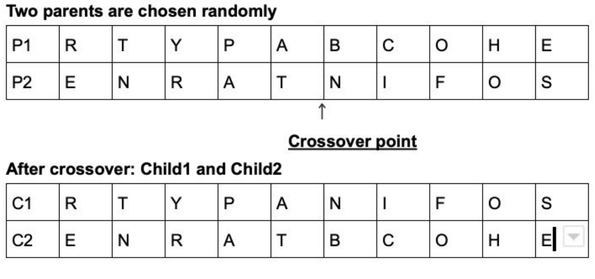

Fig.2 Single point crossover

Mutation: The key idea (Fig.3) is to insert random genes (alphabet) in offspring (keyword) to maintain the diversity in population to avoid the premature convergence (trapped in a local minimum).

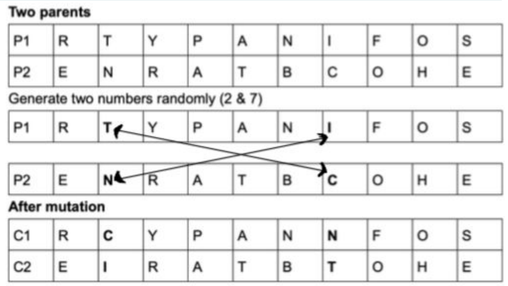

Fig.3 Mutation process

Fitness Measure: The fitness function (Fig.4) is the technique adopted to differentiate candidate keys i.e to compare n-gram statistics of the decrypted message with those of the language (which are assumed known). Fitness scores using fitness function are assigned to each individual to determine the suitability of a proposed keywords. The individual (keywords) having optimal fitness score (or near optimal) are sought.

Fig.4 Fitness Function

$$\mathrm{I} = \alpha.\sum_{i\in A}|K_i^u - D_i^u| + \beta.\sum_{i,j\in A}|K_{i,j}^b - D_{i,j}^b| + \gamma.\sum_{i,j,k\in A}|K_{i,j,k}^t - D_{i,j,k}^t|$$

Here, A is language alphabet ([A to Z]), K is known language statistics, D is decrypted message statistics, u, b & t are the unigram, bigram and trigram statistics respectively. The values of **α,** β & γ (**α+**β+γ = 1) allow different weights to each of the 3 n-gram types.

Key length: It is guessed using classic Kasiski method and validated using the formula (L) where, n is total alphabets in ciphertext, I is Index of Coincidence, and i is index of freq. [A to Z]

$$L = \frac{0.027n}{(n-1)I + 0.065 - 0.038n} \quad , \qquad \mathrm{I} = \sum_{i=1}^{26} \frac{n_i(n_i-1)}{n(n-1)}$$

Results: In the study, 0.2 mutation rate with fewer generations (5 to 50 depending on key-size) gives the best results for key lengths 1 to 15 of the population size 30.

Conclusion: GA successfully cracks the cipher with reduced time complexity compared to other human techniques such as classic frequency analysis and brute force. GA can be further optimized using the robust randomizer, excellent fitness measure, tuning parameters and NLP technique for better performance.