ANALYSIS OF THE ZODIAC 340-CIPHER

A Project Report
Presented to
The Faculty of the Department of Computer Science
San Jose State University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

By Thang Dao
December 2007

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

Professor Mark Stamp

Professor Michael Beeson

Dr. Stanley Herwitz (SAVDS Inc.)

APPROVED FOR THE UNIVERSITY

ABSTRACT

ANALYSIS OF THE ZODIAC 340-CIPHER
by Thang Dao

Computers have advanced to the stage that an inexpensive personal computer can perform millions of arithmetic calculations in less than a second. However, given such powerful machines, several mysteries have still remained unsolved due to their complexity. The Zodiac 340-letter cipher is one such mystery [1].

Zodiac, the serial killer who terrorized Northern California in the late 1960s, sent four ciphers to local newspapers [2]. The first cipher was separated into three different parts and each part was sent to Vallejo Times-Herald, the San Francisco Chronicle, and the San Francisco Examiner. The combination of all three parts formed a 408-letter cipher (Z408), which was decrypted one week after it was received [1,2]. Zodiac also sent a 340-letter cipher (Z340) that remains unsolved to this day.

The main purpose of this project is to determine whether the method used in the Z340 was a homophonic substitution, which is an improved version of the well-known simple substitution cipher. A homophonic substitution employs a "one-to-many mapping" technique, as opposed to the one-to-one mapping of a simple substitution [3]. Due to the complexity of the homophonic substitution, an exhaustive solution to the Z340 is not possible in a feasible amount of time. I propose an approach to implement an automated solution to a homophonic substitution based on a hill-climb technique [4]. My software package will be used to attempt to solve the Z340. Even if the software fails to solve the Z340, useful conclusions could be drawn. The objective is to reduce the number of methods that could have been used to encrypt the message. This research project also would provide evidence that the Z340 may not be a legitimate ciphertext message.

ACKNOWLEDGEMENTS

I would like to my gratitude to the followings persons. Without them, my research study could never have been completed:

My advisor, Professor Mark Stamp, for his resources, invaluable insights, and patience.

My committee member, Dr. Stanley Herwitz, for his motivation and editing skills.

My committee member, Professor Michael Beeson, for his useful comments.

And especially, to my lovely wife, Trannie Dao, for her encouragement, motivation, patience, and support.

TABLE OF CONTENTS

LIST OF TABLES, LISTINGS, AND FIGURES

Tables:

Listing:

Figures:

1. Introduction

1.1. The problem

The capability to perform millions of arithmetic calculations in short time periods has value in virtually every aspect of modern life in the industrial world. Three-dimensional video gaming, real-time simulation, autonomous flight of unmanned aerial vehicles, economic analysis, military applications, crime solving, and DNA research all involve complex arithmetic calculations that can now be performed using relatively inexpensive personal computers.

In the world of crime solving, the challenge of gaining an understanding of the criminal mind is a fundamental challenge. Some criminals leave no trace of their criminal activity, while other criminals leave a trace and try to challenge the minds of law enforcement detectives as well as the general public. An example of a criminal leaving a "trace" is a criminal that creates a cipher. Ciphers are defined as messages written in a secret code.

One specific example is the creator of the Zodiac 408-cipher (Z408), which was first sent to a local newspaper with a cover letter explaining that he was an actual killer. In the case of the Z408, the message was decrypted to provide insight into the twisted mind of a killer who was involved with a series of murders.

The creator of the Z408, however, also created several subsequent ciphers, which have yet to be decoded. The objective of my research was to develop a software program for operation on a standard PC as part of an effort to decode this unsolved mystery. The focus of my research was on one of the subsequent ciphers; specifically, Zodiac 340-cipher (Z340).

1.2. Prior work

Code-breakers have attempted to solve the Z340 for the past forty years with no success. During these forty years, code-breakers have investigated the Z340 from multiple perspectives: (1) as a homophonic cipher similar to the Z408; (2) as a polyalphabetic cipher, an improvement from the Z408 encryption method; (3) as a double transposition columnar, another improvement from the Z408 encryption method; and (4) as a one-time pad, the unbreakable encryption method when used properly. All of these methods have failed to deliver any meaningful conclusion. More recently, several new investigative approaches have been proposed such as: (1) the Z340 is actually a completely meaningless message; (2) the Z340 was written backwards; (3) the Z340 was written using the Zodiac circle that was segmented into 12 equals 30-degree slices [5]; or (4) the Z340 was written in a rotation of 90^{o}, 180^{o}, or 270^{o}. These investigative approaches are available for general use [6]. None of these approaches, however, have successfully decoded Z340.

1.3. Proposed solution

The premise of my research is that Zodiac did not develop a completely new system for the Z340. My research study, therefore, focuses on the homophonic substitution. I contend that it would be unlikely for Zodiac to have employed the one-time pad, the double transposition columnar or any polyalphabetic substitution method given the complexity of these methods. On the other hand, although the homophonic substitution does not provide the same level of security as the one-time pad, the double transposition columnar or the polyalphabetic substitution, the homophone substitution still may be considered very difficult to decrypt.

My proposed solution is to deliver an automated software to decrypt homophonic substitution ciphers based on the hill-climb algorithm. The Z340 is treated as a homophonic substitution cipher. The automated software developed for this research project was used to decrypt the Z340.

2. Zodiac Ciphers

2.1. Introduction to Zodiac ciphers

The creator of the Zodiac ciphers referred to himself as the Zodiac. During his time, Zodiac sent four ciphers to local newspapers. Zodiac sent four different ciphers over the period July 31, 1969 to June 26, 1970. Zodiac always included a cover letter with his ciphers to voice his demands and challenges to the general public.

The first cipher was sent to three different local newspapers: the Vallejo Times-Herald, the San Francisco Chronicle, and the San Francisco Examiner. The other three ciphers were all sent to the San Francisco Chronicle. The key point is that only the first cipher (i.e., the Z408) was successfully decoded. The other three ciphers remain unsolved. Among the remaining three unsolved ciphers, only the Z340 has gained special interest from code-breakers. Code-breakers have never attempted to decode the other two unsolved ciphers using conventional encryption methods due to the brevity of the cipher messages. The 3rd cipher message included only 13 cipher letters (Z13) while the 4th cipher included only 32 cipher letters (Z32).

The three parts of the Z408 are shown in Section 2.2.2, 2.2.3, and 2.2.4, respectively. The Z340 is shown in Section 2.3.2. The Z13 and Z32 are shown in Section 2.4.2 and 2.5.2 respectively. The cover letters of all four ciphers are shown in Appendix C.

2.2. Z408

2.2.1. Z408 time line

On July 31, 1969, Zodiac sent his first cipher: the famous three-part cipher Z408. The

Z408 was separated into three different parts; The first part was sent to Vallejo Times-Herald; the second part was sent to the San Francisco Chronicle; and the third part was sent to the San Francisco Examiner. On August 8, 1969, one week after the Z408 was published on these three local newspapers, the Z408 was decrypted by Donald and Bettye Harden, residents of Salinas California. The Z408 was sent to the local newspapers to take credit for the shooting deaths of two individuals at Lake Herman Road and two other individuals at Blue Rock Springs Golf Course.

2.2.2. Z408 part 1, sent to Vallejo Times-Herald

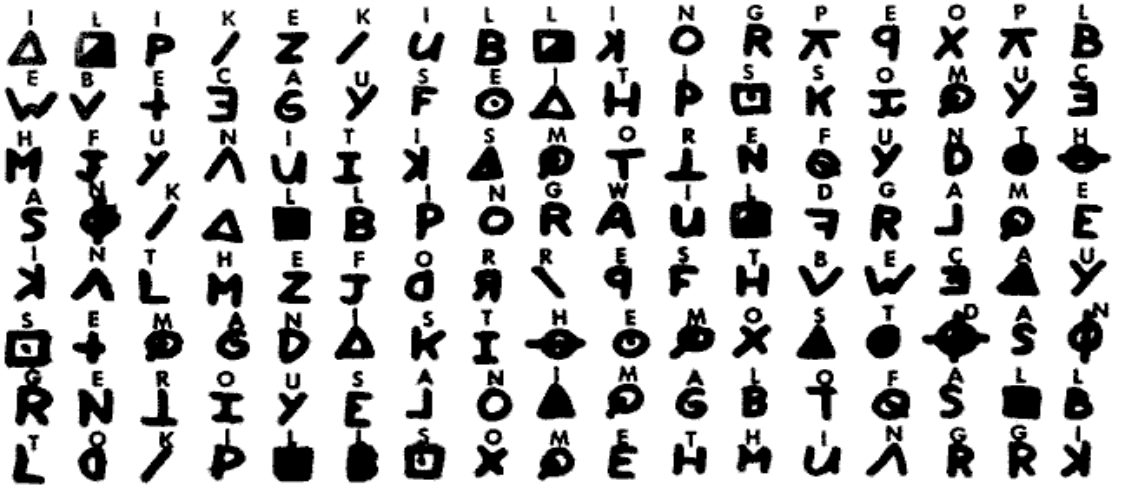


Figure 1: Z408 Part 1 [1]

2.2.3. Z408 part 2, sent to San Francisco Chronicle

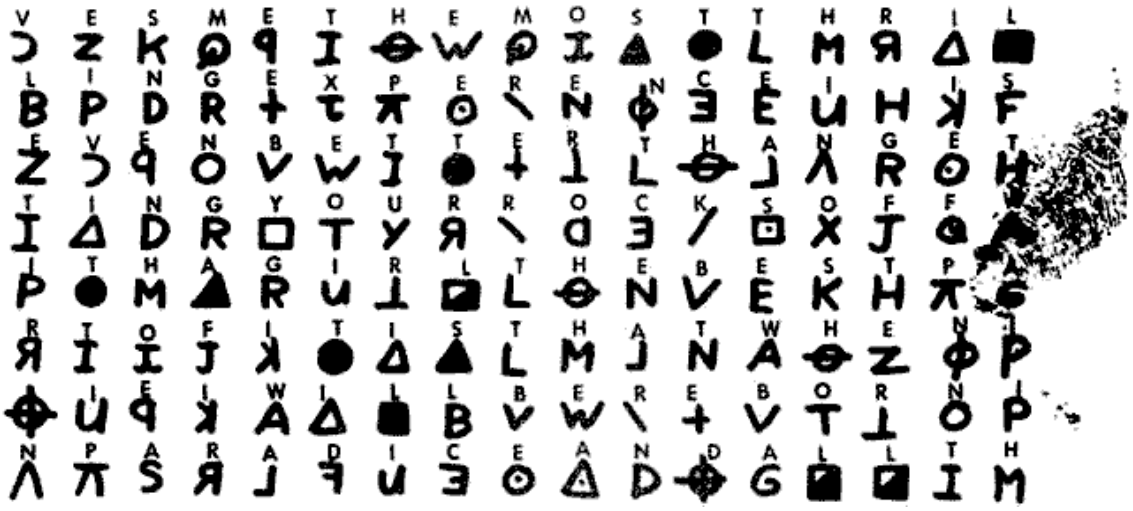


Figure 2: Z408 Part 2 [1]

2.2.4. Z408 part 3, sent to San Francisco Examiner

Figure 3: Z408 Part 3 [1]

The message read:

"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 dangerous 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 that when I die I will be reborn in paradice and all the I have killed will become my slaves I will not give you my name because you will try to slow down or stop my collecting of slaves for my afterlife"

2.3. Z340

2.3.1. Z340 time line

Three months later, on November 8, 1969, Zodiac sent Z340, his second cipher to the San Francisco Chronicle. Code-breakers have not been able to produce any meaningful details using some of the following potential encryption conventions: one-time pad; double transposition columnar; polyalphabetic substitution; and homophonic substitution. Several other possible solutions, which disregarded all encryption conventions, have been proposed. An example of one of these unconventional solutions was proposed on May 22, 2007 by Christopher Farmer, MS, National Security [7]. Mr. Farmer relied heavily on the Japanese play "Mikado" in his argument. His solution, however, is more of an argumentative solution, rather than a logically proven solution. None of the proposed solutions, including Mr. Farmer's solution, have ever been verified.

2.3.2. Z340, sent to the San Francisco Chronicle.

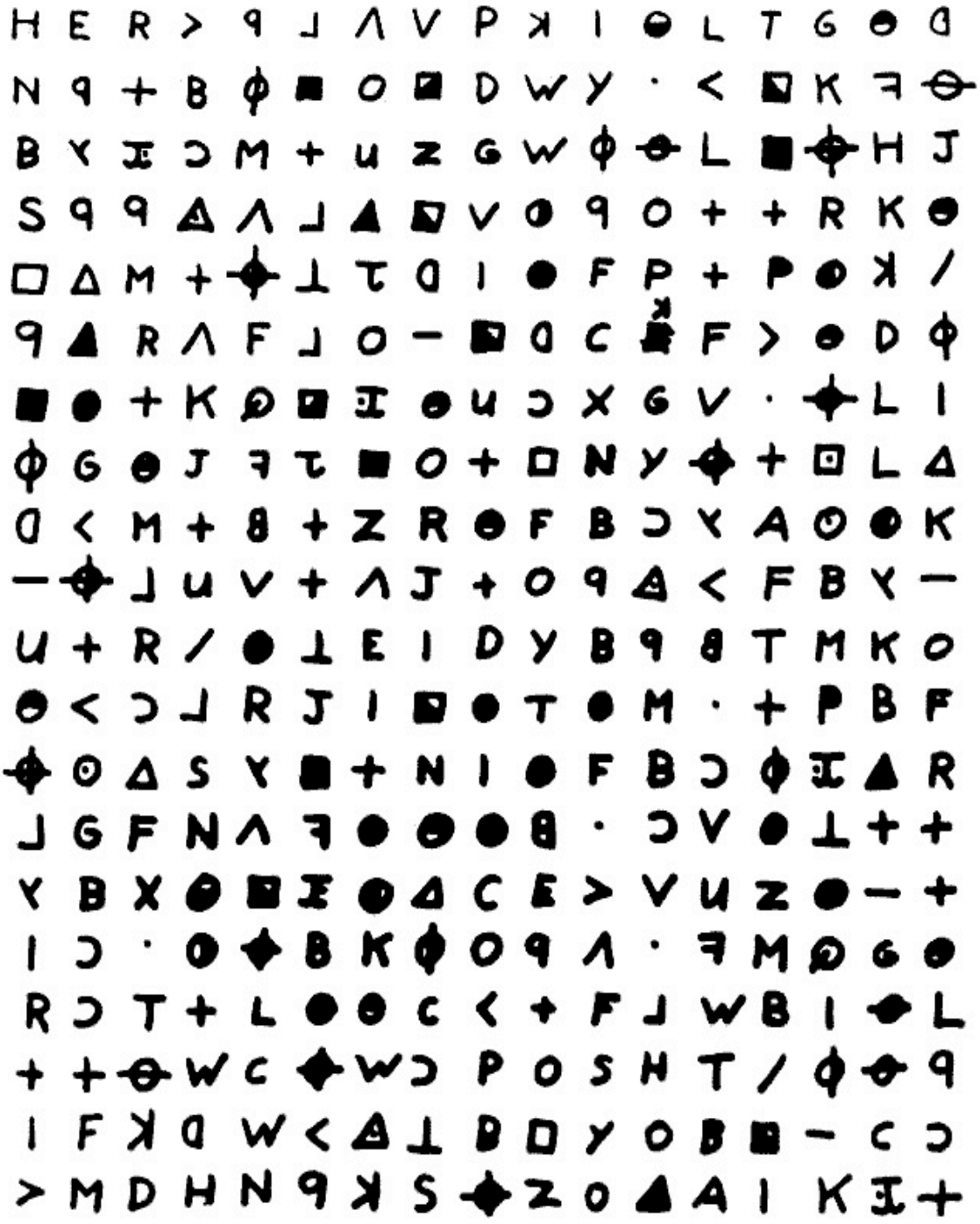


Figure 4: Z340 [1]

2.4. Z13

2.4.1. Z13 time line

On April 20, 1970, Zodiac sent a letter to the San Francisco Chronicle. The letter included 13 cipher letters (Z13). This letter was considered to be Zodiac's attempt to reveal his name. Due to the brevity of this cipher, no attempt to solve this letter has ever been proposed using any conventional encryption method. As was the case for the Z340, several possible solutions have been proposed. Mr. Farmer, using his numerological creative thinking, had hinted at some potential clues such as "A Train 8 Blvd;" A Train H Blood M;" "813 Mt. Diablo Blvd;" and "Mt. Diablo CT Street." [8]. None of these proposed translations, however, have ever been verified.

2.4.2. Z13, sent to San Francisco Chronicle

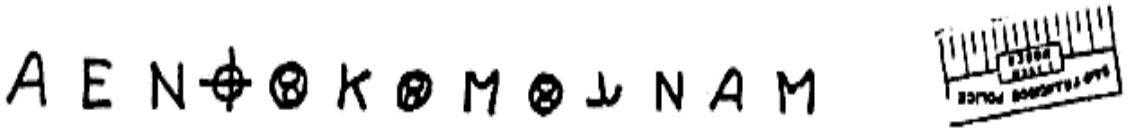


Figure 5: Zodiac's "My name is ..." cipher [1]

2.5. Z32 – Zodiac 4th cipher

2.5.1. Z32 time line

On June 26, 1970, Zodiac sent his last cipher letter to the San Francisco Chronicle. In the letter, Zodiac evidently was upset because no one followed his demand of wearing Zodiac buttons (see Appendix C). Zodiac also took credit for the murder of Sgt Richard Radetich on June 19, 1970.

2.5.2. Z32, sent to San Francisco Chronicle

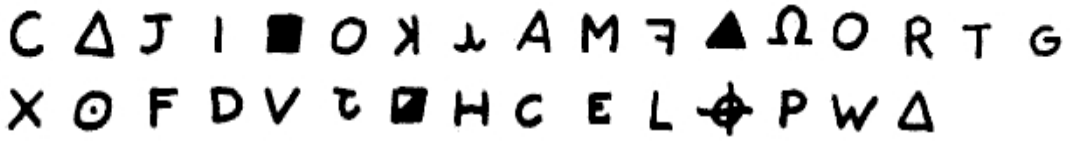


Figure 6: Zodiac's "Button" cipher [1]

3. Analysis of Zodiac Ciphers: Z408 and Z340

3.1. Z408

3.1.1. Z408 decryption method

The Z408, which was sent on August 1st, 1969, was encrypted as a homophonic substitution cipher. One week later, on August 8th, 1969, Donald and Bettye Harden, residents of Salinas California, successfully decrypted the cipher [2].

The Harden's method was based on their deduction that the message would contain the words “kill” and “I.” The reasoned that a killer having the capability to create a cipher message would have a significant ego. The mappings of the words “kill” and “I” were used to successfully decrypt the cipher.

3.1.2. Z408 cipher alphabet

For my analysis, I translated the cipher symbols into a numeric form. Table 1 is a matrix that corresponds to the Z408. For example, in the first column of Z408, the symbol corresponds to the number 1, while the symbol and the symbol correspond to the number 14 and the number 27 respectively.

Table 1: Z408 in numeric form

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

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

3.1.3. Z408 plaintext – ciphertext mappings

Table 2: Z408 plaintext-ciphertext mappings

Letter	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**
Count	24	9	10	17	52	11	12	16	43	0	6	33	16
Mapping	1, 18, 31, 39, 43, 49	15	17	42, 50	5, 12, 14, 16, 21, 34, 44	21, 28, 35	10	27, 38	1, 3, 6, 8, 19		4	2, 7, 52	26

Letter	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
Count	23	28	7	0	19	23	35	10	6	9	1	8	0
Mapping	9, 29, 36, 40	13, 25, 30, 32, 46	11		33, 47, 48	20, 23, 24, 31, 44, 49	22, 25, 30, 34, 37, 45	19	51	1, 41	28	53	

3.1.4. Z408 known errors

Table 3 shows the inconsistency in the cryptogram and how some of the mappings overlap. I hypothesize that the errors, along with the misspellings and the last meaningless eighteen characters, were intentional to make the analysis more difficult.

Table 3: Z408 inconsistency

Error #	Zodiac alphabet #	English Letter Mapping	First appearance
1	1 	A	111
		I	1
		W	345
2	19 	I	406
		U	23
3	21 	E	25
		F	116
4	25 	O	31
		T	271
5	28 	F	36
		X	159
6	30 	O	106
		T	40
7	31 	A	84
		S	42
8	34 	E	46
		T	233
9	44 	E	68
		S	108

3.1.5. Z408 letter frequencies and theoretical keyspace

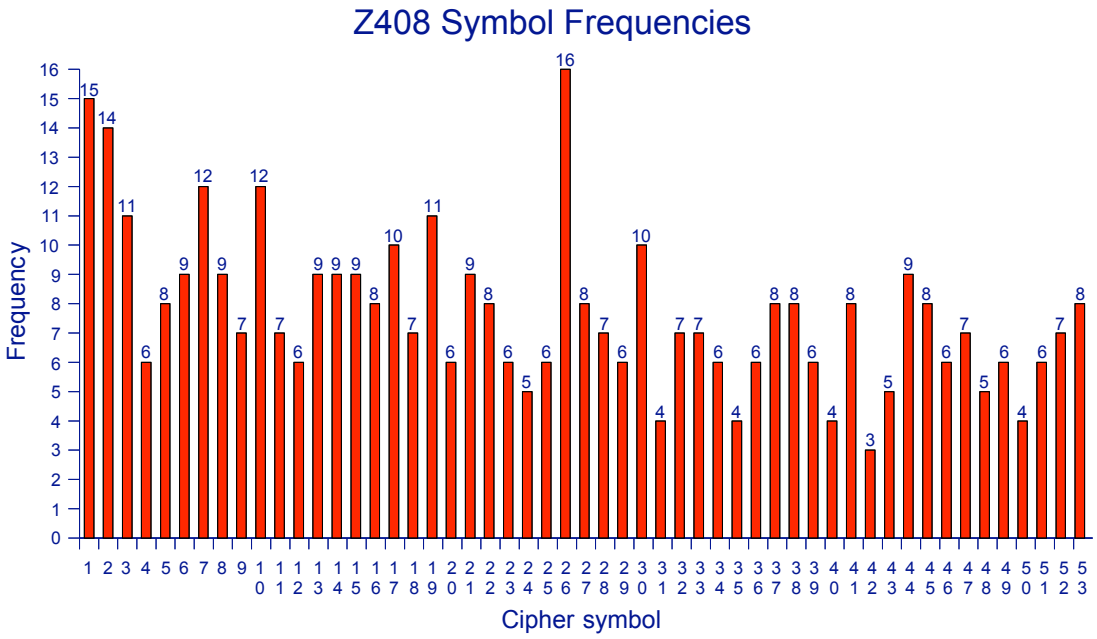


Figure 7: Z408 symbol frequencies

The total keyspace for the this cipher is $26^{408} \sim 2^{4.7 \times 408} = 2^{1917.6}$

3.2. Z340

3.2.1. Z340 cipher alphabet

Table 4 is a matrix that corresponds to the Z340 using the same approach as the table 1. For example, in the first column of Z340, the symbol corresponds to the number 1, while the symbol and the symbol correspond to the number 18 and the number 20 respectively.

Table 4: Z340 in numeric form

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
18	05	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	9	52	10	53
5	44	3	7	51	6	23	54	30	17	55	10	51	4	16	25	21

22	50	19	31	56	24	57	16	38	36	58	15	8	28	40	13	11
21	15	16	41	32	49	22	23	19	46	18	27	40	19	59	13	47
17	29	37	19	60	19	39	3	16	51	20	36	34	61	62	52	31
54	40	6	38	8	19	7	41	19	23	5	43	29	51	20	34	54
38	19	3	53	50	48	2	11	25	27	20	5	60	14	37	31	23
16	29	36	6	3	41	11	30	50	14	50	37	28	19	9	20	51
40	62	47	42	34	22	19	18	11	50	51	20	36	21	57	44	3
6	15	51	18	7	32	50	16	50	60	28	36	8	50	48	19	19
34	20	58	12	30	35	52	47	55	02	04	08	38	39	50	54	19
11	36	28	45	40	20	31	21	23	05	07	28	32	37	56	15	16
3	36	14	19	13	50	16	55	29	19	51	6	26	20	11	33	13
19	19	33	26	55	40	26	36	9	23	42	1	14	53	21	33	5
11	51	10	17	26	29	43	48	20	46	27	23	20	30	54	55	36
4	37	25	1	18	5	10	42	40	39	23	44	61	11	31	57	19

3.2.2. Z340 letter frequencies and theoretical keyspace

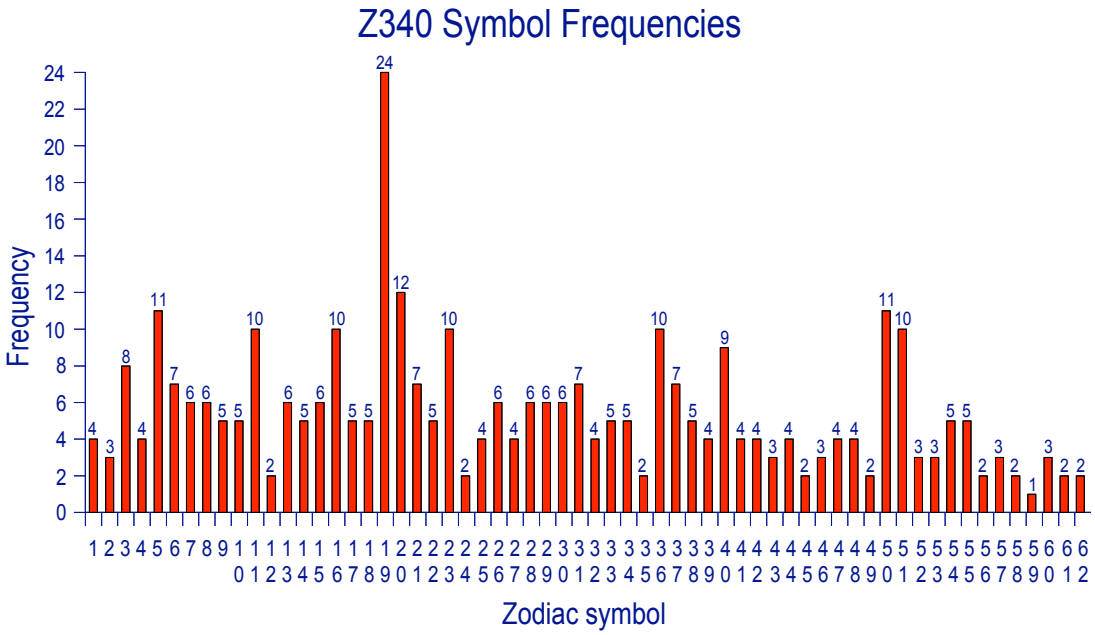


Figure 8: Z340 symbol frequencies

The total keyspace for the this cipher is $26^{340} \sim 2^{4.7 \times 340} = 2^{1598}$

3.3. Comparison between Z408 and Z340

3.3.1. Cipher class

Despite the inconsistency in its mapping, the Z408 employed the technique of homophonic cipher, which is an improved version of the classic simple substitution. I discuss homophonic substitution ciphers in Section 5.

The Z408 may have been considered too difficult to break from Zodiac's perspective because the total number of possible keys was so high. On the other hand, Z408 was decrypted only a week after the Zodiac sent the cipher to the local newspapers [2]. When Z340 was sent three months later, the symbols were similar to the previously decrypted Z408-cipher; however, it was not clear whether Zodiac had reused the homophonic substitution technique. Other possible encryption methods included, but were not limited to, one-time pad, double transposition columnar and polyalphabetic substitution.

3.3.2. Zodiac alphabet and a special character

When Z408 and Z340 are compared, it is evident that Zodiac introduced an additional 9 symbols. The number of symbols increased from 53 symbols in Z408 to 62 symbols in Z340. In addition, the symbology was further confused by using the symbol + (corresponding to our numeric "19") 24 times in the Z340, which is two times greater than the frequency of the 2nd most frequently used symbol (corresponding to numeric our "20").

In homophonic substitution, the frequency of each letter is balanced in order to prevent any statistically-based approach. In addition, in his earlier Z408, there was no such standout symbol. The most frequently used symbol (corresponding to our numeric "26"), which appeared 16 times, was only used 1 time more than the 2nd most frequently used symbol + (corresponding to our numeric "1"). In this analytical framework, the symbol (corresponding to our + numeric "19") is one of the main challenges in deciphering Z340. Some of the hypotheses regarding symbol include: (1) the notion that symbol corresponded to the space between words, and (2) the possibility that the symbol was a deliberate insertion of a meaningless character. None of these hypotheses have been validated. In my research study, I assumed the special symbol (corresponding + to the our numeric "19") was one of the English twenty-six letters.

4. Z340 Possible Encryption Methods

4.1. Rationale for selecting possible encryption methods

In the 1960s, more advanced and sophisticated encryption methods using the power of a computer such as a Data Encryption Standard (DES, 1976) or Advanced Encryption

Standard (AES, 2001) were not available. In addition, Zodiac did not have access to powerful machines such as Enigma, SIGABA, Typex, or Purple to employ more complicated mechanical encryption techniques. The available encryption methods, which were possible using a manual pencil-and-paper approach, were: (1) simple substitution; (2) advanced methods based on simple substitution (homophonic substitution and polyalphabetic substitution); (3) double transposition columnar; and (4) one-time pad.

Of these four methods, the simple substitution is the easiest to employ, but it also is the easiest to break. If a simple substitution was employed for an English message, the cipher alphabet would consist of 26 cipher letters. The Z340, in contrast, consists of 62 cipher letters. For this reason, the simple substitution is not not a valid encryption method for the Z340.

Another possibility would be the double transposition columnar method. The transposition encryption involves changing the order of the cipher letters. In the decryption process, the reverse operation is performed. In a double columnar transposition, the original message is written out in a matrix form that is a series of rows with fixed length (i.e., fixed number of columns). The order of rows and columns are changed based on predefined keys. The length of these keys corresponds to the number of rows and columns. The corresponding ciphertext is then written out column by column.

In the double transposition columnar method, the message letters are only changed in terms of their sequential order; however, the number of letters in the original alphabet and the cipher alphabet remain the same. As is the case for the simple substitution method, the English alphabet only has 26 letters, while the Z340 cipher alphabet consists of 62 cipher letters. I contend that the double transposition columnar is not a valid encryption method for the Z340.

The list of real candidates of Z340 includes homophonic substitution, polyalphabetic substitution, and one-time pad. However, given the complexity of polyalphabetic substitution and one-time pad, I contend that it would be unlikely for Zodiac to have employed these two methods. My research study, therefore, focuses on the homophonic substitution.

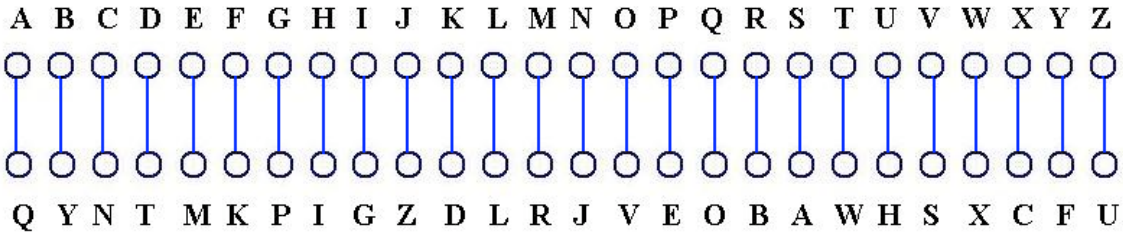
To understand the homophonic substitution and polyalphabetic, an understanding of simple substitution and its weaknesses is required. A detail explanation of simple substitution and its weaknesses is shown in Section 4.2. One-time pad encryption method is described in detail in Section 4.3.1. Polyalphabetic substitution encryption method is described in Section 4.3.2. The description and detail analysis on the strengths and weaknesses of homophonic substitution encryption method are shown in Section 5 and 6.

4.2. Simple substitution

4.2.1. Definition

The simple substitution class of cipher operates on single letter in a message. Each plaintext letter is mapped to one and only one ciphertext symbol. No ciphertext symbol is mapped to by two or more plaintext letters. Figure 9 is an example of a simple substitution mapping:

Plaintext



Ciphertext

Figure 9: An example of a simple substitution mapping

With lowercase letters representing the plaintext letters and uppercase letters representing the ciphertext letters, the alphabets shown in Figure 9 can be written as:

+ Plaintext alphabet: abcdefghijklmnopqrstuvwxyz

+ Ciphertext alphabet: QYNTMKPIGZDLRJVEOBAWHSXCFU

Using the ciphertext alphabet, the message

"the quick brown fox jumps over the lazy dog"

can be encrypted as

"WIM OHGND YBVXJ KVC ZHREA VSMB WIM LQUF TVP"

In order to decrypt the ciphertext, the reverse operation should be performed. For example, the ciphertext message

"AQJ ZVAM AWQWM HJGSMBAGWF"

can be decrypted as

"san jose state university"

However, a simple substitution is vulnerable to multiple methods of decryption. The two most effective approaches to determine the mappings involve using: (1) a large dictionary; or (2) statistics [9,10]. Although simple substitution is easy to use, it also is easy to break. I will discuss the dictionary-based attack on Section 4.2.2. and the statistics-based attack on Section 4.2.3.

4.2.2. Dictionary-based attacks

4.2.2.1. Overview

The method takes advantage of an existing English dictionary. Typically, a dictionary used for cryptanalysis contains approximately, but is not limited to, 100,000 to 120,000 words. This approach is to use trial and error on some consecutive letters of the ciphertext (i.e., from letter i to letter j, $C_{i..j}$). The attack method searches for similar words in the dictionary that have similar pattern to $C_{i..j}$.

For example, when the attacker sees in the ciphertext the 5-letter word *ADLLE,* his job is to decide the correct word in the dictionary that can replace the 5-letter cipher word *ADLLE*. The attacker needs to try all possible words in the dictionary and decide the best matched word. For example, some possible words in the dictionary that can replace *ADLLE* are *HELLO*, *DENNY*, or *LARRY*. The attack continues until no more possible words can be found. The most suitable matched decrypted text is recorded as the most likely candidate for the original message.

It is possible that not all of the cipher letter is decrypted. If the most suitable matched decrypted text is in proper English, the attacker's job is done; otherwise, the attacker will need to manually edit the decrypted text to see if it makes sense or to retrieve some information (e.g., some particular mappings between cipher letter – English letters). The attack would be restarted using those mappings as hints.

4.2.2.2. Disadvantages

This attack method assumes the cipher being mono-alphabetic. If a cipher is a poly-alphabetic substitution, this attack method cannot re-use any of its successful mapping in the attacking process because polyalphabetic substitution employs "many-to-many" technique. The "many-to-many" mapping technique allows a cipher letter to represent many plaintext letters while a plaintext letters can also be represented by many cipher letters.

Another disadvantage of the dictionary-based attack would be the misspelled English words. The attack itself relies heavily on the existing dictionary. Thus, if the cipher contains non-standard or deliberately misspelled words (e.g., EXPEREENCE instead of

EXPERIENCE), the attack would fail to produce a sensible and readable corresponding plaintext.

4.2.3. Statistics-based attacks

The attack method takes advantage of English letter frequencies. In the language language, the letter E is frequently used, while the letter X or Z is relatively rare. For this reason, cryptanalysts often use these frequencies to assist the analysis of the ciphertext.

Cryptanalysts usually focus on the result of the combination of two and three consecutive letters (bigram and trigram). For example, some of the most frequently used combination of two consecutive letters are "*he*", "*nt*", "*of*", and "*io*" while some of the most frequently used combination of three consecutive letters are "*the*", "*ing*", "*ion*", and "*nce*".

4.3. Z340 possible encryption methods

4.3.1. One-time pad

Several improved versions of simple substitution have been created to increase the security of the ciphers. The most secure method is to use a one-time pad in which the key is only used one time. Without knowing the correct key, the ciphertext is theoretically unbreakable as the ciphertext could have different meanings based on different keys.

Cryptanalysis of ciphertext encrypted with one-time pad becomes possible only when the key has been reused. The disadvantage of one-time pad is that the key itself has the length of the original message, and the key has to be transported securely before the person receiving the message can decrypt it. Although the use of one-time pad has this disadvantage, it was used by the Soviet Union before, during, and after World War II. The United States and the British secret code-breakers, however, were able to decrypt and translate these secret messages from the Soviet Union (i.e., VENONA Project) because the Soviets mistakenly used the key more than once [11]. A simple example taken from the book *Information Security: Principles and Practice* is described follow to demonstrate the importance of the key in the one-time pad decryption [12]. Suppose the following letters are encoded to:

e=000 h=001 i=010 k=011 l=100 r=101 s=110 t=111

and the encryption is:

ciphertext = plaintext $\oplus$ key

thus the decryption is:

plaintext = ciphertext $\oplus$ key

For example, the message "*heilhitler"* can be decrypted using the key "*trsrtlerse"*

	h	e	i	l	h	i	t	l	e	r
Plaintext	001	000	010	100	001	010	111	100	000	101
Key	111	101	110	101	111	100	000	101	110	000
Ciphertext	110	101	100	001	110	110	111	001	110	101
	s	r	l	h	s	s	t	h	s	r

When decrypting with correct key "*trsrtlerse"*, the correct message will be retrieved

	s	r	l	h	s	s	t	h	s	r
Ciphertext	110	101	100	001	110	110	111	001	110	101
Key	111	101	110	101	111	100	000	101	110	000
Plaintext	001	000	010	100	001	010	111	100	000	101
	h	e	i	l	h	i	t	l	e	r

However, when decrypting with a different key "*rtsrtlerse*", the original message has a totally different meaning:

	s	r	l	h	s	s	t	h	s	r
Ciphertext	110	101	100	001	110	110	111	001	110	101
Fake Key	101	111	000	101	111	100	000	101	110	000
Plaintext	011	010	100	100	001	010	111	100	000	101
	k	i	l	l	h	i	t	l	e	r

4.3.2. Polyalphabetic substitution

Another significant improvement is the polyalphabetic substitution. This class of substitution uses multiple substitution alphabets instead of simple substitution. It means that each plaintext letter can be represented by any symbol in the cipher alphabet. In addition, each symbol in the cipher alphabet can be mapped to different letters in the plaintext. Although polyalphabetic substituting provides the necessary security on plaintext, it also is difficult to use.

A famous simplified version of polyalphabetic substitution is the Vigenère cipher, which is based on the classic Caesar ciphers [12]. Each plaintext letter is shifted down N places

in the alphabet (e.g., if N = 3, then A → D, B → E, C → F).

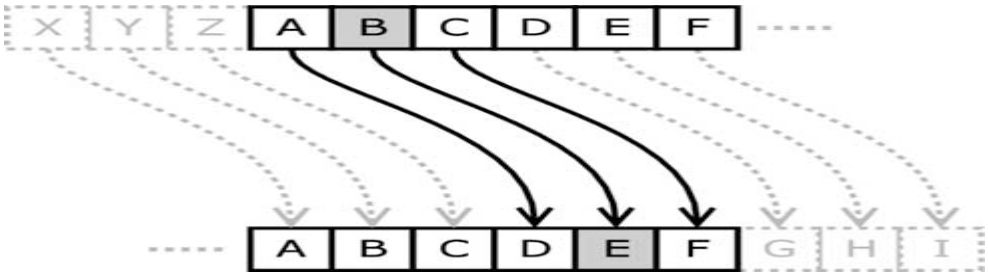


Figure 10: An example of Caesar cipher [13]

To use Vigenère cipher, the keyword is used and repeated until it matches the length of the text. Each letter of the keyword is one Caesar cipher. Encrypting a message using Vigenère cipher means encrypting the message using a series of different Caesar ciphers. An example with keyword "CIPHER" is set forth below:

Plaintext	A	T	T	A	C	K	A	T	D	A	W	N
Key	C	I	P	H	E	R	C	I	P	H	E	R
Ciphertext	C	B	I	H	G	B	C	B	S	H	A	E

5. Description of Homophonic Substitution Cipher

The homophonic substitution method employs the "one-to-many mapping" technique. This technique means that means a plaintext letter can be mapped to multiple symbols in the cipher alphabet, while a symbol in the cipher alphabet can only represent one letter in the plaintext alphabet. This technique increases the number of symbols in the cipher alphabet and balances the frequencies of all the symbols in the cipher alphabet. As a result, a "brute force attack" is not possible due to the dramatically increased keyspace. Moreover, approaches using dictionary and normal frequency analysis are simply not practical. Figure 11 shows an example of homophonic substitution mappings.

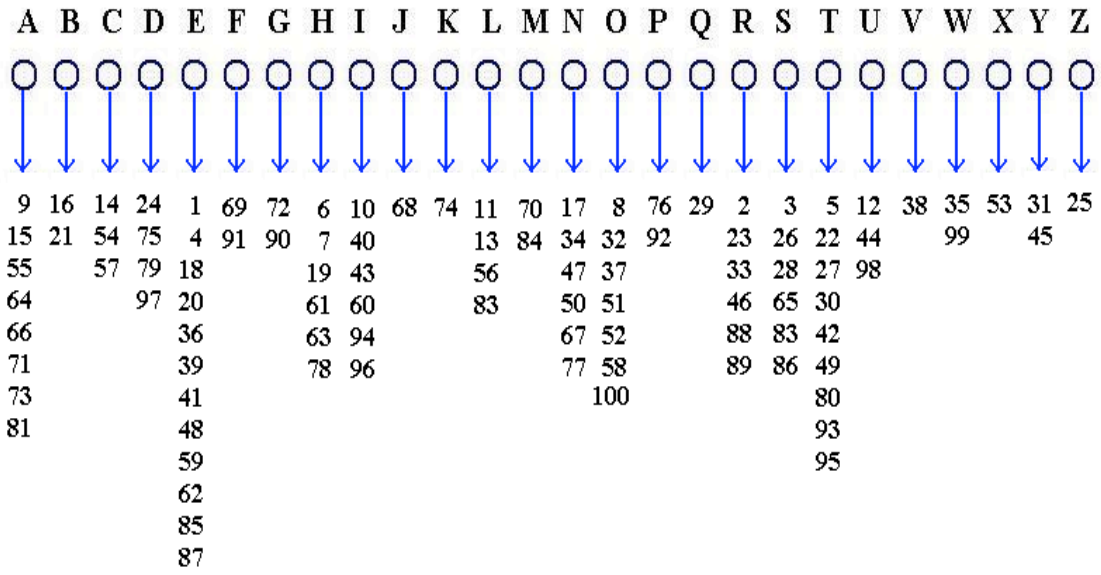


Figure 11: An example of a homophonic substitution

Given the increased number of letters in the alphabet, it is generally agreed that it is easier to represent the ciphertext symbols as numbers. In the example shown in Figure 8, twenty-six English letters are mapped into 100 ciphertext numbers ranging from 1 to 100. The message

"The quick brown fox jumps over the lazy dog"

now can be encrypted in multiple ways. Five possible ways of encrypting this message are as follows.

Possibility #1:
5 7 36 29 44 43 54 74 69 58 53 68 12 84 76 28 51 38 1 46 30 63 85 11 66 25 45 24 52 90

Possibility #2:
27 61 41 29 12 40 14 74 91 8 53 68 44 84 76 26 37 38 85 46 93 7 39 56 81 25 31 75 51 72

Possibility #3:
93 63 39 29 12 10 57 74 91 8 53 68 98 70 92 86 58 38 4 89 80 63 18 56 81 25 45 24 32 72

Possibility #4:
49 6 18 29 44 40 14 74 69 52 53 68 12 84 92 65 58 38 20 23 80 7 18 83 71 25 31 79 58 72

Possibility #5:
42 78 48 29 12 94 54 74 69 37 53 68 12 70 92 65 32 38 41 2 27 61 87 11 64 25 45 97 58 90

6. Analysis of Homophonic Substitution

6.1. Theoretical keyspace

A homophonic cipher with N characters will have a theoretical keyspace of 26^N compared to the theoretical keyspace of simple substitution being *only* 26! ~ 4.033 x 10^{26}. This keyspace is computed without considering special characters such as spaces and hyphens. Given the unmanageable size of this keyspace, an exhaustive key search is not feasible. For example, for a relatively short message with N = 50, an exhaustive key search using a personal computer (which can test approximately 10^6 keys/second) would take 26^{50} keys/ 10^6 keys/second = 5.6 * 10^{64} seconds = **1.8 * 10^{57} years.** However, if the message is encrypted using simple substitution, an exhaustive key search can be done in *only* 26! keys / 10^6 keys/second = 4.03 * 10^{20} seconds = **1.28 * 10^{13} years**. The difference between the homophonic substitution key space and the simple substitution key space increases exponentially as the message becomes longer.

6.2. Dictionary-based attacks

While the simple substitution is susceptible to the dictionary-based attack, the homophonic substitution encryption method virtually has no vulnerability against the dictionary-based type of attack. The sole reason is the "one-to-many" mapping technique. The "one-to-many" technique allows an English letter to be represented by multiple cipher letters. For example, for a combination of five consecutive cipher letters *ADLLE*, using a dictionary-based attack on a simple substitution would suggestive that the group *ADLLE* represents a word *HELLO*, *DENNY*, or *LARRY* based on the pattern of the group *ADLLE*. The same group *ADLLE*, using a dictionary-based attack on a homophonic substitution would suggest any of the following combinations of five consecutive English letters: *LILLE*, *MOMMY*, or *DADDY*. As shown in the example for the homophonic substitution, cipher letter A and L can represent the same English letter, which is not the case with simple substitution. For this reason, with the homophonic substitution, when a word in the dictionary is used, the word does not assist in the decoding. Therefore, a dictionary-based attack cannot work on the homophonic substitution.

6.2. Statistics-based attacks

One of the most effective solutions for deciphering simple substitution ciphers is to employ cipher symbol frequency analysis. Based on the statistics of the symbol frequencies, the attacker is able to make guesses of possible mappings of plaintext-ciphertext. The homophonic substitution, however, has increased the number of symbols in the cipher alphabet. The typical homophonic substitution cipher employs the cipher alphabet of 100 letters represented numerically (i.e., 1 to 100).

Each English letter is mapped to a certain number of symbols in the cipher alphabet

based on their own frequencies. For example, in English letter frequency, letter A is about 8.2%; thus, letter A is mapped to 8 different symbols in the cipher alphabet, while letter E is mapped to 12 different cipher symbols. With this technique, the symbols in the cipher alphabet will have a much more balanced frequency.

The attacker normally will not see symbols in the cipher alphabet standing out in a homophonic cipher. In addition, with the frequencies being similar, the attacker has a much difficult time to group appropriate cipher symbols to their correct mappings. In the example shown in Figure 11, the attacker has to determine cipher symbol 1, 4, 18, 20, 36, 39, 41, 48, 59, 62, 85, and 87 all map to E. The cipher alphabet, however, is not restricted to 100 symbols. The cipher alphabet size varies based on the intentions of the sender. In Zodiac case, 53 symbols were used in his original alphabet for Z408, while 62 symbols were used in the Z340 that followed.

When considering the frequency of consecutive letter combinations, the standard approach involves the analysis of only one, two or three consecutive letter combinations (i.e., a "*monograph*" analysis, a "*bigraph*" analysis or a "*trigram*" analysis). For my research project, I modified this approach to include a greater number of consecutive letter combinations.

7. Modifications to Normal Frequency-based Attack

7.1. Greedy algorithm variants

With the keyspace being 26^N, which is unmanageable for performing an exhaustive key search, the most effective approach is to use a greedy algorithm. A greedy algorithm is any algorithm which, at any point in the search, always follows the path to the local optimum [14]. This path, however, does not guarantee an optimal solution. There are many forms of greedy algorithm. "Hill-climb" is the variation that fits the attack of homophonic substitution (Section 8).

7.2. Employing higher level N-graphs

Frequency-based attack on simple substitution ciphers is a form of the greedy algorithm method. The frequency-based attack, however, does not require a letter combination exceeding a 3-graph ("trigraph".) A trigraph score table is comprised of all of the combinations of any 3 consecutive English letters. For example, in English, the combination of "*aba*" has the frequency of 0.0075% while the combination of "*epa*" has the frequency of 0.1086%. In order to solve the homophonic substitution, the cryptanalyst needs to employ much higher N-graphs tables that increase the probability of distinguishing meaningful text from meaningless text.

For my research project, the method used to identify more precise relationship levels

among the cipher alphabet's letters involved the following higher N-graph scores:

- "*tetragraph*" defined as a combination of 4 consecutive letters
- "*pentagraph*" defined as a combination of 5 consecutive letters
- "*hexagraph*" defined as a combination of 6 consecutive letters
- "*heptagraph*" defined as a combination of 7 consecutive letters

8. Hill-climb Algorithm

8.1. Introduction

Hill-climb algorithm (HCA) is a variant of a greedy algorithm in which the search for the global optimum is illustrated as an action to climb multiple hills to reach local optima [15]. Among these local optima, the best result is considered the global optimum. Depending on certain constraints, the number of local optima varies from one to infinity. For example, the function $f(x)= -x^2$ has only one local optimum at $x=0$, which is also its global optimum. On the other hand, the function $f(x) = \text{cosine}(x)$ has infinite local optima at $x = 2k\pi$ with $k \in Z$: set of integer numbers. The process of searching for all the local optima, therefore, usually never ends automatically. Instead, the search is only terminated by the user.

8.2. A single local optimum search

At any stage of a particular local optimum search, the current node, which is a combination of all cipher symbols' mappings, tries to evaluate all possible adjacent nodes. The current node then proceeds to the adjacent node that corresponds to the best result. The other adjacent nodes are ignored and never considered again in this current local optimum search. When there are no remaining nodes available for further consideration, the search terminates and the current node is considered the local optimum. The algorithm records this optimum and starts climbing another hill. Listing 1 shows the pseudocode of a single local optimum search.

Listing 1: HCA: a single local optimum search

```
localBestScore = 0
HCA(node: currentNode)
begin
        currentNodeBestScore = -INF
        bestNeighbor = NULL
        neighborSet = findAllNeighbors(currentNode)

        for (all node in neighborSet)
        begin
```

```
                score = calculateScore(currentNode, node)
                if (score > currentNodeBestScore)
                begin
                        currentNodeBestScore := score
                        bestNeighbor := node
                end
        end

        if (bestNeighbor != NULL)
        begin
                updateLocalBestScore()
                HCA(bestNeighbor)
        end
end
```

The variable "*localBestScore*" stores the optimum value of a single HCA local optimum search after the method "*HCA(node x)*" terminates. This value is compared to the "*globalBestScore*", which stores the global optimum value. If the value of the variable "*localBestScore*" is more optimal, the algorithm will update the value of the variable "*globalBestScore*" appropriately. Otherwise, the algorithm discards this local optimum search.

The general HCA would be operating in the framework of an infinite number of runs for a local optimum search. Figure 12 displays a function that consists of two local optima. Both of these two local optima can be reached by choosing two different starting nodes.

$$z = e^{-(x^2 + y^2)} + 2e^{-((x-2)^2 + (y-2)^2)}$$

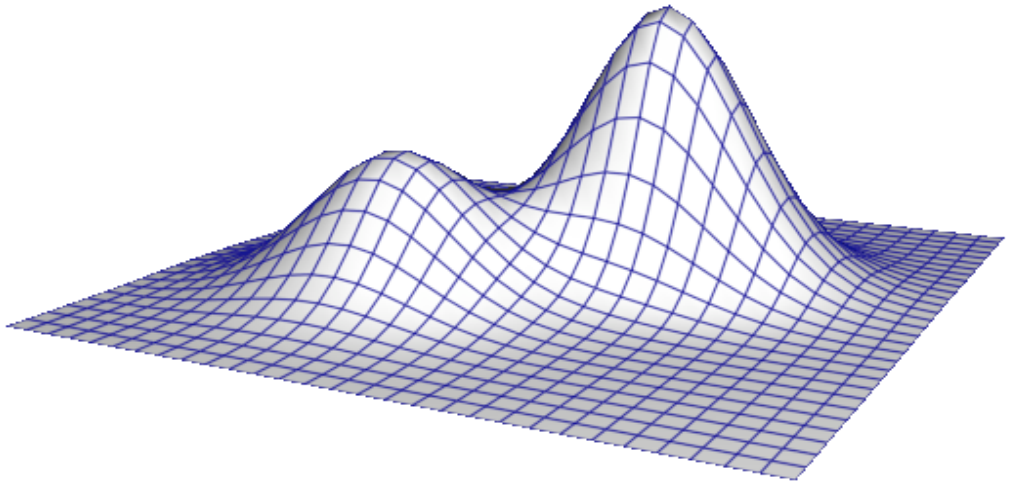


Figure 12: An example of two local optima (illustrated by Apple Grapher software in 3D mode)

8.3. HCA advantages

The HCA can reach the best local optimum, or sometimes global optimum, much faster because the algorithm bypasses multiple ineffective paths in its searches. The algorithm also explores only a subset of the total solution space; thus it can finish the search much faster.

Another advantage of the HCA is the use of multi-directional search. Other regular frequency-based algorithms are linear and single-directional when performing the search. As a result, the HCA has more coverage in the solution space. Figure 13 shows an example, in which the HCA can reach the global optimum much faster than regular methods. Without HCA, the search will follow a rather circuitous path corresponding to the black curves shown in Figure 13. With HCA, the search will follow the more direct blue path.

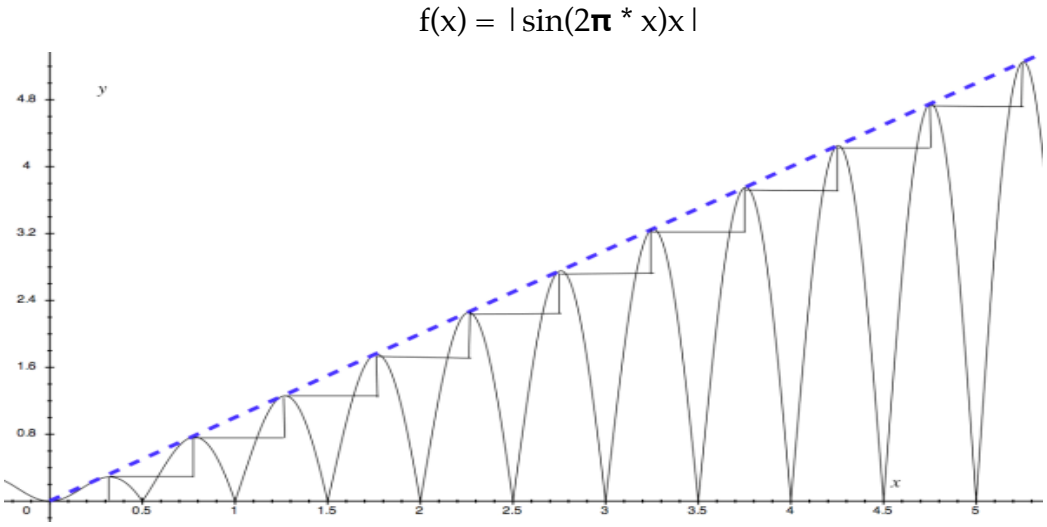


Figure 13: An example of the HCA advantages (illustrated by Apple Grapher software in 2D mode)

8.4. HCA disadvantages

Although the HCA can reduce the total paths and total time of the search, it also faces a greater chance of omitting paths that can lead to the global optimum. Another disadvantage of the HCA is the convergence of search paths. The problem occurs when different paths all visit a certain point "*t*". From point "*t*" onward, the local optimum search computations are repeated as many times as the number of joined paths.

Figure 14 demonstrates the disadvantage of the HCA. Instead of choosing the best

adjacent node, which does not lead to the global optimum, the search can follow a different path, which leads to the global optimum.

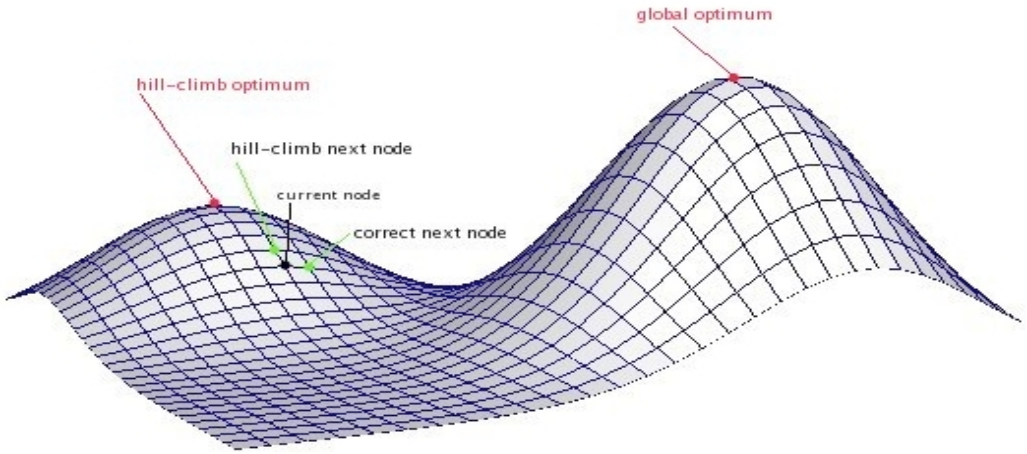


Figure 14: An example of the HCA disadvantages (illustrated by Apple Grapher software in 3D mode)

9. Structure Design of HCA on Homophonic Substitution

9.1. Selected starting nodes

The first phase of the HCA is to select starting nodes corresponding to each local optimum search. With a typical homophonic cipher, the starting nodes are selected on the basis of the number of the symbols corresponding to each particular English letter. For example, if a cipher alphabet consists of 100 symbols, then:

English letter E has a frequency of 12%, then 12% of 100 symbols (12 symbols) are mapped to E
English letter A has a frequency of 8%, then 8% of 100 symbols (8 symbols) are mapped to A
English letter C has a frequency of 3%, then 3% of 100 symbols (3 symbols) are mapped to C

This cipher frequency approach assumes that the frequencies of all cipher symbols are balanced. Therefore, the mappings are decided based on the "*quantity*" (i.e., the number of cipher symbols are considered). This cipher frequency approach does not work on modified homophonic substitution cipher because the frequencies of the cipher symbols vary. In the case of the Zodiac's Z340, the frequency of all 60 cipher symbols is not

consistent. Some of the cipher symbols have much higher frequencies than the others.

Thus, in my research project, I implement another approach to generate starting nodes. In order to determine the mapping of an English letter **α**, I choose a small subset of the cipher alphabet with one condition: the sum of all the frequencies of the letters in the subset is similar to the frequency of the letter **α** with some reasonable error rate **ε**. This cipher frequency approach is based on the "quality" (i.e., the actual frequency of each cipher symbol is considered.) For example, if a cipher alphabet consists of 100 symbols, and:

Cipher symbol 10 has a frequency of 10%
Cipher symbol 13 has a frequency of 1%
Cipher symbol 15 has a frequency of 5%
Cipher symbol 24 has a frequency of 1%
Cipher symbol 90 has a frequency of 5%

then, there are two combinations out of these five cipher symbols that can be mapped to English letter E because in each combination, the frequencies of all cipher symbols add up to 12%.

Combination 1: symbol 10 + symbol 13 + symbol 24
Combination 2: symbol 13 + symbol 15 + symbol 24 + symbol 90

Figure 15 shows the English letters frequencies to help demonstrate my method of generating starting nodes. Listing 2 is the pseudocode of the generating starting nodes phase.

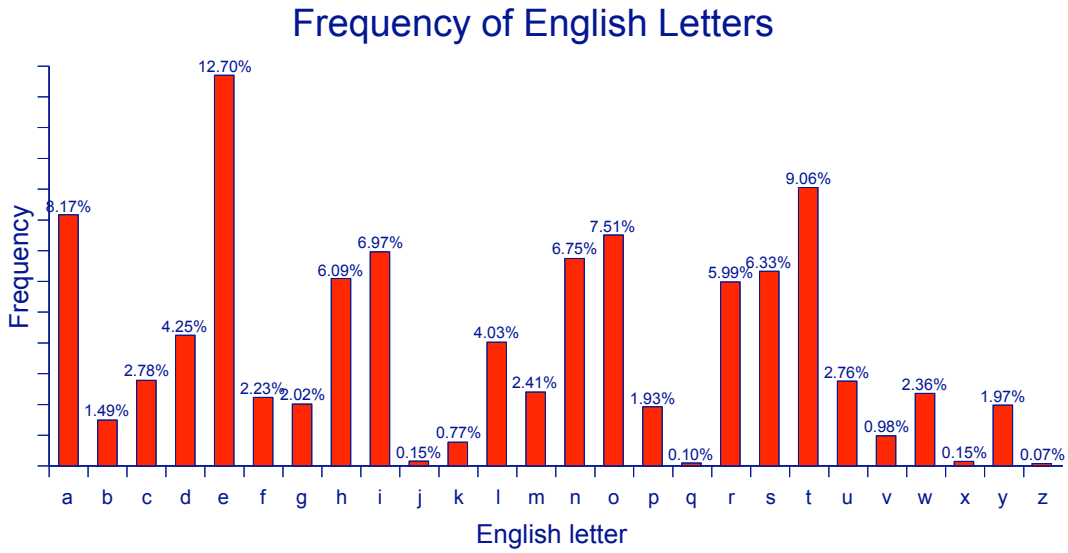


Figure 15: English letter frequencies

Listing 2: HCA 1st phase: selecting starting nodes

```
// β: represent a letter in the cipher alphabet
// α: represent a letter in the English alphabet
// ε: represent the allowable error rate
// mapped[]: to indicate which cipher letter has been mapped
selectStartingNodes()
begin
        createCipherLetterCount()
        createExpectedEnglishLetterFreq()
        for (β: all cipher symbols set)
                mapped[β] = -1

        for (α: all English letter set)
        begin
                count = 0
                ε = 1
                for (number of trials)
                begin
                        //pick a random letter in the cipher alphabet
                        β = pick_random(CIPHER_ALPHABET)
                        if (mapped[β] > -1) continue   //already mapped, continue

                        //check to see if go over error limit ε
                        count += cipherLetterCount[β]
                        if (count > expectedFreq[α] + ε)
                                count -= cipherLetterCount[β]
                        else
                        begin
                                mapped[β] =  α
                                //satisfied? if yes, break
                                if (count +  ε > expectedFreq[α]) break
                        end

                        //this error ε rates does not work? increase it
                        if (trial % EPS_LIMIT == 0) ε++
                end
        end

        //If there is any cipher letter which is unmapped,
        //randomly pick an English letter to map it to.
        for (β: all cipher letters)
                if (mapped[β] == -1)
                        mapped[β] = pick_random(ENGLISH_ALPHABET)
end
```

```
//Count the appearances/frequency of each cipher letter
createCipherLetterCount()
begin
        for (β: all cipher letters)
                cipherLetterCount[β] = 0
        for (i = 0; i < ciphertext.length; ++i)
                cipherLetterCount[ciphertext[i]]++
end

//monograph: English letter monograph
createExpectedEnglishLetterFreq()
begin
        for (α: all English letter set)
                expectedFreq[α] =
                        (int) (monograph[α] * ciphertext.length / 100.0 + 0.5)
end

//randomly pick out a number in [1, upper]
int pick_random(int upper)
begin
        return ((int)(rand() * 1.0 / ((double)RAND_MAX + 1.0)) * upper) + 1)
end
```

9.2. Finding and swapping adjacent nodes

Using the starting nodes generated by Listing 2, the HCA evaluates all adjacent nodes to determine where to proceed. The HCA design on homophonic substitution ciphers, however, cannot include all adjacent nodes in the evaluation because of the unmanageable total keyspace. A homophonic substitution cipher alphabet consists of N letters. At any stage of the search, the current node has close to N * (N-1) / 2 adjacent nodes. Table 5 shows the total keyspace after the search has performed ten swaps. The focus is on N = 60 as the number of cipher symbols in Z340 is 63.

Table 5: HCA: total keyspace after ten swaps

#	Total possibilities	N = 50	N = 60	N = 80	N = 100
1	$N * (N-1) / 2$	$1.2 * 10^3$	$1.7 * 10^3$	$3.1 * 10^3$	$4.9 * 10^3$
2	$N^2 * (N-1)^2 / 4$	$1.5 * 10^6$	$3.1 * 10^6$	$9.98 * 10^6$	$2.4 * 10^7$
3	$N^3 * (N-1)^3 / 8$	$1.8 * 10^9$	$5.5 * 10^9$	$3.1 * 10^{10}$	$1.2 * 10^{11}$
4	$N^4 * (N-1)^4 / 16$	$2.2 * 10^{12}$	$9.8 * 10^{12}$	$9.97 * 10^{13}$	$6.0 * 10^{14}$

5	$N^5 * (N-1)^5 / 32$	$2.75 * 10^{15}$	$1.7 * 10^{16}$	$3.15 * 10^{17}$	$2.97 * 10^{18}$
6	$N^6 * (N-1)^6 / 32$	$3.38 * 10^{18}$	$3.07 * 10^{19}$	$9.96 * 10^{20}$	$1.47 * 10^{22}$
7	$N^7 * (N-1)^7 / 32$	$4.14 * 10^{21}$	$5.44 * 10^{22}$	$3.15 * 10^{24}$	$7.28 * 10^{25}$
8	$N^8 * (N-1)^8 / 32$	$5.0 * 10^{24}$	$9.63 * 10^{25}$	$9.94 * 10^{27}$	$3.6 * 10^{29}$
9	$N^9 * (N-1)^9 / 32$	$6.2 * 10^{27}$	$1.7 * 10^{29}$	$3.14 * 10^{31}$	$1.78 * 10^{33}$
10	$N^{10} * (N-1)^{10} / 32$	$7.6 * 10^{30}$	$3.02 * 10^{32}$	$9.93 * 10^{34}$	$8.8 * 10^{36}$

The keyspace size increases exponentially after each swapping; therefore, an exhaustive swapping for each stage of the search is not feasible. In order to improve the time and speed of the HCA, the algorithm must evaluate only a subset of all the adjacent nodes.

My first experimental approach involved defining adjacent nodes. Two nodes N_1 and N_2 are adjacent if they satisfy the following conditions:

1. These two nodes only differ in 2 symbols out of a total of N positions: p_1, p_2
2. $N_1[p_1] = N_2[p_2]$ and $N_1[p_2] = N_2[p_1]$
3. These two cipher symbols map to different English letters
4. These two cipher symbols have similar frequencies.

The HCA used the adjacent nodes definition to evaluate possible swap at any stage of the search. Listing 3 demonstrates the methodology of finding two swappable nodes.

Listing 3: HCA: finding swappable nodes

```
//freq[]: store the frequency of cipher letters
//mapped[]: store the mapping of cipher letters→ English letters
hillclimbSwap(currentCipherLetter)
begin
        if (currentCipherLetter > total_cipher_alphabet) return
        update_bestScore()    //update the best decrypted plaintext

        for(cipherLetter2 = currentCipherLetter+1 to total_cipher_alphabet)
        begin
                if (swappable(currentCipherLetter, cipherLetter2))
                begin
                        swap(currentCipherLetter, cipherLetter2)

                        //calculate the new score with this new node
                        calculateScore()
                        if (new score is better) hillclimbSwap(cipherLetter2)
```

```
                    swap(currentCipherLetter, cipherLetter2)    //swap back
                end
            end

            //done with this current cipher letter,move on to next letter
            hillclimbSwap(currentCipherLetter + 1)
    end

    //check to see if these 2 cipher letters can be swapped, using the conditions:
    //1. mapped to different English letter
    //2. similar frequencies
    swappable(cLetter1, cLetter2)
    begin
            if (fabs(freq[cLetter1] – freq[cLetter2]) > ACCEPTABLE_EPS) return false
            if (map[cLetter1] == map[cLetter2]) return false;
            return true;
    end
```

9.3. Score calculation formula

In order to grade each node, the algorithm employs a formula to compute the node's score. Without knowledge of the actual message, the only statistic available for the algorithm is the frequency statistics. It is generally agreed that a "*bigraph*" and "*trigraph*" (a group of two and three consecutive letters) are enough for any frequency analysis. Therefore, in my first experimental approach, the algorithm only employs these two frequency statistics. The algorithm also needs to determine the weight of each of these two frequencies. Due to the fact that a correct "*trigraph*" is more difficult to derive than a correct "*bigraph*", I assigned a weighted score of 4 for a "*trigraph*" frequency compared to a weighted score of 1 for a "*bigraph*" frequency. The formula employed in the first experimental run was:

score = (biscore) + (triscore << 2)

The Listing 4 shows the method that calculates the score for each node

Listing 4: HCA: score calculation formula

```
    //Apply the current key and calculate the score
    calculateScore()
    begin
            for (i = 0 to ciphertext.length – 1)
                    plaintext[i] = map[ciphertext[i]]
            newScore = (calcBiScore(plaintext) +calcTriScore(plaintext) << 2)
    end

    //[i][j] – the frequency of letter i followed by letter j
```

```
calculateBiScore(plaintext)
begin
        score = 0
        for (i = 0 → plaintext.length – 2)
                score += biscore[plaintext[i]][plaintext[i+1]]
        return score
end

//[i][j][k] – the frequency of letter i followed by letter j followed by letter k
calculateTriScore(plaintext)
begin
        score = 0
        for (i = 0 → plaintext.length – 3)
                score += triscore[plaintext[i]][plaintext[i+1]][plaintext[i+2]]
        return score
end
```

10. Test Suite 1 and Results

10.1. Test Suite 1

10.1.1. Original message and its corresponding ciphertext

The original objective of my research project was to create a message with similar length and similar cipher alphabet size to the Z340. For this purpose, I selected some text consisting of 327 alphabetic letters from the famous book "Alice in Wonderland:"

> "*There was nothing so very remarkable in that; nor did Alice think it so very much out of the way to hear the Rabbit say to itself "Oh dear! Oh dear! I shall be too late!" (when she thought it over afterwards it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but, when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on*" [16]

After removing spaces and other non-alphabetic characters, this message reads:

> "*therewasnothingsoveryremarkableinthatnordidalicethinkitsoverymuchoutofthe waytoheartherabbitsaytoitselfohdearohdearishallbetoolatewhenshethoughtitover afterwardsitoccurredtoherthatsheoughttohavewonderedatthisbutatthetimeitall seemedquitenaturalbutwhentherabbitactuallytookawatchoutofitswaistcoatpocket andlookedatitandthenhurriedon*"

I generated a ciphertext based on this message using a cipher alphabet with 60 different symbols, numbered 1 through 60. Table 6 shows the ciphertext in numeric form.

Table 6: Test Suite 1: ciphertext in numeric form

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

10.1.2. Plaintext letter frequencies

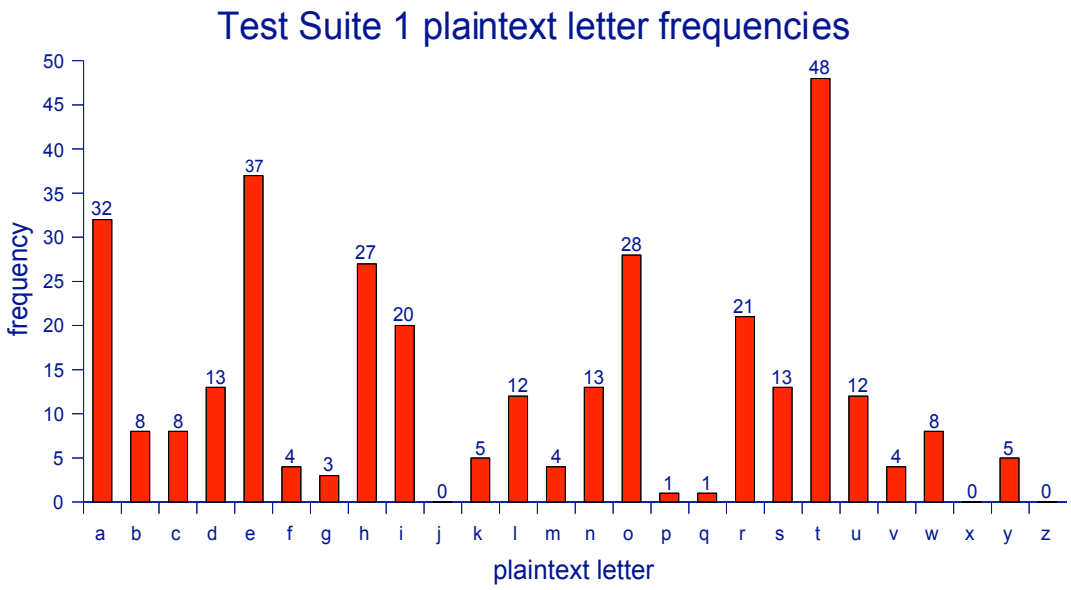


Figure 16: Test Suite 1plaintext letter frequencies

10.1.3. Cipher symbol frequencies

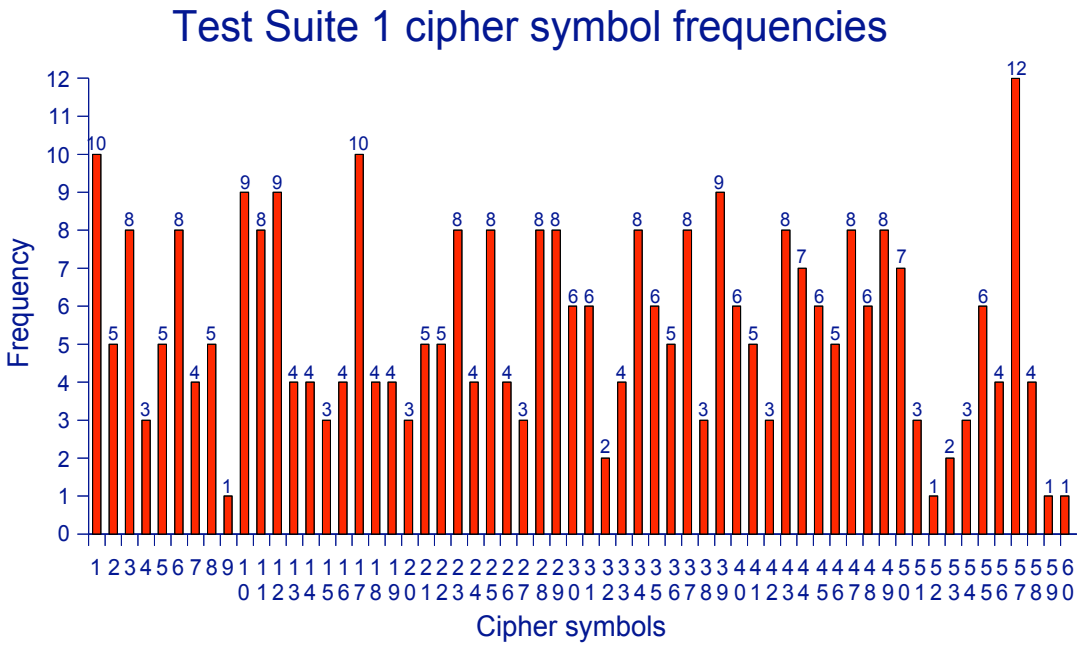


Figure 17: Test Suite 1 cipher symbol frequencies

10.1.4. Test Suite 1 actual plaintext-ciphertext mappings

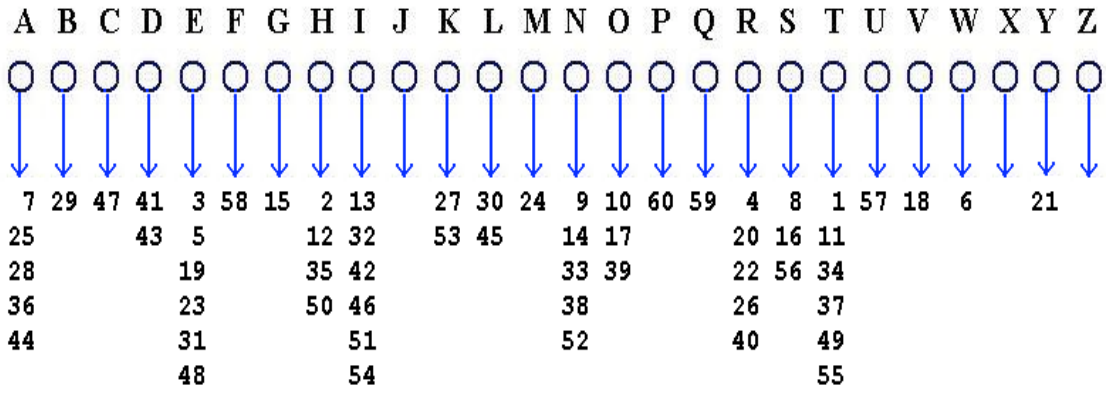


Figure 18: Test Suite 1 actual mappings

10.2. 1st experimental run results

The results were lower then expected. The average success rate (i.e., the percentage of encrypted cipher symbol were decrypted correctly) was only 8%. Table 7 below shows some of the results obtained by using my original HCA.

Table 7: Test Suite 1 results using original HCA

Decrypted text	Success rate (%)	Time (sec)
wsehetrodtatasaiplctyemeneuhddiiltaiertoedldgsmmpvaeafesdhariwsblilnwavnhiannneochyeminhirwrunstddubeattgamrerstesaoippoiheodrfheuaepttttehevntlievuaiepswarhacnnsryneoerbhhwiropceihcdipepiegibwyneabesdoaheaitettpstbirpochyfheteedetsreiepdpdolethefmlepiltcntimdheimitpsvneuaesfibdiaigeufatetaspsyowplfiroethdoinheimuniswoeeceag	9.48	21.45
esehetrodtatasaiplctyhreneuhddiitlgihmtledoadsfmpvaeabesdhamiesrlioneavnhiannnelchyhmfnwimerunstadurhattdamremsthsaiippliheodroweuaepttttttewevntoievuafepseamhacnnsmynelerrhheiropcefhcafpepihdireynearesdlahegitettpstrirpochyoweteedetsreiepdpalletheoroepiotcnlimdheimitpsvneuaesoiraiaideuoatelaspsyleploirohthdiinweifunfseieecead	9.78	43.22
llatitqthdstrpoaerhhalegrrdfeunnatwelodtaecuepbntqyessrlhfswellvrocrlsqqmeoeergthmalndriooladqpmuudvlstheonaeolmllosoeetntrthasiidyatttdihgiiqqhcngqdydrellswmnerepoaratravfgleqtehidghudtrtaleovlaresvelhtymrwoiitdelhveqethgasiimngheipareaehtutrrimnsecntochhrtonumnnnedelqendoglsevuesnegdsrmntolepatltrseatlimusneinobdrdllsiihioe	7.34	41.23
beghedmrrduwyietelwngeavztefamooootuendagavclichauohuwterfudrbeilsvzbuupsuettzvawsgehozisnbrepitcmeieuwnlehrhneteeeiseeaohtrrryieeogawwdnnvieupnvovueooteebudsanztingzgatrifebrmreweoewcoatatelsibgzhuiheraosttsneddeenirmerwegyiethvrhnirtrgeracaltnshyavhasvnwzoshmshohrdeeutheeveyricruolveyythoeeeigabalyurrensmiotihscezoebieeweel	7.64	65.43

10.3. Discussion

Although the HCA restricted the condition that allowed the swapping between two cipher symbols (i.e., adjacent nodes), the algorithm still had to explore an unmanageable keyspace. A second shortcoming was the method of selecting the starting nodes, which is the most important phase in the HCA. In the first experimental run, the frequencies of the English letters was used as a guide. The Test Suite 1 message was chosen randomly, and the message did not mirror the frequencies of the English letters. Consequently, the selected starting nodes were on completely different paths. And these paths did not lead to the local optimum. A third shortcoming was the use of "*bigraph*" and "*trigraph,*" which was found to be inadequate for identifying the meaningful text. In Section 10, I introduce the technique that I applied to my method for accelerating the HCA.

11. Hill-climb Algorithm Optimization

11.1. Randomization algorithm

To improve the performance of the HCA, I added the randomization algorithm to the current linear method of attack. The restricted swappable condition was removed. The removal of the swappable condition means that, at any stage of the local optimum search, the algorithm generates two randomized cipher symbols. As long as these two cipher symbols are different from each other and map to different English letters, these two cipher symbols can be swapped regardless of their current mappings. The significant improvements included: (1) accelerating the single local optimum search; and (2) broadening the keyspace coverage compared to the earlier linearity-based approach. Listing 5 demonstrates the use of randomization in the algorithm.

Listing 5: HCA: apply randomization

```
//#CipherLetters = # of letters in the cipher alphabet
//mapped[x] = true if x has already mapped, false otherwise
//The constant MAX_TRIAL is preset at 1,000,000 swaps.
hillclimbSwap()
begin
        update_BestScore()              //update the global best score
        for (trial = 0 to MAX_TRIAL)
        begin
                //Randomly pick out 2 letters in the cipher alphabet
                cipherLetter1 = pick_random(#CipherLetters)
                cipherLetter2 = pick_random(#CipherLetters)

                //Check too see if cipherLetter1 and cipherLetter2 map to same
                //English letter
                if (mapped[cipherLetter1] == mapped[cipherLetter2]) continue
```

```
                    //Swap these 2 and check to see if it improves the score
                    swap(cipherLetter1, cipherLetter2)
                    calculateScore()            //calculate the score of this node
                    if (new score is better)
                        hillclimbSwap()         //go on with this new key

                    //swap back
                    swap(cipherLetter1, cipherLetter2)
                end
            end
```

10.2. Improved score calculation formula

In my second attempt, I decide to include much higher level of letters relationship besides the "*bigraph*" and "*trigraph*". These are "*tetragraph*", "*pentagraph*", "*hexagraph*", and "*heptagraph*" (see Section 6.2.) This methodology significantly increases the probability of deriving meaningful English text. The final formula employed in my HCA is:

*score = biscore + (triscore * 2) +(tetrascore * 4) + (pentascore * 6) + (hexascore * 7) + (heptascore * 8)*

12. Test Suite 1 Results using Optimized Hill-climb Algorithm

12.1. Definition of a Crib

A *Crib* is defined as a known plaintext (i.e., a word in the original message). In the process of analyzing the ciphertext, code-breakers can assume that some of the suspected words are in the original message. Thus, code-breakers can guess some of the plaintext-ciphertext mappings based on the crib(s) and they use those mappings in their cryptanalysis.

12.2. Report format

The format used in the statistical reports presented in Section 11-13 is a combination of:

Crib: a list of crib(s) used in the experimental run
Cipher symbols covered: # of letters in the cipher alphabet covered by the crib(s)
Ciphertext letters covered: # of letters in the ciphertext covered by all the letters of all the crib(s)
Total possible keys: the possible keys minus all the letters from all the crib(s)
Result file: the file that stores the result of the actual run on this test
"*success rate*": is defined as the percentage of the encrypted symbology that has been decrypted correctly

"*": indicates weak weight distribution of each N-graphs as the text with the higher score in the attack has much less matching to the original message (See Section 10.2 for weight distribution)
"**": indicates the maximum score obtained in this experimental run.

12.3. 2nd Experimental run – Test 1 – No crib used

Crib: [none]
Cipher alphabet covered: 0
Ciphertext letters covered: 0
Total possible keys: $26^{60} = 7.91 * 10^{84}$
Result File: TS1_NoCrib.txt

Table 8: Test Suite 1 experimental test 1 result

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
6132	ttheieorjroaciiuaiwilahenemstdnbdeabedsalyialhenesoukbnfhgqp	3304769	5065	0.47	8.257
Text: ttheieorjroaciiuaiwilahenemstdnbdeabedsalyialhenesoukbnfrinaleheasheagethebleransanahesttoorblesbernlgatinnasalineyfsolltinardatheandushesshianhesiningehaeailuyoreeheawieashattbtrahahisoorsbineadlhahistnacutheatoahthenbenddfwhewlqhcthinthandtheeaninaheattbeseehalllearmseseeashergheresofoesaeprekiendidasmiioehtodioandtheachiai					
6120*	etiacsadbdmeoneinackeiswsesateryoetheftounhrldaingopjladhiqf	3032393	1839	12.33	7.034
Text: etiacsadbdmeoneinackeiswsesateryoetheftounhrldaingopjladdariewhatthenintisheederaoatieattomdheetledilinthisiteucsandgalltcanderesseioigingthetadntaiksiesosrkuinmdaahaichnngsoethedeinhegmmdgharsnousoshaeaeoitheremesedwryeseedciwcuqhoeinsehisethestinaeserttleaanhrlleendsasanatthndidedsaodmatrefdajcesohentschaedeafhmerftheooihnn					
6135	nespostiqadsiomlengbtrebewcaldofarhahuervinalettheayfounetkj	3257522	9010	5.64	17.737
Text: nespostiqadsiomlengbtrebewcaldofarhahuervinalettheayfounanortbetheerethessatrasoaruhswalladiatheoritlteenteresvoepinetlllloueadanesstalesheeemhuehentbethersabvlidatteprgnheeerneanisseemeddaeanoseaverenanusillerandsenebofreddngsbgvkeinsoeneredlehshtousewallohatheallteacasahtheehaterisaandteahjatforeandeeconthentundsoueewrisneo					

16121 **	snedisgsinthedisamjolrelardatheci eoueitscaneapominoqfhtihlkb	3559585	1900	151.69	24.159
Text: snedisgsinthedisamjolrelardathecieoueitscaneapominoqfhtinmerllhoothealinesulenhe astoerattotsulethesmalannmarthciadaingaatitanheseshmisneinthiotpitmmoaleesseocs atnoohdrjnianessnussheahinttnnumesaicesenasthestheesthespleceahhijeljckhesedashr ahthesomdtherettheaoiheaaleandasaioothinlpessaoitoteebnofieainhatdingepsginthein hrseenad					

This 2nd experimental run was the best result obtained from Test Suite 1 without any crib used. Due to the nature of randomization, the algorithm obtained different values with each run. The algorithm could only manage to achieve a highest success rate of 24%. The average success rate was only 17.5%.

However, these results from the 2nd experimental run were expected given the brevity of the message and the irregular frequencies of the English letters. The next phase of the analysis involved including some additional cribs to further evaluate the algorithm's performance.

12.4. 2nd Experimental run – Test 2 – 1 known crib

Crib: [alice]
Cipher symbols covered: 5
Ciphertext letters covered: 34 (out of 327 ~ 10.398%)
Total possible keys: $26^{55} = 6.66 * 10^{77}$
Result file: TS1_1crib.txt

Table 9: Test Suite 1 experimental test 2 result

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
4152	nnalergokothereaohpfnoedachalm ersehisitaousalicendjqgotbhitp	3289744	1076	1.01	35.474
Text: nnalergokothereaohpfnoedachalmersehisitaousalicendjqgotboheondhchtheoinnarineo heaathacalljtoinstoeoelionseaothoealubdgllletoomanerhesadandthehtinthefaisearafoa utocchlopsnodeanninohaohedttodiherosoeaesanthealheanthenidereammbpadpothenar anhoamlhsrherthecallosacnhallnsooharanchthnoiieorajbtctaspocgeeassmothesgsingist heinhcaeasor					
598	sthaleagqserifandotliseaddprloumi nthnchrighaliceetbjkiofeomd	3218772	1713	13.43	18.043

Text. sthaleagqserifandotliseaddprlouminthnchrighaliceetbjkiofsousiaecthendoethehinsrurr othdrllbeghinhingelodthedshrildagftallllodsoaseereinthetheatoiehoeldonerealingescce asthedtersthsgrhdeateesthouediierehrsorinlenaseresiaumndoofthatimeishfdsesdolenet eforedallinrceeallindsprerectheesoingerbfechandscklndihodhplhanisacheructedrihhdf					
599*	theringijoindisotherlahmnrugldim dataschedlealicestoqkoafefbp	3173618	4486	2.91	30.275
Text. theringijoindisotherlahmnrugldimdataschedlealicestoqkoafohialmectheatfshenalaonig eatergllоiialshoaielftheenahndinrlftgllliatodathnnedotesthestaishhernfshenardolioccer aeestthethatinetestiiotahintddhehegtandoleaatinhtimimanddfeemedbedteinteandlesnte ianhrallosgcsealllstougngscthesofiaingofichaspockiandedthuiegsitgceinicheredeeti					
1988 **	smedstahbothengsingralecaroailef nthespiroadaliceshnjkntonlqb	3626749	5082	916.03	52.599
Text. smedstahbothengsingralecaroailefnthespiroadaliceshnjkntoonelacnchintilsmeteatohea rtheraiintheasinthellimdealihosadaohallistiolasethensheshinghtisineralsertarosatoccnd lgdsihersmeshheinghttohenetinoeredasthesintasthesiceftallogecgoqnesenasnlalinsthen theraiinsacsnallasiooataschinsolithtanotciasbockstandliiosdasisapdthepmnrreedin					

The algorithm improved the success rate with "alice" as its only crib to 52.99%. The result is promising as in general, in any cryptanalysis process, code-breakers can guess a couple of cribs with high success rate.

12.5. 2nd Experimental run – Test 3 – 2 known cribs

Crib: [alice, rabbit]
Cipher symbols covered: 10
Ciphertext letters covered: 65 (out of 327 ~ 19.878%)
Total possible keys: $26^{50} = 5.61 * 10^{70}$
Result file: TS1_2cribs.txt

Table 10: Test Suite 1 experimental test 3 result

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
8508	ishaleandatthepeellaiioanrcabghori handemlbealiceerikofoosdjg	2669223	2735	1.37	23.853
Text. ishaleandatthepeellaiioanrcabghorihandemlbealiceerikofooalhiiaschesiedesheaiiatham					

ohhrabbitnainefineldeseenietllnaborallbloeagaioetererheresphoieeleandnomeaalebtacc saileeeromisainthesprttaralheerlomoeaiothebsiaittoiiahoinggolhalljshihenisingbsnehee otorabbfnacesallineacaeaecheseadiineaiotceangacolinregeecleaniiadetthdssrmhheee					
8244	ngapssadjhthereathalniecareaalech emadforeoialiceneiqknofhipb	3206161	16367	20.10	36.086
Text: ngapssadjhthereathalniecareaalechemadforeoialiceneiqknofhheinchcmohetingasanehh earomaraaaitdandonedelitgieaiohesapofeallasothlaneshehaeaneohemoinohelaidersalea othcchpiainteerngandhatheettheahestheereianoheaaheanthenicecealllfaacaephenaranhi alahdsmeroheraaandacnhallndtheasancmohnhiiedsaiftcoadbhckseahiltoesiadinafithefg hrreaitr					
9171*	siecaeahjothesggiracamesarmable andhelierdinalicedhiqkbtonfop	3171115	5358	54.22	51.376
Text: siecaeahjothesggiracamesarmableandhelierdinalicedhiqkbtooremasnchendifdieeeadoh eartherabbithealebdhelfiineamehdaaciohallbatiolaseehenghedhenghtiderecaflereacdgit occncmandihersieshheinghttohereeinderenasthegbndasthesiseadalloaesadonesesasnma lbnlehestherabbblacdnallaliomaeadchendofidheaiotcealpockadannliemanalisaintheiinr reenis					
4415 **	shelisahjothergdoromaledarpablea nnhennereodaliceshiqkicfeibg	3209636	3495	231.22	54.434
Text: shelisahjothergdoromaledarpableannhennereodaliceshiqkicforeladecheenoisheseanoh earcherabbitheaneinheliohdealeheialofhallbicoolaseshendhesheeghciseremainersamed otoccellodsohersheshheoeghttoheresoneeredaschedbenasthesideanallfoedoebeeserasel albenshercherabbinacseallanoopasascheesoiinhsaiftceangockinandloepidanisandthenh erreedor					

The success rate of 54.4% was the highest that the HCA achieved with "Alice" and "Rabbit" as the cribs. The average success rate of the attack using these two cribs was about 37%, a 15% improvement compared to the minimum coverage. The results were quite good as total solution space was 10^{70}.

12.6. 2nd Experimental run – Test 4 – 3 known cribs

Crib: [alice, rabbit, watch]
Cipher symbols covered: 15
Ciphertext letters covered: 90 (out of 327 ~ 27.522%)
Total possible keys: $26^{45} = 4.72 * 10^{63}$
Result files: TS1_3cribs.txt

Table 11: Test Suite 1 experimental test 4 result

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
8533	shemowooketheacghriinaedarmablepidhadfernanalicethiqjpoleagb	3102805	7401	0.18	46.789
Text: shemowooketheacghriinaedarmablepidhadfernanalicethiqjpolereandecheedhathewandehearoherabbitoandepdoelahhneaaehnoamalhollboohelasewheighetheechoitereiaaderwaingateccemainthhershasohehechtteharewhinerenasohegbedasthesidepdallliedingeeseaaseaalbedwheaoherabbpdacteallndhemawatcheeteaidowailtceadbecjodainlhemonodisofnthefherreenha					
7113	nhebiwggjothedosoredalemarfaboeainheloereanalicethackidsefcp	3174961	5005	15.78	56.269
Text: nhebiwggjothedosoredalemarfaboeainheloereanalicethackidsorelamecheenoftheweanoheardherabbatgealeingelfohnealeheiabashgllbidoooanewheishetheeohditeredaflerwadesatocceblentohernhengheoeohttoherewoieerenandhesbenanthenimeanaooseemeeceenedanelaobelwheddherabbilacteallaloofawatcheetofingwaastcealpockinainooefinglingontheoherreenod					
8942 *	nhelpwadjothingoshimiremareableanehanderofealicethikbbocraqp	2949172	1373	125.66	5.81
Text: nhelpwadjothingoshimiremareableanehanderofealicethikbbocoherimrcheresathewaieohearoherabbitdainebedelasheearehopalfchallbposolanewhenohetherghoitehemaanerwamooftoccrlrietshernhandhesrghttohahewsnoereeanohiobreanthenimeaeallciemioqrinenanrralbrnwhenoherabbbnactrallinsoeawatchertoaiedwaictceanpocbpeanelseepeaninadethedhrrrieesn					
6481 **	nherewahjothingnoromalenarmablebpchenderposalicethigkidfafqg	3102216	91342	325.73	61.162
Text: nherewahjothingnoromalenarmablebpchenderposalicethigkidforelanacheacoftheweacoheardherabbitheaneichelfohsealehpearofhallbedoolanewhepnhetheaghditeremafnerwampnotoccarlostohernhenhheoaghttoherewopperesandhinbacantheninebcallfoenopqainenanalalbanwhendherabbinactaallanoomawatcheatofichwaiftceangockecapsloemesaninadsthedharrieson					

The success rate of 61.162% was the highest that the HCA achieved with three known cribs. The average success rate of the HCA using three cribs was about 48%, a 21% improvement compared to the minimum coverage. The results, again, were quite

satisfactory because the total solution space was 10^{63}. However, this test run was performed for the purpose of evaluating the HCA only. The probability of code-breakers guessing three cribs correctly in place was quite low. For example, the Test Suite 1 message has a short length 327 characters. For crib "alice", there are (327 – 5 = 322) possibilities to test for "alice". For crib "rabbit", there are (327 – 6 = 321) possibilities. For crib "watch", there are (327 – 5 = 321) possibilities. The total number of possibilities is close to 322 * 321 * 321 ~ 32 * 10^6. For each possibility, the HCA had to select its starting nodes. Consequently, the HCA could not afford to cover all the selected starting nodes. The longer message, the total number of starting nodes increase exponentially. Therefore, in the Test Suite 2, the experimental runs with three known cribs were not included.

12.7. Discussion

With the additions of Randomization Algorithm and higher N-graphs table, the HCA improved the performance in all the experimental runs. However, due to the fact that Test Suite 1 message was chosen randomly, the frequencies of the letters in the original message did not correspond to the frequencies of the English letters. The irregular letter frequencies was the main obstacle that had prevented the HCA to achieve more success rate. In Section 2, a Test Suite with longer message is used to evaluate the performance of the HCA.

13. Test Suite 2 and Results

13.1. Test Suite 2

13.1.1. Original message

In Test Suite 1, the message is so short that the frequency of the letters does not correspond to the normal frequency of English letters. The HCA, however, depends on the frequencies to generate the starting nodes. It is clear that the HCA performs more effectively when the frequency of letters in the message approaches the frequency of English letters. The purpose of the Test Suite 2 was to further test the HCA. The Test Suite 2 message contains 8634 characters, which is the first chapter of the book "Alice in Wonderland" (see Appendix B). The encrypted version of the Test Suite 2 message contains 64 cipher symbols, named 1 to 64.

13.1.2. Plaintext letter frequencies

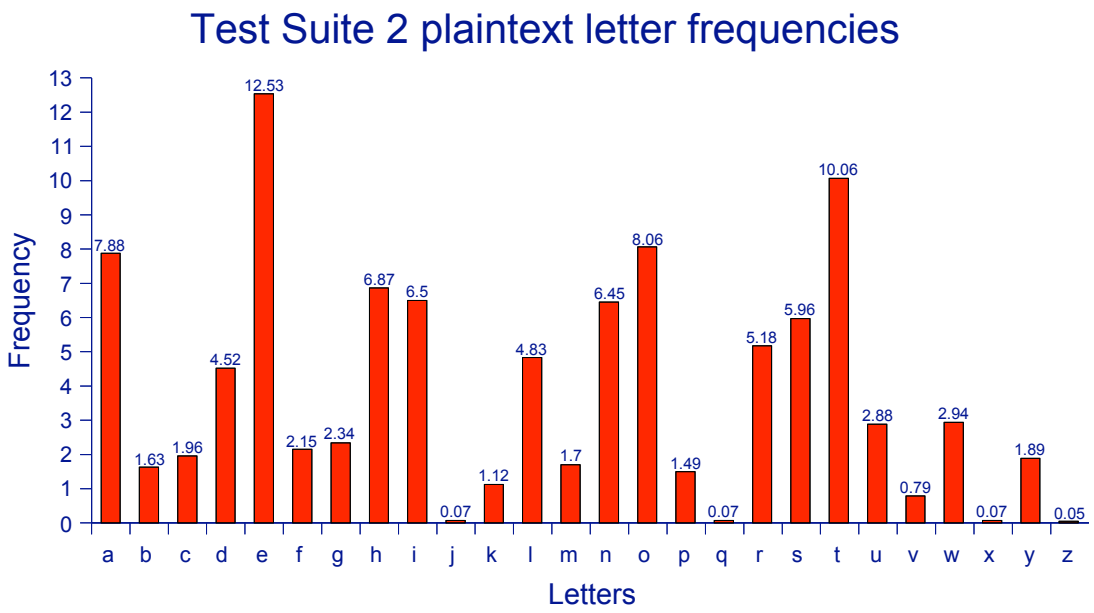


Figure 19: Test Suite 2 plaintext letter frequencies

13.1.3. Ciphertext letters frequency

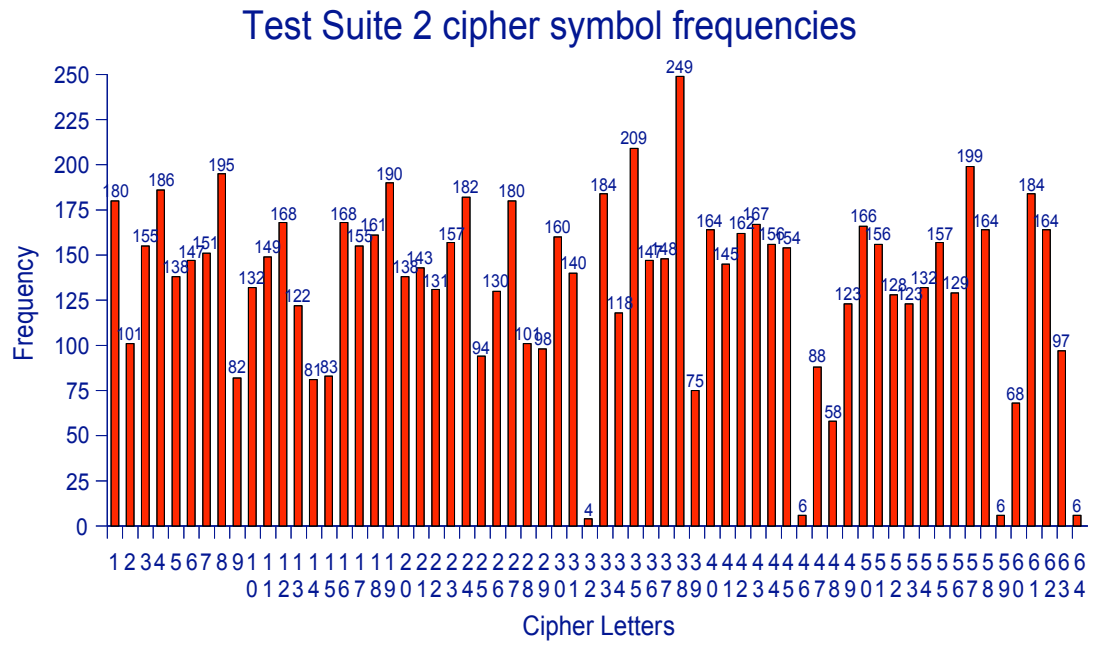


Figure 20: Test Suite 2 cipher symbol frequencies

13.1.4. Test Suite 2 actual mappings

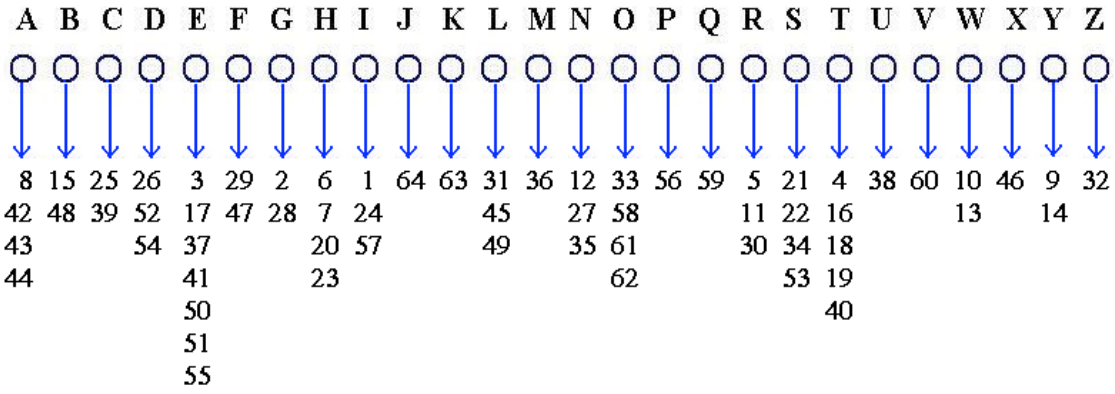


Figure 21: Test Suite 2 actual mappings

13.2. Results

13.2.1. 3rd experimental run – Test 1 – No crib used

Crib: None
Cipher symbols covered: 0
Ciphertext letters covered: 0
Total possible keys: $26^{64} = 3.616 * 10^{90}$
Result file: TS2_1crib.txt

Table 12: Test Suite 2: 3rd experimental run: test 1 result

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
93	igetrhhaomsnwybtetthsshacdndorlki anlenyteaaalxfuoeeasdeoiiupaapq	82148505	5927	95.98	69.331
3786	iaetshhaoarnwobtetthashicdngarlkei nmeapneaaanufbleedsdelioyhoemp	84696642	8462	135.11	74.531
2832 *	ioetrhheawrnwamtetthsshicdngfrlkoi nneacseaaaluubleeaadeaiokpooiy	81062207	11755	185.50	78.284
4530 **	icetrhhairrnwybtetthsshicdngfrlkosn haapteaaalumbleeasdeniobmooib	84953627	10155	251.32	83.021

Due to the length of the message, all of the results for the 3rd experimental runs are saved in separate text files for evaluating and validating purposes. The result were impressive

because the HCA achieved the highest success rate 83% in a little over four minutes. The average success rate for this test was 72%.

13.2.2. 3rd Experimental run – Test 2 – 1 known crib

Crib: [alice]
Cipher symbols covered: 5
Ciphertext letters covered: 775 (out of 8634 ~ 8.976%)
Total possible keys: $26^{59} = 3.044 * 10^{83}$
Result file: TS2_1Crib.txt

Table 13: Test Suite 2: 3rd experimental run: test 2 result

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
8389	ioetrhhaossnwyctetthsshicdngarlbon nheacteaaalffoleedaaeliiumaaip	86396000	5046	81.46	73.361
8389	iaetrhhaiannwoctetthsshicdngarlbos nheacteaaalffyleedsaeliiumaoop	86827290	7233	117.56	77.114
8389	iaetrhhaiannwaatetthsshicdngfrlbosn heacteaaalufcleeasdelioymooip	87378236	11714	190.45	82.326
2041 **	iaetrhhaiarnwabtetthsshicdngfrlkesn meacteaaalufbleeasdeaioulooip	86592909	6547	246.46	85.256

The highest success rate of this experimental run was 85.256%. The average success rate of this experimental run was 74%. The improvement over the experimental run with no crib was low. I did not include the experimental test run with two cribs because: (1) there were approximately 74 millions possible combinations of any two cribs which was difficult to evaluate all of them; (2) the success rate was impressive in both cases: no crib and one crib.

13.3. Discussion

The HCA was started generating meaningful results for messages having letter frequencies similar to those in the English alphabet. Regardless of having or not having cribs, the average success rate of the HCA was an impressive 75%. Some results were as high as 85%. As a result, I conclude that the HCA is able to decrypt those homophonic ciphers comprised of letter frequencies similar to those in the English alphabet. However, given the fact that the Z340 is such a short message, it is not clear whether the original message contains the needed letter frequencies.

14. Applying the Optimized HCA to Z340

14.1. Test run 1 – No crib

Table 14: Z340 Test run 1

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
1747	sphintlerdtonthefaeatthkaelarrslinosestnaioeibooohemieachcjdag	4265024	12941	8.03	N/A
Text: sphintlerdtonthefaneatthkaelarrslianoseesthetintnsainnoltereinheehseboeenooftherermdinehletherfadeieatthesckhesscheanntthealothebalnejnofreedetheeasnagmsentseelaehnoreanesehihoptalandteshersthatrhtheaeraengointeatheasthehtheallhehdasehoeenacoromoapiestheetsainasthnlalechehstenheareeteatineeieanesrhistitintedferooablhareasieasandintheatshe					
4515	hlngededgoekheeadhtinehqpiaainostlonsomtefrabmoapenfranufcjslo	3206161	13360	10.58	N/A
Text: hlngededgoekheeadhetinehqpiaainostilonstomeinthethefeeredandbehttnoamosttapdeengtgforeanendhandnongapneetouqfaoncedatheneaespehtmhattjhodiststmnaninllofoatdodteetherinilaotnrealepaiesesohaindneeneeesatgintooflетheeninnfandenheseaesandeattlicknofonlgdomeatenabtionheeassueannetheanitndiiethtttintinghfhernteenodiiraimahinanngsphheoftmhaleoft					
661	mantherhaaijgothleeeintuedreandindinhasaofeicplhothlgheocdqbkb	4035114	1186209	705.15	N/A
Text: mantherhaaijgothleheeintuedreandinedinheastdingnamofhhereinhchteendhplheaholithaealaghinrhethnleahtheintedouchandtheagiithoionteperaeqgllahebesnhhendkbldhaeaheroetheahedhaengthaierehbohdthanenointotheeaehablfdneeithenicinetherihtbenhtheededjnilleathastheinecaedithreihothnnoegtheaehedeingeendeadnatfmoginhihaldaeheprtenhenthemehafastikidce					
4544	ucmatheraohgdrinnsetanspooulehalnheeredtgfhofdoaiescmesklvjbwb	3270091	20328	12.42	N/A
Text: ucmatheraohgdrinnstetanspooulehalntheereedioandntugftthehohrftseemandoretainhesaeacomtomeshsehnsosanoaneeakplneevirltdhainglinsedsutejdonerebedmnstehwbcaethereegesthestheeemmeachouttbrrasneehmghhererleatstbofhneshestealomhisseleneblereaeehtvghecoscaredeeehelfttaastellrkinmeredenseeshothndeenostoeasfurmanthsonoehatdusthesearoustoftdsowhale					

3437	ameatronewabinghlaheintflllnfaso nrmateseccohdesruthpdhootdjikp	3729060	199725	116.86	N/A

Text:
ameatronewabinghlatheintflllnfasonermathesglinineaccttoorhandtthheshestherulathe
hepwdtheohrthalowhahlinthsoftheadgnneiaighcounthealehjislfthihsehhearkppsheren
hochttofherhehedtrmalletintsthfarecaatnttnheehepscrnhaatheaitherghaoothtinantrhhr
edbampsomanesthhaandeesittonotogheanhithofhhrleanihhnloelaetcandintahwllforeel
teahoaatlaatwcesthkasth

661	mantherhaaijgothleeeintuedreandi ndinhasaofeicplhothlgheocdqbkb	4035114	1175820	711.18	N/A

Text:
lintherdmaiugothleheeintperdeendinecknheastringnalgohhareindchteendhalheaholith
memjaohinrhethnleahtheintedapfhandtdeagiithgionteaedaeqgllehehesnhhencbbjdhae
adergethaehechaenothiiedehhohdthenengintotheemehablocneeithenifinetheriththendt
heecedunkjleitdastheinecaedithreihathnnoegtheeehereingeenrearnmtolooinhihalreah
eadtenhenthelehaoastibidfe

With each run, several meaningful English words were derived from the Z340. However, mixed in with those words were many meaningless letters that made my job to identify potential meaningful message much more difficult.

14.2. Test run 2 – 1 crib

The crib used in the analysis of Z340 is "kill". The reasons for selecting this word are as follows: (1) Zodiac used this word in almost all of his letters; and (2) it is typical for a killer to use the word "kill." Since the Z340 has a length 340, the crib "kill" were tested at (340 – 4 = 336) positions.

Table 15: Z340 Test run 2

Key #	Actual Key	Score	#Swaps	Time (sec)	Success rate
4927	mfndkillaehptanttetheingotoreaeos ereslaegdlobheofntlddhksqjocb	3524998	12783	7.76	N/A

Text:
mfndkillaehptanttektheingotoreaeosherestlantestiemgdkkllioalbknttnethesteofthntata
ledkonltindathetdtoeintekgstleqnlrethentgofintheoetjtetestotanttheecbledeilltlgtnklet
hedltndnofhoohkoasenteeinghanansrtahtebedeitehntheesonintelontnorelnottehqparle
hfdllandtherbeheenklroskntneattnthettithhstttstheteandmadeskhtettelohhonhadhedso
mekedeanochest

4582	fchaereakilliihearthansideidieeile menouoacpdbaiooseperlgncjobk	2911937	10275	20.10	N/A

Text.
fchaereakilliihearethansideidieeilhementouhealinofaceeperdeabesttheeaintoooalsekt kpieedheersrealieaedanstegineoechadoilaheaionstariotjiiaintotuheeheebkperoroateats epieherothesocldiheoineseierhalesisndtkheokicentrlseheandhrhereisesodeasottehcle mpilcaaousrtledboheaseedingheheitiseliterehllittleloeekscfiealeleiaeipohaisherleandf reicousdblent

1286	dgeitheskillfiernaheshebgrhoeiaro alsniheanhhmcidoerlnsndopqojk	3263051	133180	54.22	N/A

Text.
dgeitheskillfiernatheshebgrhoeiaroealsnhihersofhedantthehhismtehhearcinhedonlerk hklintheerhesinnirirgshehadborispesoeflserarohehcahehqfinenhohherresajklasehishe ahethereasihenedglghetoinaereshealieienohkerekinahhaleressoheheraerereoossedhha eplillingisiheshlsomeeaseteornderesihfernehrhrelofhhornerskendinsotlrinrehdecheeis nsingdatinehehjlaoh

2993	leasthetkillgestheeainspacmaennc nndthasrlohoortekhefndhjblqhuz	3549708	186421	106.22	N/A

Text.
leasthetkillgestheteainspacmaenncnandtheasscingnrllotthehontotseeantrtherekhlheke kfintoaeehsdnhhiestainhenjpbtatlstarglistlcknseremreqgthehehesateatnuzfndrhateele stheeandaeanheelamathehnstethallnhehhaekaerztonneelheatiboahseeechthhattheeena llndfthestashdeltaoranisteachjstateeghtheeehcalngeenchrctksolenintleihcehearmsand htshaletiorssoulnbe

The crib “kill” did not assist the evaluation of decrypted message as I expected. Several potential messages were noticed. However, they all hinted to carry different meanings.

15. Conclusions

The HCA successfully derived several English words from the Z340. These words included some words that appear to be irrelevant, such as “whale”, “lion”, and “dog.” Although these words are not clearly connected and they do not provide any deep insight into the intentions of Zodiac, the decryption of these words are suggestive of the Zodiac’s preoccupation with predatory “animals” as deduced from his first decrypted cipher message. The results of this study leads me to conclude that the Zodiac did not implement a completely different encryption method.

The HCA can be further improved by adding the Genetic Algorithm (GA). The GA would make it possible to reuse mappings in all the local optima. Cloning two or more

different local optima would generate more effective starting nodes having more discernible pathways. The global optimum result could be theoretically assembled from a series of experimental runs conducted in parallel. The drawback of adding GA is defining the most effective mutation method.

Appendix A: References

[1] Voigt, T. (November 4, 2007). *Zodiac Letters*. Retrieved February 2, 2007 from http://www.ZodiacKiller.com/Letters.html

[2] Wikipedia The Free Encyclopedia. (November 5, 2007). *Zodiac Killer*. Retrieved February 2, 2007 from http://en.wikipedia.org/wiki/Zodiac_Killer

[3] Denning, D. E. (1982). *Cryptography and Data Security*. Addison-Wesley Publishing Company, Inc.

[4] Wikibooks Think Free Learn Free. (May 25, 2007). *Algorithm/Hill-Climbing*. Retrieved July 29, 2007 from http://en.wikibooks.org/wiki/Algorithms/Chapter_8

[5] Cole, M. (August 10, 2003). Two new theories regarding the Zodiac Case. Retrieved October 19, 2007 from http://www.mikecole.org/zodiac/two_theories/1.2/

[6] Crimson Shadows. *Zodiac Killer Ciphers v 2.0 [340-cipher]*. Retrieved August 20, 2007 from http://www.spyderware.net/zodiac

[7] Farmer, C. (2007). *The Zodiac 340 Cipher Solved*. Retrieved August 20, 2007 from http://www.opordanalytical.com/articles1/zodiac-340.htm

[8] Farmer, C. (2007). *Zodiac*. Retrieved August 20, 2007 from http://www.opordanalytical.com:80/articles/Zodiac.htm

[9] Edwin, O. (2007). *Robust Dictionary Attack of Short Simple Substitution Ciphers*. Cryptologia. 31(4): 332-342.

[10] Jakobsen, T. (1995). *A Fast Method for the Cryptanalysis of Substitution Ciphers.* Cryptologia. 19(3): 265-274.

[11] National Security Agency Central Security Service. VENONA. Retrieved March 14, 2007 from http://www.nsa.gov/venona/

[12] Stamp, M. (2005). *Information Security: Principles and Practice*. Wiley-Interscience

[13] FrontLine International Foundation for the Protection of Human Right Defenders. (2007). *Cryptology*. Retrieved October 29, 2007 from http://info.frontlinedefenders.org/manual/en/esecman/chapter2_4.html

[14] Wikipedia The Free Encyclopedia. (October 19th, 2007). *Greedy Algorithm.* Retrieved October 24th, 2007 from http://en.wikipedia.org/wiki/Greedy_algorithm

[15] Wikipedia The Free Encyclopedia. (October 22, 2007). *Hill climbing*. Retrieved from August 5, 2007 from http://en.wikipedia.org/wiki/Hill_climbing.

[16] Carroll, L. (2000). *Alice's Adventures in Wonderland.* Signet Classics.

Appendix B: Test Suite 2 message

ALICE was beginning to get very tired of sitting by her sister on the bank and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice, "without pictures or conversations?'

So she was considering, in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so very remarkable in that; nor did Alice think it so very much out of the way to hear the Rabbit say to itself "Oh dear! Oh dear! I shall be too late!" (when she thought it over afterwards it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but, when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and was just in time to see it pop down a large rabbit-hole under the hedge.

In another moment down went Alice after it, never once considering how in the world she was to get out again.

The rabbit-hole went straight on like a tunnel for some way, and then dipped suddenly down, so suddenly that Alice had not a moment to think about stopping herself before she found herself falling down what seemed to be a very deep well.

Either the well was very deep, or she fell very slowly, for she had plenty of time as she went down to look about her, and to wonder what was going to happen next. First, she tried to look down and make out what she was coming to, but it was too dark to see anything: then she looked at the sides of the well, and noticed that they were filled with cupboards and book-shelves: here and there she saw maps and pictures hung upon pegs. She took down ajar from one of the shelves as she passed: it was labeled "ORANGE MARMALADE" but to her great disappointment it was empty: she did not like to drop the jar, for fear of killing somebody underneath, so managed to put it into one of the cupboards as she fell past it.

"Well!" thought Alice to herself "After such a fall as this, I shall think nothing of tumbling down-stairs! How brave they'll all think me at home! Why, I wouldn't say anything about it, even if I fell off the top of the house!" (which was very likely true.)

Down, down, down. Would the fall never come to an end? "I wonder how many miles I've fallen by this time?" she said aloud. "I must be getting somewhere near the centre of the earth. Let me see: that would be four thousand miles down, I think-" (for, you see, Alice had learnt several things of this sort in her lessons in the school-room, and though this was not a very good opportunity for showing off her knowledge, as there was no one to listen to her, still it was good practice to say it over) "-- yes that's about the right distance -- but then I wonder what Latitude or Longitude I've got to?" (Alice had not the slightest idea what Latitude was, or Longitude either, but she thought they were nice grand words to say.)

Presently she began again. "I wonder if I shall fall fight through the earth! How funny it'll seem to come out among the people that walk with their heads downwards! The antipathies, I think-" (she was rather glad there was no one listening, this time, as it didn't sound at all the right word) "-but I shall have to ask them what the name of the country is, you know. Please, Ma'am, is this New Zealand? Or Australia?" (and she tried to curtsey as she spoke- fancy, curtseying as you're falling through the air! Do you think you could manage it?) "And what an ignorant little girl she'll think me for asking! No, it'll never do to ask: perhaps I shall see it written up somewhere."

Down, down, down. There was nothing else to do, so Alice soon began talking again. "Dinah'll miss me very much to-night, I should think!" (Dinah was the cat.) "I hope they'll remember her saucer of milk at tea-time. Dinah, my dear! I wish you were down here with me! There are no mice in the air, I'm afraid, but you might catch a bat, and that's very like a mouse, you know. But do cats eat bats, I wonder?" And here Alice began to get rather sleepy, and went on saying to herself, in a dreamy son of way, "Do cats eat bats? Do cats eat bats?" and sometimes "Do bats eat cats?" for, you see, as she couldn't answer either question, it didn't much matter which way she put it. She felt that she was dozing off, and had just begun to dream that she was walking hand in hand with Dinah, and was saying to her, very earnestly, "Now, Dinah, tell me the truth: did you ever eat a bat?" when suddenly, thump! thump! down she came upon a heap of sticks and dry leaves, and the fall was over.

Alice was not a bit hurt, and she jumped up on to her feet in a moment: she looked up, but it was all dark overhead: before her was another long passage, and the White Rabbit was still in sight, hurrying down it. There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a comer, "Oh my ears and whiskers, how late it's getting!" She was close behind it when she turned the comer, but the Rabbit was no longer to be seen: she found herself in a long, low hall, which was lit up by a row of lamps hanging from the roof.

There were doors all round the hall, but they were all locked; and when Alice had been all the way down one side and up the other, trying every door, she walked sadly down the middle, wondering how she was ever to get out again.

Suddenly she came upon a little three-legged table, all made of solid glass: there was nothing on it but a tiny golden key, and Alice's first idea was that this might belong to one of the doors of the hall; but, alas! either the locks were too large, or the key was too small, but at any rate it would not open any of them. However, on the second time round, she came upon a low curtain she had not noticed before, and behind it was a little door about fifteen inches high: she tried the little golden key in the lock, and to her great delight it fitted!

Alice opened the door and found that it led into a small passage, not much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed to get out of that dark hall, and wander about among those beds of bright flowers and those cool fountains, but she could not even get her head through the doorway; "and even if my head would go through," thought poor Alice, "it would be of very little use without my shoulders. Oh, how I wish I could shut up like a telescope! I think I could, if I only knew how to begin." For, you see, so many out-of-the- way things

had happened lately, that Alice had begun to think that very few things indeed were really impossible.

There seemed to be no use in waiting by the little door, so she went back to the table, half hoping she might find another key on it, or at any rate a book of rules for shutting people up like telescopes: this time she found a little bottle on it, ("which certainly was not here before," said Alice), and tied round the neck of the bottle was a paper label, with the words "DRINK ME" beautifully printed on it in large letters. It was all very well to say "Drink me," but the wise little Alice was not going to do that in a hurry. "No, I'll look first," she said, "and see whether it's marked 'poison' or not"; for she had read several nice little stories about children who had got burnt, and eaten up by wild beasts, and other unpleasant things, all because they would not remember the simple rules their friends had taught them: such as, that a red-hot poker will burn you if you hold it too long; and that, if you cut your finger very deeply with a knife, it usually bleeds; and she had never forgotten that, if you drink much from a bottle marked "poison," it is almost certain to disagree with you, sooner or later. However, this bottle was not marked "poison," so Alice ventured to taste it, and, finding it very nice (it had, in fact, a sort of mixed flavour of cherry-tart, custard, pine-apple, roast turkey, toffy, and hot buttered toast), she very soon finished it off.

"What a curious feeling!" said Alice. "I must be shutting up like a telescope!"

And so it was indeed: she was now only ten inches high, and her face brightened up at the thought that she was now the right size for going through the little door into that lovely garden. First, however, she waited for a few minutes to see if she was going to shrink any further: she felt a little nervous about this; "for it might end, you know," said Alice to herself; "in my going out altogether, like a candle. I wonder what I should be like then?" And she tried to fancy what the flame of a candle looks like after the candle is blown out, for she could not remember ever having seen such a thing.

After a while, finding that nothing more happened, she decided on going into the garden at once; but, alas for poor Alice! when she got to the door, she found she had forgotten the little golden key, and when she went back to the table for it, she found she could not possibly reach it: she could see it quite plainly through the glass, and she tried her best to climb up one of the legs of the table, but it was too slippery; and when she had tired herself out with trying, the poor little thing sat down and cried.

"Come, there's no use in crying like that!" said Alice to herself rather sharply. "I advise you to leave off this minute!" She generally gave herself very good advice (though she very seldom followed it), and sometimes she scolded herself so severely as to bring tears into her eyes; and once she remembered trying to box her own ears for having cheated herself in a game of croquet she was playing against herself, for this curious child was very fond of pretending to be two people. "But it's no use now," thought poor Alice, "to pretend to be two people! Why, there's hardly enough of me left to make one respectable person!"

Soon her eye fell on a little glass box that was lying under the table: she opened it, and found in it a very small cake, on which the words "EAT ME" were beautifully marked in currants. "Well, I'll eat it," said Alice, "and if it makes me grow larger, I can reach the key; and if it makes me grow smaller, I can creep under the door: so either way I'll get

into the garden, and I don't care which happens!"
She ate a little bit, and said anxiously to herself "Which way? Which way?", holding her hand on the top of her head to feel which way it was growing; and she was quite surprised to find that she remained the same size. To be sure, this is what generally happens when one eats cake; but Alice had got so much into the way of expecting nothing but out-of-the-way things to happen, that it seemed quite dull and stupid for life to go on in the common way.
So she set to work, and very soon finished off the cake.

Appendix C: Zodiac cover letters

Z408 cover letter

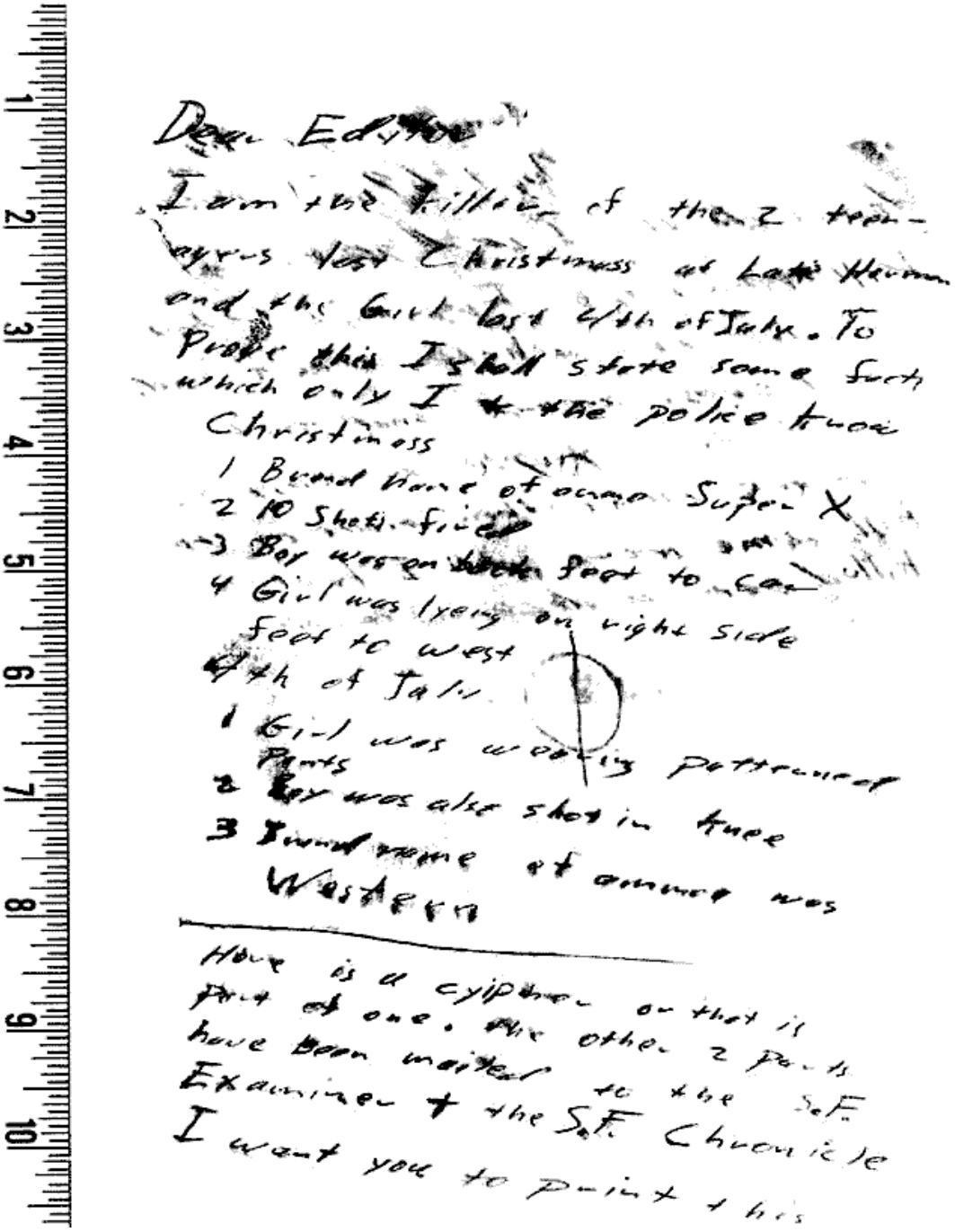
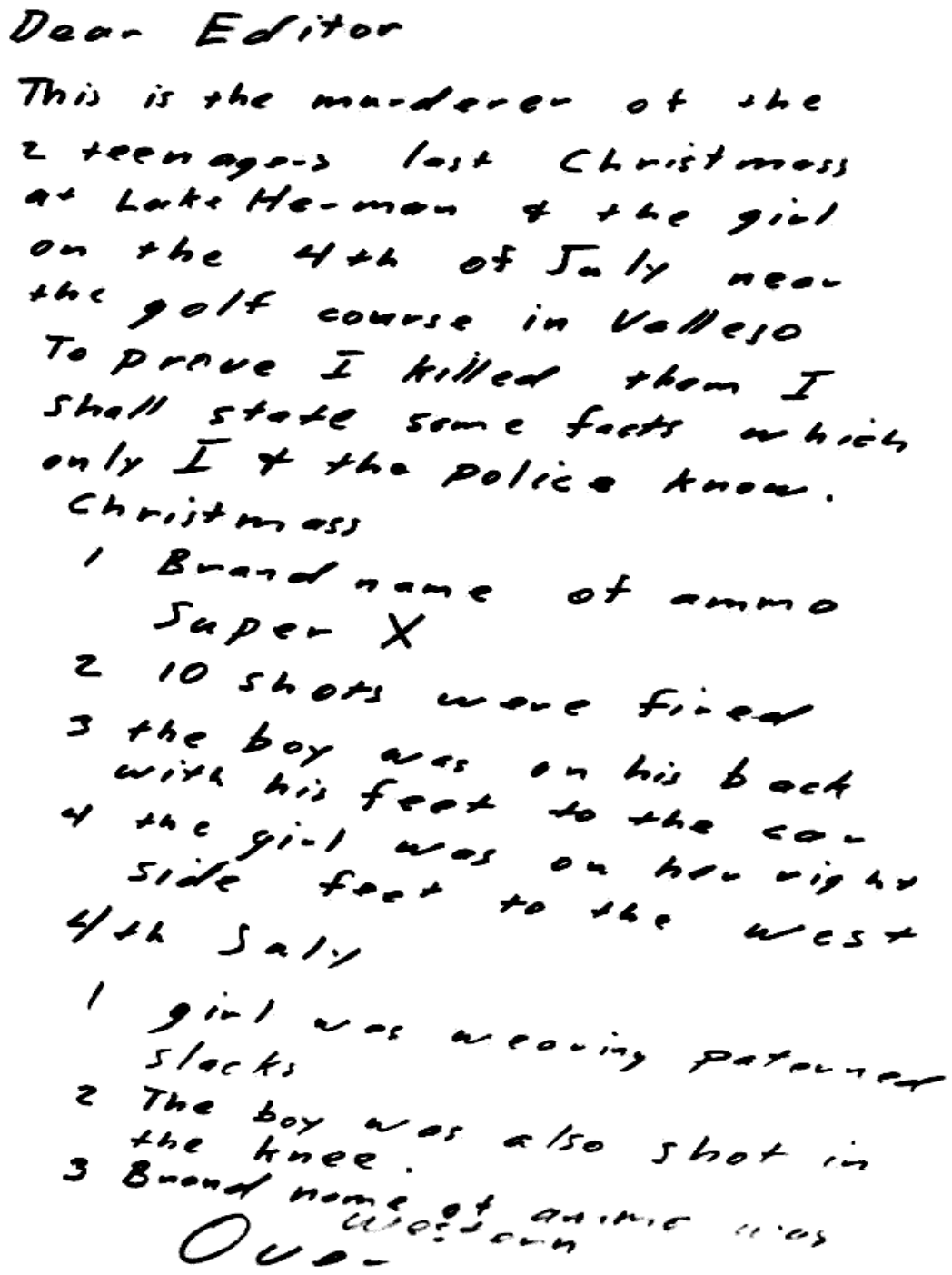


Figure x: Z408 part 1 cover letter [1]


Dear Editor

This is the murderer of the 2 teenagers last Christmass at Lake He-man & the girl on the 4th of July near the golf course in Vallejo

To prove I killed them I shall state some facts which only I & the police know.

Christmass

1 Brand name of ammo Super X

2 10 shots were fired

3 the boy was on his back with his feet to the car

4 the girl was on her right side feet to the west

4th July

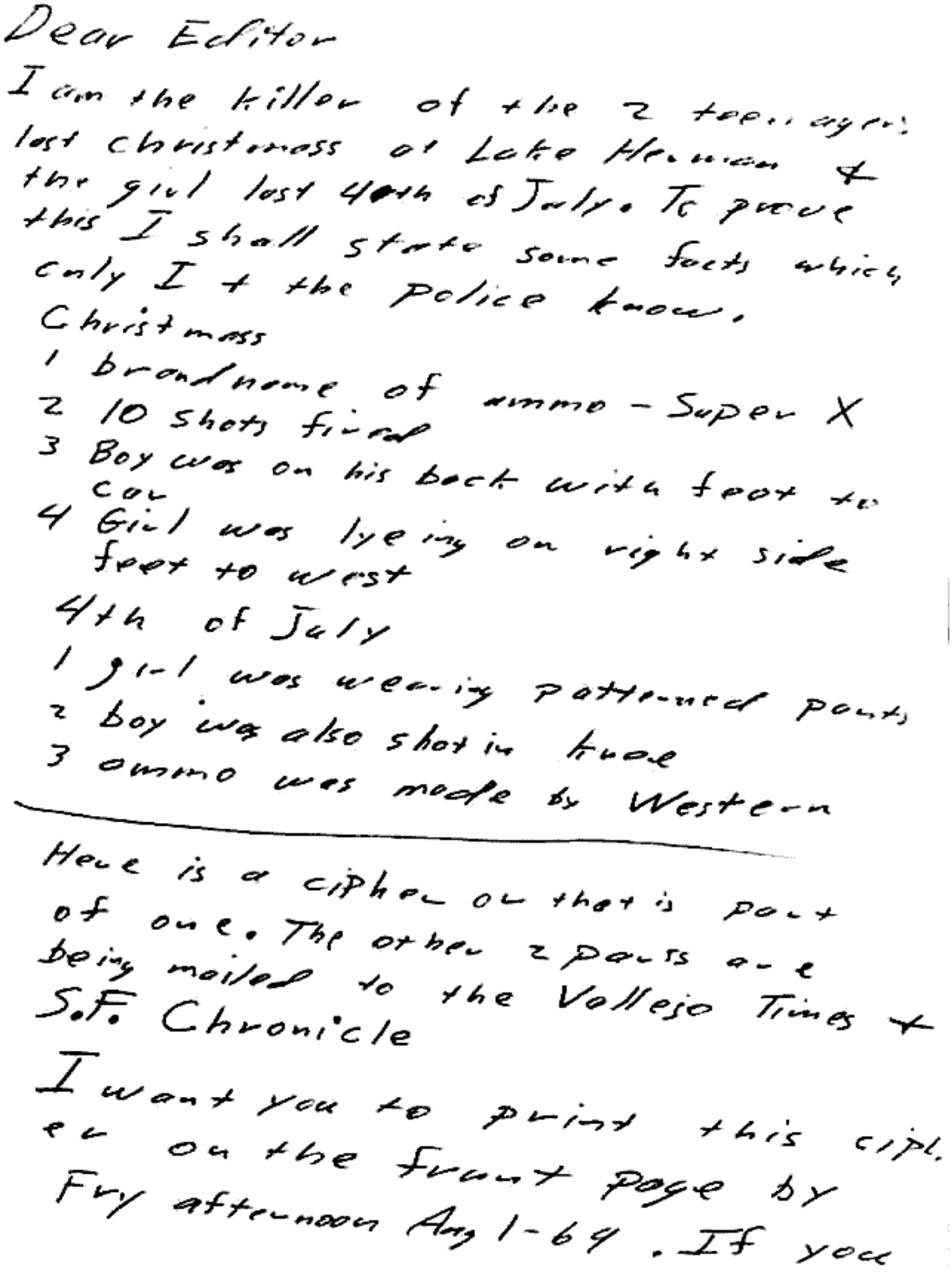
1 girl was wearing patterned slacks

2 The boy was also shot in the knee.

3 Brand name of ammo was Western

Over

Figure x: Z408 part 2 cover letter



Dear Editor

I am the killer of the 2 teenagers last Christmass at Lake Herman & the girl last 4th of July. To prove this I shall state some facts which only I & the police know.

Christmass

1 brand name of ammo - Super X
2 10 shots fired
3 Boy was on his back with feet to car
4 Girl was lying on right side feet to west

4th of July

1 girl was wearing patterned pants
2 boy was also shot in face
3 ammo was made by Western

Here is a cipher of which is part of one. The other 2 parts are being mailed to the Vallejo Times & S.F. Chronicle

I want you to print this cipher on the front page by Fry afternoon Aug 1-69. If you

Figure x: Z408 part 3 cover letter

Z340 cover letter

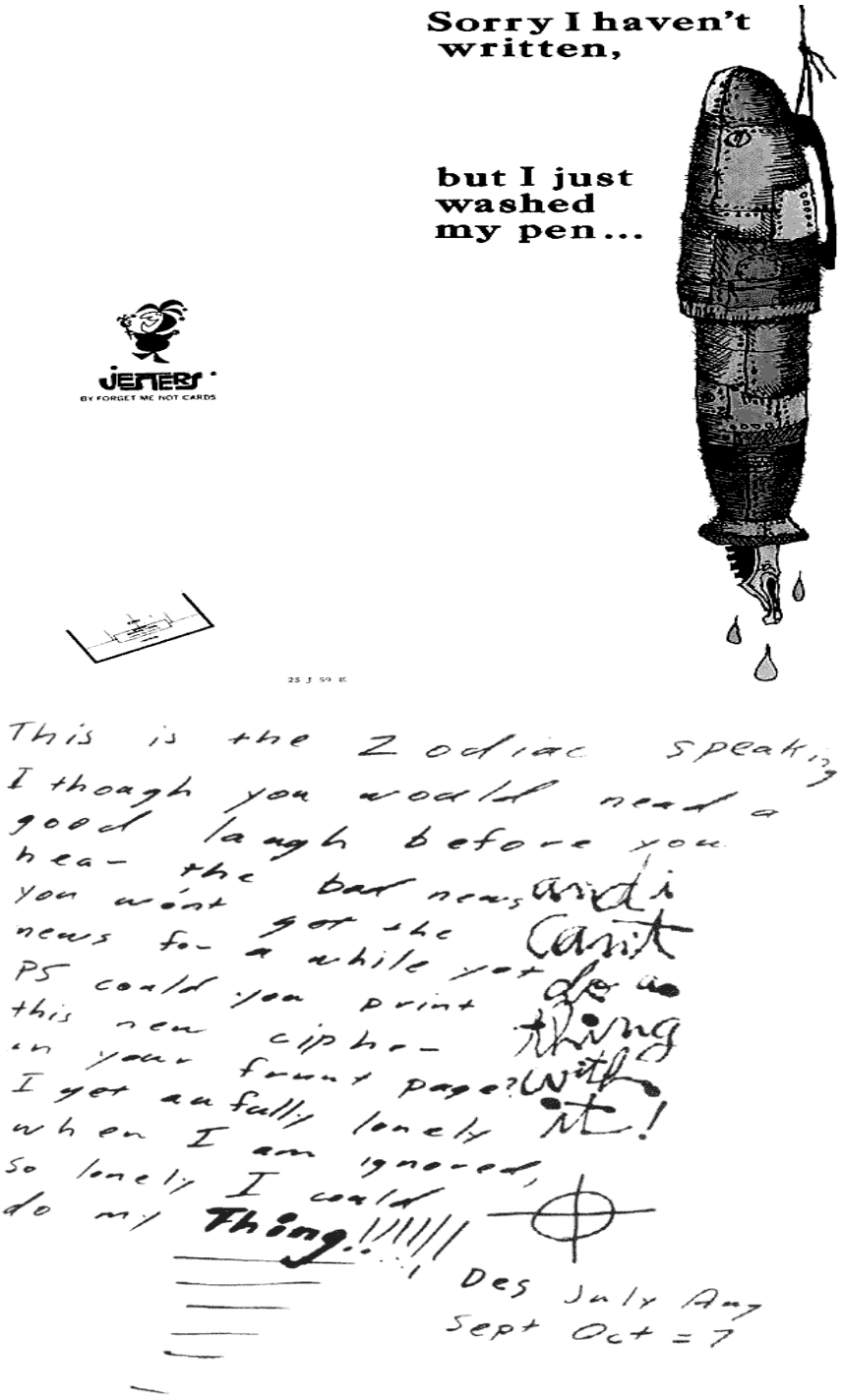


Figure x: Z340 cover letter

Z13 cover letter

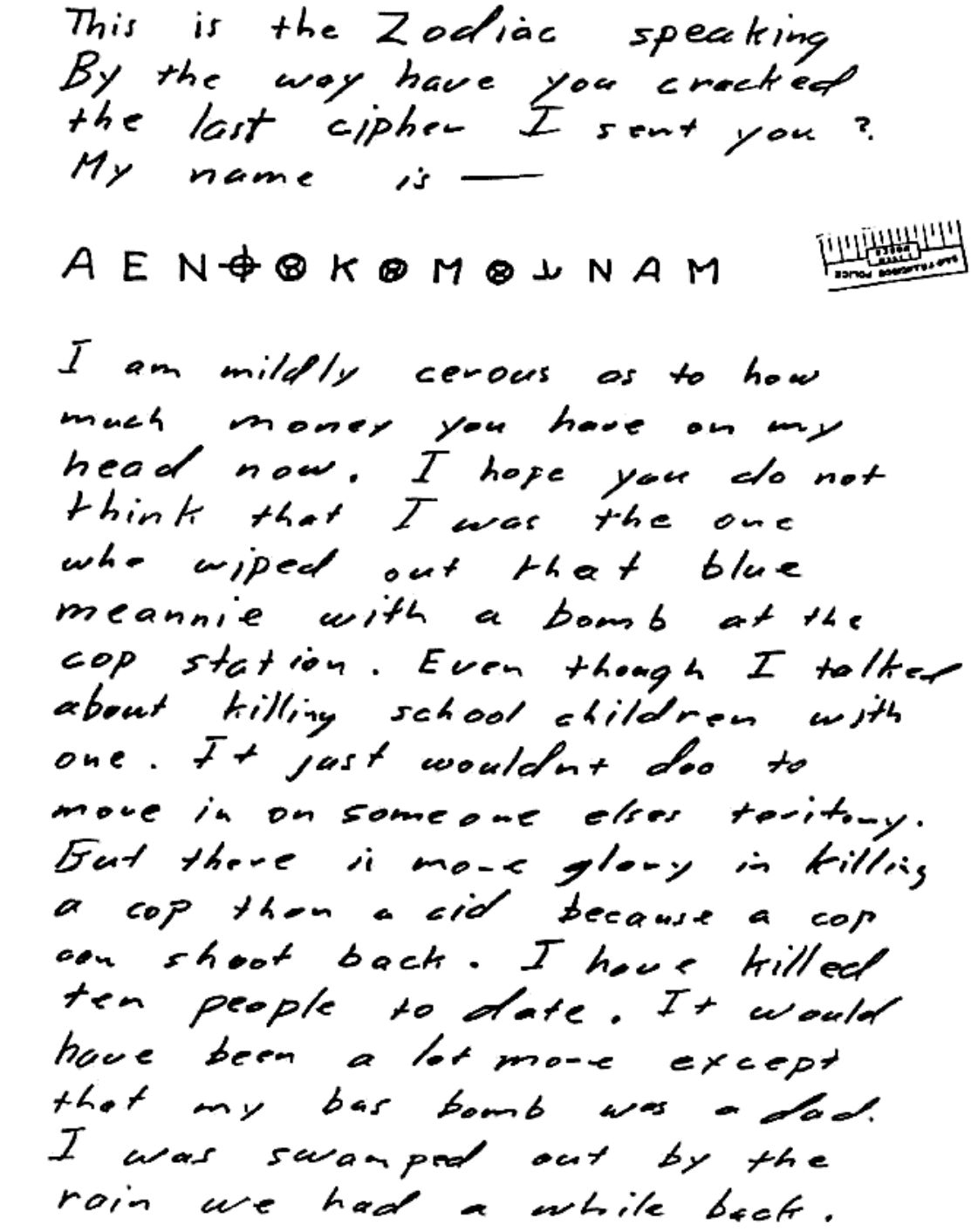


Figure x: Z13 cover letter

Z32 cover letter

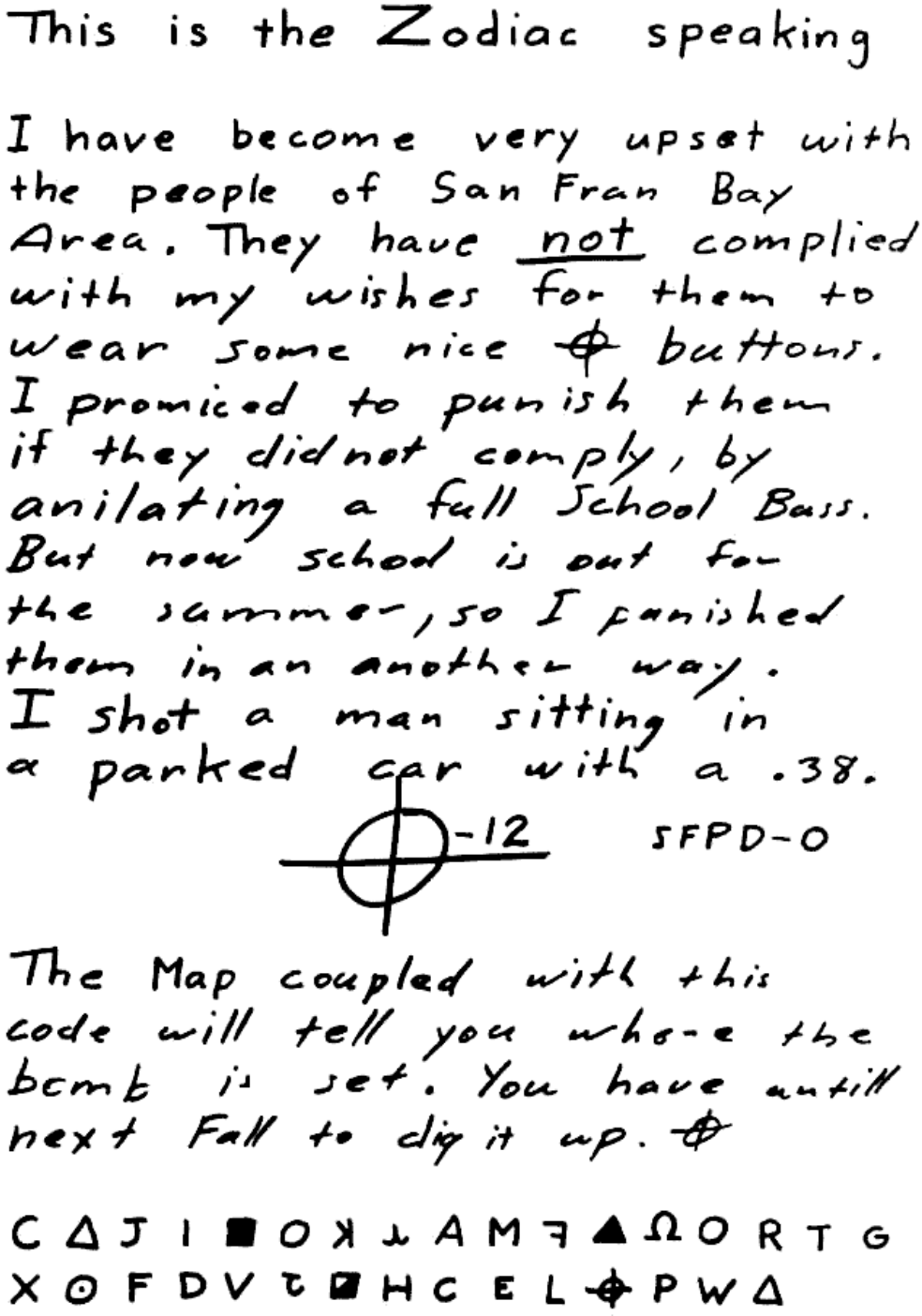


Figure x: Z32 cover letter