**An investigation into the Zodiac Ciphers using Genetic Algorithms**

Final Report for CS39440 Major Project

*Author*: Martin Key ([mak18@aber.ac.uk](mailto:yourid@aber.ac.uk))

*Supervisor*: Dr. Richard Jensen ([mkj@aber.ac.uk](mailto:mkj@aber.ac.uk))

30 April 2014

Version 0. (Draft)

s

This report is submitted as partial fulfilment of a BSc degree in

Computer Science (G400)

Department of Computer Science

Aberystwyth University

Aberystwyth

Ceredigion

SY23 3DB

Wales, UK

**Declaration of originality**

In signing below, I confirm that:

* This submission is my own work, except where clearly indicated.
* I understand that there are severe penalties for plagiarism and other unfair practice, which can lead to loss of marks or even the withholding of a degree.
* I have read the sections on unfair practice in the Students’ Examinations Handbook and the relevant sections of the current Student Handbook of the Department of Computer Science.
* I understand and agree to abide by the University’s regulations governing these issues.

Signature ……………………………………………. (Martin Key)

Date …………………………………………………

**Consent to share this work**

In signing below, I hereby agree to this dissertation being made available to other students and academic staff of the Aberystwyth Computer Science Department.

Signature ……………………………………………. (Martin Key)

Date …………………………………………………

# Abstract

# Contents

Abstract 3

Contents 4

Figures 5

1 Introduction 6

1.1. Overview 6

1.2. The Problem 6

1.3. Aims 6

2 Background 7

2.1. Substitution Ciphers 7

2.2. Zodiac Ciphers 9

2.3. Genetic Algorithms 13

3 Analysis 19

3.1. Z408 19

3.2. Z340 19

4 Design 21

4.1. Overview 21

4.2. Management 21

4.3. Population Generation 21

4.4. Chromosomes 21

4.5. Fitness 21

4.6. Selection 21

4.7. Genetics 21

4.8. Termination 21

5 Implementation 22

5.1. Management 22

5.2. Population Generation 22

5.3. Chromosomes 22

5.4. Fitness 22

5.5. Selection 22

5.6. Genetics 22

5.7. Termination 22

6 Testing 23

6.1. Unit 23

6.2. Functional 23

7 Investigation Methods 24

7.1. Substitution 24

7.2. Z408 24

7.3. Z340 24

8 Investigation Results 25

8.1. Substitution 25

8.2. Z408 25

8.3. Z340 25

9 Conclusions 26

9.1. Substitution 26

9.2. Z408 26

9.3. Z340 26

10 Critical Evaluation 27

11 Bibliography 28

# Figures

[Figure 1 - Generalised homophonic substitution cipher to directly counter plain text letter frequency[26] 9](#_Toc386644629)

[Figure 2 - Generalised homophonic substitution cipher[26] 9](#_Toc386644630)

[Figure 3 - Z408 Cipher Part 1[19] 10](#_Toc386644631)

[Figure 4 - Z408 Cipher Part 2[19] 10](#_Toc386644632)

[Figure 5 - Z408 Cipher Part 3[19] 11](#_Toc386644633)

[Figure 6 - Z340 Cipher[19] 12](#_Toc386644634)

[Figure 7 - Z13 Cipher[19] 13](#_Toc386644635)

[Figure 8 - Z32 Cipher[19] 13](#_Toc386644636)

[Figure 9 - An example of chromosomes encoded as binary strings[28]. 15](#_Toc386644637)

[Figure 10 - A visualisation of Roulette Wheel Selection 17](#_Toc386644638)

[Figure 11 - A visualisation of Stochastic Universal Sampling 17](#_Toc386644639)

# Introduction

## Overview

The murders of David Faraday and Betty Lou Jensen on Friday Dec. 20, 1968[16] began the shocking killing spree of a man naming himself "The Zodiac". Over the next 5 years, there would be another 3 deaths attributed to this mystery man, and a possible connection to 4 other victims[19]. Throughout this time however, 'The Zodiac' would continue to write letters and postcards to a number of newspapers across California claiming a body count of up to 37 murders by the end of correspondence in January of 1974.

Included in a number of these letters were ciphers, messages encoded by 'The Zodiac' which may contain clues as to other victims, or even the identity of 'The Zodiac' themselves. There were a total of 4 ciphers included, generally known as the Z408, the Z340, the Z13 and the Z32, referring to the number of characters included in each. The first cipher, the Z408, has been solved, however the further 3 have yet to be solved, further investigation into the ciphers is shown in section 2.1.

## The Problem

The problem encountered is that where the Z408 was solved at the time of the original investigation, the Z340 cipher has yet to be solved despite a large amount of study of the cryptogram, by both academics and amateurs. During the last 40 years, the attempts to solve the Z340 have assumed that the cipher is a 'homophonic' cipher, as the Z408, a 'polyalphabetic' cipher, a 'double transposition columnar' cipher and a 'one time pad'. However none have been able to provide a suitable solution. Other attempts have attempted to find solutions through use of anagrams and arbitrary expanding letters into full words. A number have even attempted to simply

to fit their solutions to predetermined messages[17].

More Recently a number of ideas have surfaced which could create a solution to the problem. These include the ideas that the cipher was actually written backwards, or at a rotation of 90°, 180° or 270[20]. The idea has even been posed that the cipher is actually meaningless, in an attempt to confuse. However, these ideas have also yet to be validated.

## Aims

The main aim of my research and software development is to attempt to solve the Z340 cipher through use of a genetic algorithm. By creating a framework to create a genetic algorithm which can attempt to solve substitution ciphers, I hope to then use the frame work to attempt to solve the Z340 cipher. Extrapolating that the Z340 cipher is also a homophonic substitution cipher, like the Z408 cipher, I can attempt to create a solution for both ciphers, as well as a simple substitution cipher. This means I can refine the method I wish to use to attempt the Z340 cipher, by first working on the substitution cipher, and then the Z408 cipher, both of which I already have the solution to.

# Background

## Substitution Ciphers

### Overview

Due to the assumptions made about the ciphers to make this a feasible investigation into the Zodiac case with genetic algorithms, there are two types of cipher which must be looked into. The simple substitution cipher, as a benchmark for investigation. And the Z408 cipher, assumed to be a homophonic substitution cipher. The assumption is made that the Z340 cipher is also a homophonic cipher. It seems unlikely that the writer of the ciphers would be able to implement more than one complicated type of cipher, especially considering the mistakes they made in implementation.

The definition of a substitution cipher in cryptography is one where the letters of the alphabet are replaced by some standard method with another 'alphabet' converting a 'plain text' into a 'cipher text'. This new alphabet can be any collection of letters, numbers, lines and dots or any other written symbol. Letters can be replaced individually or in groups of any number of letters. The standard process by which letters are replaced is what decides what permutation of substitution cipher it is. A cipher which replaces one letter with another symbol is a 'simple substitution' cipher, one that works on groups of letters is a 'polygraphic' cipher. One which has a fixed alphabet size the same as the plain text alphabet is known as a 'monoalphabetic' cipher, and one which uses multiple cipher text alphabets is known as a polyalphabetic cipher[22]. A Homophonic cipher is one where single plain text characters map to more than one cipher text character in one alphabet.

What follows is more information on the two types of substitution cipher used in this investigation, the simple substitution cipher and the homophonic substitution cipher.

### Simple Substitution Cipher

As previously mentioned, the simple substitution cipher is one where each plain text letter is replaced by another symbol. Each letter of the alphabet is associated with the letter of a second alphabet[21]:

Plain text letters: abcdefghijklmnopqrstuvwxyz

Cipher text letters: zyxwvutsrqponmlkjihgfedcba

Convention says that the plain text should be written in lower case, and the cipher text in capitals, or in this case small capitals.

The cipher alphabet is usually rotated, reversed or scrambled in a more complicated way. In the case above, the alphabet has been reversed, making it an 'Atbash' cipher. If the alphabet is rotated it is referred to as a 'Caesar' cipher. Something like this:

Plain text letters: abcdefghijklmnopqrstuvwxyz

Cipher text letters: lmnopqrstuvwxyzabcdefghijk

The more complicated scrambling usually involves the use of a key word, which has repeated letters removed, and then has the remaining alphabet letters added in the usual fashion. But the alphabet can be scrambled in any number of more complicated ways, for which only the sender, and the intended recipient would have the key.

If we take the second example of a cipher alphabet and encrypt a set of words, say 'Simple substitution cipher', this will give us the cipher text:

simple substitution cipher.

dtxawp dfmdetefetzy ntaspc[18].

This provides us with an encrypted message, however it is obvious where words end, and punctuation gives away where the sentence ends, therefore punctuation would be removed, and the cipher would be split into uniform 5 letter groups, giving:

dtxaw pdfmd etefe tzynt aspc

All of this however does not create a particularly secure cipher. Though brute force is still not a good option, due to the fact that with a 25 letter alphabet, that creates a key space of 403,291,461,126,605,635,584,000,000[21] which is far too large to check every permutation in real time. However, simple substitution ciphers can be solved easily with the method of frequency analysis[21]. This is the process of finding patterns in the frequency of characters in a cipher, for example, the letter 'e' is the most common letter in the English language, appearing 11.1607%[22] of the time. From this we can expect that in any cipher, of long enough length to be statistically significant, the letter which appears approximately 11.1% of the time is the letter 'e'.

As well as simply looking for single letters, the English language has many sets of letters which often appear together, for example the collection of letters 'th' appear very often. The frequency of these groups of letters are described using n-grams. An n-gram is a way of storing an object as a vector[25], mostly used for analysing language, they can be used for words, syllables or letters. Analysis of the frequency of these n-grams can also be useful in the cryptanalysis of substitution ciphers.

Though the examples here have been shown using the Latin alphabet, the cipher alphabet can be any set of symbols, for example the pigpen alphabet, which was devised using lines and dots on a grid which looks something like this[18]:

http://rumkin.com/tools/cipher/media/pigpen/D.gifhttp://rumkin.com/tools/cipher/media/pigpen/T.gifhttp://rumkin.com/tools/cipher/media/pigpen/X.gifhttp://rumkin.com/tools/cipher/media/pigpen/A.gifhttp://rumkin.com/tools/cipher/media/pigpen/W.gif http://rumkin.com/tools/cipher/media/pigpen/P.gifhttp://rumkin.com/tools/cipher/media/pigpen/D.gifhttp://rumkin.com/tools/cipher/media/pigpen/F.gifhttp://rumkin.com/tools/cipher/media/pigpen/M.gifhttp://rumkin.com/tools/cipher/media/pigpen/D.gif http://rumkin.com/tools/cipher/media/pigpen/E.gifhttp://rumkin.com/tools/cipher/media/pigpen/T.gifhttp://rumkin.com/tools/cipher/media/pigpen/E.gifhttp://rumkin.com/tools/cipher/media/pigpen/F.gifhttp://rumkin.com/tools/cipher/media/pigpen/E.gif

### Homophonic Substitution Cipher

A homophonic substitution cipher is a substitution cipher in which a given character may have any number of different representations[26]. This allows the encoding to disguise the structural properties of the plain text and reduce the effectiveness of frequency analysis in the cryptanalysis of this type of cipher by making the single letter frequency of the cipher text constant. The elimination of frequency analysis as an efficient method of solving the cipher fixes one of the major problems with the simple substitution cipher. This use of multiple cipher text characters creates a key which looks something like this[26]:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z |
| c | r | y | p | t | o | g | r | a | m | 5 | 6 | 7 | 8 | 9 | b | d | e | f | h | i | j | k | l | n | q |
| 1 |  |  |  | 2 |  |  |  | 3 |  |  |  |  |  | 4 |  |  |  |  |  | s |  |  |  |  |  |
| u |  |  |  | v |  |  |  | w |  |  |  |  |  | x |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  | z |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

This example shown previously is not an effective homophonic cipher however, according to Fred Stahl in his paper "A homophonic cipher for computational cryptography"[26] to completely level the frequency range between letters you would require a homophonic cipher key described below.

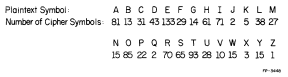


Figure - Generalised homophonic substitution cipher to directly counter plain text letter frequency[26]

However, even with the use of multiple cipher text characters for each plain text letter, there are still problems which need to be taken into account. Bi-grams and tri-grams, two and three letter n-grams respectively, will still have frequencies which will be recognisable by a cryptanalyst. To counter this, the number of cipher text alphabet must again be altered to take into account these frequencies. By reducing the restriction on keeping the letter frequency curve flat, these other abnormal frequencies can be taken into account, resulting in a cipher like that shown below, also from Fred Stahl's paper.

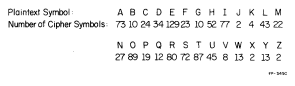


Figure - Generalised homophonic substitution cipher[26]

Due to the large variation in cipher text letters, the length of the cipher now comes into play. The longer the cipher, the more of these variations can be used, therefore the more secure the cipher becomes.

## Zodiac Ciphers

### Overview

Between 1967 and 1974, the Zodiac killer provided more than 20 written communications to police officials, usually in the form of letters and postcards written to newspapers. However this also included letters to a prominent lawyer at the time, Melvin Belli, and to an independent TV station, KCAL-TV, channel 9. During the time the Zodiac was corresponding with the authorities, 4 ciphers were sent to the press, always contained within a larger covering letter. The covering letters usually contained claims about his kills, threats about further attacks and demands on the public. The first cipher, the Z408, was sent as 3 different parts to the Vallejo Times-Herald, the San Francisco Chronicle, and the San Francisco Examiner. The further three were all received to the San Francisco Chronicle on the being the Z408, the Z13 and the Z32.

Of the three, the only one to be solved was the Z408, it was solved by Donald and Bettye Harden less than a week after it was received by the newspaper. Other than this however only the Z340 has any possibility of ever being solved. Both the Z13 and the Z32, containing 13 and 32 characters respectively, have too few characters for a statistically significant solution to be found, there are too many possible combinations of characters. The fact that the smaller ciphers are unlikely to be solved has not however stopped people attempting to propose keys for the ciphers, however none have ever been verified.

### Z408 Cipher

Received on the 31st July, 1969, the Z408 cipher is the longest and the only solved cipher sent by the Zodiac, while he was active. It was sent as three separate sections to three different newspapers:

1. Figure 3 - Part 1 received by the Vallejo Times-Herald
2. Figure 4 - Part 2 received by the San Francisco Chronicle
3. Figure 5 - Part 3 received by the San Francisco Examiner

The cipher was solved by a married couple of teachers, Donald and Bettye Harden, less than a week after the ciphers were originally published, and the solution published in the San Francisco Chronicle on 9th August, 1969. However it recently came to light that the authorities also received a solution to the cipher from a person claiming to only be a 'concerned citizen' on 10th August, 1969. Described as “substantially accurate”[27] by the FBI, it is still unknown who sent this second solution. All three parts of the cipher are shown below.

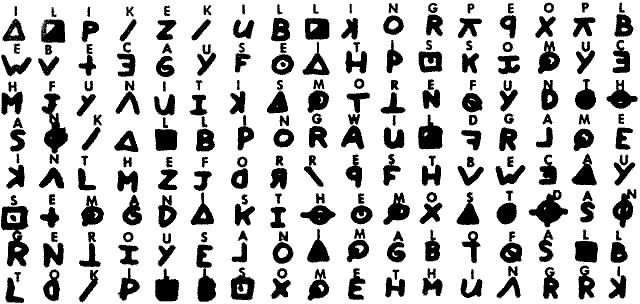


Figure - Z408 Cipher Part 1[19]

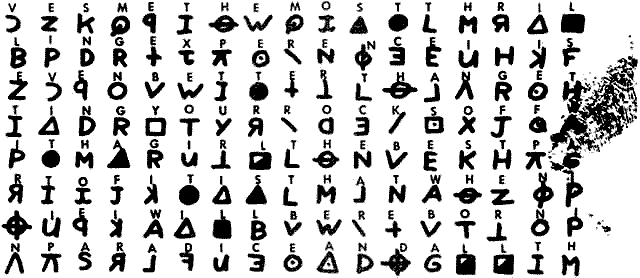


Figure - Z408 Cipher Part 2[19]

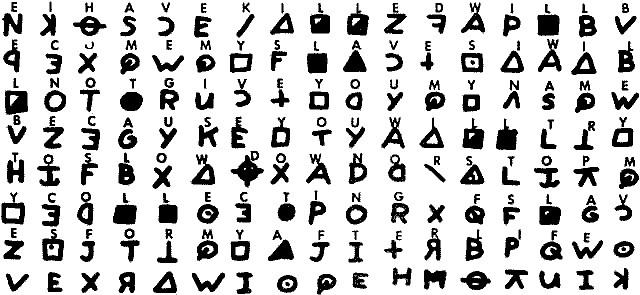


Figure - Z408 Cipher Part 3[19]

The solution proposed by the Hardens is shown below, with the original spelling mistakes left in. A number of reasons for these errors have been proposed, ranging from the Zodiac losing his place in the enciphering process, using the wrong symbols intentionally to confuse, or simply the fact that the Zodiac was not a particularly good at spelling. Any one of these explanations could be true, or any combination thereof.

"I LIKE KILLING PEOPLE BECAUSE IT IS SO MUCH FUN IT IS MORE FUN THAN KILLING WILD GAME IN THE FORREST BECAUSE MAN IS THE MOST DANGEROUE ANAMAL OF ALL TO KILL SOMETHING GIVES ME THE MOST THRILLING EXPERENCE IT IS EVEN BETTER THAN GETTING YOUR ROCKS OFF WITH A GIRL THE BEST PART OF IT IS THAE WHEN I DIE I WILL BE REBORN IN PARADICE AND ALL THEI HAVE KILLED WILL BECOME MY SLAVES I WILL NOT GIVE YOU MY NAME BECAUSE YOU WILL TRY TO SLOI DOWN OR ATOP MY COLLECTIOG OF SLAVES FOR MY AFTERLIFE. EBEORIETEMETHHPITI"[17]

### Z340

Following the Z408, the Z340 cipher was sent to the San Francisco Chronicle and published on 8th November, 1969. Unlike the Z408 cipher, this was sent in its entirety to a single newspaper. Though there have been many attempts to solve the cipher, none of the proposed solutions have ever been verified, even up to the present. Attempts have been made based on existing accepted cryptology documentation, but others have attempted to solve the cipher in unconventional methods, some even think the cipher was never meant to be solved and is merely meaningless. The entire Z340 cipher is shown over the page.

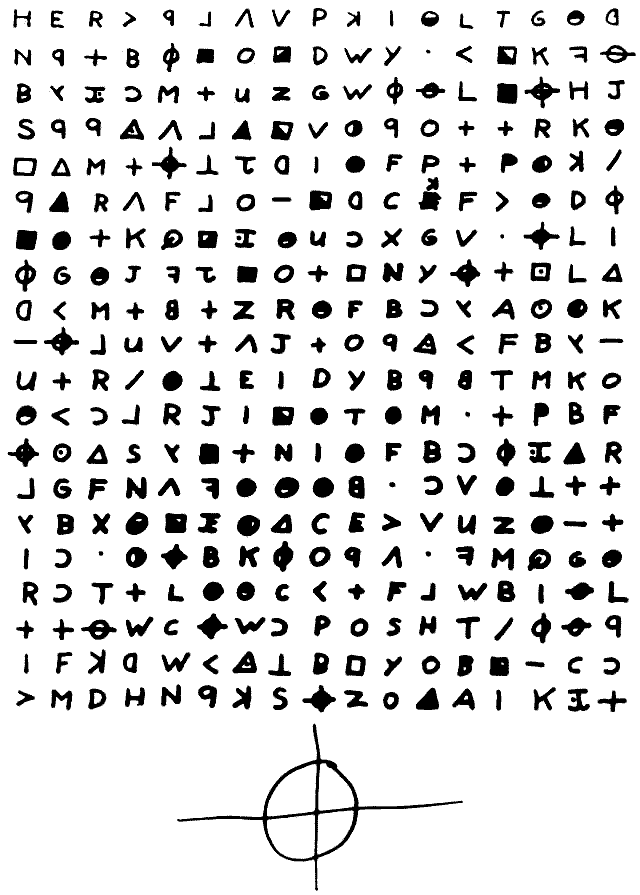


Figure - Z340 Cipher[19]

### Z13 Cipher

Received and again published in the San Francisco Chronicle on the 20th April 1970, the Z13 is the shortest of the 4 ciphers, and was included in a letter which hinted at this cipher containing the Zodiac's real identity. The cover letter contained the line "My name is"[19] above the cipher itself, and with the length of the cipher this has led to many suggestions that the cipher is actually made up of the sentence "My name is" followed by the identity of the killer. Though this is one interpretation of the cipher there have been a number which have taken different ideas in order to find a solution.

However, due to the length of the cipher, it is unlikely any solution will be verified.

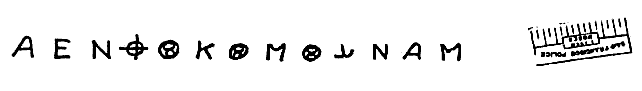


Figure - Z13 Cipher[19]

### Z32 Cipher

The final cipher, received and published in the San Francisco Chronicle on 26 June 1970 was the Z32 cipher. This cipher could be something to do with the location of a bomb the Zodiac threatened was set, and would go off the following Autumn, or could be something to do with his annoyance with people who were not complying to his previous demands. But again, due to the length of the cipher, it is unlikely that any solution will be verified.



Figure - Z32 Cipher[19]

## Genetic Algorithms

### Overview

The genetic algorithm, is a subsection of an area of artificial intelligence called evolutionary computing. This area of artificial intelligence is interested in taking inspiration from the natural world, genetics and the ideas of natural selection and was first introduced by I. Richenberg in his 1960 paper "Evolutionary strategies"[28]. The definition of natural selection is the process whereby organisms in the natural world are more likely to survive and breed if they are better adapted to their surroundings, meaning the genes for the better adaptations are passed on to the next generation. Genetic Algorithms attempt to take these ideas and use them in the area of computer science. The terminology revolving around the ideas have been taken directly from the biological science, therefore there are a few terms which require definitions.

* Population - A group of candidate solutions, each individual is represented by a chromosome.
* Chromosome - A member of a population, an encoding of the solution made up of a series of alleles. Often represented by a binary string.
* Allele - A single piece of information about a solution. In the example of a binary string, each binary digit would be an allele.
* Fitness - The measure of how close a solution is to being correct. In biological terms, how adapted an individual is to its surroundings.
* Selection - The process of choosing those solutions in a population which will be allowed to reproduce to create the next generation.
* Reproduction - The process by which members of the current population interact and are altered in order to create the new generation population. Made up of a number of genetic operators, usually crossover and mutation.
* Crossover - The process of creating 2 offspring by combining the alleles of the current population.
* Mutation - The process of randomly altering a single chromosome in order to help the algorithm not to stall in at a local high fitness, rather than the global high fitness.

The overview of the genetic algorithm is really quite simple, consisting of a generation of a randomised population at the commencement of the algorithm then simply looping through the process of selecting a new population, performing reproduction and then calculating the fitness of the new populations. This looping process will be repeated until a certain termination condition is completed, which could simply be a time limit or a certain number of populations, or it could be something more complicated like calculating whether the fitness of the population is still improving after each generation.

This outline of the process looks something like this:

Generate initial population

Calculate population fitness

Perform Selection

Replace old population with new population

Perform Crossover

Perform Mutation

Place in new population

Test termination condition

Finish

Each of these sections is covered in slightly more detail in the following sections, population generation in 2.3.2, fitness in 2.3.3, selection in 2.3.4, Crossover and Mutation in 2.3.5 and termination in 2.3.6.

The most common use of a genetic algorithm is for an optimisation problem, though the range of what is being optimised can be very large. For example genetic algorithms have been used to optimise a paper production process, or to optimise the schedule of production lines, even optimise the design of engines for commercial passenger aircraft[9]. The use of genetic algorithms has become very widespread in a number of industries. In the case of this research it could be said that we are trying to optimise the key for a cipher, rather than having to attempt every possible permutation in a brute force attack.

The genetic algorithm can however also be used in conjunction with other areas of artificial intelligence. A specific example would be that a genetic algorithm can be used in both the construction and training of artificial neural networks[10].

### Population Generation

Population generation and the creation of chromosomes, also referred to as problem encoding, is one of the 2 areas of a genetic algorithm which are problem specific[6]. The other being the fitness evaluation function.

The aim of encoding the problem in chromosomes is to contain some information about the solution it represents. The most used encoding of a chromosome is a binary string[28], which could look something like this:

|  |  |
| --- | --- |
| Chromosome 1 | 1101100100110110 |
| Chromosome 2 | 1101111000011110 |

Figure - An example of chromosomes encoded as binary strings[28].

Each of these chromosomes could represent the solution in a number of ways, each bit of the string could represent some binary characteristic of the solution, or the whole string could represent a number. There are disadvantages to all actions however, and in this case for example, using a bit string to represent a decimal number, the numbers 15 and 16 are neighbours in integer space. However in the binary representations, 01111 and 10000, are not neighbours in the bit space. This could create problems in terms of attempting combination in crossover[10]. Problems arise in the case of a problem with a discrete number of solutions. If there is exactly 1200 solutions to a problem, this requires a bit string of at least 11 bits to cover. But this leaves a range of 2048 possible encoding, 848 bit patterns which are not used. This can be solved using a default worst possible encoding, or maybe encoding a more likely solution multiple times, to increase likelihood[6].

The binary string is not the only way to encode, chromosomes could take almost any encoding, depending entirely on the problem which is being solved.

The initial population generation is one of the more important features of any genetic algorithm, and there are a large number of factors which could be taken into account and could have an influence on the success of the genetic algorithm. Some of these factors could be the fitness function, the number of individuals, the problem difficulty, the diversity of the population and the search space[14]. Any one of these factors could heavily influence how the initial population is created, or they could simply be taken into account, for example if the population is created in a random, or sudo-random fashion.

It is also shown that the creation of a good initial population has a positive effect on the likelihood that there will be a good solution found[29], as well as the reverse , that a bad initial population makes it difficult for a good solution to be produced by the genetic algorithm[30].

Though the effect of population size on the results of a genetic algorithm are very subjective, experimentation by M. Odetayo into differing population sizes[15] has established some results countering previous work in the area claiming the most efficient initial population size was 100. Depending on the requirement of the genetic algorithm, a population size of 100 on average sampled the least amount of points, had the shortest possible time and number of generations, this population size also had the highest averages in terms of time taken and generations. A population size of 300 however had the shortest average time taken and number of generations required[15]. Again however as this data was taken from one specific use of a genetic algorithm, this data could still be very subjective depending on data used and the problem encoding.

### Fitness

The second area of a genetic algorithm which will depend on the problem being assessed, the fitness function, will probably be the most complicated part of the algorithm. The assessment of the population can be done in many different ways, it could be that the success is simulated in some way, or that the evaluation may be performance based, representing only an approximation of how fit a solution is[6].

Though the fitness function of a genetic algorithm is very subjective and will reflect the problem, one thing a that will be constant will be the need for speed in processing. Due to what are effectively large population sizes, for which the fitness function must be run for each member of the population, the evaluation must be completed quickly. For example for a population of 200 strings, if evaluation takes an hour, it takes over a year to complete 10000 evaluations. This would only be approximately 50 generations[6].

### Selection

Selection is the process by which the next generation is chosen, those which will then be put through the process of crossover and mutation. Darwin's theory of natural selection says that the best should always be those that survive to create offspring[28], however there is a huge range of methods by which the next generation of chromosomes. There are three well known functions which I will focus on, these are roulette wheel selection, stochastic universal sampling and tournament selection.

Roulette wheel selection, or fitness proportionate selection, if the process by which chromosomes are chosen depending on the fitness of the population. The idea of this type of selection is that a random value is selected between 0 and the sum of the fitness of the population. This random value will refer to the sum of the fitness value up to a certain chromosome in the population[28]. This means that there is a higher chance of selecting chromosomes with a higher fitness value, as they take up more of the total fitness value. More higher fitness value chromosomes will be selected, however more identical chromosomes will be selected. This can mean that the variation in the population will reduce rapidly, which can lead to a local maximum value unless mutation is high enough to counter this reduction in genetic variation. A small population size, or one chromosome with a very high fitness can be a serious problem with this type of selection.

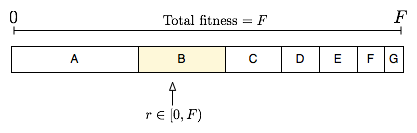


Figure - A visualisation of Roulette Wheel Selection

Stochastic universal sampling is a very similar selection process as roulette wheel selection. Again the process involves calculating the total fitness of the population. However with that value we now divide that total fitness by the amount of chromosome we want to keep, call this value 'n'. By taking a random value between 0 and 'n' we can then sample the population for the amount of chromosomes we need at equal intervals. This selection function gives more possibility of selecting the lower fitness chromosomes, which can in turn keep more genetic variation in the population. This can keep some good parts of a solution which happen to be contained within a less good chromosome in the population for crossover.

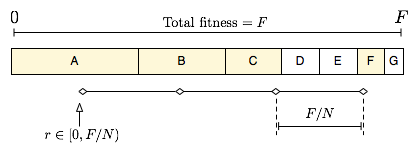


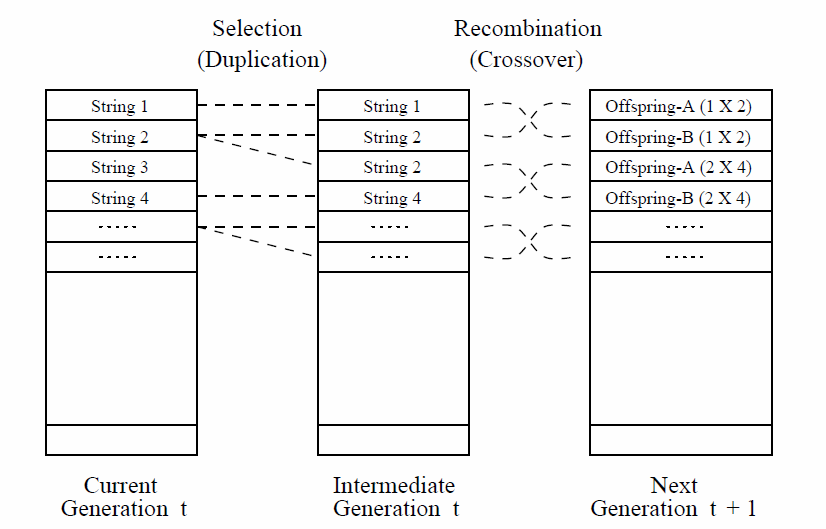
Figure - A visualisation of Stochastic Universal Sampling

The third process of selection, tournament selection is a different process of selection, however is extremely simple. The process involves taking a set number of chromosomes from the population, comparing them, and selecting the chromosome with the highest fitness. The process is repeated until enough chromosomes have been selected[31]. Tournaments of two chromosomes are those most used, however tournaments can be completed of any number of chromosomes.

Another idea which can be involved with the selection process is elitism. This is a process which can increase the success of a genetic algorithm by way of preventing the loss of the best chromosome[28]. Prior to the selection process the best, or a group of the best chromosomes, are added directly to the new population, then the selection process is completed in the normal fashion.

### Genetic Operators

After the selection process there are two genetic operators which are completed, first crossover, followed by mutation.



### Genetic Algorithms for Cryptanalysis

# Analysis

## Z408

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 4 | 6 | 7 | 2 | 8 | 9 | 10 | 11 | 12 | 13 | 11 | 7 |
| 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 1 | 22 | 3 | 23 | 24 | 25 | 26 | 19 | 17 |
| 27 | 28 | 19 | 29 | 6 | 30 | 8 | 31 | 26 | 32 | 33 | 34 | 35 | 19 | 36 | 37 | 38 |
| 39 | 40 | 4 | 1 | 41 | 7 | 3 | 9 | 10 | 42 | 6 | 2 | 43 | 10 | 44 | 26 | 45 |
| 8 | 29 | 46 | 27 | 5 | 28 | 47 | 48 | 49 | 12 | 20 | 22 | 15 | 14 | 17 | 31 | 19 |
| 23 | 16 | 26 | 18 | 36 | 1 | 24 | 30 | 38 | 21 | 26 | 13 | 31 | 37 | 50 | 39 | 40 |
| 10 | 34 | 33 | 25 | 19 | 45 | 44 | 9 | 31 | 26 | 18 | 7 | 32 | 35 | 39 | 41 | 7 |
| 46 | 47 | 4 | 3 | 41 | 7 | 23 | 13 | 26 | 45 | 22 | 27 | 6 | 29 | 10 | 10 | 8 |
| 51 | 5 | 24 | 26 | 12 | 30 | 38 | 14 | 26 | 25 | 31 | 37 | 46 | 27 | 48 | 1 | 41 |
| 7 | 3 | 36 | 10 | 16 | 52 | 11 | 21 | 49 | 34 | 40 | 17 | 45 | 6 | 22 | 8 | 20 |
| 5 | 51 | 12 | 9 | 15 | 14 | 30 | 37 | 16 | 33 | 46 | 38 | 33 | 29 | 10 | 21 | 22 |
| 30 | 1 | 36 | 10 | 53 | 32 | 19 | 48 | 49 | 47 | 17 | 4 | 23 | 13 | 28 | 35 | 42 |
| 3 | 37 | 27 | 1 | 10 | 6 | 33 | 2 | 46 | 38 | 34 | 15 | 45 | 24 | 22 | 11 | 18 |
| 48 | 30 | 25 | 28 | 8 | 37 | 1 | 31 | 46 | 27 | 44 | 34 | 42 | 38 | 5 | 40 | 3 |
| 50 | 6 | 12 | 8 | 42 | 1 | 41 | 7 | 15 | 14 | 49 | 16 | 15 | 32 | 33 | 9 | 3 |
| 29 | 11 | 39 | 48 | 44 | 43 | 6 | 17 | 21 | 54 | 36 | 50 | 18 | 2 | 2 | 30 | 27 |
| 34 | 8 | 38 | 39 | 51 | 45 | 4 | 1 | 2 | 2 | 5 | 43 | 42 | 3 | 41 | 7 | 15 |
| 12 | 17 | 13 | 26 | 14 | 26 | 53 | 20 | 41 | 31 | 51 | 16 | 23 | 1 | 42 | 1 | 7 |
| 2 | 9 | 32 | 37 | 10 | 6 | 51 | 16 | 53 | 47 | 19 | 26 | 53 | 29 | 39 | 26 | 14 |
| 15 | 5 | 17 | 18 | 19 | 24 | 45 | 53 | 32 | 19 | 42 | 1 | 2 | 41 | 46 | 33 | 53 |
| 22 | 25 | 20 | 17 | 13 | 1 | 50 | 13 | 42 | 36 | 47 | 19 | 54 | 46 | 25 | 11 | 26 |
| 53 | 17 | 47 | 41 | 41 | 21 | 17 | 37 | 3 | 9 | 10 | 13 | 35 | 20 | 2 | 18 | 51 |
| 5 | 23 | 28 | 32 | 33 | 26 | 53 | 31 | 28 | 30 | 16 | 48 | 7 | 3 | 35 | 14 | 21 |
| 15 | 45 | 13 | 48 | 1 | 14 | 30 | 21 | 26 | 45 | 22 | 27 | 38 | 11 | 6 | 30 | 8 |

## Z340

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 18 | 5 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| 20 | 34 | 35 | 36 | 37 | 19 | 38 | 39 | 15 | 26 | 21 | 33 | 13 | 22 | 40 | 1 | 41 |
| 42 | 5 | 5 | 43 | 7 | 6 | 44 | 30 | 8 | 45 | 5 | 23 | 19 | 19 | 3 | 31 | 16 |
| 46 | 47 | 37 | 19 | 40 | 48 | 49 | 17 | 11 | 50 | 51 | 9 | 19 | 52 | 53 | 10 | 54 |
| 5 | 44 | 3 | 7 | 51 | 6 | 23 | 55 | 30 | 17 | 56 | 10 | 51 | 4 | 16 | 25 | 21 |
| 22 | 50 | 19 | 31 | 57 | 24 | 58 | 16 | 38 | 36 | 59 | 15 | 8 | 28 | 40 | 13 | 11 |
| 21 | 15 | 16 | 41 | 32 | 49 | 22 | 23 | 19 | 46 | 18 | 27 | 40 | 19 | 60 | 13 | 47 |
| 17 | 29 | 37 | 19 | 61 | 19 | 39 | 3 | 16 | 51 | 20 | 36 | 34 | 62 | 63 | 53 | 31 |
| 55 | 40 | 6 | 38 | 8 | 19 | 7 | 41 | 19 | 23 | 5 | 43 | 29 | 51 | 20 | 34 | 55 |
| 38 | 19 | 3 | 54 | 50 | 48 | 2 | 11 | 25 | 27 | 20 | 5 | 61 | 14 | 37 | 31 | 23 |
| 16 | 29 | 36 | 6 | 3 | 41 | 11 | 30 | 50 | 14 | 53 | 37 | 28 | 19 | 52 | 20 | 51 |
| 40 | 63 | 47 | 42 | 34 | 22 | 19 | 18 | 11 | 50 | 51 | 20 | 36 | 21 | 58 | 44 | 3 |
| 6 | 15 | 51 | 18 | 7 | 32 | 50 | 16 | 53 | 61 | 28 | 36 | 8 | 53 | 48 | 19 | 19 |
| 34 | 20 | 59 | 12 | 30 | 35 | 53 | 47 | 56 | 2 | 4 | 8 | 38 | 39 | 50 | 55 | 19 |
| 11 | 36 | 28 | 45 | 40 | 20 | 31 | 21 | 23 | 5 | 7 | 28 | 32 | 37 | 57 | 15 | 16 |
| 3 | 36 | 14 | 19 | 13 | 12 | 63 | 56 | 29 | 19 | 51 | 6 | 26 | 20 | 11 | 33 | 13 |
| 19 | 19 | 33 | 26 | 56 | 40 | 26 | 36 | 9 | 23 | 41 | 1 | 14 | 54 | 21 | 33 | 5 |
| 11 | 51 | 10 | 17 | 26 | 29 | 43 | 48 | 20 | 46 | 27 | 23 | 20 | 30 | 55 | 56 | 36 |
| 4 | 37 | 25 | 1 | 18 | 5 | 10 | 42 | 40 | 39 | 23 | 44 | 62 | 11 | 31 | 58 | 19 |

# Design

## Overview

Due to the format of a genetic algorithm lending itself to a modular implementation this made dividing the implementation easy. As previously mentioned, the genetic algorithm is made up of 7 major areas, a central management process, a method of generating a population, the chromosomes which make up the population, a method of calculating fitness, a method of selecting the next generation, a method of altering the population in a 'genetic' level and a function to calculate when to stop the algorithm. This gave an easy way of creating a high level design and framework which could then be used to generate multiple implementations of each module, allowing for investigation into differing methods of creating a solution, as well as allowing the application of the same modules to both simple substitution and homophonic substitution ciphers.

## Management

## Population Generation

## Chromosomes

## Fitness

## Selection

## Genetics

## Termination

# Implementation

## Management

## Population Generation

## Chromosomes

## Fitness

## Selection

## Genetics

## Termination

# Testing

## Unit

## Functional

# Investigation Methods

## Substitution

## Z408

## Z340

# Investigation Results

## Substitution

## Z408

## Z340

# Conclusions

## Substitution

## Z408

## Z340

# Critical Evaluation

# Bibliography

|  |  |
| --- | --- |
| [1] | J. M. Carrol and S. Martin, “The Automated Cryptanalysis of Substitution Ciphers,” *Cryptologia,* vol. 10, no. 4, pp. 193-209, 1986. |
| [2] | R. Spillman, M. Janssen, B. Nelson and M. Kepner, “Use of A Genetic Algorithm in the Cryptanalysis of Simple Substitution Ciphers,” *Cryptologia,* vol. 17, no. 1, pp. 31-44, 1993. |
| [3] | B. Delman, *Genetic Algorithms in Cryptography,* Rochester Institute of Technology, Rochester: Unpublished Masters Thesis, 2004. |
| [4] | D. Thá̆ng, “Analysis of the Zodiac 340-Cipher,” Unpublished Master's Thesis, San Jose State University, California, 2008.  As an in depth investigation into the structure of the Zodiac ciphers I used this paper extensively in my research about how to model the ciphers computationally, as well as looking into ways I could take into account the errors made by the zodiac in the ciphers. If I could explain how the mistakes were made I could possibly work around them, or alter my computational models in such a way to make the genetic algorithm succeed. The use of the hill climb algorithm in an attempt to solve a homophonic cipher was not directly related to my work, but the optimisation ideas may have been a useful in attempting methods of optimising the genetic algorithm. |
| [5] | “Feature Driven Development Processes,” [Online]. Available: http://www.nebulon.com/articles/fdd/download/fddprocessesUSLetter.pdf. [Accessed 4 February 2014].  In describing the methods of using FDD this was a very useful document at the beginning of my investigation, however as it became more apparent that FDD was not the right choice of process for a project of this type, it became less useful. |
| [6] | D. Whitley, “A Genetic Algorithm Tutorial,” *Statistics and Computing,* vol. 4, no. 2, pp. 65-85, 1993. |
| [7] | A. Clarke and E. Dawson, “A Parallel Genetic Algorithm for Cryptanalysis of the Polyalphabetic Substitution Cipher,” *Cryptologia,* vol. 21, no. 2, pp. 129-138, 1997. |
| [8] | R. Toemeh and S. Arumugam, “Applying Genetic Algorithms for Searching Key-Space of Polyalphabetic Subsititution Ciphers,” *International Arab Journal of Information Technology,* vol. 5, no. 1, pp. 87-91, 2008. |
| [9] | D. Whitley, “Genetic Algorithms and Evolutionary Computing,” in *Van Nostrand's Scientific Encyclopedia*, 2002. |
| [10] | D. Whitley, “Genetic Algorithms and Neural Networks,” in *Genetic Algorithms in Engineering and Computier Science 3*, 1994, pp. 203-216. |
| [11] | R. A. J. Matthews, “The Use of Genetic Algorithms in Cryptanalysis,” *Cryptologia,* vol. 17, no. 2, pp. 187-201, 1993. |
| [12] | B. L. Miller and D. E. Goldberg, “Genetic Algoritms, Tournament Selection and the Effects of Noise,” *Complex Systems,* vol. 9, no. 3, pp. 193-212, 1995. |
| [13] | P. K. Basavaraju, “Heuristic Search Cryptanalysis of the Zodiac 340 Cipher,” Unpublished Master's Thesis, San Jose State University, California, 2009. |
| [14] | P. A. Diaz-Gomez and D. F. Hougen, “Initial Population for Genetic Algorithms: A Metric Approach,” in *GEM*, 2007, pp. 43-49. |
| [15] | M. O. Odetayo, “Optimal Population Size for Genetic Algorithms: An Investigation,” *Genetic Algorithms for Control Systems Engineering, IEE Colloquium on,* vol. 2, no. 1, 1993. |
| [16] | C. Symons, Solving the Zodiac, 1st ed., C. Symons, 2009.  A concise representation of all of the evidence presented in the original investigations of the Zodiac case. Containing all of the case notes and other police documentation this was a little in depth for my uses, but contained useful information on dates and for cross-referencing information from other sources. |
| [17] | “Zodiac Ciphers Wiki,” 2014. [Online]. Available: http://zodiackillerciphers.com/wiki/index.php. [Accessed April 2014].  Though a community driven resource this wiki contained a large amount of useful information relating to both the case and the ciphers themselves. Much of the information could be cross-referenced with other sources and those sources cited by users were very useful, but the site itself also contained insights not found elsewhere. |
| [18] | T. Atkins, “Substitution Cipher,” 2014. [Online]. Available: http://rumkin.com/tools/cipher/substitution.php. [Accessed 20 April 2014].  A useful tool for creating simple substitution ciphers, includes all the features I required for creating ciphers for both the background section and for use in experimentation. |
| [19] | T. Voigt, “Zodiac Killer,” 1998. [Online]. Available: http://zodiackiller.com/. [Accessed 26 April 2014].  The most complete source of scans of the Zodiac correspondence I could find. The scans were used in both creating computer based models of the ciphers and for illustration in this report. |
| [20] | M. Cole, “Two new theories regarding the Zodiac Case,” 10 August 2003. [Online]. Available: http://mikecole.org/zodiac/two\_theories/1.2/two\_theories.1.2.pdf. [Accessed 21 April 2014].  An Example of where new ideas have come forward more recently about the Zodiac ciphers. Though I did not take into account any of the ideas specifically put down in this article, it was useful in the thought process of how I would attempt the ciphers myself. |
| [21] | C. Christensen, “Simple Substitution Ciphers,” Northern Kentucky University, Cincinnati, 2006.  A useful source of the history and description of simple substitution ciphers, as well as containing much information about the currently used methods of cryptanalysis of this type of cipher. Also contained useful examples of substitution ciphers. |
| [22] | G. J. Simmons, “Encyclopædia Britannica - Substitution Cipher,” Encyclopædia Britannica, Inc., 14 May 2013. [Online]. Available: http://www.britannica.com/EBchecked/topic/571059/substitution-cipher. [Accessed 18 April 2014].  A brief, simple definition of a substitution cipher, useful as a confirmed source of the information, as many other sources do not cite for this information. |
| [23] | “What is the frequency of the letters of the alphabet in English?,” Oxford University Press, [Online]. Available: http://www.oxforddictionaries.com/words/what-is-the-frequency-of-the-letters-of-the-alphabet-in-english. [Accessed 28 April 2014].  A reputable source for the figures of letter frequency in the English language. |
| [24] | P. Norvig, “Natual Language Corpus Data,” 13 March 2012. [Online]. Available: http://norvig.com/ngrams/. [Accessed 27 March 2014].  A source of large scale examples of bi and tri-grams for the English language from use as a spell checking example for the English language. |
| [25] | “N-gram tutorial,” Missouri Uniersity of Science and Techonology, [Online]. Available: http://web.mst.edu/~tauritzd/ngram/tutorial.html. [Accessed 24 March 2014].  A clear description of the definition and major uses of n-grams. |
| [26] | F. A. Stahl, “A homophonic cipher for computational cryptography,” in *National Computer Conference and exposition*, New York, 1973.  A useful paper on the creation of homophonic ciphers for use in computational cryptography. An in depth look at the history and creation of the homophonic cipher was a useful addition to my knowledge about this type of cipher. As well as having examples of cipher creation using both the Latin alphabet and in binary. The use of binary to create homophonic ciphers could have been useful if I had modelled the Zodiac ciphers in a different fashion. |
| [27] | “Who is the “Concerned Citizen”?,” 29 December 2012. [Online]. Available: http://www.zodiackillerciphers.com/?p=224. [Accessed 29 March 2014].  A blog post containing information about the lesser known solver of the Z408 cipher. As I used the Harden cipher solution in my investigation, though interesting this was mainly a side note in my research. |
| [28] | M. Obitko, “Introduction to Genetic Algorithms,” 1998. [Online]. Available: http://www.obitko.com/tutorials/genetic-algorithms/index.php. [Accessed 17 January 2014]. |
| [29] | E. K. Burke, S. Gustafson and G. Kendall, “Diversity in genetic programming: An analysis of measures and correlation with fitness,” *IRRR Transactions on Evolutionay Computation,* vol. 8, no. 1, pp. 47-62, 2004. |
| [30] | F. G. Lobo and C. F. Lima, “A review of adaptive population sizing schemems in genetic algorithms,” in *Proceedings of the Genetic and Evolutionary Computation Conference*, Washington, D.C., 2005. |