Polyalphabetic Ciphers

- One of the main problems with simple substitution ciphers (monoalphabetic ciphers) is that they are extremely vulnerable to frequency analysis. Given a sufficiently large ciphertext sample, it can easily be broken by mapping the frequency of its letters to the known frequencies of the language text.

- A very effective way to defeat frequency analysis is to flatten out the normal frequency distribution of letters by using more than one alphabet to encrypt the message, i.e., a polyalphabetic ciphers.

- A **polyalphabetic substitution cipher** involves the use of two or more cipher alphabets. Instead of there being a one-to-one relationship between each letter and its substitute, there is a one-to-many relationship between each letter and its substitutes.

The Vigenère Cipher

- Historically, the Vigenère Cipher was one of the great breakthroughs in the development of cryptography. The birth of this cipher can be traced back to the work of the Italian genius Leon Alberti.

- Born in 1404, Alberti was one the leading figures of the Renaissance; a painter, composer, poet and philosopher. He was also the author of the first scientific analysis of perspective, a treatise on the housefly, and a funeral oration for his dog.

- Although he ended up making one of the most significant breakthroughs in cryptography in over 10 centuries, he never implemented his system into a full-fledged cipher.

- Subsequently several cryptographers of the time took Alberti's seminal work and began to implement ciphers based on it. These included Johannes Trithemius (Steganography), a German abbot born in 1462, Giovanni Porta, an Italian scientist born in 1535, and Blaise de Vigenère, a French diplomat born in 1523.

- The Vigenère Cipher, proposed by Blaise de Vigenère from the court of Henry III of France in the sixteenth century, is a polyalphabetic substitution.

- It is thought to have remained unbroken until Charles Babbage broke it in the 19th century. It was called "le chiffre indéchiffrable", French for "the indecipherable cipher".

- Alberti's cipher used 2 cipher alphabets. However, the Vigenère cipher uses 26 distinct cipher alphabets. The 26 cipher alphabets are arranged in a tableau referred to as a Vigenère Square, which is used for encryption and decryption.

- The square has a plaintext alphabet followed by 26 cipher alphabets, each one shifted by one more letter with respect to the previous one. Hence, row number 1 represents a cipher alphabet with a Caesar shift of 1, row number 2 represents a cipher alphabet with a Caesar shift of 2, and so on.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	: Plaintext
1	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	
2	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	
3	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	
4	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	
5	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	
6	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	
7	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	
8	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	
9	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	
10	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	
11	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	
12	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	
13	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	
14	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
15	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
16	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
17	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
18	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
19	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
20	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
21	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
22	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	
23	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
24	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
25	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
26	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	

- The strength is this cipher is that there are 26 distinct cipher alphabets that can be used to encrypt a message. Thus, a plaintext character can encrypted using any one of 26 cipher alphabets.

- For example, if we use cipher alphabet 21, “z” will be encrypted as “u”, but if we use cipher alphabet 4, then “z” becomes “d”. In this way, the Vigenère cipher uses different rows to encrypt different letters in the message.

- To decrypt the ciphertext, the recipient will need to know which row of the Vigenère square was used to encrypt which letter. Therefore, there must be an a priori agreed upon on the row-switching sequence. This is achieved by using a keyword or key phrase.

Encryption:

- As an example we will encrypt a message: “**beware the ides of march**”, using a short keyword: “**folly**”.

- The keyword is spelt out above the message, repeated as many times as necessary so that each letter in the message is associated with a letter from the keyword.

- The table below illustrates this:

Keyword	f	o	l	l	y	f	o	l	l	y	f	o	l	l	y	f	o	l	l	y
Plaintext	b	e	w	a	r	e	t	h	e	i	d	e	s	o	f	m	a	r	c	h
Ciphertext	g	s	h	l	p	j	h	s	p	g	i	s	d	z	d	r	o	c	n	f

- To encrypt the first letter “**b**”, we identify the key letter above it, “**f**”, which in turn identifies a specific row in the Vigenère Square, **row 5**, the cipher alphabet row. Next we locate where the column headed by “**b**” intersects row 5, and locate the ciphertext “**g**”.

- This process is repeated for every letter in the plaintext. The complete ciphertext is shown in the table.

Decryption:

- Decryption of the ciphertext is equally straightforward. We simply repeat the encryption process in reverse. Write the keyword repeatedly above the ciphertext similar to the encryption table as follows:

Keyword	f	o	l	l	y	f	o	l	l	y	f	o	l	l	y	f	o	l	l	y
Ciphertext	g	s	h	l	p	j	h	s	p	g	i	s	d	z	d	r	o	c	n	F
Plaintext	b	e	w	a	r	e	t	h	e	i	d	e	s	o	f	m	a	r	c	h

- This time we use the **keyword letter** to pick a column of the table and then trace down the column to the row containing the **ciphertext letter**.

- The **index of that row** is the plaintext letter.

- To decrypt the first letter “**g**”, we identify the key letter above it, “**f**”, and then move down the column until we reach down to the letter “**g**”. The index of that row (**row 1**) is the letter “**b**”, which is the plaintext letter.

- Repeat this process for the rest of the ciphertext.

- The resilience of the Vigenere cipher against frequency analysis attacks becomes apparent when we examine the ciphertext. There are four “**e**”s in the plaintext message and they have been encrypted by “**s**”, “**j**”, “**p**”, and “**s**” respectively.

- The fact that a letter which appears in the plaintext can represent a different plaintext on each occasion results in a lot of ambiguity for the analyst.
- The more letters in the Vigenère key, the stronger the encrypted message will be against a bruteforce attack. The choice of “folly” is a weak choice because it only has five letters. A key with only five letters has $26^{5\,(}$11,881,376) possible combinations.

- An average desktop will be able to attempt them all in 2 or 3 hours by first decrypting the message with the key “aaaa” and check if the resulting decryption was in English. Then it can attempt “aaaab”, “aaaac”, and so on, working through all of the alphabetical letter combinations until it settles on “folly”.

- The following table shows the number of possible keys based on the key length (26^n; where n = key length):

Key Length	Possible Keys
1	26
2	676
3	17,576
4	456,976
5	11,881,376
6	308,915,776
7	8,031,810,176
8	208,827,064,576
9	5,429,503,678,976
10	141,167,095,653,376
11	3,670,344,486,987,776
12	95,428,956,661,682,176
13	2,481,152,873,203,736,576
14	64,509,974,703,297,150,976

- Once we get beyond 12-letter keys we start to get into quadrillions (10^{15}) of keys, which makes the code-breaking effort completely out of reach for any consumer grade machine architectures.

- Nevertheless, The Vigenère Cipher today is just moderately good. No serious cryptologist would use it for secure information transmission.

- The code example provided illustrates the implementation of the Vigenère Cipher.

Breaking the Vigenère Cipher

- To most cryptanalysts the Vigenère Cipher appeared to be unbreakable, primarily due to its use of up to 26 different cipher alphabets. Over time, the Vigenère cipher became known as 'Le Chiffre Undechiffrable', or 'The Unbreakable Cipher'.

- It wasn't until 1854, over two hundred years later, that the Vigenère Cipher was finally cracked by the British cryptographer Charles Babbage. Babbage applied a very careful statistical analysis on the structure of groups of letters and a great deal of hard work. He never published his work in his lifetime, and it was over a hundred years later, in the 1970's, that his technique was finally made public.

- Later studies revealed he used a method that was later published by the early 20th-century mathematician Friedrich Kasiski. "**Kasiski Examination**" is a process used to determine how long the Vigenère key used to encrypt a ciphertext was. After this is determined, frequency analysis can be used to break each of the subkeys.

- A Vigenère key does not have to be a word like "**folly**". It can be any combination of letters, such as "**dkguhelokori**". In fact, the use of dictionary words is anathema as far as securing information is concerned.

- The word "cryptography" is a 12-letter key that is easier to remember than "dkguhelokori" even though they have the same number of letters. But a cryptanalyst would typically start with the assumption that the cryptographer is being lazy by using an English dictionary word for the Vigenère key.

- There are 95,428,956,661,682,176 possible 12-letter keys, but there are only about 1,800 12-letter words in the previous dictionary file. For a 12-letter English word key, it would be easier to brute-force that ciphertext than it would be to brute-force the ciphertext from a 3-letter random key.

- The cryptographer is helped by the fact that the cryptanalyst does not know how many letters long the Vigenère key is. But the cryptanalyst could try all 1-letter keys, then all 2-letter keys, and so on.

Babbage’s Method

- The most important component of Babbage’s work is centered on the observation that whole words will be encrypted into different ciphertext depending on the word’s position relative to the keyword.

- For example, if we use a four-letter keyword, then the word “**the**” will be enciphered as: **aph**, or **bki** depending on its relative position to the keyword.

- There will only four ways to encrypt the word “the”. This means that if the message contains several instances of “the”, then it is highly likely that one or more of the encipherments will be repeated in the ciphertext.

- Consider the following example that uses the keyword “**hide**” to encrypt a plaintext string:

Keyword	**h**	**i**	**d**	e	h	i	d	e	h	**i**	**d**	**e**	h	i	d	e	h	**i**	**d**	**e**	h	i	d	e
Plaintext	**t**	**h**	**e**	s	u	n	a	n	d	**t**	**h**	**e**	m	a	n	i	n	**t**	**h**	**e**	m	o	o	n
Ciphertext	**a**	**p**	**h**	w	b	v	d	r	k	**b**	**k**	**i**	t	i	q	m	u	**b**	**k**	**i**	t	w	r	r

- The word “**the**” is encrypted as “**aph**” in the first instance, the word “**bki**” in the second and third instances.

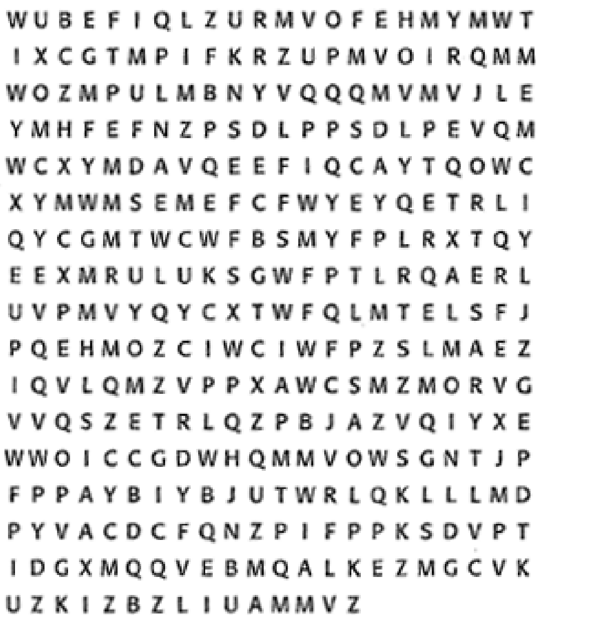
- Notice that the duplicate “**bki**” in the third instance is displaced by **8 positions** relative to the “**bki**” in the second instance. **Eight** is a multiple of the length of the keyword which is **four**.

- The second “the” was encrypted according to its relationship to the keyword (“**the**” is directly below “**ide**”), and by the time the algorithm reached the third “the”, the keyword has cycled twice and it is also positioned under “**ide**” again. Hence the reason for the duplicate “**bki**”.

- Babbage was able to define a series of simple steps that would allow a cryptanalyst to break the Vigenère Cipher.

- To illustrate his techniques we will examine an example taken from “**The Code Book**” by Simon Singh (pages 70 – 77).

- The following ciphertext was encrypted using the Vigenère Cipher, but nothing is known about the original message or the keyword.



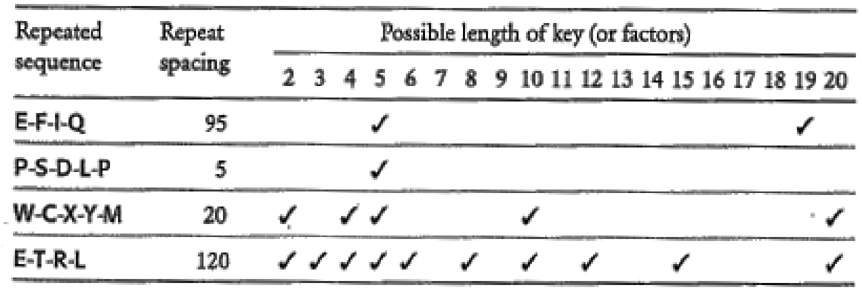
```
W U B E F I Q L Z U R M V O F E H M Y M W T
I X C G T M P I F K R Z U P M V O I R Q M M
W O Z M P U L M B N Y V Q Q Q M V M V J L E
Y M H F E F N Z P S D L P P S D L P E V Q M
W C X Y M D A V Q E E F I Q C A Y T Q O W C
X Y M W M S E M E F C F W Y E Y Q E T R L I
Q Y C G M T W C W F B S M Y F P L R X T Q Y
E E X M R U L U K S G W F P T L R Q A E R L
U V P M V Y Q Y C X T W F Q L M T E L S F J
P Q E H M O Z C I W C I W F P Z S L M A E Z
I Q V L Q M Z V P P X A W C S M Z M O R V G
V V Q S Z E T R L Q Z P B J A Z V Q I Y X E
W W O I C C G D W H Q M M V O W S G N T J P
F P P A Y B I Y B J U T W R L Q K L L L M D
P Y V A C D C F Q N Z P I F P P K S D V P T
I D G X M Q Q V E B M Q A L K E Z M G C V K
U Z K I Z B Z L I U A M M V Z
```

Figure 13 The ciphertext, enciphered using the Vigenère cipher.

- The first step is to identify sequences of letters that appear more than once in the ciphertext. For long sequences, the most likely reason for repetitions is that the same sequence of letters in the plaintext has been encrypted using the same segments of the keyword.

- The following table summarizes all of all such repetitions as well as the intervals between the repetitions. For example, the sequence **EFIQ** appears in the first line of the ciphertext and then in the fifth line, shifted by 95 letters.

Table 8 Repetitions and spacings in the ciphertext.



Repeated sequence	Repeat spacing	Possible length of key (or factors)																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
E-F-I-Q	95				✓														✓	
P-S-D-L-P	5				✓															
W-C-X-Y-M	20	✓		✓	✓					✓										✓
E-T-R-L	120	✓	✓	✓	✓	✓		✓		✓		✓			✓					✓

- The main objective now is to determine the length of the keyword. The data in table provides very good clues in order to estimate this. What we need to do is to use the spacing between the sequence repetitions and calculate the **factors** of the spacing.

- For example, the sequence **WCXYM** repeats itself after **20 characters**, and 1, 2, 4, 5, 10 and 20 are all factors of 20 (they divide into 20 without zero remainders).

- These factors suggest the following possibilities:

  (1). The key is 1 letter long and repeated 20 times between encryptions
  (2). The key is 2 letters long and repeated 10 times between encryptions
  (3). The key is 4 letters long and repeated 5 times between encryptions
  (4). The key is 5 letters long and repeated 4 times between encryptions
  (5). The key is 10 letters long and repeated 2 times between encryptions
  (6). The key is 20 letters long and repeated once between encryptions

- The first possibility is very unlikely due to the fact that a one-letter key would result in a monoalphabetic cipher.

- In this way we collect all of the repeated sequence intervals and list their factor of the intervals. The table summarizes all of the factors. It becomes clear that the factor 5 is the predominant value for all the repeated sequence intervals.

- At this point it is very reasonable to start the cryptanalysis process with the assumption that the keyword is of length 5.

- The next step is to break the cipher by determining the actual letters of the keyword. Let us denote the keyword as $\mathbf{L_1,L_2,L_3,L_4,L_5}$, where $\mathbf{L_1}$**,** represents the first letter of the keyword, $\mathbf{L_2}$ the second, and so on.

- Now, the letter $\mathbf{L_1}$ represents the first row of the Vigenère square used to encrypt the first plaintext letter, $\mathbf{L_2}$ is the next row used to encrypt the second plaintext letter and so on until we reach the last letter of the keyword.

- However, the sixth letter of the plaintext will be encrypted using $\mathbf{L_1}$ again (remember, 5 letter keyword), and the seventh plaintext letter using $\mathbf{L_2}$ and so on, the cycle repeats itself.

- At this point we know that the first letter of the keyword specifies one of the row of the Vigenère square which was used to encrypt the **1st, 6th, 11th, 16th, …..nth** letters (remember, 5 letter keyword) of the ciphertext.

- Now, if we examine the **1st, 6th, 11th, 16th, …..nth** letters of the full ciphertext, we can perform a frequency analysis of the sequence and compare it to a standard English plaintext sample containing the same number of letters as the ciphertext, there is a good probability that the cipher alphabet can be determined.

- The following is a frequency distribution of the sequence of letters in the **1st, 6th, 11th, 16th, …..nth** positions of the full ciphertext, which are **W, I, R, E…….:**

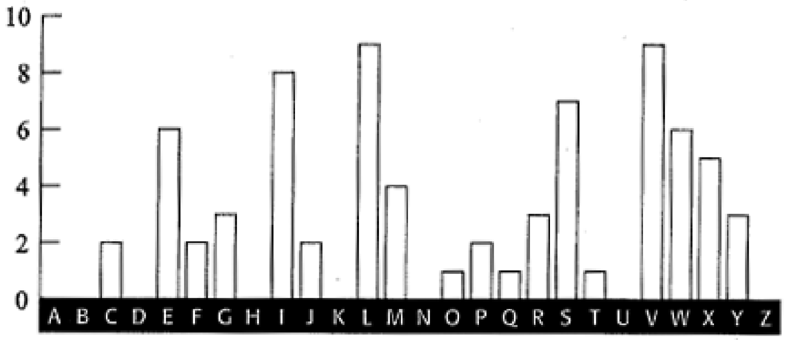


Figure 14 Frequency distribution for letters in the ciphertext encrypted using the L_1 cipher alphabet (number of occurrences).

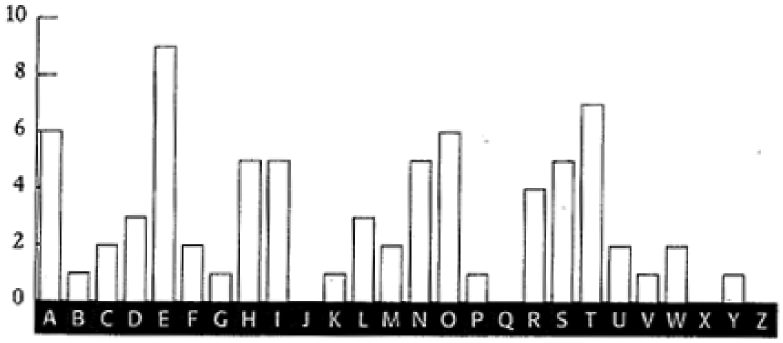


Figure 15 Standard frequency distribution (number of occurrences based on a piece of plaintext containing the same number of letters as in the ciphertext).

- Also shown is a frequency distribution of an English plaintext sample containing the same number of letters as the ciphertext.

- Keep in mind that the cipher alphabet in the Vigenère square is simply the standard alphabet shifted by an offset between 1 and 26.

- This means that the ciphertext frequency (Figure 14) distribution should have statistical features similar to the standard distribution (Figure 15), except that those statistical patterns will be shifted by some offset.

- The standard distribution has a series of peaks, plateaus, and valleys which we can use to identify similar peaks, plateaus, and valleys in the ciphertext distribution.

- For example, notice the three **peaks** at **R,S,T,** followed by a **valley** ranging from **U to Z** is a very distinctive pattern. If we look for a similar pattern in the ciphertext distribution, the closest similar pattern is noticeable at the **peak** sequence **V,W,X**, followed by a **valley** from **Y to D**.

- Now, if we shift the ciphertext distribution back until the characteristics of the two frequency distributions line up, we observe that the letters encrypted using **L_1** have been shifted four places, or stated another way, there is a very strong possibility that **L_1** defines a cipher alphabet which begins with the letter **E**.

- The following histograms illustrate a very strong correlation between the major peaks, implying a that assuming E to be the first letter in the keyword is a good assumption to proceed with:

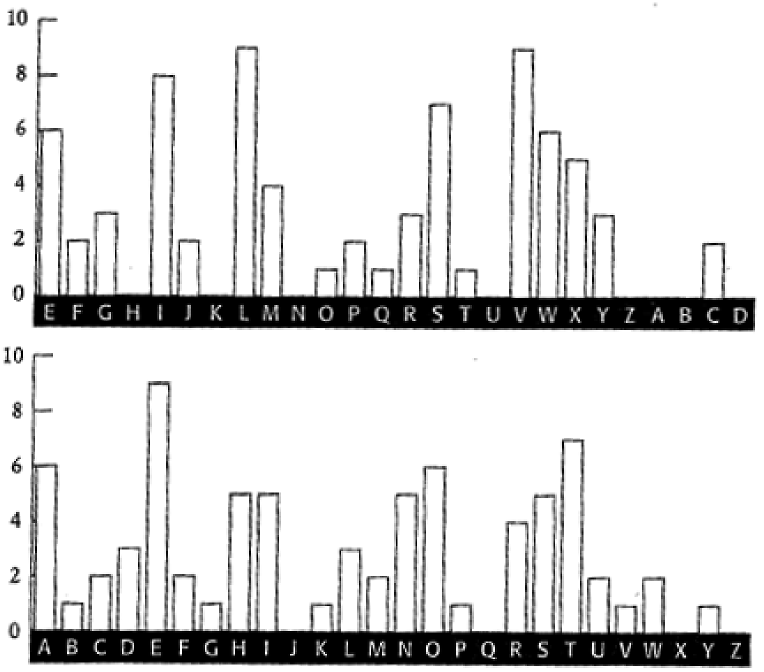


Figure 16 The L_1 distribution shifted back four letters (top), compared with the standard frequency distribution (bottom). All major peaks and troughs match.

- Next, we perform a similar analysis for the second keyword letter, **L_2** and generate a frequency distribution of the **2nd, 7th, 12th, 17th, …..nth** positions of the full ciphertext as shown below:

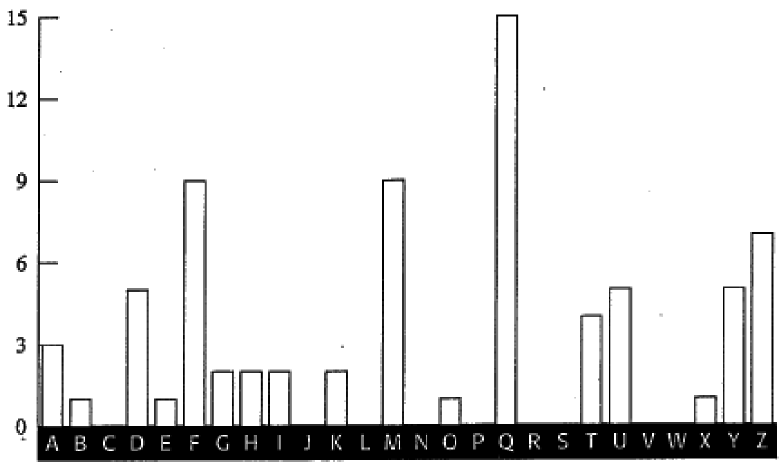


Figure 17 Frequency distribution for letters in the ciphertext encrypted using the L_2 cipher alphabet (number of occurrences).

- This distribution is a bit more challenging to analyze, since there are no obvious matching characteristics similar to the last case. However, in cryptanalysis work intuition and imagination play a very essential part.

- Let us assume that the ciphertext sequence consisting of three major peaks at D,E,F, followed by a plateau is a characteristic similar to what we observed in the analysis for **L_1**, i.e., it is match for the plaintext distribution sequence of three **peaks** at **R,S,T,** followed by a **valley** ranging from **U to Z.**

- The obviously missing part is the peak at E in the ciphertext distribution. For now, let us assume that we can attribute that to a statistical anomaly and proceed with our assumption.

- Just as before we shift the ciphertext distribution back until the characteristics of the two frequency distributions line up as shown:

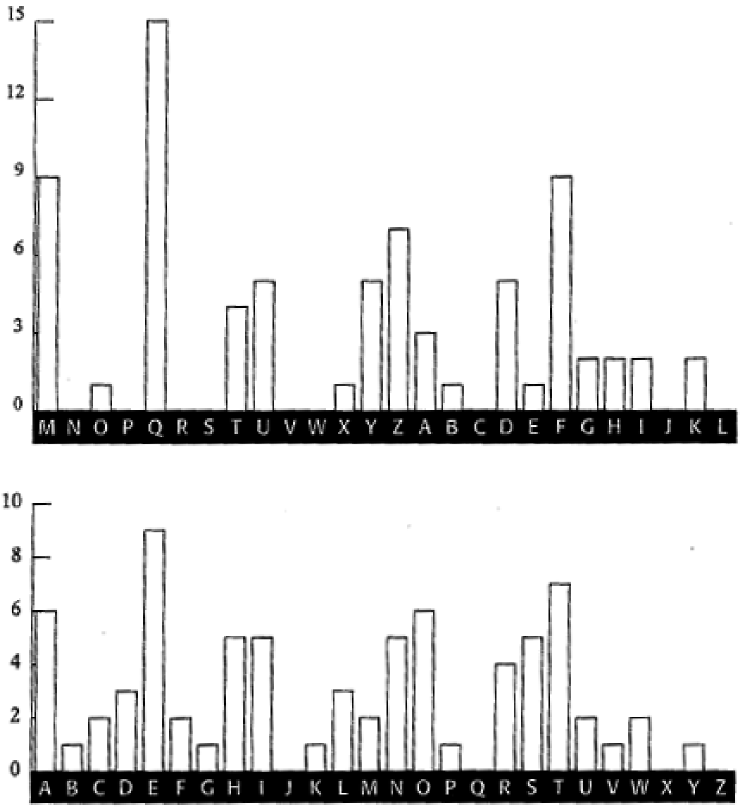


Figure 18 The L_2 distribution shifted back twelve letters (top), compared with the standard frequency distribution (bottom). Most major peaks and troughs match.

- We now observe a strong correlation between the major peaks of the two distributions, thus giving weight to our earlier assumption. The graphs suggest that all letters encrypted using $\mathbf{L_2}$ have been shifted 12 positions, meaning that $\mathbf{L_2}$ defines a cipher alphabet which has the sequence **M,N,O,P…**, and that $\mathbf{L_2}$ is the letter **M**.

- We continue with a similar analysis for the rest of the letters of the keyword until we have a good “guess” at the keyword. In this case it turns out to be “**EMILY**” (a dictionary word). The next step would be to decrypt the ciphertext using the keyword and see if it produces coherent English words.

- The plaintext consisted of verses from an Alfred Tennyson poem (try it).

- Unfortunately Babbage having cracked the Vigenère cipher, never published his work and it never received public recognition until scholars in the 20th century examined Babbage's notes in detail.

- Meanwhile his technique was independently discovered by Friedrich Wilhelm Kasiski, a retired Prussian army officer in 1863.

The Kasiski/Kerckhoff Method

- Vigenere-like substitution ciphers were generally considered unbreakable for almost 300 years until 1863 when a Prussian major named Kasiski proposed a method for breaking a Vigenere cipher.

- The first part of Kasiski Examination is to find every repeated set of letters at least three letters long in the ciphertext. These are significant, because they could indicate that they were the same letters of plaintext encrypted with the same subkeys of the key.

- Kasiski's technique for finding the length of the keyword was based on measuring the distance between repeated sequences in the ciphertext. By measuring and factoring the distances between recurring sequences, Kasiski was able to guess the length of the keyword.

- Once the length of the keyword is known, the ciphertext can be broken up into that many simple substitution cryptograms. That is, for a keyword of length 9, every 9-th letter in the ciphertext was encrypted with the same keyword letter.

- Given the structure of the Vigenère square, this is equivalent to using 9 distinct simple substitution ciphers, each of which was derived from 1 of the 26 possible Caesar shifts given in the tableau.

- The pure Kasiski method proceeds by analyzing these simple substitution cryptograms using frequency analysis and the other standard techniques.

- A variant of this method, proposed by the French cryptographer Kerckhoff, is based on discovering the keyword itself and then using it to decipher the cryptogram.

- In Kerckhoff's method, after the message has been separated into several columns, corresponding to the simple substitution cryptograms, one tallies the frequencies in each column and then uses frequency and logical analysis to construct the key.

- For example, suppose the most frequent letter in the first column is 'K'. We would hypothesize that 'K' corresponds to the English 'E'. If we consult the Vigenere tableau at this point, we can see that if English 'E' were enciphered into 'K' then row G of the table must have been the alphabet used for the first letter of the keyword. This implies that the first letter of the keyword is 'G'.

- The problem with this "manual" approach is that for short messages there are often several good candidates for English 'E' in each column. This requires the testing of multiple hypotheses, which can get quite tedious and involved.

- A more sensitive test to discover the alphabet used by each letter of the keyword is using a chi-square statistical test. Recalling that each row of the Vigenere tableau is one of the 26 Caesar shifts, we can use the chi-square test to determine which of the 26 possible shifts was used for each letter of the keyword. This modern day version of the Kerckhoff method is a much more effective technique.

- You will be required to design and implement an application to break a basic Vigenère Cipher as long as the keyword is a dictionary match. As mentioned earlier, the use of very long keywords or phrases makes the process far too difficult for the average consumer machine architecture.