An investigation into the Zodiac Ciphers using Genetic Algorithms

Report Name	**Outline Project Specification**
Author	**Martin Key (mak18)**
Supervisor	**Richard Jensen (rkj)**
Module	**CS39440**
Degree Scheme	**G401 (Computer Science)**
Date	**10 February 2014**
Revision	**1.0**
Status	**Release**

Contents

1 Project Description

My intention with this project is to look into the way that a genetic algorithm (GA) could be used in the cryptanalysis of classical substitution ciphers. Specifically I will attempt to write a modular GA to tackle homophonic ciphers in general with an easy way to plug in different fitness functions, genetic operators and populations. This will allow the GA to function not only in the area in which I wish to use it, but hopefully to be used in other examples of homophonic ciphers as well.

The cipher, or ciphers, I am specifically interested in are those written by the self styled 'Zodiac' serial murderer from northern California in the late 1960s and 70s. The murderer sent a number of ciphers to local papers at the time, and I hope to apply the written GA to these ciphers. The first, sent in 3 parts to different papers and known as the Z408 due to its length of 408 characters, was solved at the time. This gives me the chance to test my GA against a cipher for which there is a known and verified solution. If the GA succeeds in matching a key to the Z408 cipher, I hope to move on to the second well known cipher, the Z340. Again named because it is made up of 340 characters, this cipher is yet to be solved, despite in depth research by a number of people. There are two other ciphers written by this author, a 13 character cipher and a 32 character cipher, however due to the short nature of these 2 it is unlikely there will ever be any verifiable solution, therefore I will not be attempting them.

2 Proposed Tasks

2.1. Tasks to complete for the Genetic Algorithm.

- Define genetic representation for the key - Due to the nature of the cipher, it doesn't use the Latin alphabet, requiring more than the standard 26 letter English alphabet. This means I need to find another way to define the cipher key candidate solutions which is understandable in a computing situation.
- Define fitness functions - I will need to define at least one fitness function for the GA, more if I have time. The fitness function defines which are the 'best' solutions in a generation, so they can be passed on to the next generation. Different fitness functions may have differing levels of success and therefore more than one fitness function may increase chances of success.
- Define genetic operators - There are a number of different types of genetic operators, normally separated into mutation and crossover categories. I will need to decide which genetic operators are most useful in improving the population. Differing genetic operators could be matched with different fitness functions to find the most successful match.
- Generate Populations - I will need some way of generating a set sized population of keys. Normally a randomly generated set, due to the homophonic nature of the cipher I will need some way of having keys generated in a suitable manner.

2.2. Tasks to complete using the Genetic Algorithm

- Apply genetic algorithm to simple defined ciphers - To test the genetic algorithm in the first instance I could write some of my own very simple ciphers with which I could test the algorithm as a whole.
- Apply genetic algorithm to Z408 cipher - As a further test of the software, I will attempt to find a key for the Z408 cipher. Having the solution already I can compare and contrast how successful differing fitness functions and genetic operators are in finding a solution.
- Apply genetic algorithm to Z340 cipher - Finally I will apply the completed algorithm to the Z340 cipher in attempt to find a valid key to the cipher.

3 Project Deliverables

3.1. Official deliverables

There are a number of official deliverables associated with the university process surrounding the Major Project, these are:

- Outline Project Specification - This document, outlining the project as I consider it, an overall framework of how I intend to go about it, the documents I intend to deliver pertaining to the project and a bibliography of current research material.
- Mid Project Demonstration - A demonstration of progress in development, including any development up to the date of demonstration, prototypes and development on final product.
- Technical Report - The final report of the research, development and investigation of the major project.
- Final Demonstration - A demonstration of the culmination of development during the major project.

3.2. Unofficial deliverables

As I am intending to attempt to follow feature driven development there will be a number of unofficial deliverables I am hoping to complete for my own benefit throughout the project, these are:

- Overall Model - A class diagram, or series of class diagrams which focus on the high level scope of the design. Includes and methods or attributes identified at this high level. Would normally be peer reviewed and refined, I will need to find some way to emulate this process.
  - approx. a week.
- Features List - The list of features which will make up the GA allowing for the prioritisation and segmentation of work over the course of the project.
  - approx. a week.
- Features Plan - The plan of the order and timing of completion of each feature, based on complexity, risk and importance. As well as taking into account the need for something to be demonstrated at the Mid Project Demo.
  - approx. a week.
- Features Design - Repeated for each of the features, a specific plan made before coding begins, refining the Overall model in one area. Defines classes, methods and parameters of the feature, as well as any other specifics which would be required for coding.
  - approx. 2/3 weeks.
- Configurable Genetic Algorithm - The hope is to deliver a GA which will be fully configurable in terms of selection, genetic operators, population and termination. The attempt at this will be included in the Technical Report and will be demonstrated at the Final Demonstration.
  - unsure of the length of time it will take to run the different attempts to crack ciphers.

4 Initial Annotated Bibliography

J. M. Carroll & S. Martin (1986). `THE AUTOMATED CRYPTANALYSIS OF SUBSTITUTION CIPHERS'.
Cryptologia 10(4):193-209.

An investigation into traditional methods of automated cryptanalysis. The use of letter frequency could be useful, but could not due to the application to polyalphabetic ciphers rather than homophonic.

R. Spillman, et al. (1993). `USE OF A GENETIC ALGORITHM IN THE CRYPTANALYSIS OF SIMPLE SUBSTITUTION CIPHERS'.
Cryptologia 17(1):31-44.

An investigation into simple monoalphabetic ciphers. Could be useful in the investigation of genetic operators, as well as fitness function, however the use of letter frequencies again may be a problem. Concludes that there is usually a need for a visual inspection is normally required, due to not finding the exact key, may need to investigate further text recognition, if I wish for full automation.

Đào, Thắng (2008). "ANALYSIS OF THE ZODIAC 340-CIPHER". Unpublished Master's Theses. Paper 3570.
San Jose State University, California.
Available at: http://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=4566&context=etd_theses&sei-redir=1

A useful investigation into the Zodiac timeline as well as into possible methods of encryption of the various Zodiac ciphers. Goes on to attempt solution with traditional cryptanalysis methods, so less useful in this area.

Bethany Delman (2004). "GENETIC ALGORITHMS IN CRYPTOGRAPHY". Unpublished Master's Theses.
Rochester Institute of Technology, Rochester, New York. Available at: https://ritdml.rit.edu/handle/1850/263

A comparison of both genetic algorithms and traditional cryptanalysis methods. Though it concludes that genetic algorithms are actually less useful that traditional methods, does have useful investigation into the development of a genetic algorithm. Does not specifically attempt homophonic cipher, but does attempt mono- and polyalphabetic ciphers.

Feature Driven Development Processes [Online]. Available at:
http://www.nebulon.com/articles/fdd/download/fddprocessesUSLetter.pdf [Accessed: 4 Feb 2014].

A guide to the specific use of each of the FDD processes, which I will hopefully be following as closely as possible. Refers to a team, therefore some alteration required.