Allineamento di sequenze

L'algoritmo di Needleman-Wunsch

Davide Bresolin
a.a 2025/26



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sommario



1 Il problema da risolvere

2 L'algoritmo che lo risolvere

3 Il programma che esegue l'algoritmo

Allineamento di sequenze

- L'allineamento di sequenze è una operazione fondamentale in bioinformatica.
- Serve a confrontare due sequenze di amminoacidi, DNA o RNA.
- Permette di individuare regioni identiche o simili tra le sequenze.
- Queste somiglianze possono indicare relazioni funzionali, strutturali o filogenetiche (evolutive).

Esempi

Come allineo `cane` e `pane`?

```
c a n e
  | | |
p a n e
```

- 1 mismatch: `c` diventa `p`
- 3 match: le altre lettere non cambiano

Esempi

Come allineo `cane` e `pane`?

```
c a n e
  | | |
p a n e
```

- 1 mismatch: `c` diventa `p`
- 3 match: le altre lettere non cambiano

Come allineo `mare` e `amare`?

```
- m a r e
  | | | |
a m a r e
```

- 1 gap: per far meglio combaciare le lettere
- 4 match: per le lettere in comune

Esempi

Come allineo `giotto` e `guitto`?

Esempi

Come allineo `giotto` e `guitto`?

```
g i o t t o
|     | | |
g u i t t o
```

- 2 mismatch
- 4 match

Esempi

Come allineo `giotto` e `guitto`?

```
g i o t t o
|     | | |
g u i t t o
```

- 2 mismatch
- 4 match

```
g u i - t t o
|   |   | | |
g - i o t t o
```

- 2 gap
- 5 match

Qual è l'allineamento migliore?

Dare un punteggio agli allineamenti



- A ciascuna coppia di simboli viene assegnato un punteggio
- La somma dei punteggi costituisce il punteggio totale dell'allineamento
- Esistono diversi modi di assegnare i punteggi
- Esempio di punteggio per le sequenze di DNA:
  - Match: $+1$
  - Mismatch: 0
  - Gap: -1
- Esempio di punteggio per le proteine:
  - Match e mismatch: da una tabella di corrispondenze
  - Gap: -5

Quale è l'allineamento migliore?

```
g i o t t o      g u i - t t o
|     | | |      |   |   | | |
g u i t t o      g - i o t t o
```

- Match: $+1$
- Mismatch: 0
- Gap: -1

Quale è l'allineamento migliore?


```
g i o t t o
|     | | |
g u i t t o
```

Punteggio: 4

```
g u i - t t o
|   |   | | |
g - i o t t o
```

Punteggio: 3

- Match: $+1$
- Mismatch: 0
- Gap: -1

Definizione del problema

Allineamento ottimo di sequenze

Date due sequenze di caratteri

Trova l'allineamento con punteggio massimo

Usando il seguente punteggio:

- Match: $+1$
- Mismatch: 0
- Gap: -1

Sommario

1 Il problema da risolvere

2 L'algoritmo che lo risolvere

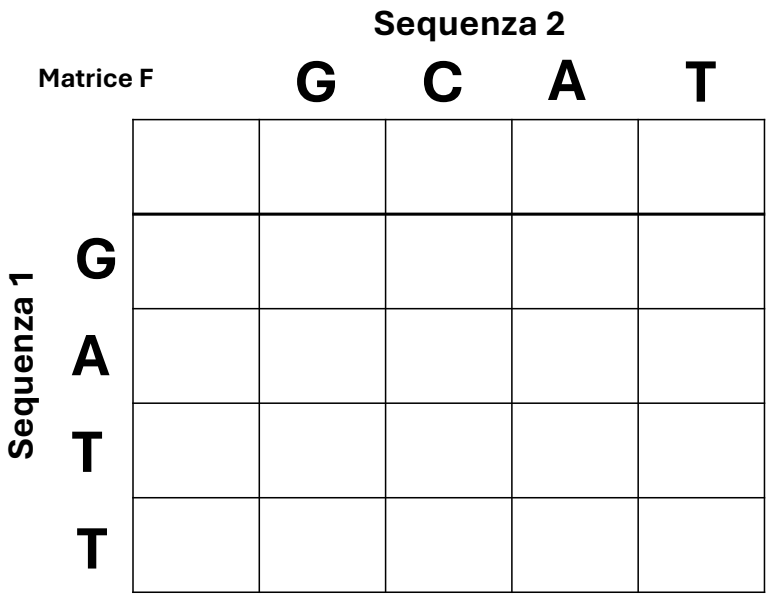
3 Il programma che esegue l'algoritmo

Algoritmo di Needleman-Wunsch

- Obiettivo: trovare (tutti!) gli allineamenti ottimi di 2 sequenze
- È un algoritmo di programmazione dinamica
- Programmazione dinamica = risolvere problemi complessi suddividendoli in sottoproblemi più semplici, memorizzando i risultati per evitare calcoli ripetuti.

Memorizzare i risultati

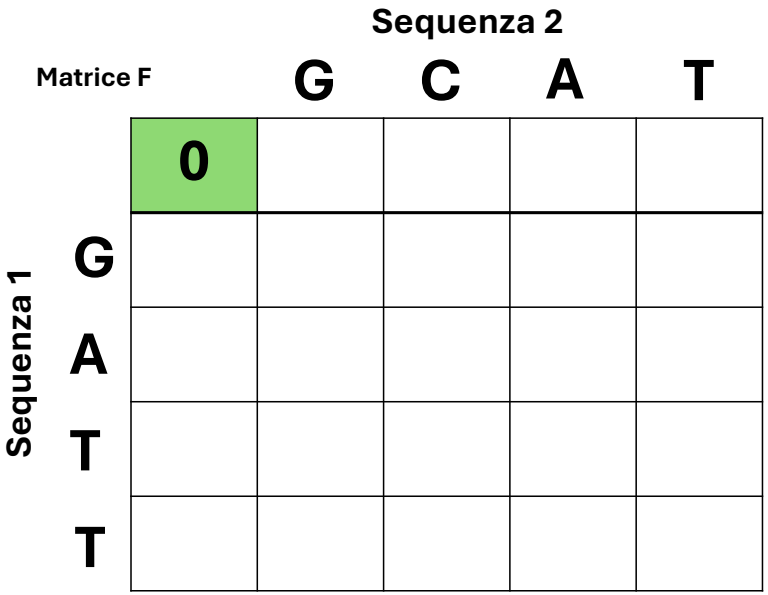
Sequenza 1: G A T T Sequenza 2: G C A T



- Punteggio:
  - Match: $+1$
  - Mismatch: 0
  - Gap: -1
- n: lunghezza di Sequenza 1
- m: lunghezza di Sequenza 2
- Crea un matrice F:
  - $n+1$ righe
  - $m+1$ colonne

Regole di calcolo

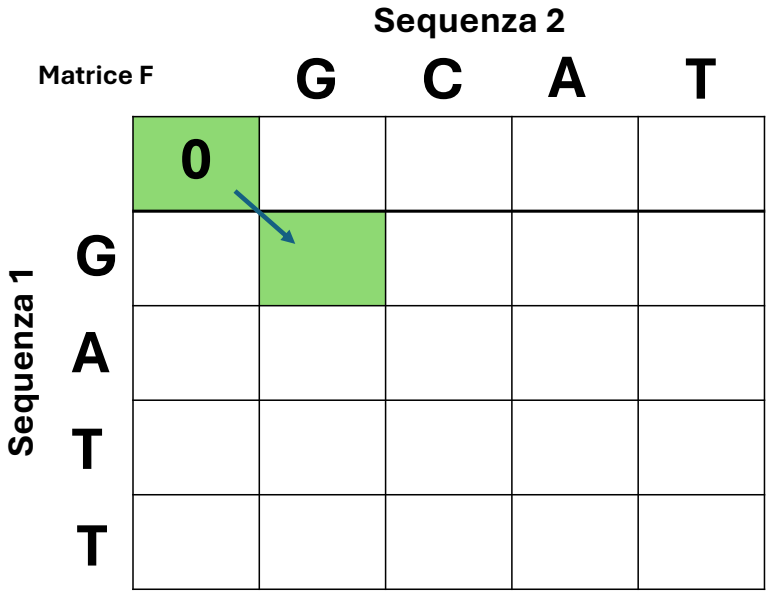
Sequenza 1: G A T T Sequenza 2: G C A T



- Calcola i punteggi per l'intera tabella, una cella alla volta
- Inizia mettendo 0 nella cella in posizione $(0, 0)$

Regole di calcolo

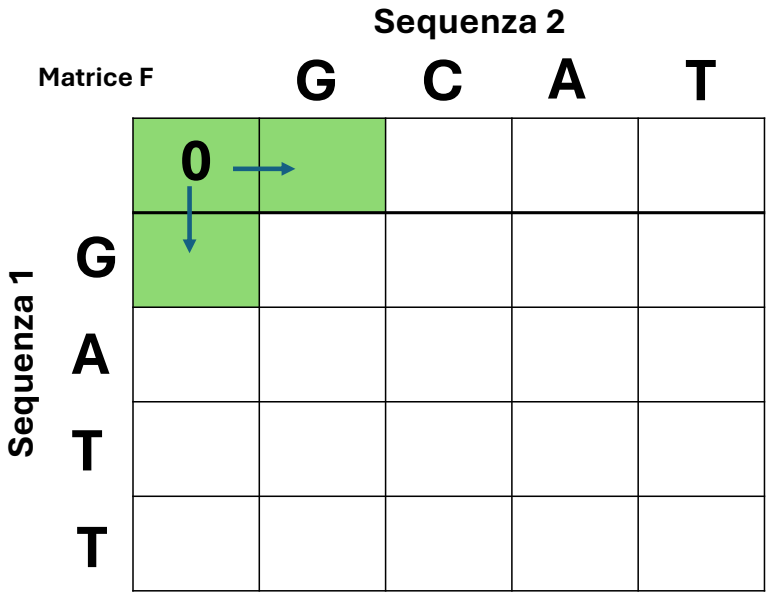
Sequenza 1: G A T T Sequenza 2: G C A T



- Calcola i punteggi per l'intera tabella, una cella alla volta
- Inizia mettendo 0 nella cella in posizione $(0, 0)$
- Un movimento in diagonale è un match o un mismatch

Regole di calcolo

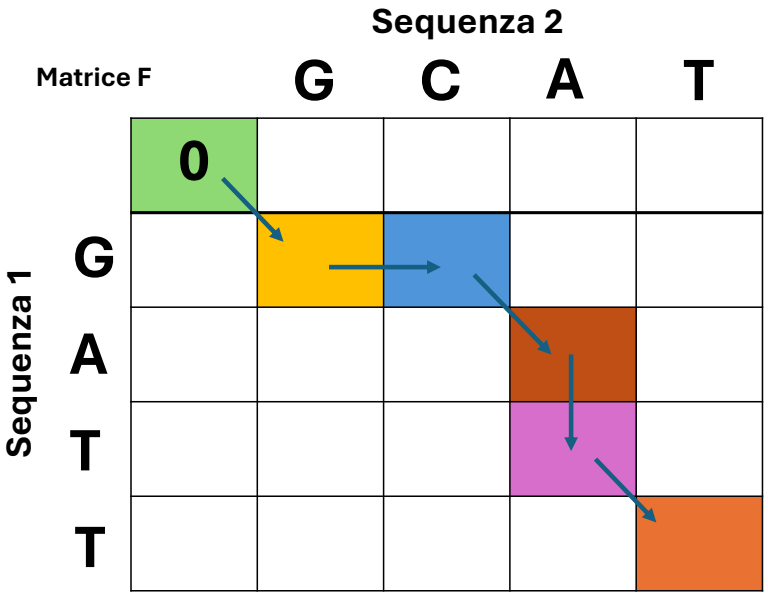
Sequenza 1: G A T T Sequenza 2: G C A T

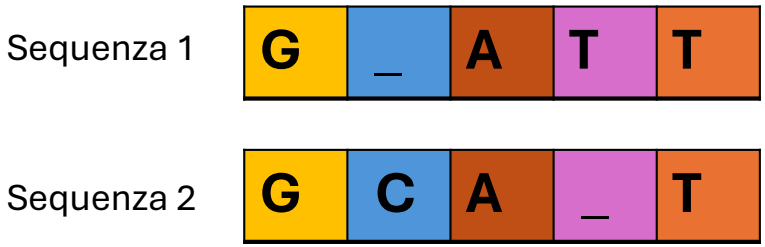


- Calcola i punteggi per l'intera tabella, una cella alla volta
- Inizia mettendo 0 nella cella in posizione $(0, 0)$
- Un movimento in diagonale è un match o un mismatch
- Un movimento in orizzontale o in verticale è sempre un gap

Rappresentare il risultato

Sequenza 1: G A T T Sequenza 2: G C A T





Problema:

come capire quale percorso seguire?

Inizializzazione

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1				
A	-2				
T	-3				
T	-4				

- Se mi muovo verso destra o verso il basso sto inserendo un gap in una delle due sequenze
- Più gap inserisco più devo cumulare le penalità (nel nostro sistema di punteggio -1)
- Riempio quindi le celle della prima riga e della prima colonna cumulando le penalità dalla cella (0,0)

Calcolo di una cella

Sequenza 1: G A T T Sequenza 2: G C A T

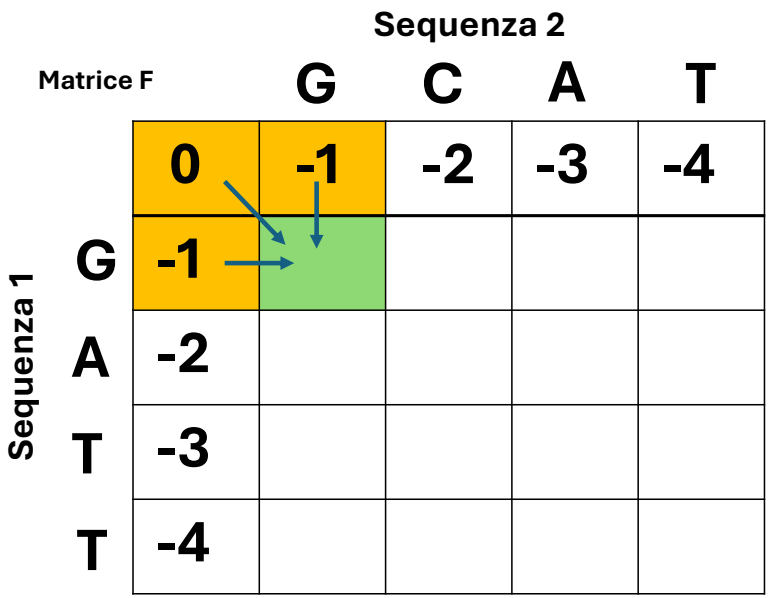
Matrice F

Sequenza 1		Sequenza 2 G	C	A	T
	0	-1	-2	-3	-4
G	-1				
A	-2		?	?	
T	-3		?		
T	-4				

- Posso riempire una cella vuota con un numero solamente quando le celle a sinistra, sopra e in diagonale sono già state riempite

Calcolo di una cella

Sequenza 1: G A T T Sequenza 2: G C A T



Per riempire una nuova cella, calcolo:

- il punteggio che avrei arrivando da sinistra (cumulo penalità gap)
- il punteggio che avrei arrivando da sopra (cumulo penalità gap)
- il punteggio che avrei arrivando dalla diagonale (cumulo punteggio match o mismatch)

Il valore della cella è il massimo dei 3 punteggi

Cella (1, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1				
A	-2				
T	-3				
T	-4				

Per la cella (1, 1):

Cella (1, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1				
A	-2				
T	-3				
T	-4				

Per la cella $(1, 1)$:

- punteggio da sinistra: $-1 - 1 = -2$

Cella (1, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1				
A	-2				
T	-3				
T	-4				

Per la cella $(1, 1)$:

- punteggio da sinistra: $-1 - 1 = -2$
- punteggio da sopra: $-1 - 1 = -2$

Cella (1, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1				
A	-2				
T	-3				
T	-4				

Per la cella $(1, 1)$:

- punteggio da sinistra: $-1 - 1 = -2$
- punteggio da sopra: $-1 - 1 = -2$
- punteggio dalla diagonale (match): $0 + 1 = 1$

Cella (1, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1			
A	-2				
T	-3				
T	-4				

Per la cella $(1, 1)$:

- punteggio da sinistra: $-1 - 1 = -2$
- punteggio da sopra: $-1 - 1 = -2$
- punteggio dalla diagonale (match): $0 + 1 = 1$

Il valore della cella è 1

Cella $(1, 2)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	?		
A	-2				
T	-3				
T	-4				

Per la cella $(1, 2)$:

Cella $(1, 2)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	?		
A	-2				
T	-3				
T	-4				

Per la cella $(1, 2)$:

- punteggio da sinistra: $1 - 1 = 0$

Cella $(1, 2)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	?		
A	-2				
T	-3				
T	-4				

Per la cella $(1, 2)$:

- punteggio da sinistra: $1 - 1 = 0$
- punteggio da sopra: $-2 - 1 = -3$

Cella $(1, 2)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	?		
A	-2				
T	-3				
T	-4				

Per la cella $(1, 2)$:

- punteggio da sinistra: $1 - 1 = 0$
- punteggio da sopra: $-2 - 1 = -3$
- punteggio dalla diagonale (mismatch): $-1 + 0 = -1$

Cella $(1, 2)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1		Sequenza 2 G	Sequenza 2 C	Sequenza 2 A	Sequenza 2 T
	0	-1	-2	-3	-4
G	-1	1	0		
A	-2				
T	-3				
T	-4				

Per la cella $(1, 2)$:

- punteggio da sinistra: $1 - 1 = 0$
- punteggio da sopra: $-2 - 1 = -3$
- punteggio dalla diagonale (mismatch): $-1 + 0 = -1$

Il valore della cella è 0

Cella $(1, 3)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	
A	-2				
T	-3				
T	-4				

Per la cella $(1, 3)$:

- punteggio da sinistra: $0 - 1 = -1$
- punteggio da sopra: $-3 - 1 = -4$
- punteggio dalla diagonale (mismatch): $-2 + 0 = -2$

Il valore della cella è -1

Cella $(1, 4)$

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2				
T	-3				
T	-4				

Per la cella $(1, 4)$:

- punteggio da sinistra: $-1 - 1 = -2$
- punteggio da sopra: $-4 - 1 = -5$
- punteggio dalla diagonale (mismatch): $-3 + 0 = -3$

Il valore della cella è -2

Cella (2, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0			
T	-3				
T	-4				

Per la cella (2, 1):

- punteggio da sinistra: $-2 - 1 = -3$
- punteggio da sopra: $1 - 1 = 0$
- punteggio dalla diagonale (mismatch): $-1 + 0 = -1$

Il valore della cella è 0

Cella (2, 1)

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0			
T	-3				
T	-4				

Per la cella (2, 1):

- punteggio da sinistra: $-2 - 1 = -3$
- punteggio da sopra: $1 - 1 = 0$
- punteggio dalla diagonale (mismatch): $-1 + 0 = -1$

Il valore della cella è 0

Ora continuate voi a riempire la tabella

Risultato finale

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

- Il valore nella cella in basso a destra è il punteggio massimo di tutti gli allineamenti
- Nell'esempio, gli allineamenti ottimi hanno punteggio 1

Risultato finale

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

- Il valore nella cella in basso a destra è il punteggio massimo di tutti gli allineamenti
- Nell'esempio, gli allineamenti ottimi hanno punteggio 1

Problema

Come costruisco gli allineamenti di punteggio massimo?

Torniamo indietro

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1		Sequenza 2 G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

- Parti dalla cella in basso a destra
- Vai all'indietro fino alla cella in alto a sinistra, spostandoti di volta in volta nella cella che ha generato il valore della cella corrente.

Torniamo indietro

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1		Sequenza 2 G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

Per la cella $(4, 4)$:

- punteggio da sinistra: $0 - 1 = -1$
- punteggio da sopra: $2 - 1 = 1$
- punteggio dalla diagonale (match): $1 + 1 = 2$

Spostati lungo la diagonale nella cella $(3, 3)$

Torniamo indietro

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

Per la cella $(3, 3)$:

- punteggio da sinistra: $0 - 1 = -1$
- punteggio da sopra: $1 - 1 = 0$
- punteggio dalla diagonale (mismatch): $1 + 0 = 1$

Spostati lungo la diagonale nella cella $(2, 2)$

Torniamo indietro

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1		Sequenza 2 G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

Per la cella $(2, 2)$:

- punteggio da sinistra: $0 - 1 = -1$
- punteggio da sopra: $0 - 1 = -1$
- punteggio dalla diagonale (mismatch): $1 + 0 = 1$

Spostati lungo la diagonale nella cella $(1, 1)$

Torniamo indietro

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1 \ Sequenza 2		G	C	A	T
	0	-1	-2	-3	-4
G	-1	1	0	-1	-2
A	-2	0	1	1	0
T	-3	-1	0	1	2
T	-4	-2	-1	0	2

Per la cella $(1, 1)$:

- punteggio da sinistra: $-1 - 1 = -2$
- punteggio da sopra: $-1 - 1 = -2$
- punteggio dalla diagonale (match): $0 + 1 = 1$

Spostati lungo la diagonale nella cella $(0, 0)$

Risultato finale

Sequenza 1: G A T T Sequenza 2: G C A T

Matrice F

Sequenza 1		Sequenza 2 G	C	A	T
		0	-1	-2	-3	-4
	G	-1	1	0	-1	-2
	A	-2	0	1	1	0
	T	-3	-1	0	1	2
	T	-4	-2	-1	0	2

Allineamento ottimo:

```
G A T T
|   |
G C A T
```

Punteggio: 2

Sommario

1 Il problema da risolvere

2 L'algoritmo che lo risolvere

3 Il programma che esegue l'algoritmo