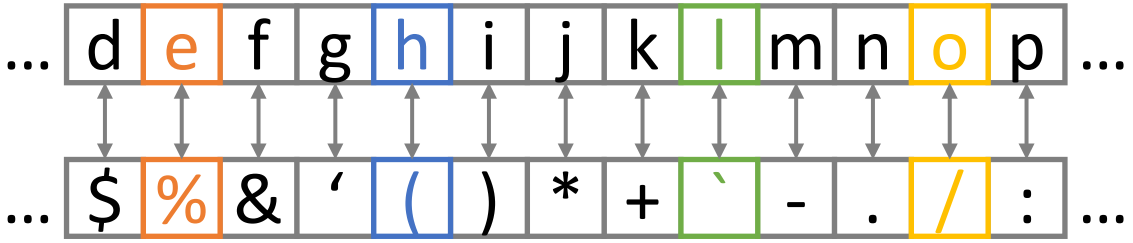
Substitution Ciphers

Elements of Applied Data Security

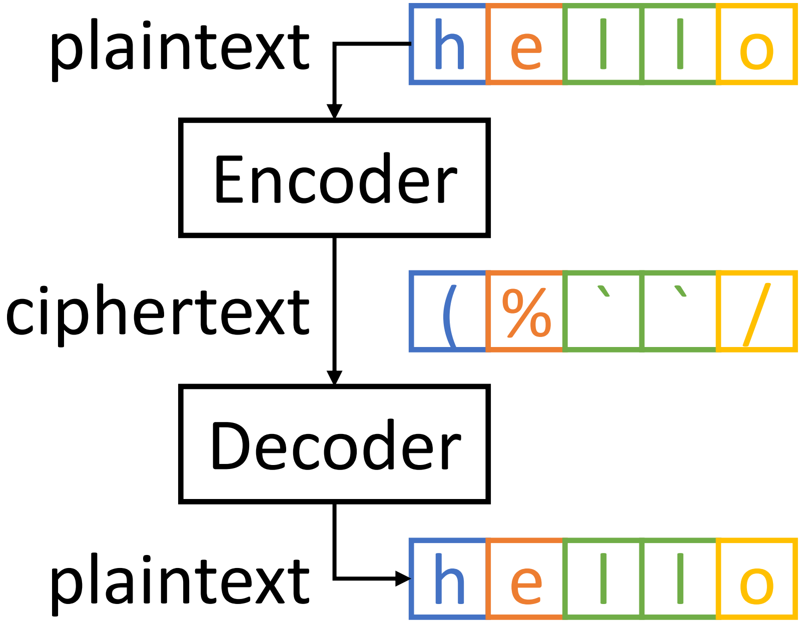
Alex Marchioni – alex.marchioni@unibo.it

Substitution Ciphers

Every plaintext character (or group of characters) is replaced with a different ciphertext symbol. The receiver deciphers the text by performing the inverse substitution.



…	d	e	f	g	h	i	j	k	l	m	n	o	p	…
…	$	%	&	‘	(	)	*	+	`	-	.	/	:	…



Substitution Ciphers

- Substitution can consider single characters (simple substitution cipher) but also group of characters (e.g., pair, triplets, and so on).
- Alphabet simple substitution Ciphers admits $26! \sim 10^{26} \sim 2^{88}$ possible encoding rules (not easy to try them all)
  - Assuming 1ns for each try, it would take $> 10^9$ years.
- However, substitution does not alter the statistics so plaintext can be deduced by analyzing the frequency distribution of the ciphertext.

Tasks

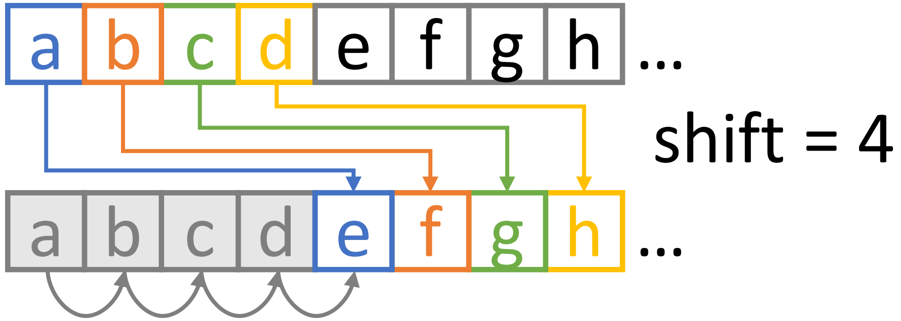
1. Breaking a Caesar Cipher
2. Breaking a Simple Substitution Cipher

Task 1: Caesar Cipher

Caesar Cipher

The method is named after Julius Caesar, who used it in his private correspondence. Each letter in the plaintext is replaced by a letter some fixed number of positions down the alphabet.

- Same characters for plaintext and ciphertext.
- Very simple encoding rule.
  Only 26 possibilities!



plaintext	h	e	l	l	o
ciphertext	l	i	p	p	s

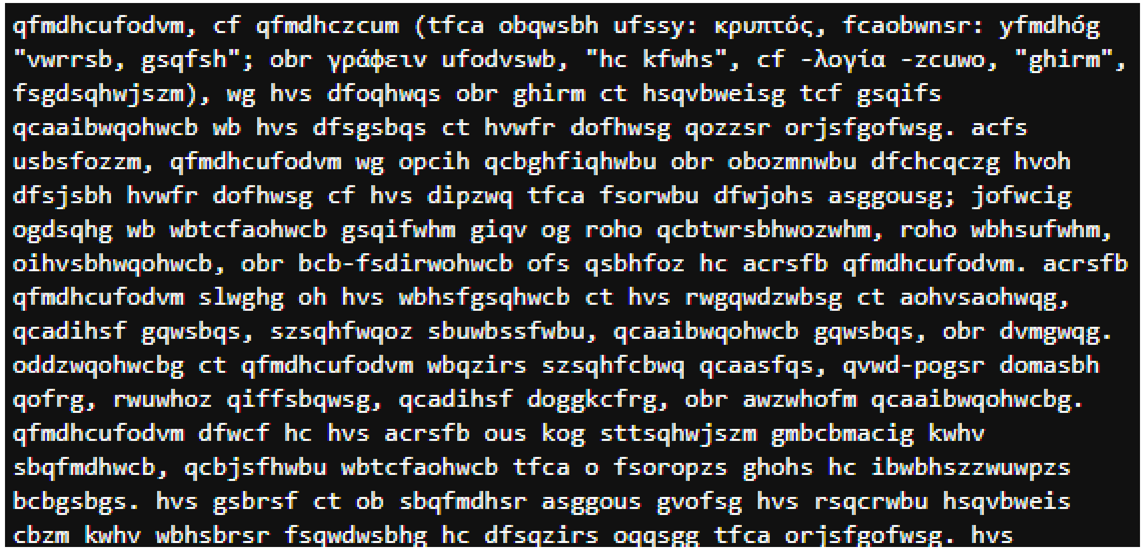
Breaking a Caesar Cipher

Two easy way to break the cipher:

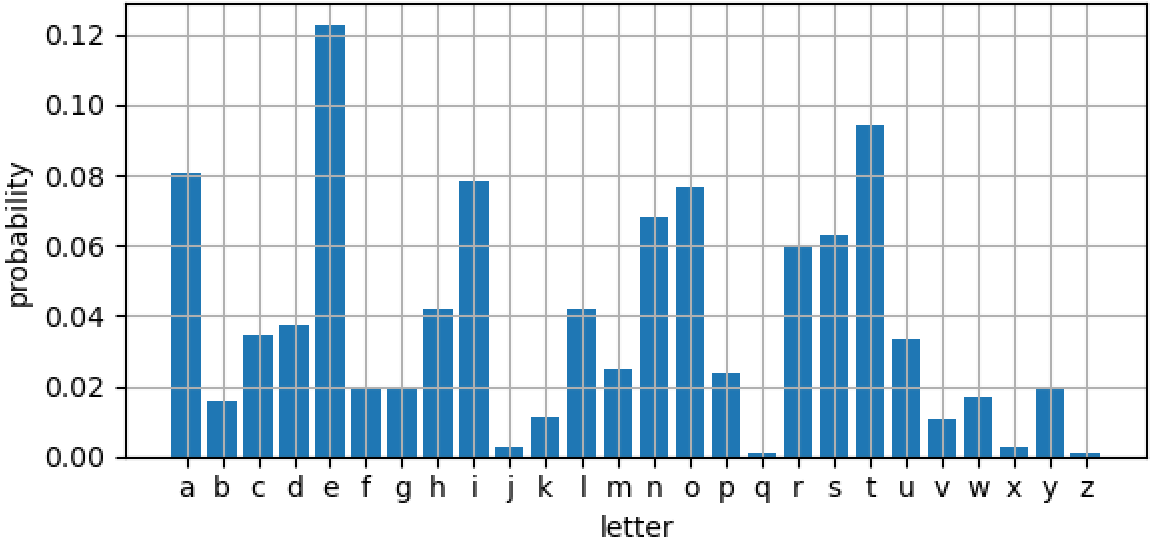
- **Brute force**:
  - Since alphabet is 26 letters long only 26 shifts are possible you can try all possibilities and check them all.
- **Frequency analysis**:
  - Knowing what is the frequency of letters (e.g., in English, letters «e», «t», «a», «i» are more common than others), it is possible to infer what shift was used by oserving the frequency of the characters in the ciphertext.

Task: Inputs

- A text file `ciphertext_caesar.txt`, containing the text of a Wikipedia page encrypted with a Caesar Cipher
  - cipher modifies only letters leaving numbers and special characters unchanged.
- The distribution of the letters in English language estimated by observing many different Wikipedia pages.



```
qfmdhcufodvm, cf qfmdhczcum (tfca obqwsbh ufssy: κρυπτός, fcaobwnsr: yfmdhóg
"vwrrsb, gsqfsh"; obr γράφειν ufodvswb, "hc kfwhs", cf -λογία -zcuwo, "ghirm",
fsgdsqhwjszm), wg hvs dfoqhwqs obr ghirm ct hsqvbweisg tcf gsqifs
qcaaibwqohwcb wb hvs dfsgsbqs ct hvwfr dofhwsg qozzsr orjsfgofwsg. acfs
usbsfozzm, qfmdhcufodvm wg opcih qcbghfiqhwbu obr obozmnwbu dfchcqczg hvoh
dfsjsbh hvwfr dofhwsg cf hvs dipzwq tfca fsorwbu dfwjohs asggousg; jofwcig
ogdsqhg wb wbtcfaohwcb gsqifwhm giqv og roho qcbtwrsbhwozwhm, roho wbhsufwhm,
oihvsbhwqohwcb, obr bcb-fsdirwohwcb ofs qsbhfoz hc acrsfb qfmdhcufodvm. acrsfb
qfmdhcufodvm slwghg oh hvs wbhsfgsqhwcb ct hvs rwgqwdzwbsg ct aohvsaohwqg,
qcadihsf gqwsbqs, szsqhfwqoz sbuwbssfwbu, qcaaibwqohwcb gqwsbqs, obr dvmgwqg.
oddzwqohwcbg ct qfmdhcufodvm wbqzirs szsqhfcbwq qcaasfqs, qvwd-pogsr domasbh
qofrg, rwuwhoz qiffsbqwsg, qcadihsf doggkcfrg, obr awzwhofm qcaaibwqohwcbg.
qfmdhcufodvm dfwcf hc hvs acrsfb ous kog sttsqhwjszm gmbcbmacig kwhv
sbqfmdhwcb, qcbjsfhwbu wbtcfaohwcb tfca o fsoropzs ghohs hc ibwbhszzwuwpzs
bcbgsbgs. hvs gsbrsf ct ob sbqfmdhsr asggous gvofsg hvs rsqcrwbu hsqvbweis
cbzm kwhv wbhsbrsr fsqwdwsbhg hc dfsqzirs oqqsgg tfca orjsfgofwsg. hvs
```



Task: Outputs

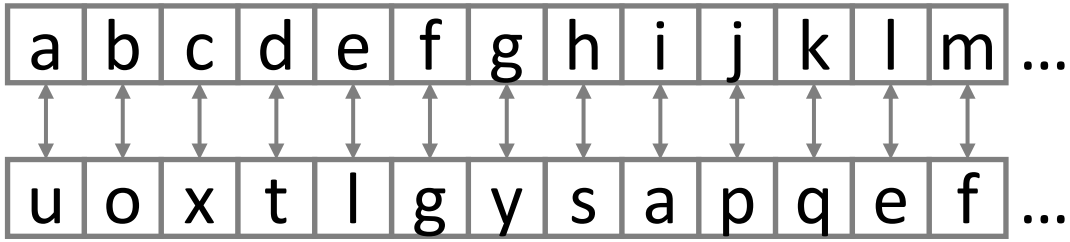
- The **substitution rule** (i.e., the key), i.e., the shift to apply to the alphabet to decrypt the ciphertext.
- The **plaintext** decrypted from the ciphertext.

Task 2: Simple Substitution

Simple Substitution Cipher

Every plaintext character is replaced with a different ciphertext character.

- As for Caesar Cipher, plaintext and ciphertext share the same set of characters (the alphabet).
- Mapping from plaintext to ciphertext can be any of the $26! \sim 10^{26} \sim 2^{88}$ possibilities



a	b	c	d	e	f	g	h	i	j	k	l	m	…
u	o	x	t	l	g	y	s	a	p	q	e	f	…

Breaking a Simple Substitution Cipher

Since nowadays machines cannot explore $26!$ candidates, **frequency analysis** must be exploited to narrow down their number.

- For reasonably large pieces of text (with enough characters to be statistically relevant), a possible procedure can be:
  - to just replace the most common ciphertext character with the most common character in the plaintext (for English text is “e”).
  - to replace the second most common ciphertext character with the second most common character in the plaintext
  - and so on

Task

- Inputs:
  - Ciphertext as a text file: `ciphertext_simple.txt`.
    - Ciphertext is a Wikipedia page where each letter is encrypted with a Simple Substitution Cipher (spaces and special characters are unchanged)
  - Letters distribution estimate from many different Wikipedia pages.

- Outputs:
  - Substitution rule
  - Plaintext