

Titolo unità didattica: Stringhe ed elaborazione di testi [09]

Titolo modulo : Operazioni elementari su stringhe [01-T]

Le stringhe di caratteri: operazioni di concatenazione, confronto, ..

Argomenti trattati:

- ✓ stringhe di caratteri su un alfabeto
- ✓ stringhe e sottostringhe
- ✓ operazioni di base su stringhe

Prerequisiti richiesti: AP-07-11-T

stringhe di caratteri

- ✓ una **stringa** su un alfabeto $\{c_j\}_{j=1,n}$ è una sequenza di caratteri dell'alfabeto:

$$c_1 c_2 c_3 c_4 c_5 \dots c_k$$

- ✓ una stringa è costituita da un numero **finito** di caratteri, che è la **lunghezza** della stringa
($c_1 c_2 c_3 c_4 c_5 \dots c_k$ ha lunghezza k)
- ✓ la **stringa vuota** è la stringa di lunghezza zero (0-stringa)
- ✓ in una stringa, ogni carattere ha una sua posizione d'ordine che si chiama **indice** della stringa
- ✓ una **sottostringa** è una *porzione* di una stringa, cioè una sequenza di caratteri **consecutivi** di una stringa
- ✓ una sottostringa di lunghezza p , il cui **primo** carattere si trova nella posizione di indice i della stringa è detta **p -sottostringa di inizio i**

stringhe di caratteri

- ✓ una stringa di lunghezza k è detta **k -stringa**
- ✓ l'insieme di tutte le 1-stringhe su un alfabeto di n caratteri ha cardinalità n
- ✓ l'insieme di tutte le 2-stringhe su un alfabeto di n caratteri ha cardinalità n^2
- ✓ l'insieme di tutte le k -stringhe su un alfabeto di n caratteri ha cardinalità n^k
- ✓ una **permutazione** di una k -stringa è una k -stringa composta dagli stessi caratteri
- ✓ l'insieme di tutte le permutazioni di una k -stringa ha cardinalità $k!$

Esempio: alfabeto $\{a,b\}$ $n = 2$

$k = 1$ $a ; b$

$k = 2$ $aa ; ab ; ba ; bb$

$k = 3$ $aaa ; aab ; aba ; abb ; baa ; bab ; bba ; bbb$

stringhe di caratteri

- ✓ una stringa di lunghezza k è detta **k -stringa**
- ✓ l'insieme di tutte le 1-stringhe su un alfabeto di n caratteri ha cardinalità n
- ✓ l'insieme di tutte le 2-stringhe su un alfabeto di n caratteri ha cardinalità n^2
- ✓ l'insieme di tutte le k -stringhe su un alfabeto di n caratteri ha cardinalità n^k
- ✓ una **permutazione** di una k -stringa è una k -stringa composta dagli stessi caratteri
- ✓ l'insieme di tutte le permutazioni di una k -stringa ha cardinalità $k!$

Esempio: alfabeto $\{a,b,c\}$ $n = 3$

$k = 1$ $a ; b ; c$

$k = 2$ $aa ; ab ; ac ; ba ; bb ; bc ; ca ; cb ; cc$

$k = 3$ $aaa ; aab ; aac ; aba ; abb ; abc ; aca ; acb ; acc ; baa ; bab ; bac ; bba ;$
 $bbb ; bbc ; bca ; bcb ; bcc ; caa ; cab ; cac ; cba ; cbb ; cbc ; cca ; ccb ; ccc ;$

operazioni di base su **stringhe** di caratteri

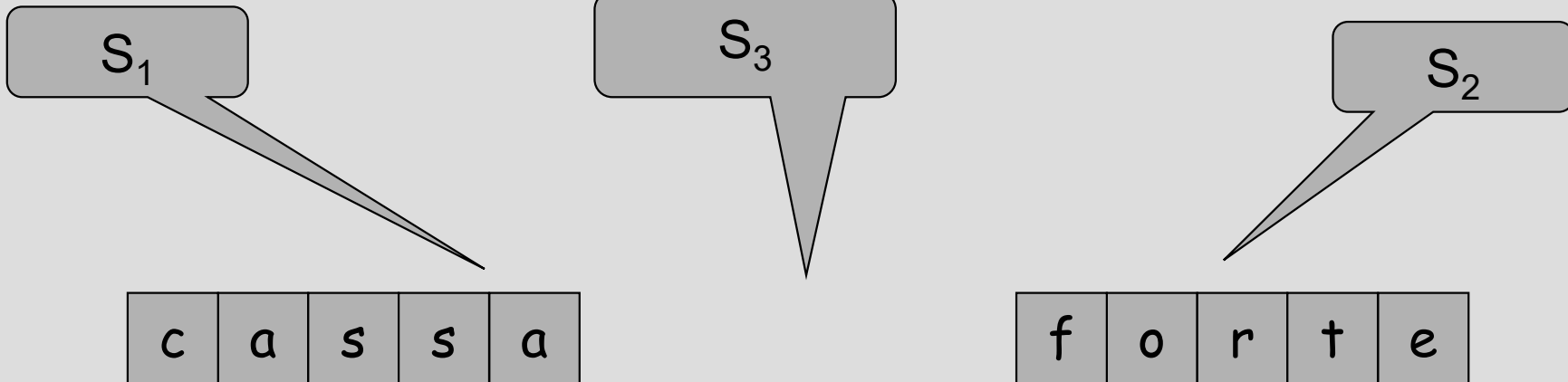
- **concatenazione**
- **ricerca di** un carattere in una stringa
- **confronto** tra due stringhe
- **matching** di due stringhe
- creazione di una **copia** di una stringa
- calcolo della **lunghezza** di una stringa
- **confronto di sottostringhe** di due stringhe
- **copia di una sottostringa** di una stringa
- **ricerca di una sottostringa** in una stringa

concatenazione di stringhe

- stringhe:

$S_1 = \text{cassa}$

$S_2 = \text{forte}$



- si ottiene una nuova stringa S_3 che è la concatenazione delle due stringhe S_1 ed S_2

$S_3 = S_1 // S_2 = \text{cassaforte}$

ricerca di un carattere in una stringa

- ricercare (**ricerca sequenziale**) un dato carattere all'interno della stringa

cassaforte

- carattere da ricercare (**chiave**): **f**

NO

c	a	s	s	a	f	o	r	t	e
---	---	---	---	---	---	---	---	---	---

SI

f	f	f	f	f	f
---	---	---	---	---	---

confronto fra due stringhe

- determinare l'uguaglianza o la disuguaglianza

$S_1 = \text{cassaforte}$

$S_2 = \text{pianoforte}$

<u>no</u>	<u>no</u>	<u>no</u>	<u>no</u>	<u>no</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>
c	a	s	s	a	f	o	r	t	e

p	i	a	n	o	f	o	r	t	e
---	---	---	---	---	---	---	---	---	---

confronto fra due stringhe

<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>
c	a	s	s	a	f	o	r	t	e

OK

c	a	s	s	a	f	o	r	t	e
---	---	---	---	---	---	---	---	---	---

confronto fra due stringhe per determinare quale delle due stringhe precede l'altra nell'ordine alfabetico

<u>si</u>	<u>si</u>	<u>si</u>	<						
c	a	s	s	a	f	o	r	t	e

OK

c	a	s	t	a		d	i	v	a
---	---	---	---	---	--	---	---	---	---

matching fra due stringhe

- determinare il numero di caratteri (di ugual posto) delle due stringhe che risultano uguali

$S_1 = \text{cassaforte}$

$S_2 = \text{pianoforte}$

<u>no</u>	<u>no</u>	<u>no</u>	<u>no</u>	<u>no</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>
c	a	s	s	a	f	o	r	t	e
p	i	a	n	o	f	o	r	t	e

punteggio del
matching
5

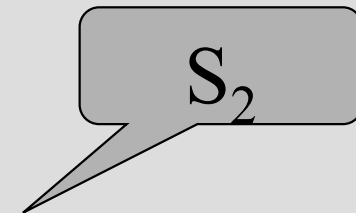
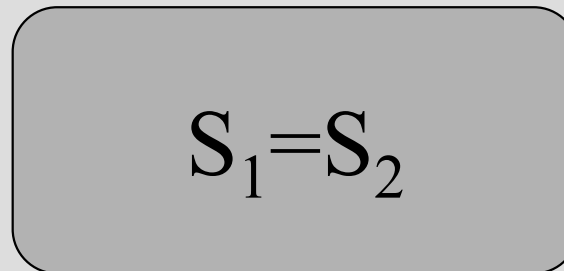
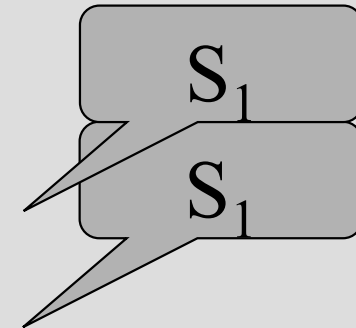
creazione di una **copia** di una stringa

- stringhe:

$S_1 = \text{cassaforte}$

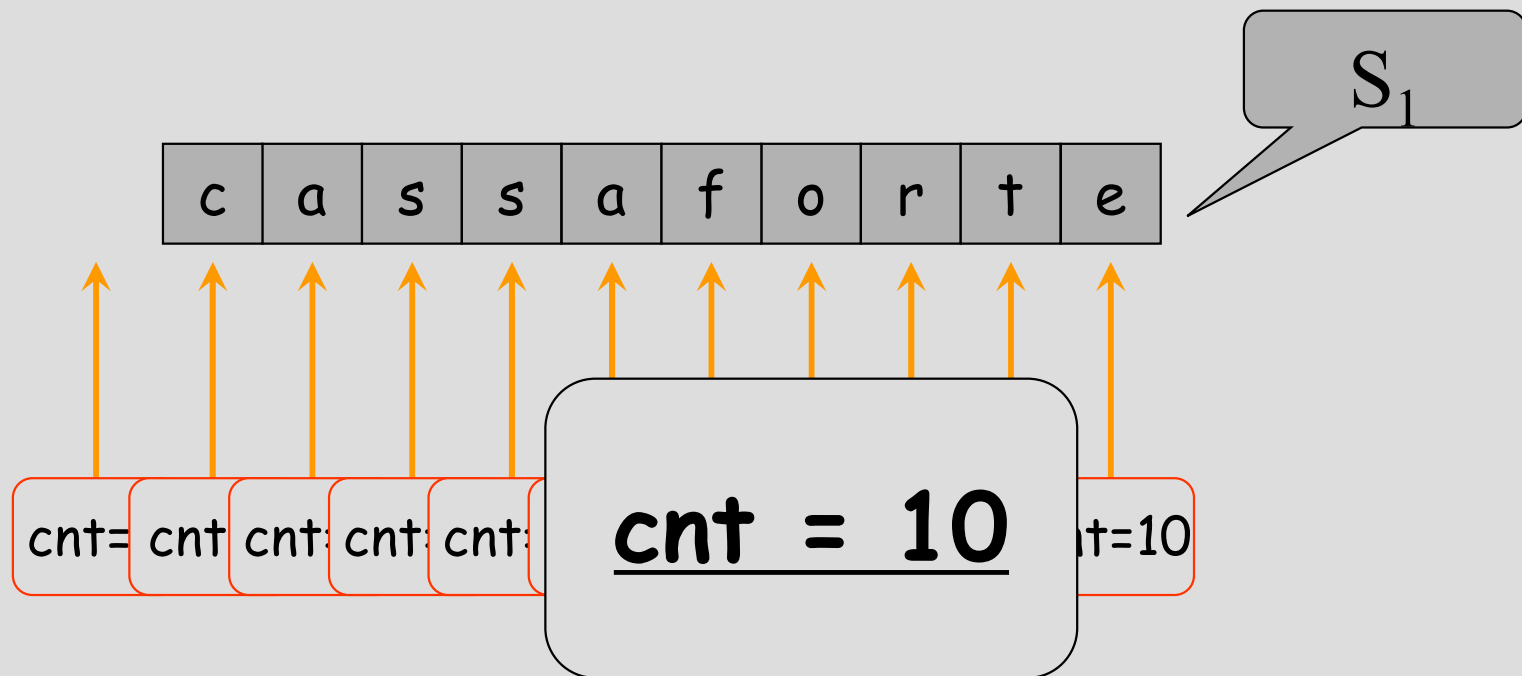
$S_2 = \text{NULL}$ (stringa vuota)

- assegnare la stringa S_1 a S_2



calcolo della **lunghezza** di una stringa

- lunghezza di una stringa, ovvero numero dei caratteri che costituiscono la stringa



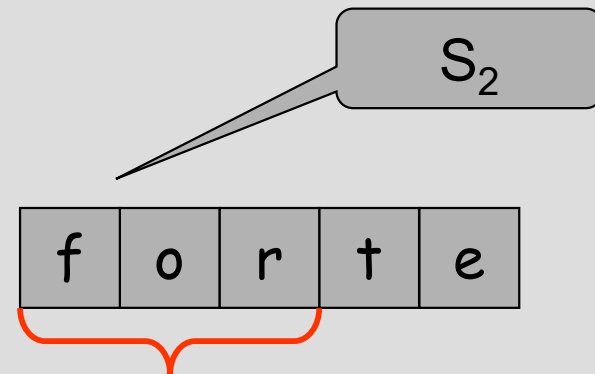
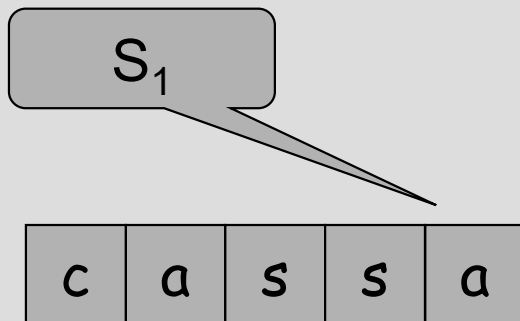
concatenazione di sottostringhe

- stringhe:

$S_1 = \text{cassa}$

$S_2 = \text{forte}$

- concatenare a S_1 la sottostringa di S_2 formata dai suoi primi m caratteri (sottostringa di lunghezza m e inizio 1)
- $m=3$



confronto di sottostringhe

- stringhe:

$S_1 = \text{cassaforte}$

$S_2 = \text{pianoforte}$

- verificare l'uguaglianza solo tra le sottostringhe di lunghezza m e inizio 1
- $m=5$

no no no no no

c	a	s	s	a	f	o	r	t	e
---	---	---	---	---	---	---	---	---	---

p	i	a	n	o	f	o	r	t	e
---	---	---	---	---	---	---	---	---	---



confronto di sottostringhe

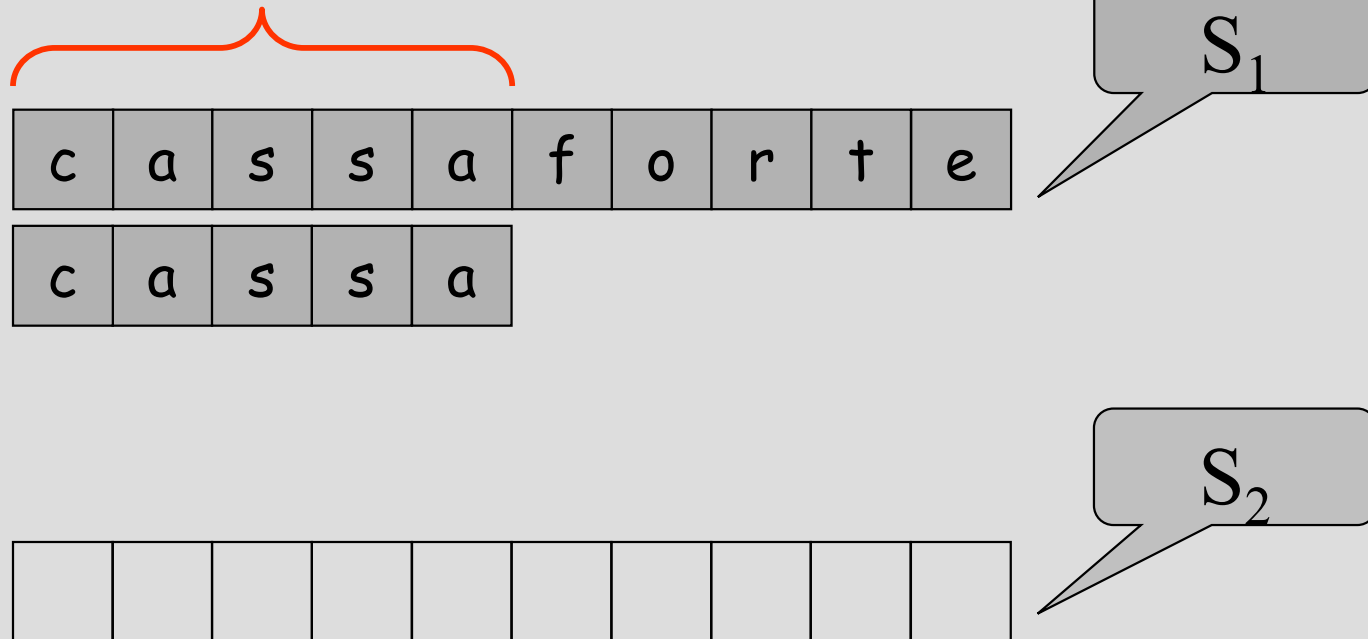
- $m = 5$
- inizio 1

<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>					
c	a	s	s	a	f	o	r	t	e

c	a	s	s	a	p	a	n	c	a
---	---	---	---	---	---	---	---	---	---

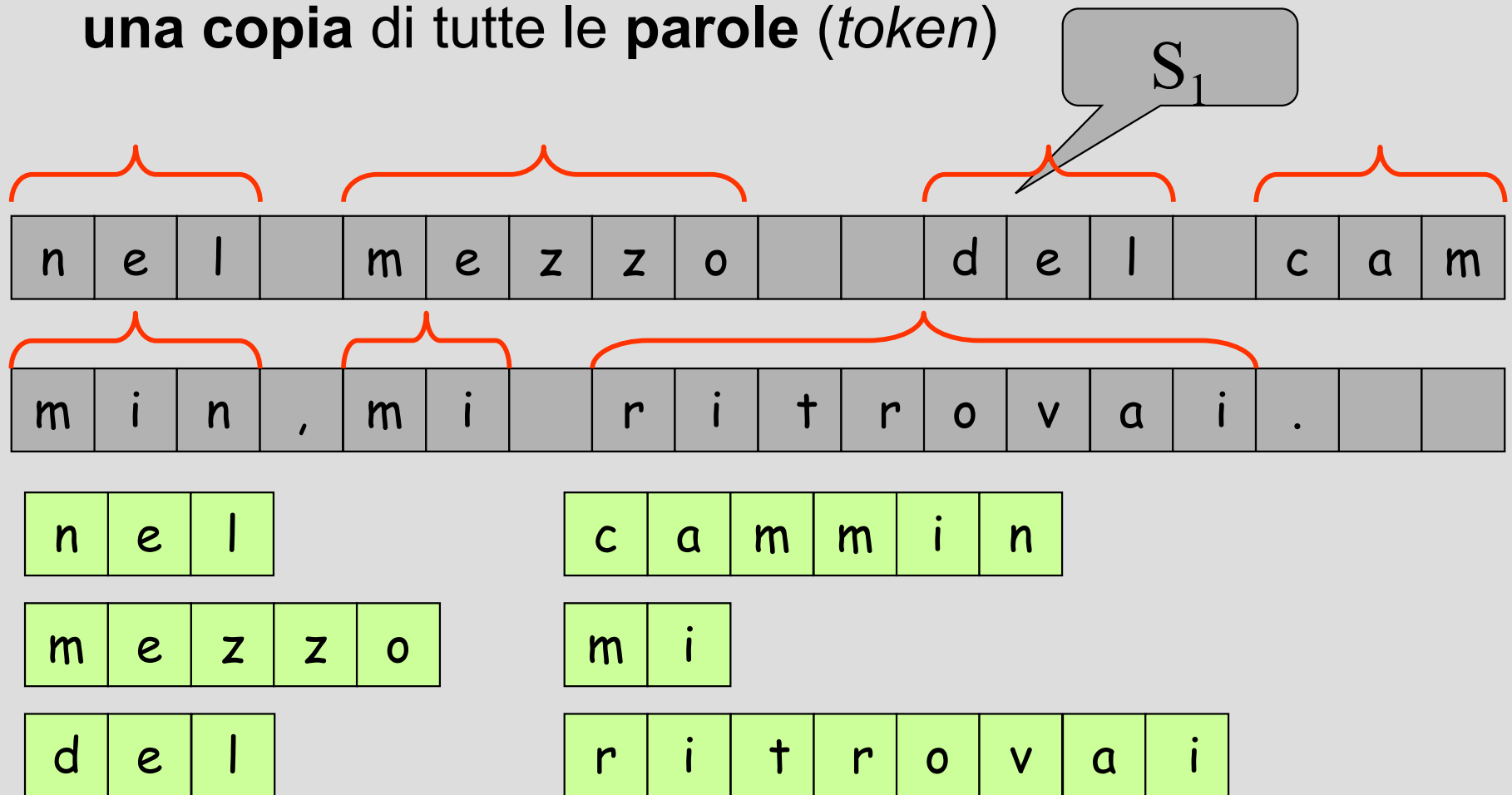
copia di una sottostringa

- data una stringa S_1 , creare una copia di una sua sottostringa di lunghezza m e inizio 1
- $m = 5$



estrazione di token da una stringa

- data una stringa S_1 , che contiene un testo, cioè un insieme di parole separate da **separatori** (spazio, segni di punteggiatura, andata a capo, etc..), **estrarre una copia** di tutte le **parole (token)**



ricerca di una sottostringa in una stringa (*string matching, pattern matching*)

- date due stringhe S_1 e S_2 , ricercare le occorrenze della stringa S_2 come sottostringa della stringa S_1

<u>no</u>	<u>no</u>	<u>no</u>	<u>no</u>	<u>no</u>	<u>si</u>	<u>si</u>	<u>si</u>	<u>si</u>
-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------

P	o	r	t	a	n	o	l	a	n	a
---	---	---	---	---	---	---	---	---	---	---

S_1

n	o	l	a
---	---	---	---

S_2

ricerca di una sottostringa in una stringa che più
si *avvicina* una stringa data
(*best matching, matching migliore*)

- date due stringhe S_1 e S_2 , trovare la sottostringa della stringa S_1 che ha il maggior numero di caratteri (di ugual posto) in comune con la stringa S_2

no no no si no si si

P o r t a n o l a n a

S_1

1 0 0 3

t o n o

S_2

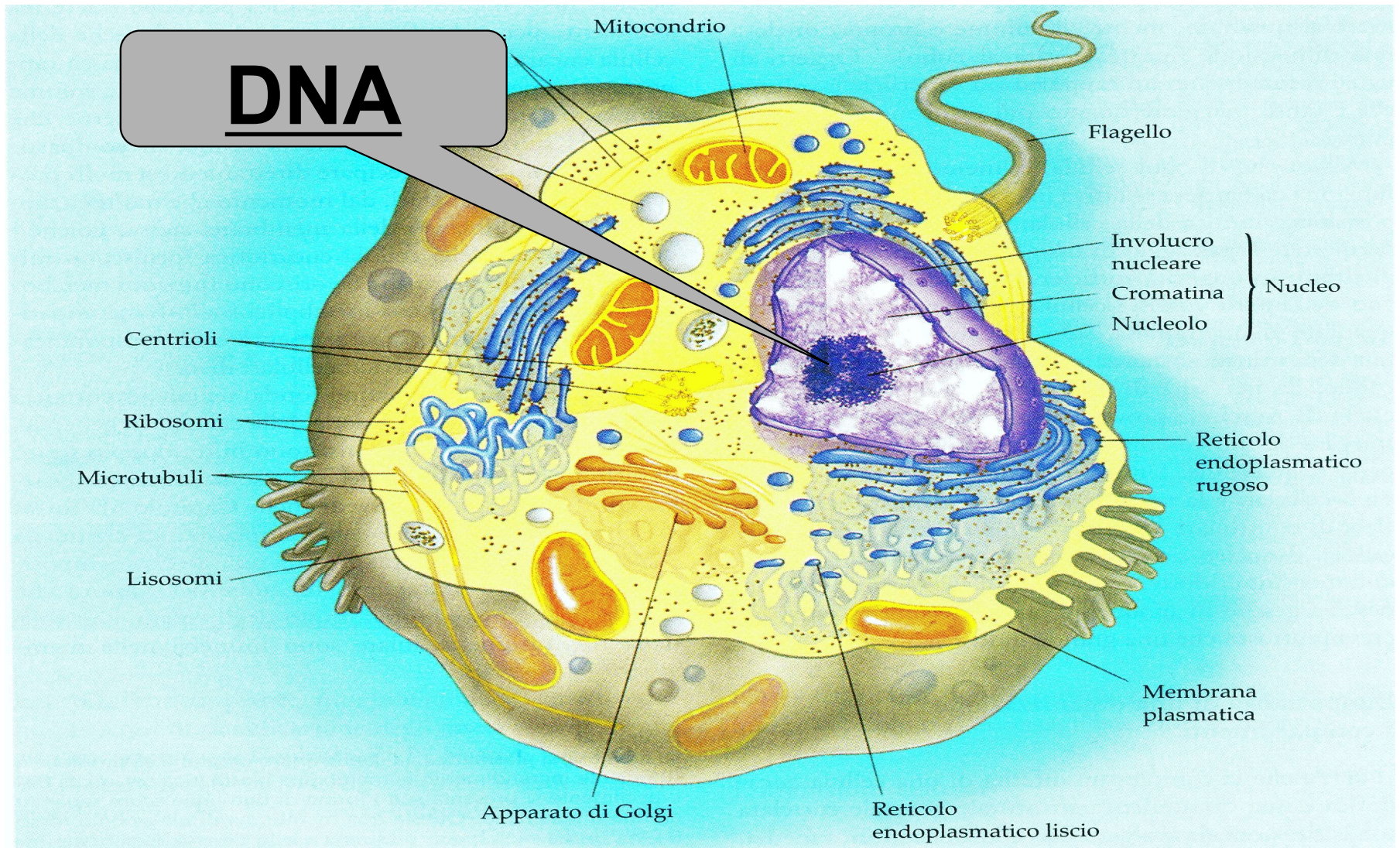
ricerca di una sottostringa in una stringa (*string matching, pattern matching*)

- ✓ sapere se una stringa (chiave) appare come sottostringa di un'altra stringa
- ✓ contare il **numero delle occorrenze** di una sottostringa in una stringa
- ✓ sapere l'**indice** di una sottostringa in una stringa
- ✓ **ricercare** una sottostringa in una stringa e **sostituirla** con un'altra sottostringa (*find/replace*)

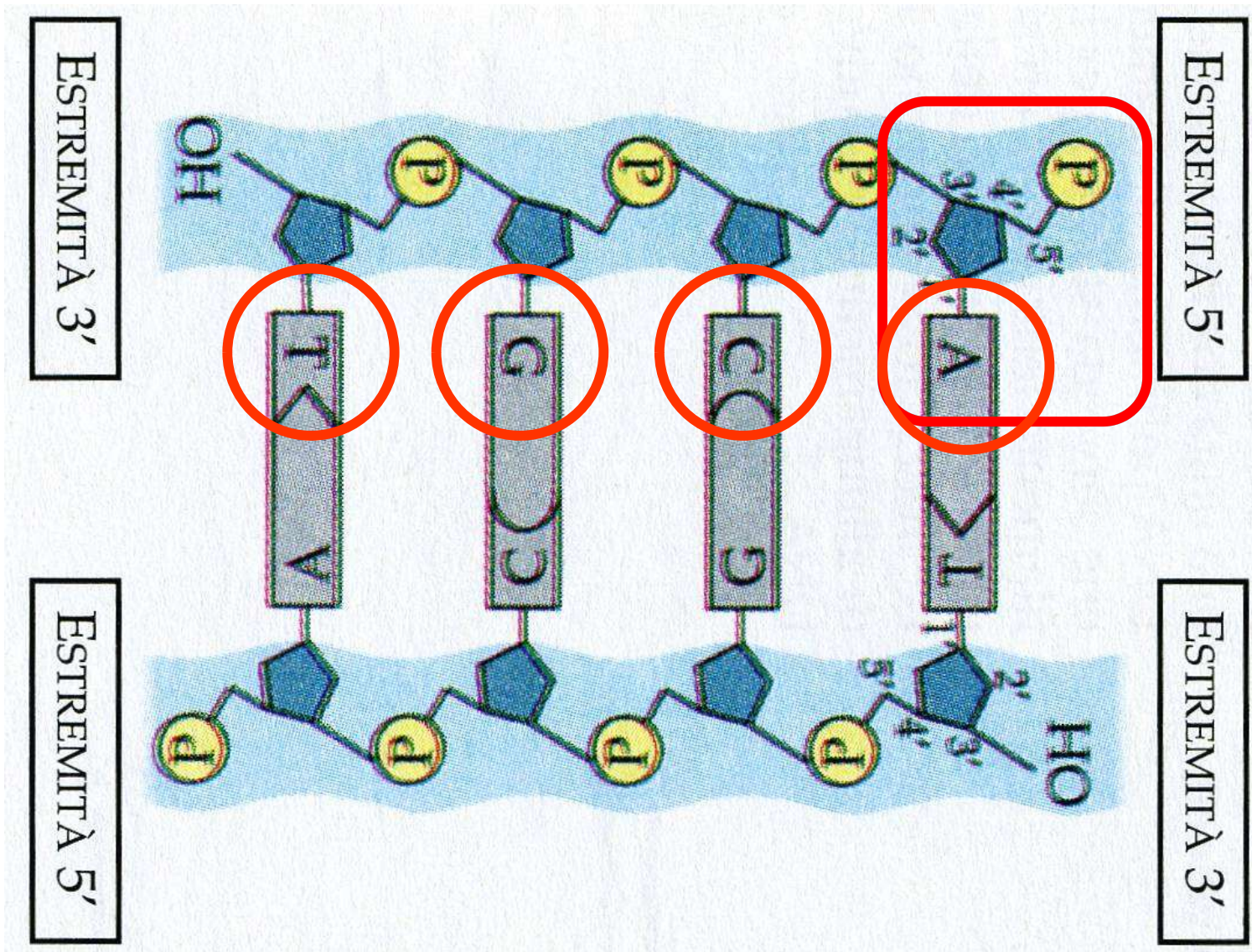
- ✓ il DNA si trova nel nucleo delle cellule
- ✓ è il *programma* che ogni cellula del nostro corpo (ce ne sono circa 10^{13}) deve continuamente eseguire
- ✓ l'elemento base del DNA è una molecola chiamata **nucleotide**
- ✓ per ogni specie o essere vivente esistono solo **4 nucleotidi**, identificati dai nomi
Adenina, Timina, Citosina, Guanina

pattern matching in Bioinformatica

cellula

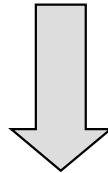


pattern matching in Bioinformatics



pattern matching in Bioinformatica

- ✓ il DNA è una lunga catena (doppia) di nucleotidi



il DNA è un **testo (stringa)** sull'alfabeto $\{A, T, C, G\}$

pattern matching in Bioinformatica

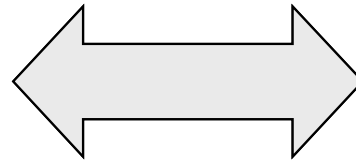
- un **gene** è un segmento di DNA, che contiene l'informazione relativa a una specifica funzionalità
- il DNA è un testo (**stringa**) contenente tutte le informazioni necessarie per la vita della cellula e dell'individuo
- l'insieme di tutte le informazioni genetiche, e quindi dell'intero DNA di un individuo o di una specie, è detto **genoma**

A	T	T	C	G	G	T	C	G	A	A	C	C	T	C	G	A	C	T
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

gene (sottostringa)

pattern matching in Bioinformatica

ricercare un
**particolare
gene** in una
**sequenza di
DNA**



ricercare un
pattern in
una
stringa