MAS 433: Cryptography

Lecture 2

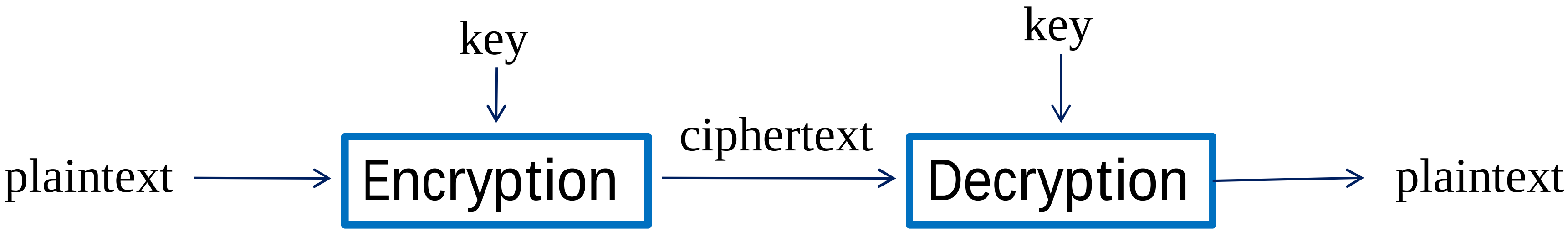
Classical Ciphers (Part 1)

Wu Hongjun

Lecture Outline

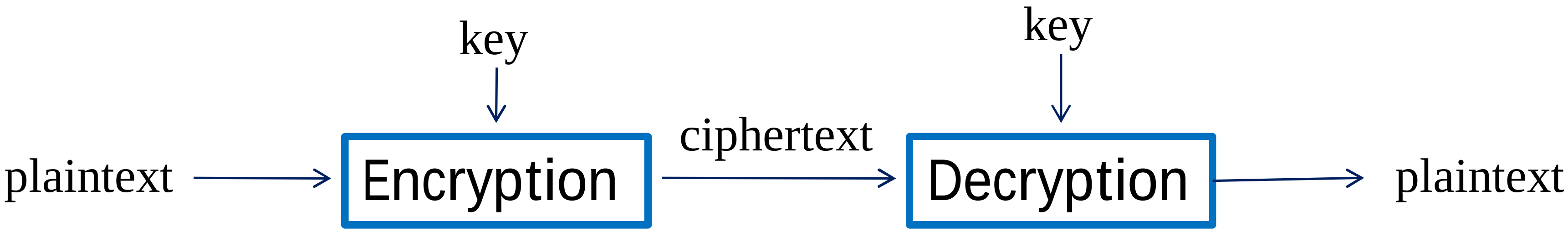
- Shift cipher (Caesar cipher)
- Substitution cipher, frequency cryptanalysis
- Vigenere Cipher
- Transposition (permutation) cipher

Encryption/Decryption



- Encryption
  - the process of transforming information to make it unreadable, except those possessing special knowledge
- Decryption
  - the inverse of encryption
- Cipher
  - the algorithm or device used for encryption/decryption
- Key: the secret information being used in a cipher
- Plaintext: the message to be encrypted
- Ciphertext: the encrypted message

Symmetric Key Cipher



- Symmetric key cipher
  - the key used for encryption is the same as that used for decryption
- Classical ciphers
  - developed before computer era
  - all the classical ciphers are symmetric key ciphers

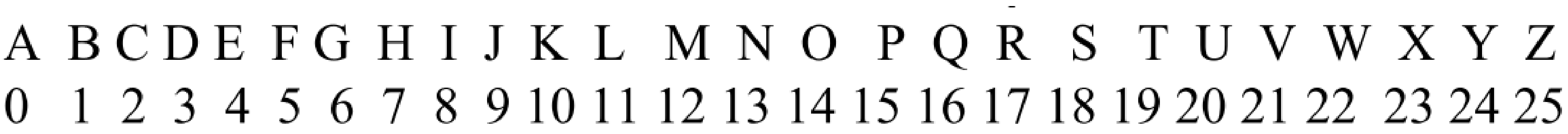
Shift cipher (Caesar cipher)

- Key
  - an integer; $1 \le K \le 25$ (for English with 26 letters)
- Encryption
  - Each letter in the plaintext P is replaced with the *K*'th letter following that letter (alphabetical order)
- Decryption
  - Each letter in the ciphertext C is replaced with the *K*'th letter before that letter

Shift cipher (Caesar cipher)

Plaintext = CRYPTOGRAPHYISFUN
K = 2
Ciphertext = ETARVQITCRJAKUHWP

Formally, let



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
0	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

then encryption: $c_i = (p_i + K) \bmod 26$
decryption: $p_i = (c_i – K) \bmod 26$

For example, to encrypt ‘Y’:
‘Y’ = 24 => 24+2 = 0 => 0 = ‘A’

Shift cipher (Caesar cipher)

- Caesar cipher
  - Shift cipher with K = 3
  - Used by Rome troops
- How about the security of shift cipher?
  - It is difficult for a person who has never heard about shift cipher to break it
  - But for a person who knows how this cipher works, shift cipher is too weak
    - Only 25 possible keys

      => try every possible key to break it (brute force)

Substitution Cipher

- An invertible secret substitution table *S* (the key)
  - Ecnryption: $c_i = S(\, p_i\,)$ mod 26
  - Decryption: $p_i = S^{-1}(\, c_i\,)$ mod 26
- Example
  - Let the secret table *S* be given as

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

  - Then
    - Plaintext: B E C A U S E
    - Ciphertext: A Z D B J S Z

Substitution Cipher

- Security of substitution cipher
  - Brute force search for the key is infeasible
    - key space size (the number of possible substitution tables) is huge: $26! \approx 4 \times 10^{26} \approx 2^{88.4}$
    - If we try one billion keys per second, it takes about 13 billion years to try all the keys
  - Thought to be unbreakable
    - until the invention of frequency analysis

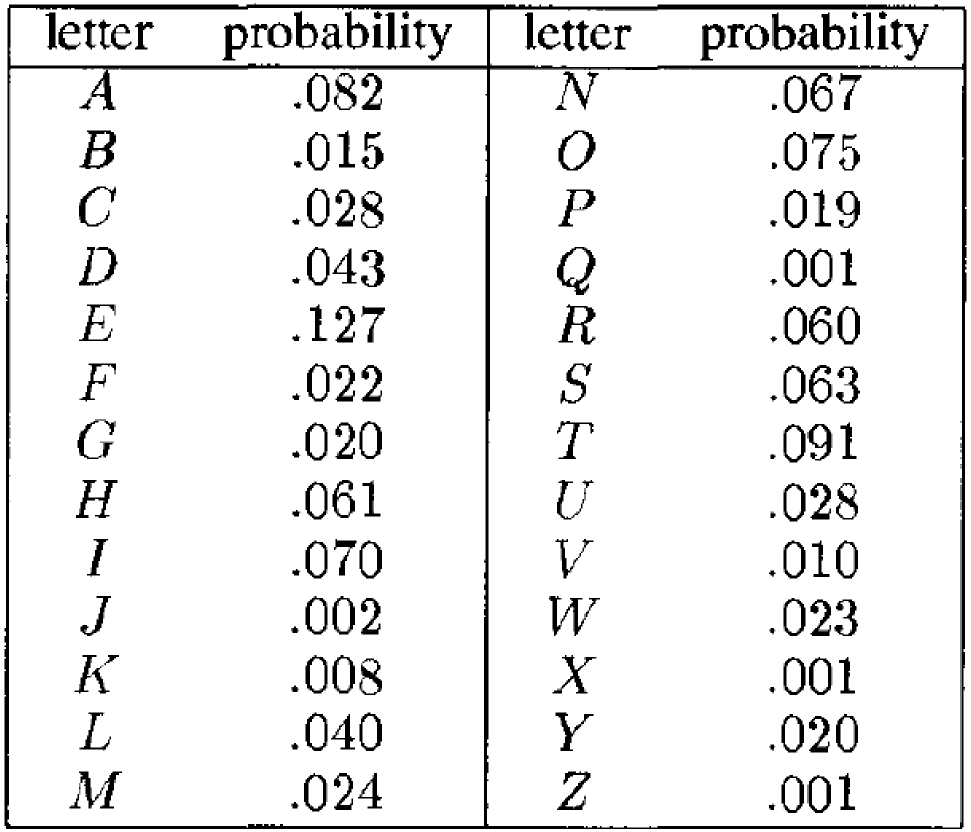
Frequency Analysis

- Invented by Arabian scientist al-Kindi in the 9th century
- Main idea
  - the encryption of substitution cipher does not randomize the frequency of occurrence of letters properly
  - calculating the frequency of occurrence of letters in ciphertext; comparing those frequency with the frequency of occurrence of letters in the language to determine the substitution table

(whenever there is non-randomness in an encryption system, there is a potential attack!)

Frequency Analysis

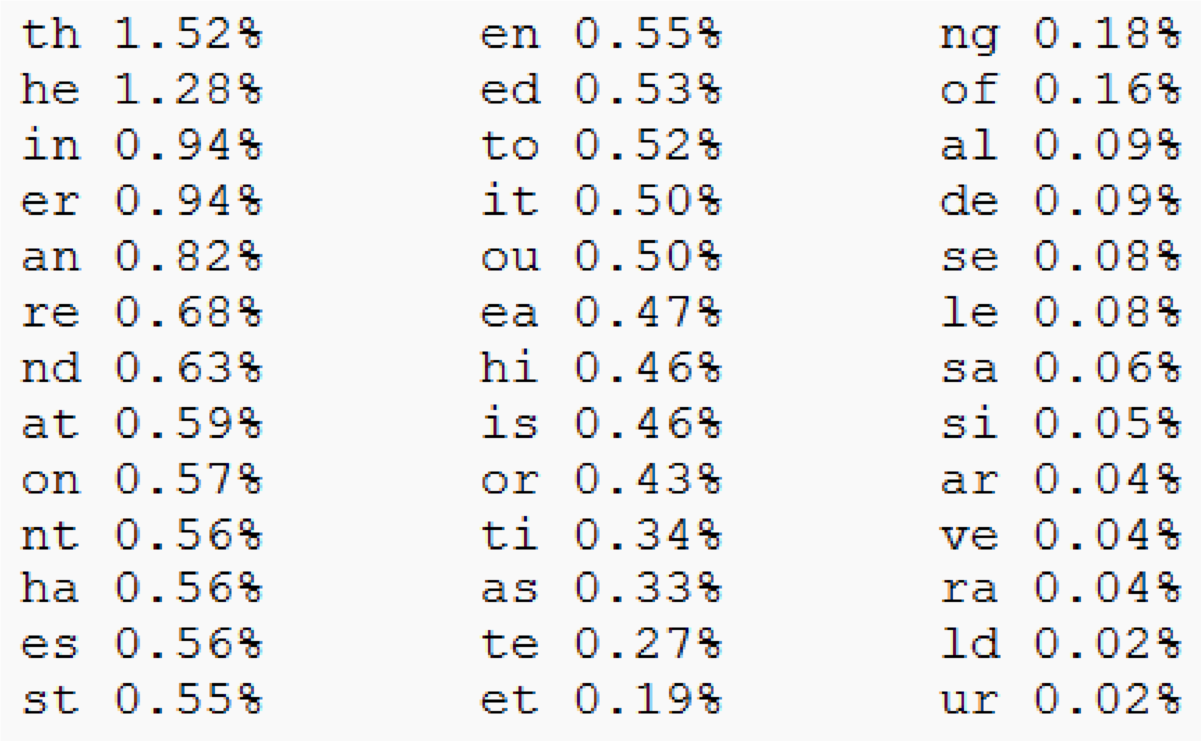
- Probabilities of occurrences of the 26 letters (English)



letter	probability	letter	probability
A	.082	N	.067
B	.015	O	.075
C	.028	P	.019
D	.043	Q	.001
E	.127	R	.060
F	.022	S	.063
G	.020	T	.091
H	.061	U	.028
I	.070	V	.010
J	.002	W	.023
K	.008	X	.001
L	.040	Y	.020
M	.024	Z	.001

Frequency Analysis

- Probabilities of occurrences of two consecutive letters, called digrams, are given as follows:



th	1.52%
he	1.28%
in	0.94%
er	0.94%
an	0.82%
re	0.68%
nd	0.63%
at	0.59%
on	0.57%
nt	0.56%
ha	0.56%
es	0.56%
st	0.55%
en	0.55%
ed	0.53%
to	0.52%
it	0.50%
ou	0.50%
ea	0.47%
hi	0.46%
is	0.46%
or	0.43%
ti	0.34%
as	0.33%
te	0.27%
et	0.19%
ng	0.18%
of	0.16%
al	0.09%
de	0.09%
se	0.08%
le	0.08%
sa	0.06%
si	0.05%
ar	0.04%
ve	0.04%
ra	0.04%
ld	0.02%
ur	0.02%

Frequency Analysis

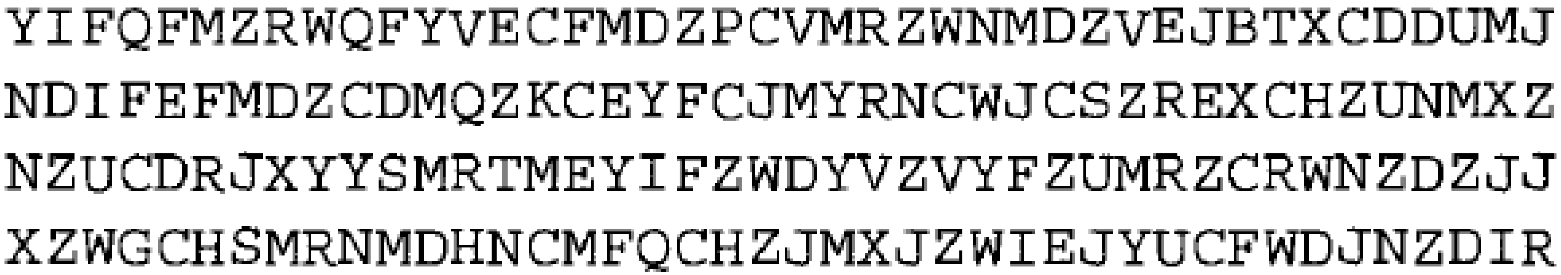
- The 16 most common trigrams in English are:

the and tha ent ing ion tio for
nde has nce edt tis oft sth men

Frequency Analysis

- Example: Given the following ciphertext encrypted with substitution cipher, how to recover plaintext?

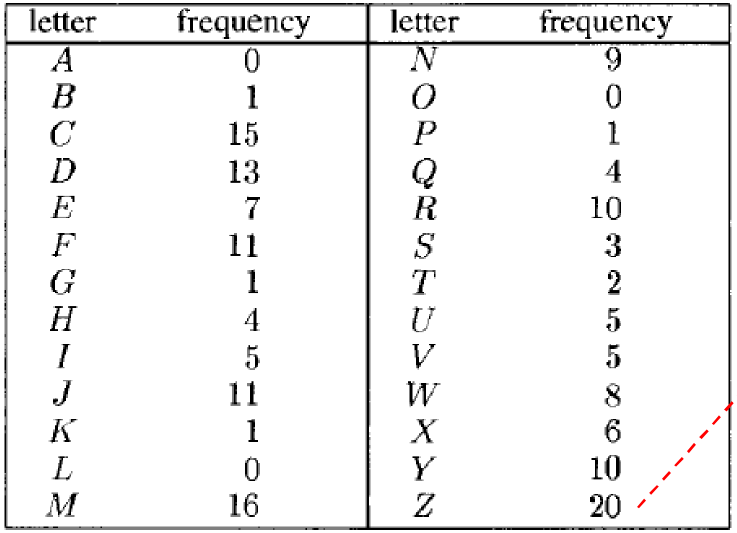
  (in this example, we use capital letter to indicate ciphertext, use small letter to indicate plaintext)



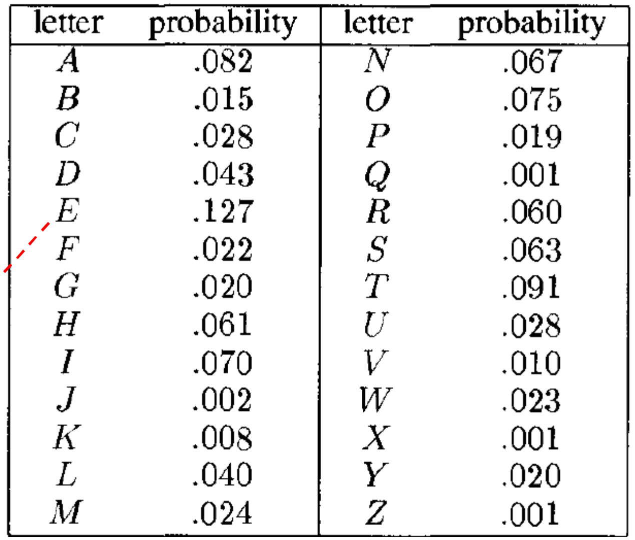
```
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

Frequency Analysis

- Step 1. Compare the frequency of letters in ciphertext with that of English:



letter	frequency	letter	frequency
A	0	N	9
B	1	O	0
C	15	P	1
D	13	Q	4
E	7	R	10
F	11	S	3
G	1	T	2
H	4	U	5
I	5	V	5
J	11	W	8
K	1	X	6
L	0	Y	10
M	16	Z	20



letter	probability	letter	probability
A	.082	N	.067
B	.015	O	.075
C	.028	P	.019
D	.043	Q	.001
E	.127	R	.060
F	.022	S	.063
G	.020	T	.091
H	.061	U	.028
I	.070	V	.010
J	.002	W	.023
K	.008	X	.001
L	.040	Y	.020
M	.024	Z	.001

‘Z’ appears most often, very likely *S*(‘e’) = ‘Z’

Frequency Analysis

- Step 2. digram in ciphertext

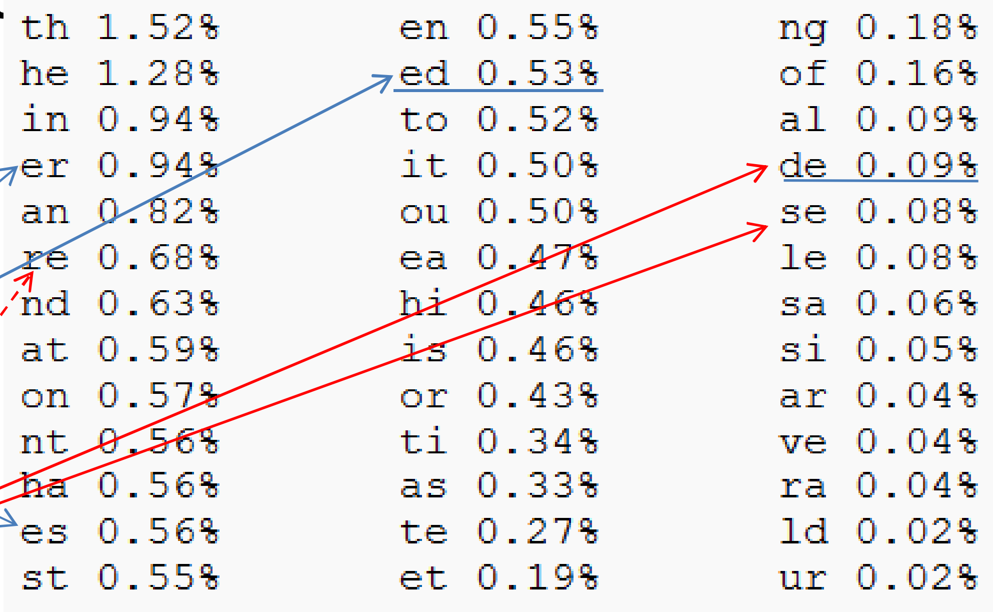
  assume *S*(‘e’) = ‘Z’

  1) ZW appears four times
     => ‘W’ may be ‘r’, ‘d’ or ‘s’

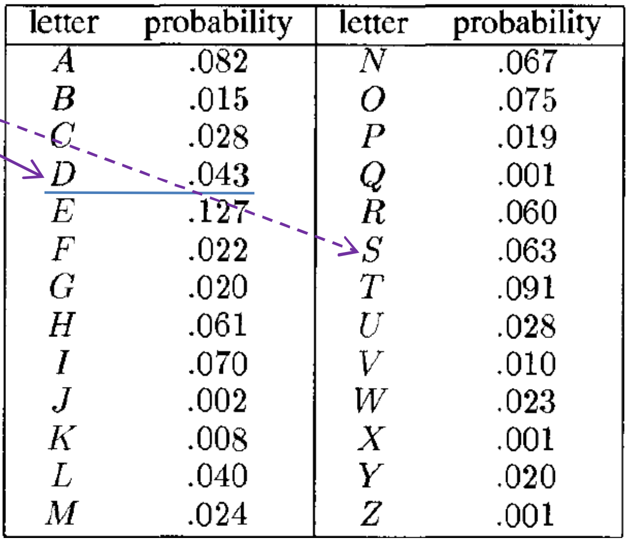
  2) WZ does not appear
     => ‘W’ is not ‘r’

  3) W appears 8 times (0.047)
     => ‘W’ is more likely to be ‘d’

  => likely, *S*(‘d’) = ‘W’



th	1.52%	en	0.55%	ng	0.18%
he	1.28%	ed	0.53%	of	0.16%
in	0.94%	to	0.52%	al	0.09%
er	0.94%	it	0.50%	de	0.09%
an	0.82%	ou	0.50%	se	0.08%
re	0.68%	ea	0.47%	le	0.08%
nd	0.63%	hi	0.46%	sa	0.06%
at	0.59%	is	0.46%	si	0.05%
on	0.57%	or	0.43%	ar	0.04%
nt	0.56%	ti	0.34%	ve	0.04%
ha	0.56%	as	0.33%	ra	0.04%
es	0.56%	te	0.27%	ld	0.02%
st	0.55%	et	0.19%	ur	0.02%



letter	probability	letter	probability
A	.082	*N*	.067
B	.015	*O*	.075
C	.028	*P*	.019
D	.043	*Q*	.001
E	.127	*R*	.060
F	.022	*S*	.063
G	.020	*T*	.091
H	.061	*U*	.028
I	.070	*V*	.010
J	.002	*W*	.023
K	.008	*X*	.001
L	.040	*Y*	.020
M	.024	*Z*	.001

Frequency Analysis

- Step 3. digram in ciphertext

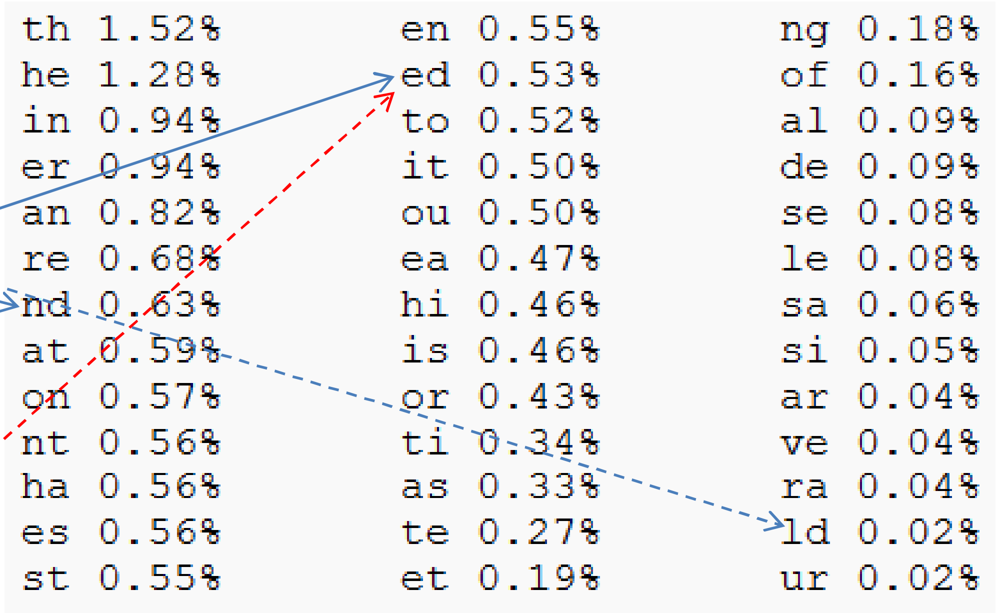
assume $S(\text{‘d’}) = \text{‘W’}$

‘RW’ appears twice

=> ‘R’ may be ‘e, n’

‘e’ is assumed to be ‘Z’

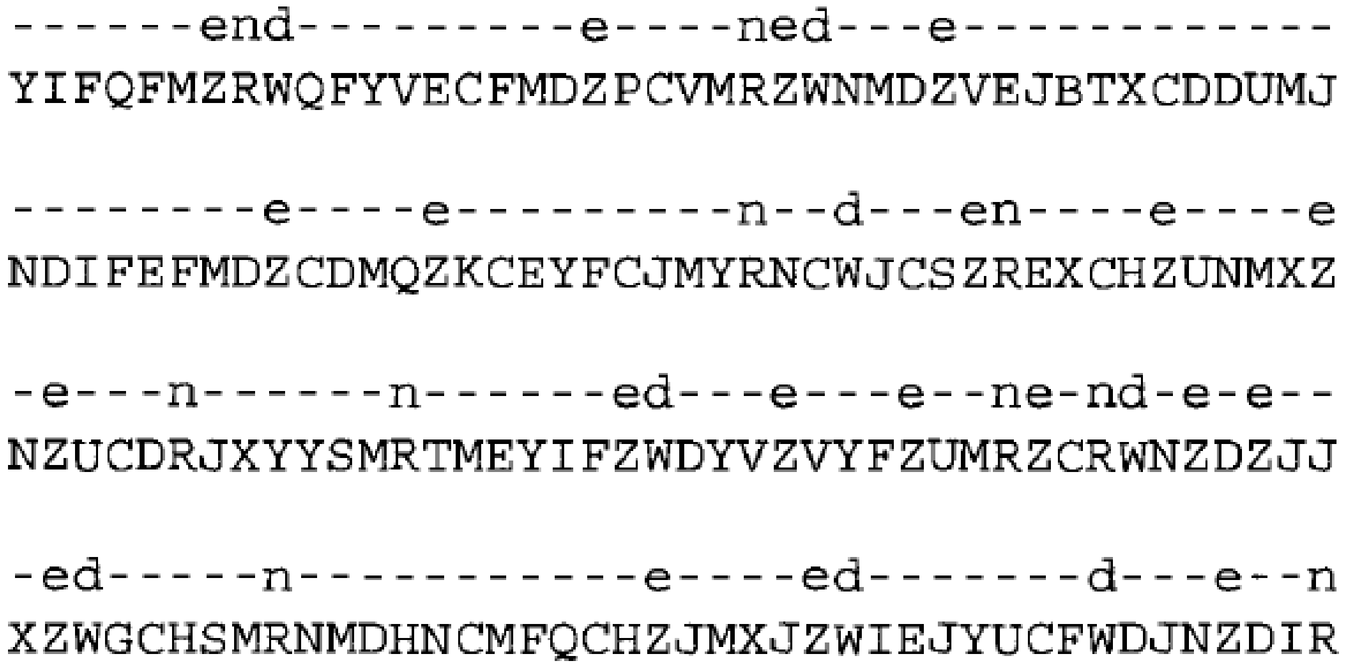
=> ‘R’ may be ‘n’



th	1.52%	en	0.55%	ng	0.18%
he	1.28%	ed	0.53%	of	0.16%
in	0.94%	to	0.52%	al	0.09%
er	0.94%	it	0.50%	de	0.09%
an	0.82%	ou	0.50%	se	0.08%
re	0.68%	ea	0.47%	le	0.08%
nd	0.63%	hi	0.46%	sa	0.06%
at	0.59%	is	0.46%	si	0.05%
on	0.57%	or	0.43%	ar	0.04%
nt	0.56%	ti	0.34%	ve	0.04%
ha	0.56%	as	0.33%	ra	0.04%
es	0.56%	te	0.27%	ld	0.02%
st	0.55%	et	0.19%	ur	0.02%

Frequency Analysis

- After the first three steps, we obtain:



```
------end---------e----ned---e------------
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ

--------e----e---------n--d---en----e----e
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ

-e---n------n------ed---e---e--ne-nd-e-e--
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ

-ed----n------------e----ed-------d---e--n
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

Frequency Analysis

- Step 4. digram in ciphertext

assume S(‘e’) = ‘Z’

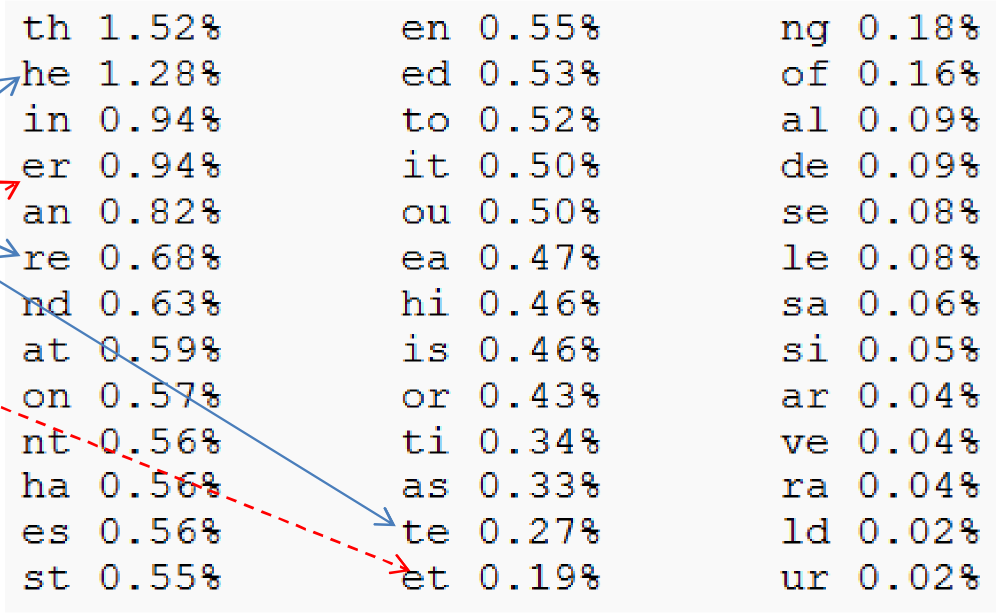
1) NZ appears three times

=> ‘N’ may be ‘h, r, t’

2) ZN does not appear

=> ‘N’ is not ‘r, t’

=> likely, S(‘h’) = ‘N’



th 1.52%	en 0.55%	ng 0.18%
he 1.28%	ed 0.53%	of 0.16%
in 0.94%	to 0.52%	al 0.09%
er 0.94%	it 0.50%	de 0.09%
an 0.82%	ou 0.50%	se 0.08%
re 0.68%	ea 0.47%	le 0.08%
nd 0.63%	hi 0.46%	sa 0.06%
at 0.59%	is 0.46%	si 0.05%
on 0.57%	or 0.43%	ar 0.04%
nt 0.56%	ti 0.34%	ve 0.04%
ha 0.56%	as 0.33%	ra 0.04%
es 0.56%	te 0.27%	ld 0.02%
st 0.55%	et 0.19%	ur 0.02%

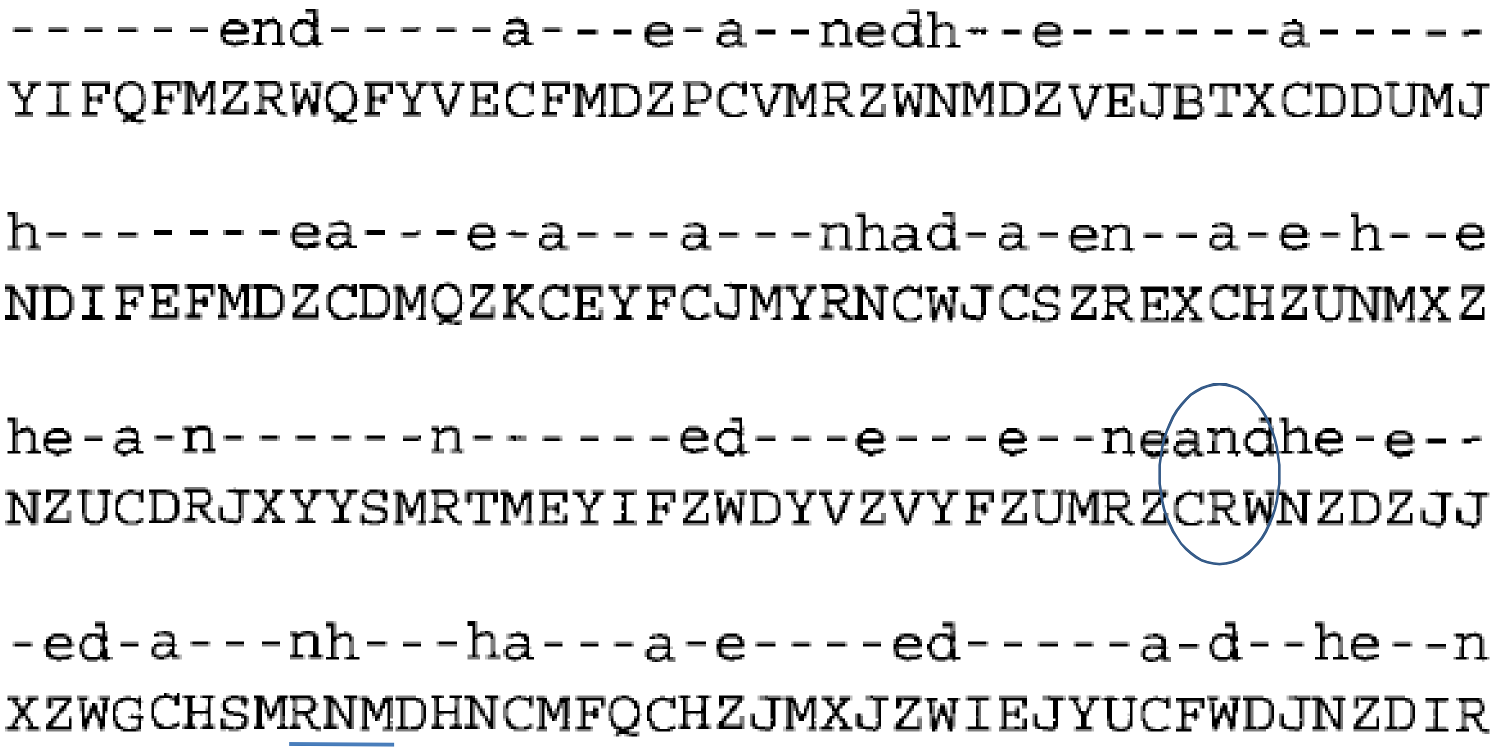
Frequency Analysis

- Step 5. Trigram in ciphertext
  - Now there is ne-ndhe in plaintext, ‘-’ denotes ‘C’ in ciphertext
  - From the distribution of trigram in English, ‘and’ appears with relatively high probability
    - Likely ‘C’ is ‘a’

=> Likely *S*(‘a’) = ‘C’

Frequency Analysis

- Now we obtain:

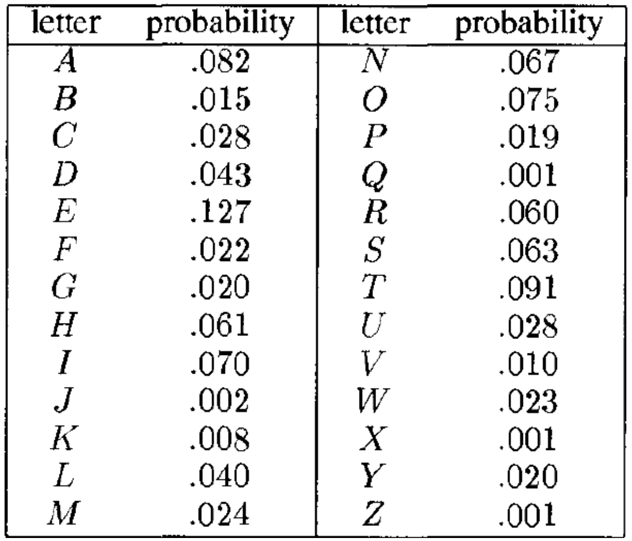


```
------end-----a---e-a--nedh--e------a-----
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ

h-------ea---e-a---a---nhad-a-en--a-e-h--e
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ

he-a-n------n------ed---e---e--neandhe-e--
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ

-ed-a---nh---ha---a-e----ed-----a-d--he--n
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

Frequency Analysis

- Step 6. Determine ‘M’
  - ‘M’ is the second most common ciphertext letter
    - ‘M’ may be ‘t,a,o,i,n,s,h,r’
  - There is ‘RNM’ in ciphertext, so it is ‘nh-’ in plaintext, suggests ‘h-’ begins a word, so ‘M’ represents a vowel
    - ‘M’ may be ‘o, i’
  - ‘CM’ appears once

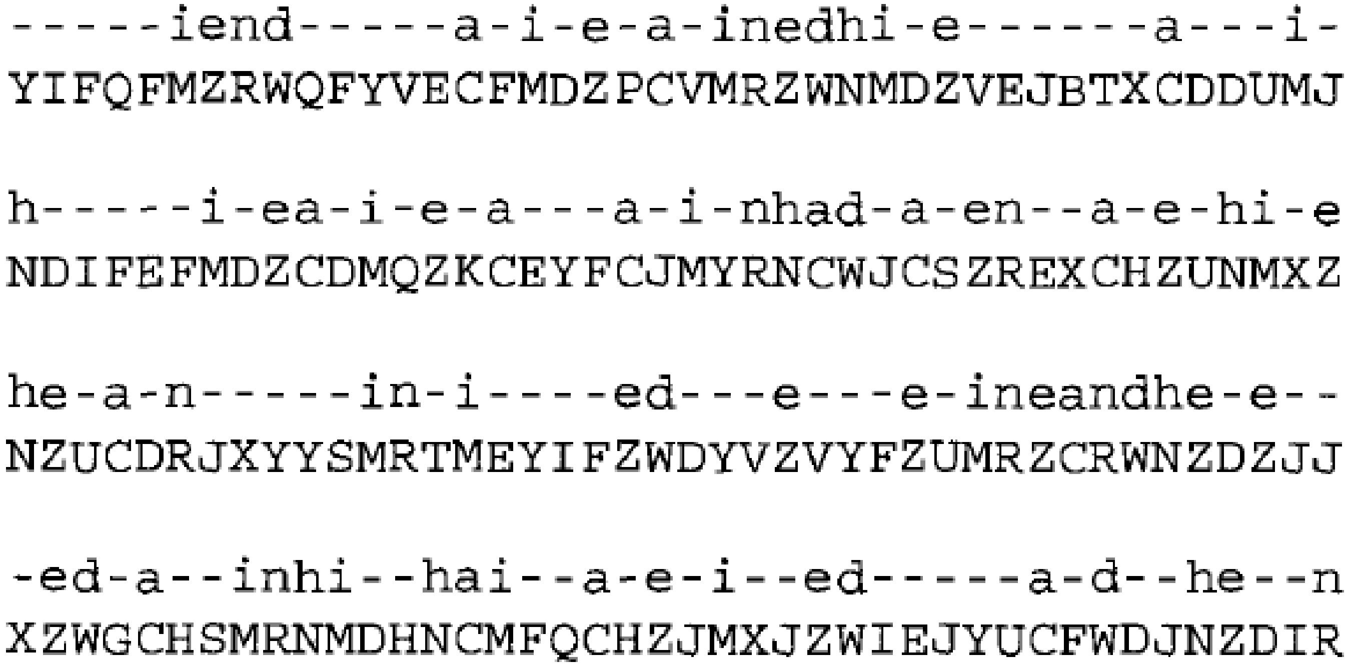
    ‘C’ is ‘a’
    - Likely ‘M’ is ‘i’ since the distribution of ‘ai’ is more than ‘ao’ in English



letter	probability	letter	probability
A	.082	*N*	.067
B	.015	*O*	.075
C	.028	*P*	.019
D	.043	*Q*	.001
E	.127	*R*	.060
F	.022	*S*	.063
G	.020	*T*	.091
H	.061	*U*	.028
I	.070	*V*	.010
J	.002	*W*	.023
K	.008	*X*	.001
L	.040	*Y*	.020
M	.024	*Z*	.001

Frequency Analysis

- Now we obtain



```
-----iend-----a-i-e-a-inedhi-e------a---i-
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ

h-----i-ea-i-e-a---a-i-nhad-a-en--a-e-hi-e
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ

he-a-n-----in-i----ed---e---e-ineandhe-e--
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ

-ed-a--inhi--hai--a-e-i--ed-----a-d--he--n
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

Frequency Analysis

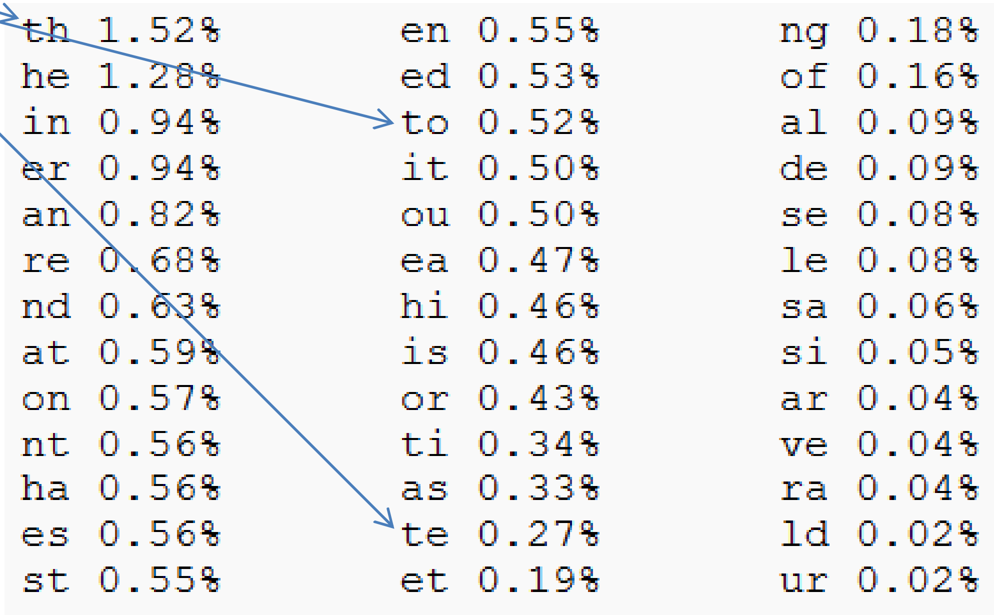
- Determine ‘J’ by considering digram
  - ‘JN’ appears twice in ciphertext, and ‘N’ is ‘h’
  - ‘th’ is the most frequent digram

  ⇒Likely ‘J’ is ‘t’

Frequency Analysis

- Determine ‘Y’ by considering digram
  - ‘JY’ appears once
  - ‘Y’ is not ‘h,e’

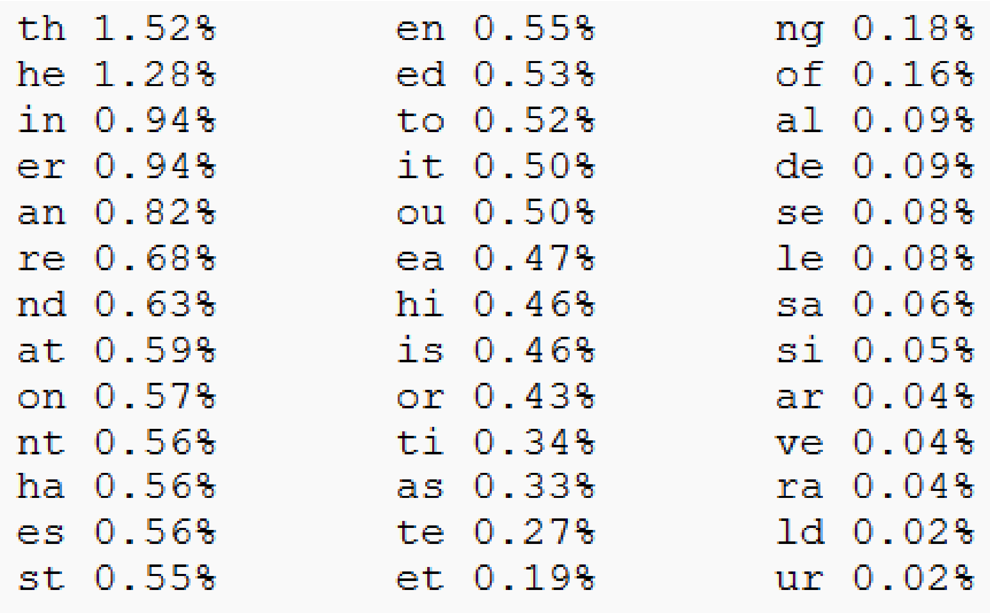
⇒Likely ‘Y’ is ‘o’



th	1.52%	en	0.55%	ng	0.18%
he	1.28%	ed	0.53%	of	0.16%
in	0.94%	to	0.52%	al	0.09%
er	0.94%	it	0.50%	de	0.09%
an	0.82%	ou	0.50%	se	0.08%
re	0.68%	ea	0.47%	le	0.08%
nd	0.63%	hi	0.46%	sa	0.06%
at	0.59%	is	0.46%	si	0.05%
on	0.57%	or	0.43%	ar	0.04%
nt	0.56%	ti	0.34%	ve	0.04%
ha	0.56%	as	0.33%	ra	0.04%
es	0.56%	te	0.27%	ld	0.02%
st	0.55%	et	0.19%	ur	0.02%

Frequency Analysis

- Determine ‘D’
  - Four occurrence of ‘MD’
  - ‘M’ is ‘i’
  - ‘D’ may be ‘n,t,s’
  - Likely ‘D’ is ‘s’



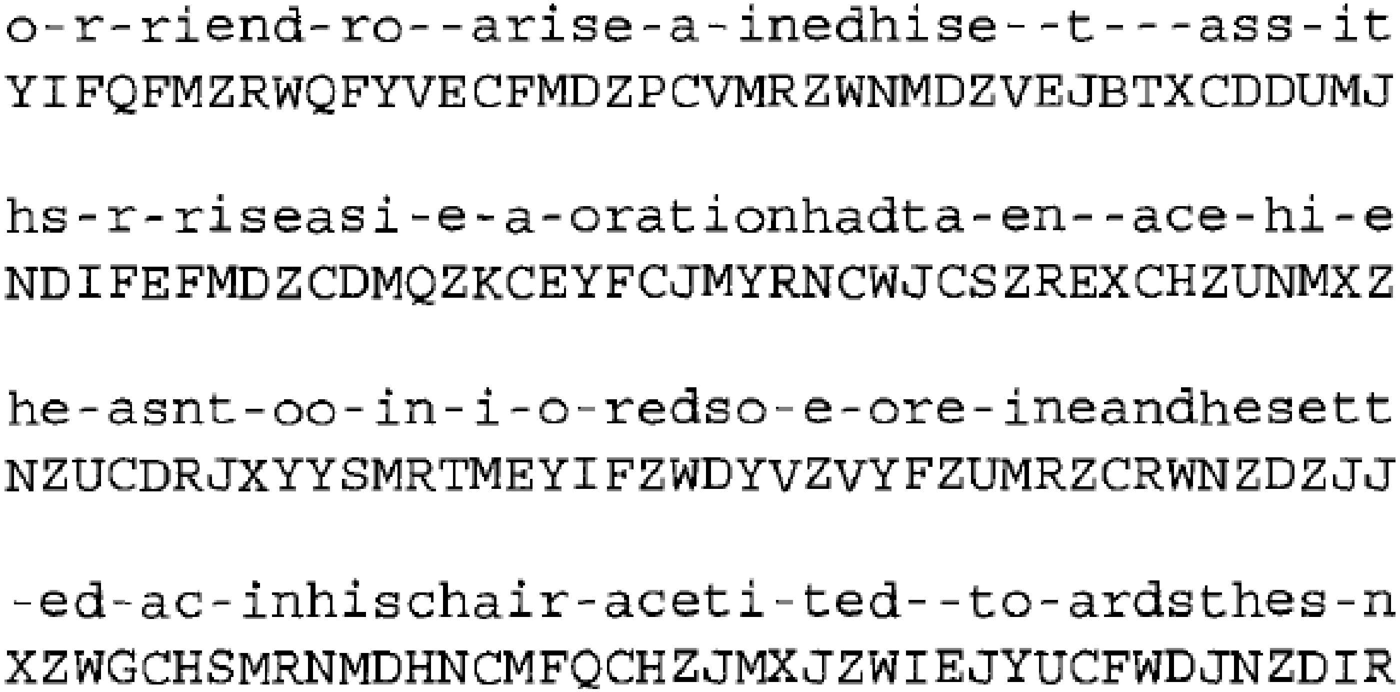
th	1.52%	en	0.55%	ng	0.18%
he	1.28%	ed	0.53%	of	0.16%
in	0.94%	to	0.52%	al	0.09%
er	0.94%	it	0.50%	de	0.09%
an	0.82%	ou	0.50%	se	0.08%
re	0.68%	ea	0.47%	le	0.08%
nd	0.63%	hi	0.46%	sa	0.06%
at	0.59%	is	0.46%	si	0.05%
on	0.57%	or	0.43%	ar	0.04%
nt	0.56%	ti	0.34%	ve	0.04%
ha	0.56%	as	0.33%	ra	0.04%
es	0.56%	te	0.27%	ld	0.02%
st	0.55%	et	0.19%	ur	0.02%

Frequency Analysis

- Determine ‘F’
  - ‘HNCMF’ in ciphertext
  - ‘chaiF’
  - Likely ‘F’ is ‘r’

Frequency Analysis

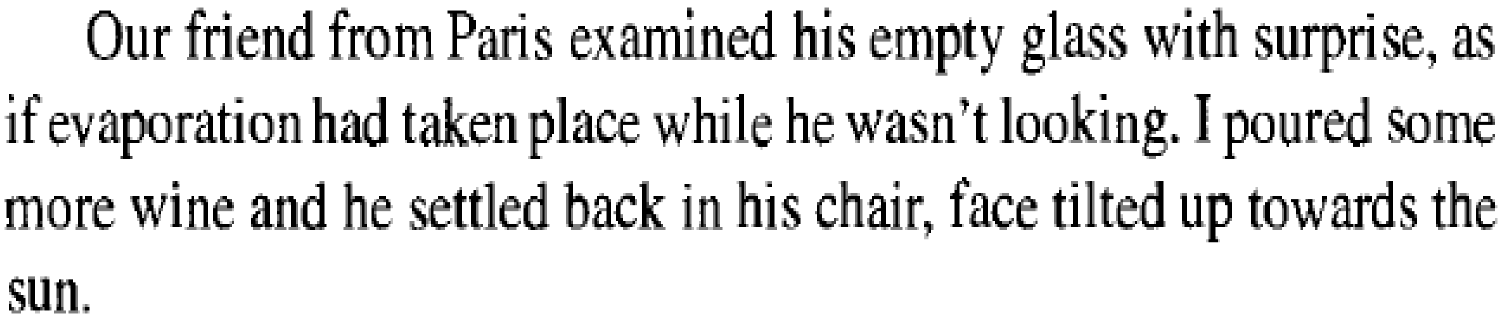
- We now obtain



```
o-r-riend-ro--arise-a-inedhise--t---ass-it
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ

hs-r-riseasi-e-a-orationhadta-en--ace-hi-e
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ

he-asnt-oo-in-i-o-redso-e-ore-ineandhesett
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ

-ed-ac-inhischair-aceti-ted--to-ardsthes-n
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

Frequency Analysis

- With further guessing, we obtain:



Our friend from Paris examined his empty glass with surprise, as if evaporation had taken place while he wasn’t looking. I poured some more wine and he settled back in his chair, face tilted up towards the sun.