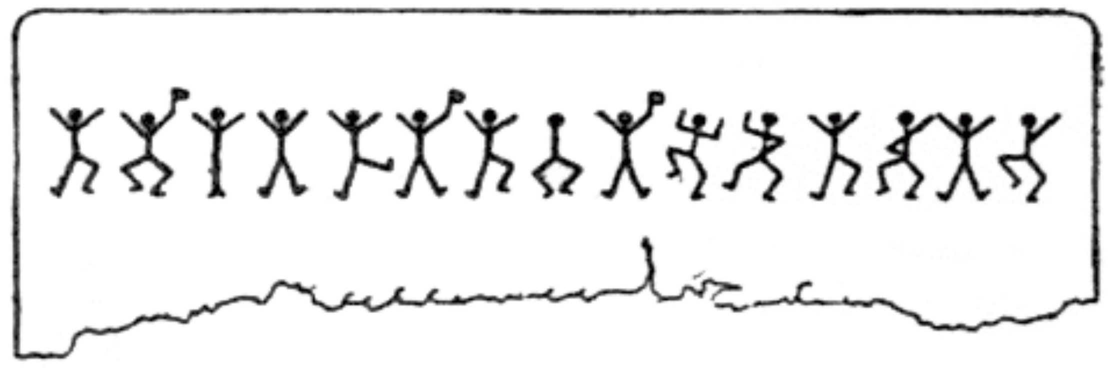


Ilya Kuzovkin

Breaking Substitution Cipher Using Genetic Algorithm

What is a substitution cipher?

Substitution cipher is an example of simple encryption scheme. Each letter in original alphabet is replaced by random symbol from another alphabet:

Alphabet	ABCDEFGHIJKLMNOPQRSTUVWXYZ
Key	KGCOPNMQBARHFZSIYDTVUJLWXE
Message	ALGORITHMICS PROJECT
Cipher	KHMSDBVQFBCT IDSAPCV

This kind of cipher is easily breakable using letter frequency analysis.

If it's so simple, why to bother?

- Although human can easily break this encryption, it is not exactly clear how machine could do the same.
- Naïve exhaustive search solution would require 26! evaluations.
- Actual encryption schemes like AES, RSA and many others use transpositions as intermediate step. Therefore if some heuristic will be able to find "correct" order, this will give out some information.

What is Genetic Algorithm?

Simulates the process of natural evolution: spices adapt in order to survive:

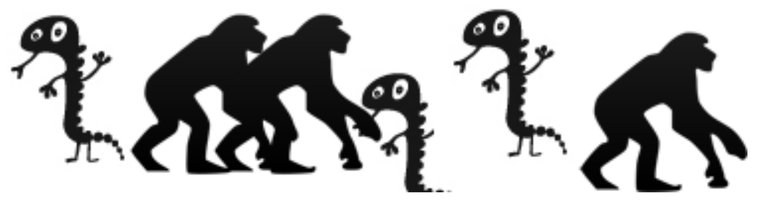
1. Initial population

2. Natural selection

3. Breeding

No picture available

4. Mutation



5. New population

Implementation details

In our case keys are individuals, and they will evolve. Natural selection's role plays objective function

$$\text{score}_{key} = \alpha \sum_{i \in A} |K_i^u - D_i^u| + \beta \sum_{i \in A} |K_i^b - D_i^b| + \gamma \sum_{i \in A} |K_i^t - D_i^t|$$

u – unigrams
b – bigrams
t – trigrams
K – n-gram distribution in decrypted message
D – actual n-gram distribution

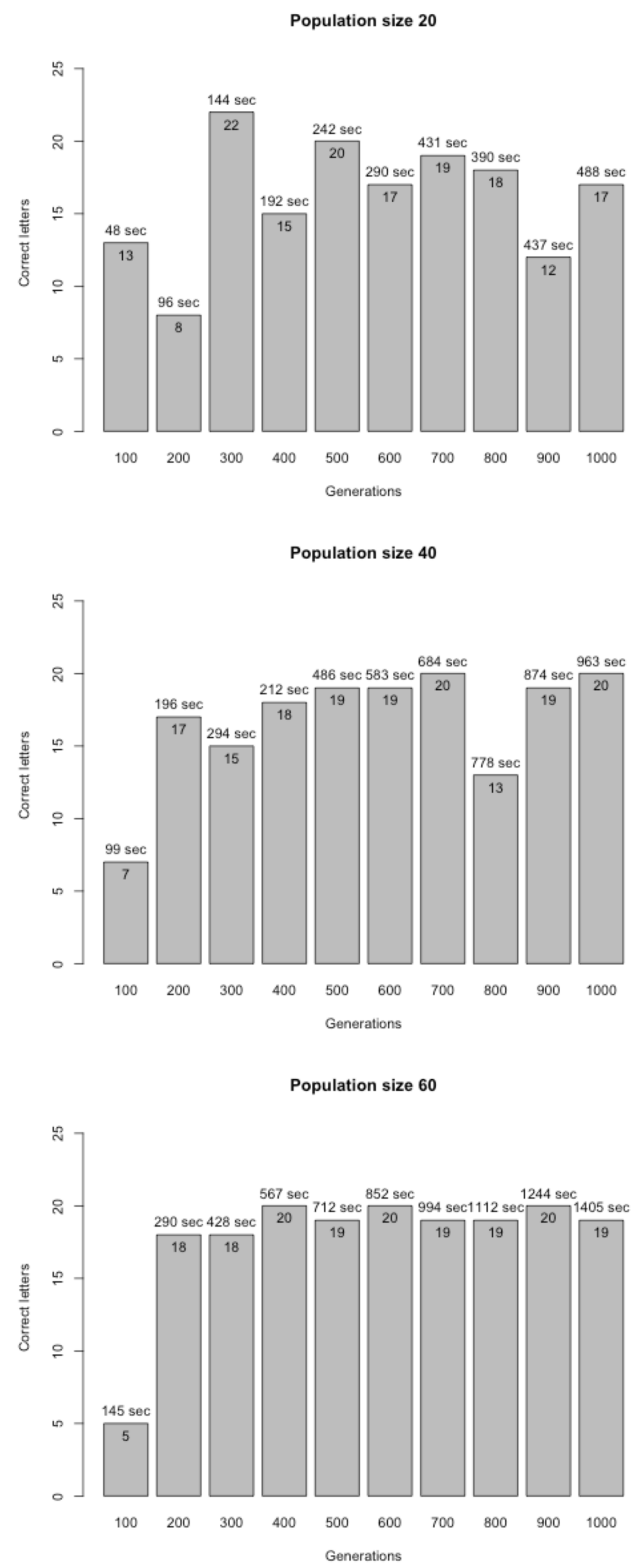
Breeding: split parents at random position.
Swap second slices to get children.

Mutating: randomly chose two chromosomes and swap them.

Algorithm

```
Initial population
For number of generations do
  decrypt cipher with each key
  compute n-gram distributions
  assign scores
  pick best half
  While not enough children do
    breed
    sometimes mutate
  new population ← children
```

Results



Number of correctly identified letters depending on the size of the initial population and number of generations. After about 500 generations result does not improve much, and later on it can even decrease. Such odd behavior is caused by the fact that we use letter frequencies in the objective function. In principle, with cipher text long enough, we could be able to get all 26 letters of the key.