Terza lezione.

1. Similarità e omologia.

2. Allineamenti di sequenze.

3. Sostituzioni e gap.

4. Algoritmo di allineamento

Similarita' e omologia

- Due sequenze sono simili se possono essere allineate in modo che molti ammino acidi corrispondenti sono identici o simili

- Tecnicamente due o piu' sequenze possono essere definite omologhe se derivano da un progenitore comune

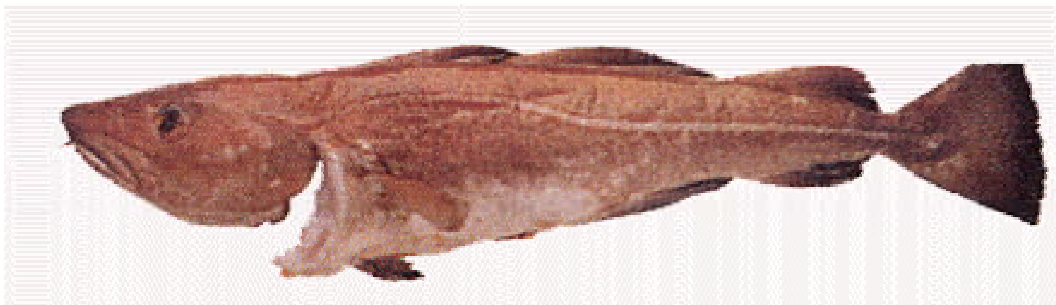
- L'omologia tra due sequenze si deduce dalla loro similarità in sequenza o funzione.

- Il concetto di similarita' si puo' estendere anche alle strutture 3D

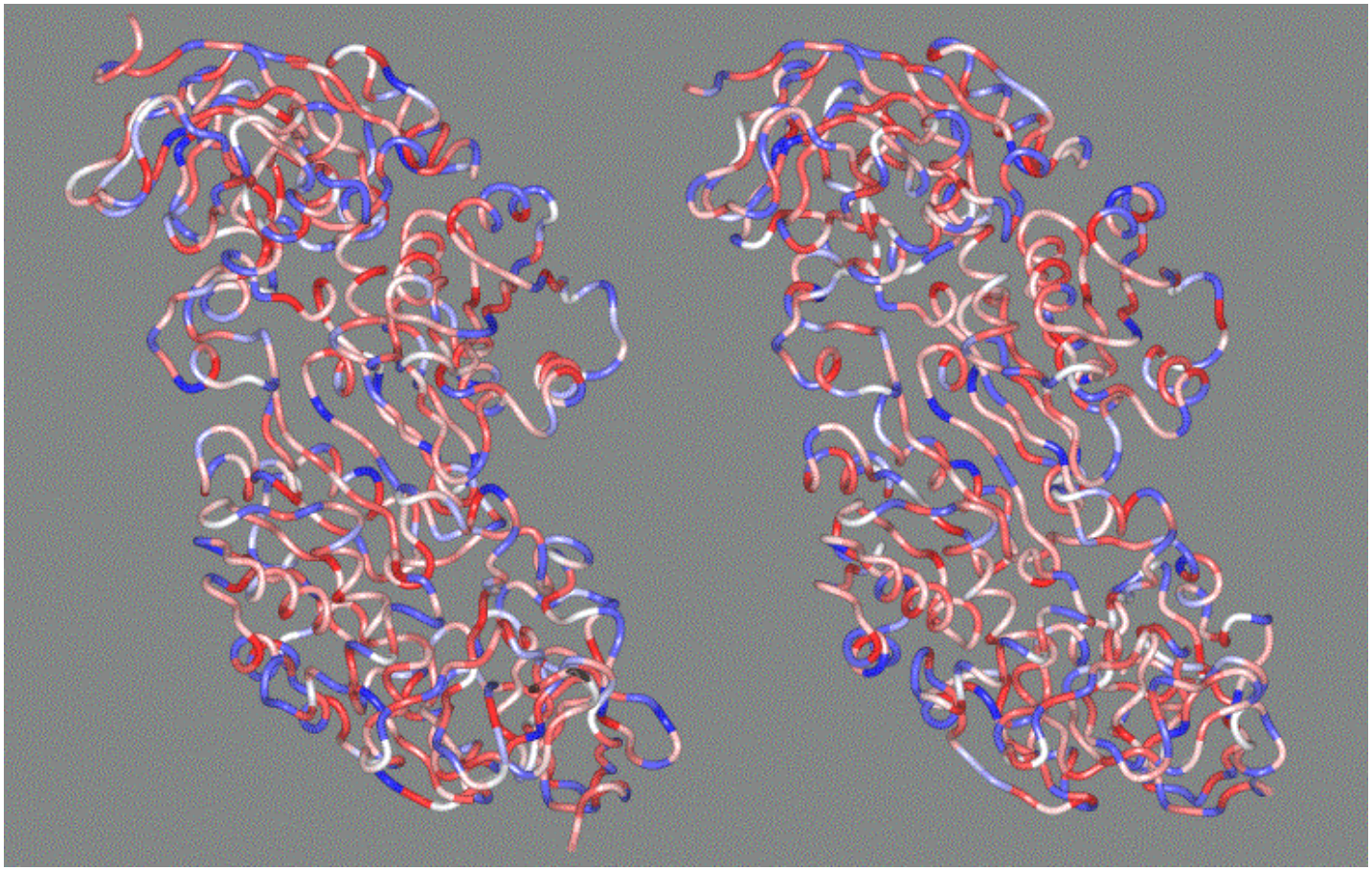
L’omologia tra due proteine/geni si deduce dalla loro similarita’ in sequenza (struttura) o funzione.

Specie = uomo

Specie = merluzzo



La differenza
e’ evidente!

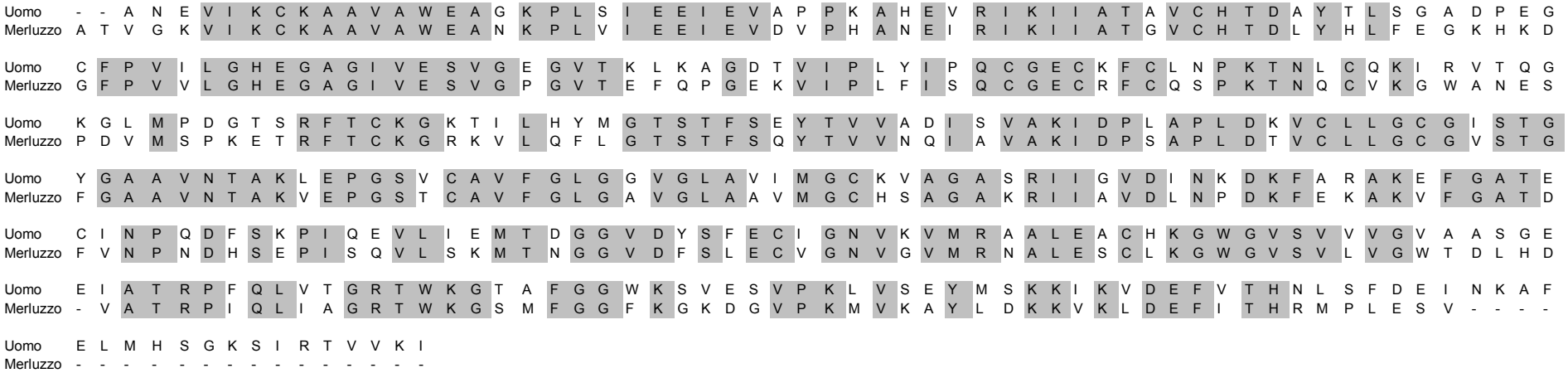


Specie = uomo

Specie = merluzzo

```
Uomo      - - A N E V I K C K A A V A W E A G K P L S I E E I E V A P P K A H E V R I K I I A T A V C H T D A Y T L S G A D P E G
Merluzzo  A T V G K V I K C K A A V A W E A N K P L V I E E I E V D V P H A N E I R I K I I A T G V C H T D L Y H L F E G K H K D

Uomo      C F P V I L G H E G A G I V E S V G E G V T K L K A G D T V I P L Y I P Q C G E C K F C L N P K T N L C Q K I R V T Q G
Merluzzo  G F P V V L G H E G A G I V E S V G P G V T E F Q P G E K V I P L F I S Q C G E C R F C Q S P K T N Q C V K G W A N E S

Uomo      K G L M P D G T S R F T C K G K T I L Q Y M G T S T F S E Y T V V A D I S V A K I D P L A P L D K V C L L G C G I S T G
Merluzzo  P D V M S P K E T R F T C K G R K V L Q Y L G T S T F S Q Y T V V N Q I A V A K I D P S A P L D T V C L L G C G V S T G

Uomo      Y G A A V N T A K L E P G S V C A V F G L G G V G L A V I M G C K V A G A S R I I G V D I N K D K F A R A K E F G A T E
Merluzzo  F G A A V N T A K V E P G S T C A V F G L G A V G L A A V M G C H S A G A K R I I A V D L N P D K F E K A K V F G A T D

Uomo      C I N P Q D F S K P I Q E V L I E M T D G G V D Y S F E C I G N V K V M R A A L E A C H K G W G V S V V V G V A A S G E
Merluzzo  F V N P N D H S E P I S Q V L S K M T N G G V D F S L E C G N V G V M R N A L E S C L K G W G V S V L V G W T D L H D

Uomo      E I A T R P F Q L V T G R T W K G T A F G G W K S V E S V P K L V S E Y M S K K I K V D E F V T H N L S F D E I N K A F
Merluzzo  - V A T R P I Q L I A G R T W K G S M F G G F K G K D G V P K M V K A Y L K K K V K L D E F I T H R M P L E S V - - - -

Uomo      E L M H S G K S I R T V
Merluzzo  - - - - - - - - - - - - - - -
```

La differenza e' evidente?



Due sequenze sono simili se possono essere *allineate* in modo che molti amminoacidi (o nucleotidi) corrispondenti sono identici o simili.

Con l'allineamento di sequenze cerchiamo la corrispondenza tra amminoacidi (o basi nucleotidiche) che più probabilmente riflette l'evoluzione di due proteine (geni)

Allineamento di sequenze

Qual è la corrispondenza tra gli amminoacidi (nucleotidi) che più probabilmente rispecchia l'evoluzione delle due proteine (geni)?

Allineamento di sequenza

Scopo: minimizzare la distanza evolutiva tra le sequenze da allineare, quindi minimizzare le differenze (che equivale a massimizzare le similarita') tra le componenti (nucleotidi o amminoacidi) delle sequenze stesse

L'***ipotesi*** di allineamento ***piu' ragionevole*** e' quella che coinvolge il numero minore di eventi di mutazione per passare da una sequenza all'altra

allineamenti di sequenze

Applicazioni

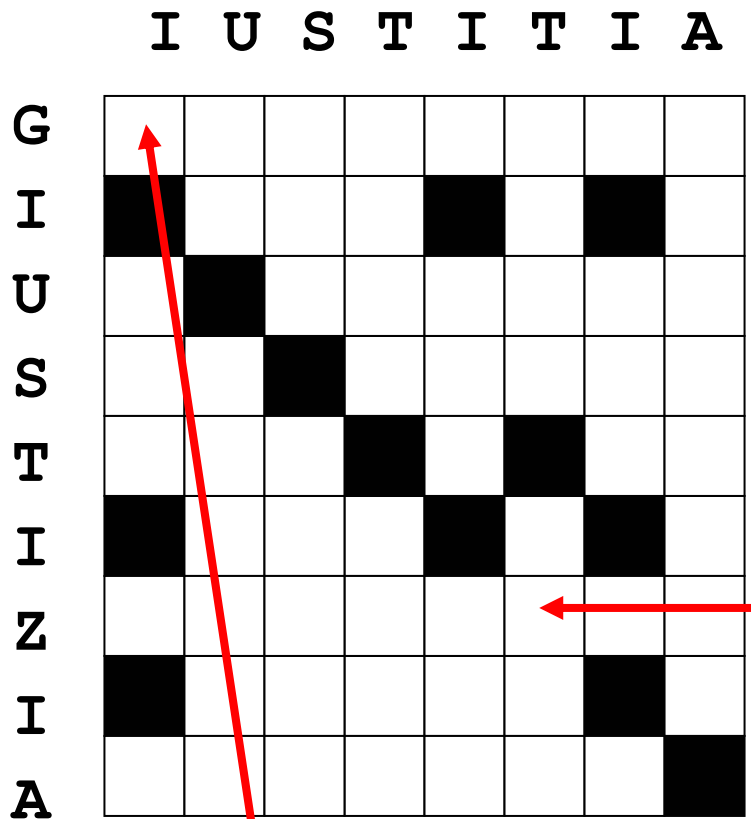
- Riconoscimento di funzione
  - basta stabilire che due sequenze sono simili
- Filogenia
  - occorre misurare la similarita' quantitativamente
- Model building
  - occorre costruire il miglior allineamento possibile

La distribuzione delle sequenze

Delle sequenze proteiche note ad oggi meno del 15% hanno una caratterizzazione funzionale sperimentale soddisfacente

Il 25% delle proteine non ha omologhe note

Il **dotplot** (disegno a punti) fornisce una visione immediata della similarita’ tra due stringhe di caratteri

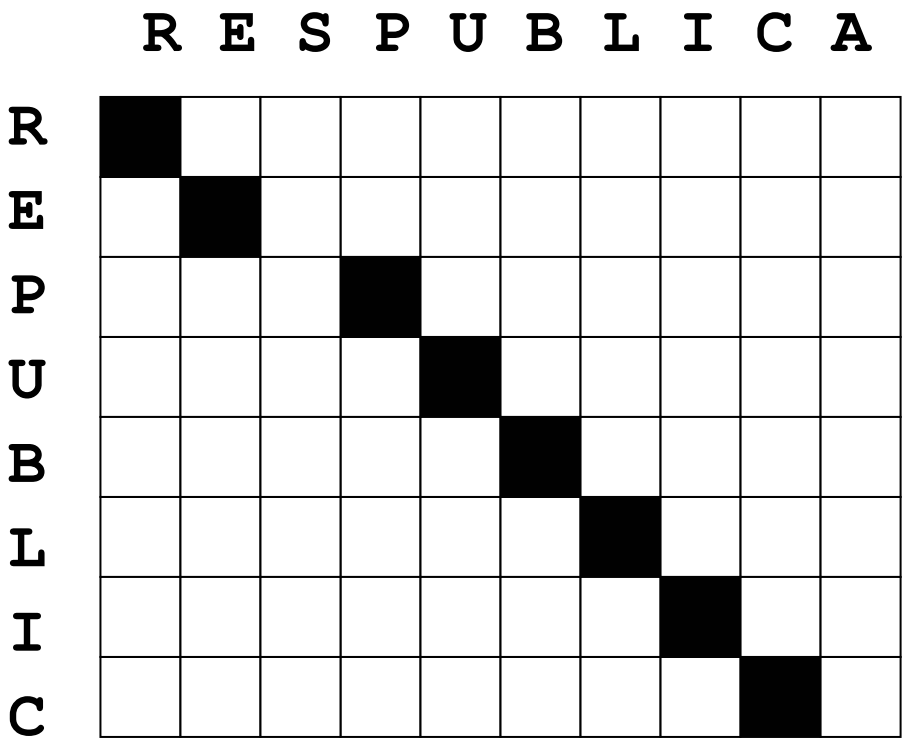


latino –IUSTITIA

italiano GIUSTIZIA

inserzione

mutazione

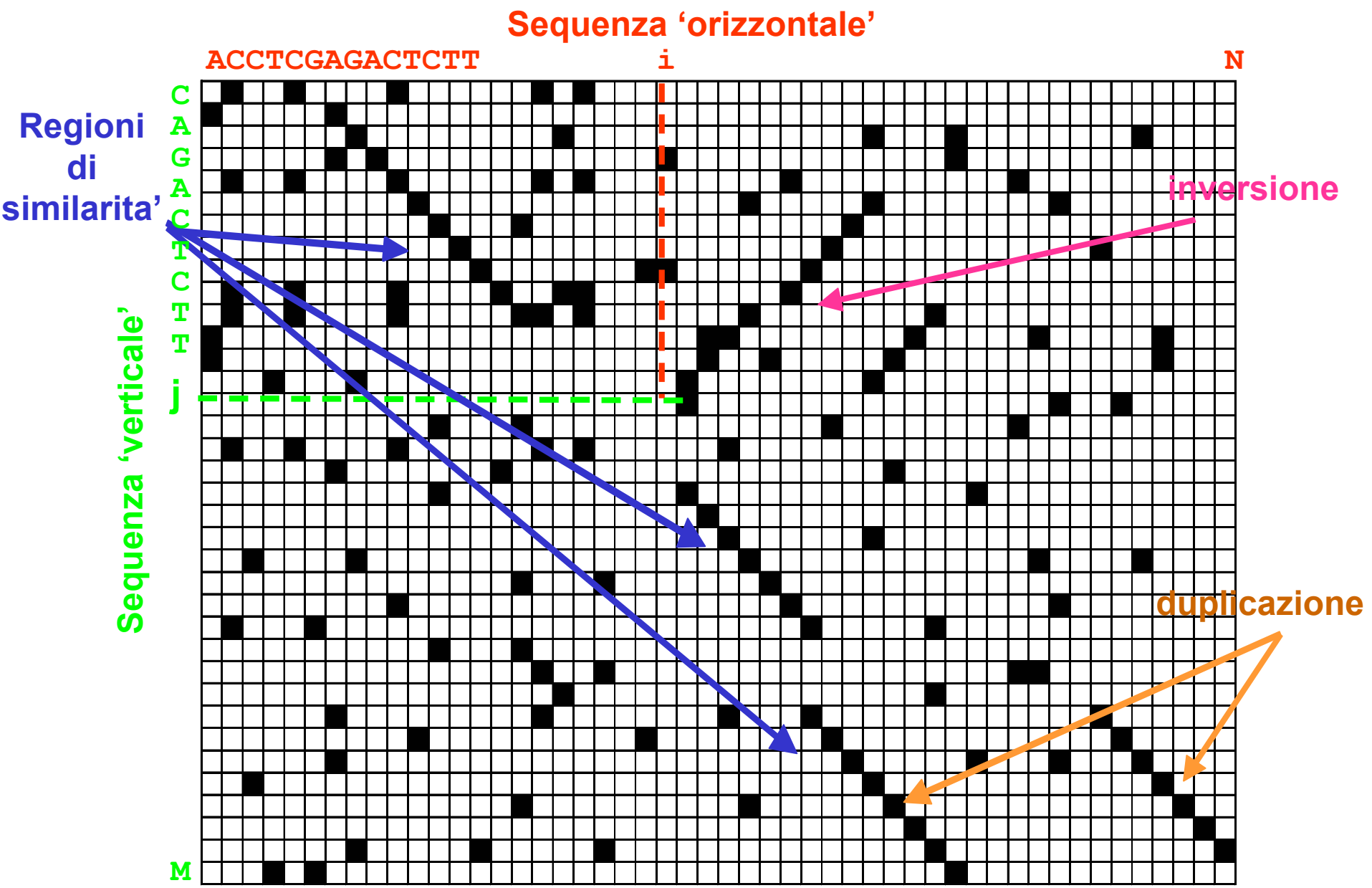


Latino **RESPUBLICA**

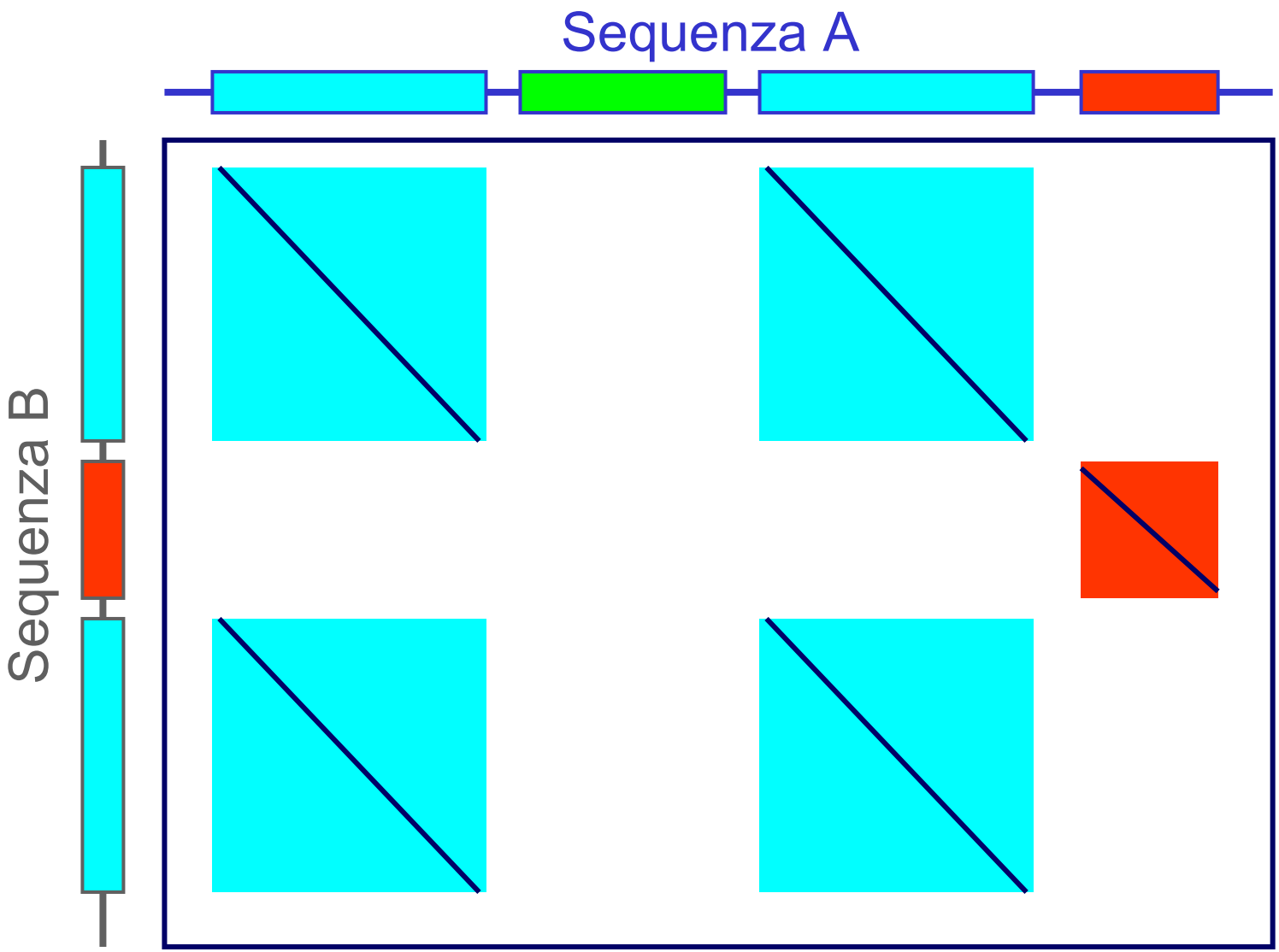
Inglese **RE-PUBLIC-**

delezioni

Dot plot


Sequenza 'orizzontale'
ACCTCGAGACTCTT
i
N
Regioni di similarita'
C
A
G
A
C
T
C
T
T
j
Sequenza 'verticale'
M
inversione
duplicazione

Dotplot di proteine domain-shuffled



Sequenza 1 (19 caratteri)

LAMIAPRIMASEQCREATA

Sequenza 2 (22 caratteri)

MIAALTRASEQDALLINEARE

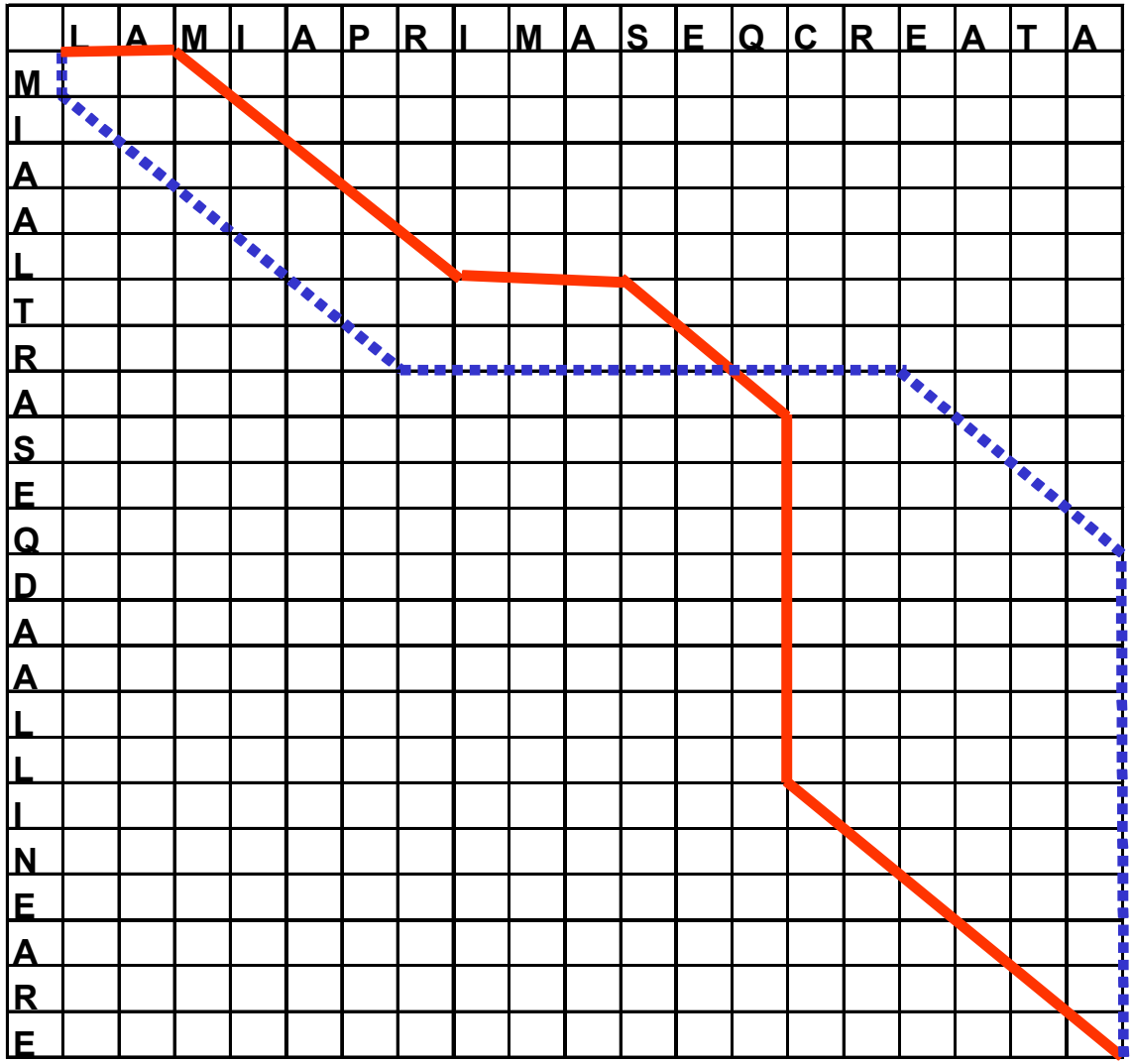
	L	A	M	I	A	P	R	I	M	A	S	E	Q	C	R	E	A	T	A
M	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
S	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

	L	A	M	I	A	P	R	I	M	A	S	E	Q	C	R	E	A	T	A
M	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
S	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

10 punti

LAMIAP---RIMASEQ-------CREATA

--MIA-ALTR--ASEQDAALLIN--EARE



LAMIAPRIMASEQCREATA----------

-MIAALTR-------ASEQDAALLINEARE

LAMIAPRIMASEQ--------CREATA

--MIAAL---TRASEQDAALLINEARE

Allineamento

1) Punteggio per la corrispondenza di aa/basi
2) Penalizzazione di inserzioni / delezioni
3) Algoritmo che effettui l'allineamento
4) Misura della significativita' dell'allineamento

Allineamento

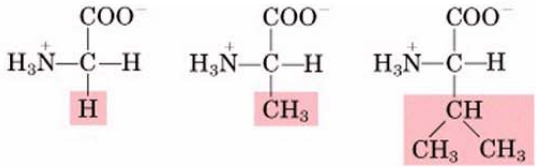
1) Punteggio per la corrispondenza di aa/basi
2) Penalizzazione di inserzioni / delezioni
3) Algoritmo che effettui l'allineamento
4) Misura della significativita' dell'allineamento

1. Punteggio (*Score*)

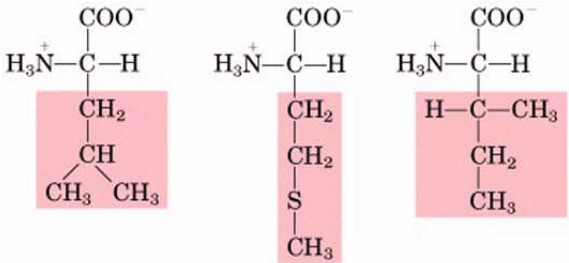
- Identita' tra basi o amminoacidi
- Proprieta' chimico-fisiche degli amminoacidi
- Numero minimo di basi che e' necessario mutare per ottenere la mutazione osservata
- Frequenze di rimpiazzo osservate in famiglie di proteine

Twenty standard Amino Acids

non polari

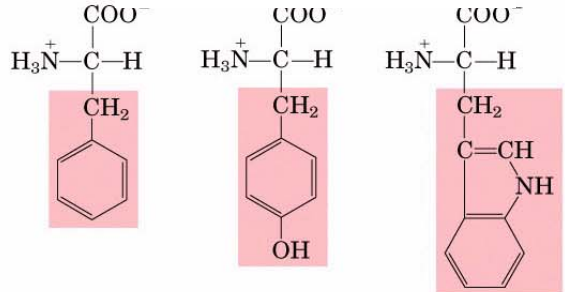


Glicina(**G**) Alanina(**A**) Valina(**V**)



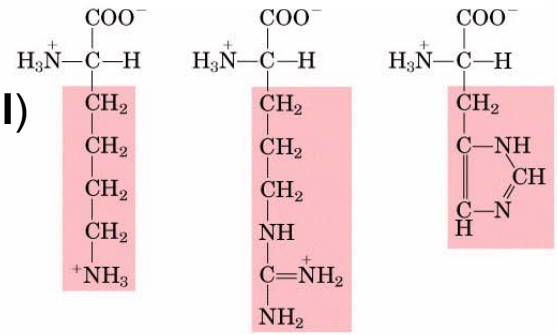
Leucina(**L**) Metionina(**M**) Isoleucina(**I**)

aromatici



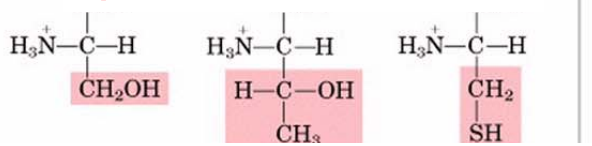
Phenilalanina(**F**) Tirosina(**Y**) Triptofano(**W**)

basici

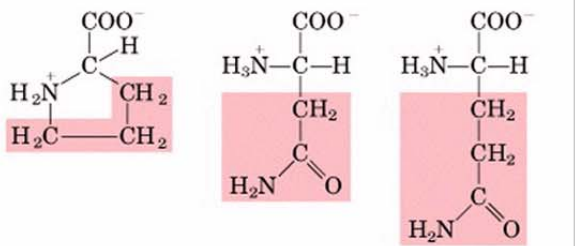


Lisina(**K**) Arginina(**R**) Histidina(**H**)

polari, non carichi

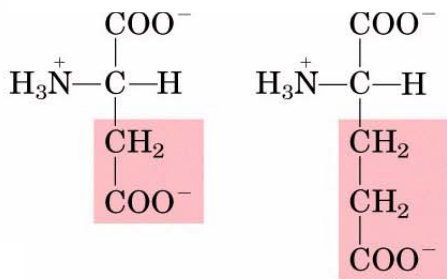


Serina(**S**) Treonina(**T**) Cisteina(**C**)



Prolina(**P**) Asparagina(**N**) Glutammina(**Q**)

acidi



Aspartato(**D**) Glutammato(**E**)

es. 1 Matrice di Sostituzione

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2																			
C	-2	12																		
D	0	-5	4																	
E	0	-5	3	4																
F	-4	-4	-6	-5	9															
G	1	-3	1	0	-5	5														
H	-1	-3	1	1	-2	-2	6													
I	-1	-2	-2	-2	1	-3	-2	5												
K	-1	-5	0	0	-5	-2	0	-2	5											
L	-2	-6	-4	-3	2	-4	-2	2	-3	6										
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6									
N	0	-4	2	1	-4	0	2	-2	1	-3	-2	2								
P	1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6							
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4						
R	-2	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6					
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2				
T	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3			
V	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4		
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10

PAM 250

Point Accepted Mutations

Matrici 20x20

es. 2 Matrice di Sostituzione

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4																			
C	0	9	-3																	
D	-2	-3	6																	
E	-1	-4	2	5																
F	-2	-2	-3	-3	6															
G	0	-3	-1	-2	-3	6														
H	-2	-3	-1	0	-1	-2	8													
I	-1	-1	-3	-3	0	-4	-3	4												
K	-1	-3	-1	1	-3	-2	-1	-3	5											
L	-1	-1	-4	-3	0	-4	-3	2	-2	4										
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5									
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6								
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7							
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5						
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5					
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4				
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5			
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4		
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

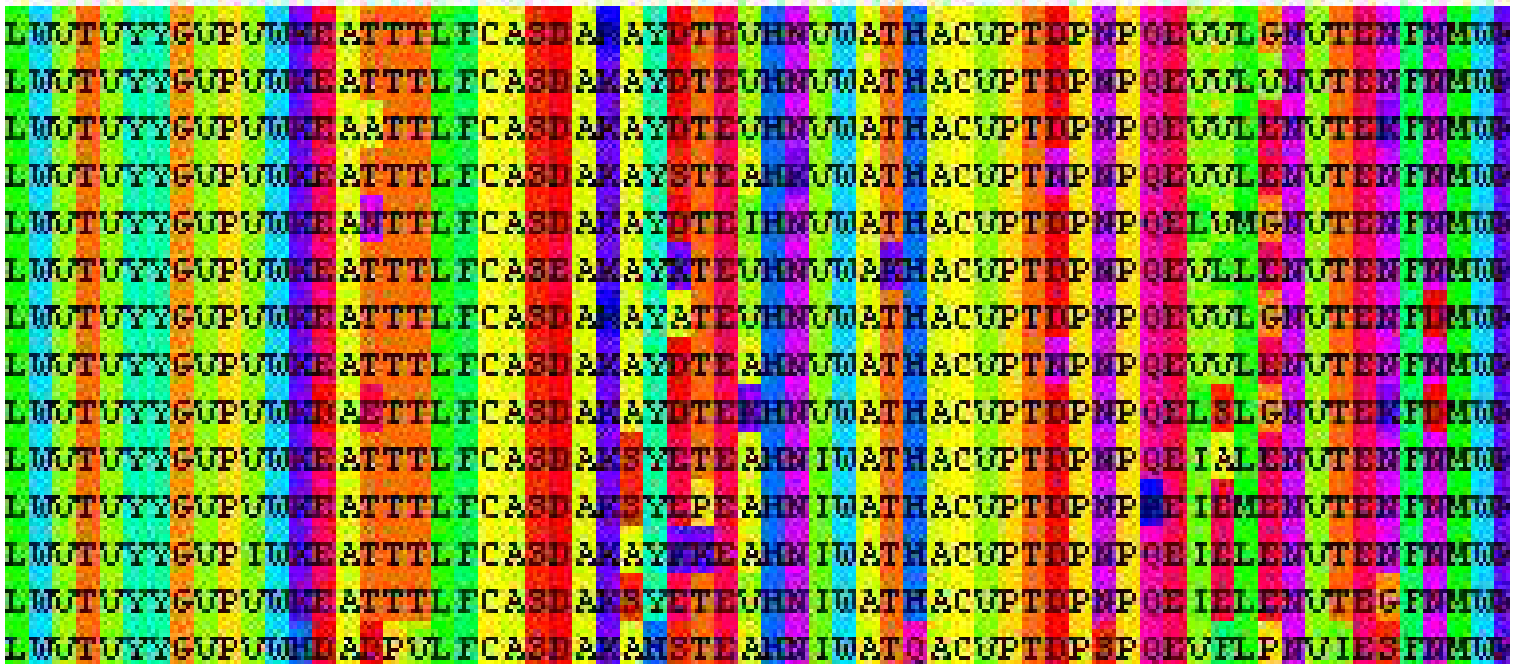
BLOSUM 62

BLOcks **SU**bstitution **M**atrix

(dalla banca dati derivata BLOCKS)

Matrici 20x20

Matrici sito specifiche, basate su regole empiriche



PAM N: Percent/Point Accepted Mutations (dove N è il numero di mutazioni accettate per 100 aa)

BLOSUM N : BLOcks SUbstitution Matrix (dove N è la % max di identità di sequenza tra le omologhe allineate)

Le matrici PAM e BLOSUM riportano i $\log_2$ di:

$$\frac{f_{ij}}{f_i \times f_j}$$

dove

f_{ij} rappresenta la frequenza con cui troviamo due amminoacidi allineati tra loro

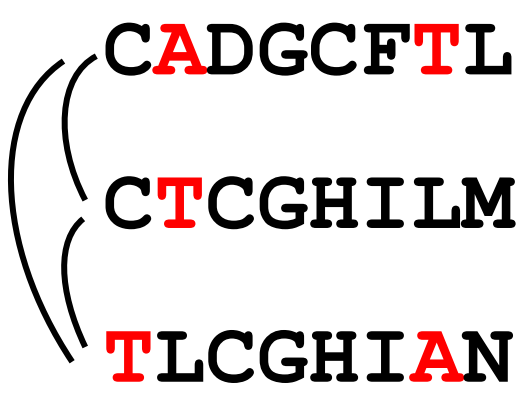
es. f_{AT} è il numero di volte in cui osserviamo una Ala allineata con una Thr diviso il numero totale di coppie in un allineamento

f_i ed f_j rappresentano le frequenze con cui compaiono due amminoacidi in un allineamento

es. f_A ed f_T sono il numero di volte che abbiamo una Ala o una Thr diviso il numero totale di amminoacidi in un allineamento

Esempio di calcolo di punteggio per la sostituzione di una Ala con una Thr

$$\frac{f_{ij}}{f_i \times f_j}$$



Tot aa = 3 x 8 = 24

Tot coppie aa allineati = 3 x 8 = 24

$f_{AT} = 2 / 24 = 0.083$

$f_A = 2 / 24 = 0.083$

$f_T = 3 / 24 = 0.12$

$f_A \times f_T = 0.010$

$$\frac{f_{AT}}{f_A \times f_T} = \frac{0.083}{0.010} = 8$$

$$\ln_2\left(\frac{f_{AT}}{f_A \times f_T}\right) = 3$$

Esempio di calcolo di punteggio per la sostituzione di una Ala con una Thr

$$\frac{f_{ij}}{f_i \times f_j}$$

Seq1 CADGCFTL

Seq2 CTCGHILM

Seq3 TLCGHIAN

Tot aa = 3 x 8 = 24

Tot coppie aa allineati = 3 x 8 = 24

Nelle 3 sequenze considerate l'allineamento A-T occorre 8 volte più che per caso

$$\frac{f_{AT}}{f_A \times f_T} = \frac{0.083}{0.010} = 8$$

$$\ln_2\left(\frac{f_{AT}}{f_A \times f_T}\right) = 3$$

Esempio di calcolo di punteggio per la sostituzione di una Ala con una Thr

$$\frac{f_{ij}}{f_i \times f_j}$$

Seq1 CADGCFTL

Seq2 CTCGHILM

Seq3 TLCGHIAN

Tot aa = 3 x 8 = 24

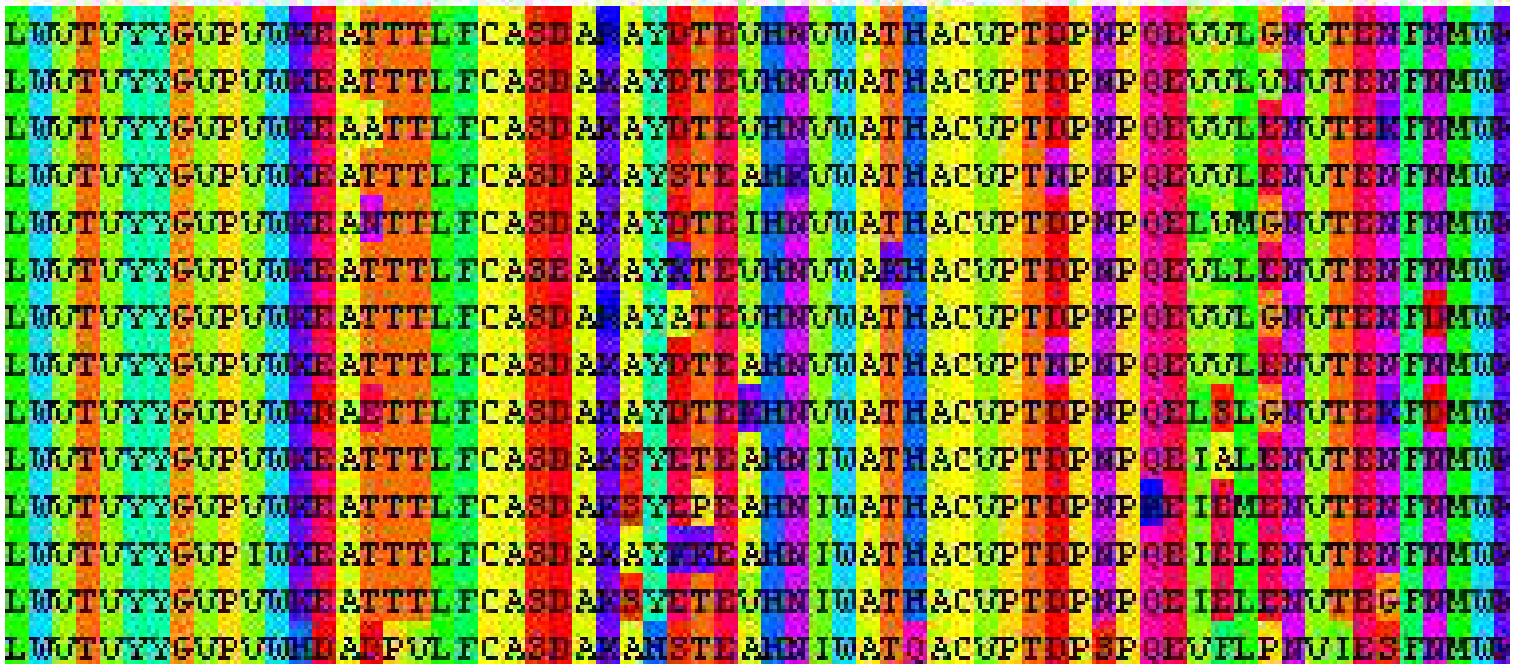
Tot coppie aa allineati = 3 x 8 = 24

In una matrice di sostituzione scriveremmo 3 all'incrocio tra Ala(A) e Thr(T)

$$\frac{f_{AT}}{f_A \times f_T} = \frac{0.083}{0.010} = 8$$

$$\ln_2\left(\frac{f_{AT}}{f_A \times f_T}\right) = 3$$

Matrici sito specifiche, basate su regole empiriche



PAM N: Percent/Point Accepted Mutations (dove N è il numero di mutazioni accettate per 100 aa)

BLOSUM N : BLOcks SUbstitution Matrix (dove N è la % max di identità di sequenza tra le omologhe allineate)

- **PAM 1** (ottenuta da allineamento di proteine con sequenze differenti per 1 aa su 100) può essere usata per generare matrici per distanze evoluzionistiche maggiori:

  moltiplicandola ripetutamente per se stessa.

  PAM2 = PAM1 * PAM1

  etc etc

- **PAM250:**
  - 2,5 mutazioni per residuo
  - equivalente al 20% matches rimanenti tra due sequenze, cioè l'80% delle posizioni amminoacidiche sono cambiate.
  - È la matrice usata di default in molti programmi di analisi.

- Le matrici **BLOSUM** sono state sviluppate per allineare sequenze scarsamente correlate. Hanno largamente soppiantato le PAM.

- Sono ottenute dalla banca dati derivata **BLOCKS** contenente allineamenti di regioni di proteine strettamente correlate, allineabili senza gap.

- **BLOSUM62:** ottenuta da allineamenti di proteine con un massimo di 62 % di identità di sequenza. Molto usata. (Corrisponde approssimativamente ad una PAM110).

es. 1 Matrice di Sostituzione

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2																			
C	-2	12																		
D	0	-5	4																	
E	0	-5	3	4																
F	-4	-4	-6	-5	9															
G	1	-3	1	0	-5	5														
H	-1	-3	1	1	-2	-2	6													
I	-1	-2	-2	-2	1	-3	-2	5												
K	-1	-5	0	0	-5	-2	0	-2	5											
L	-2	-6	-4	-3	2	-4	-2	2	-3	6										
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6									
N	0	-4	2	1	-4	0	2	-2	1	-3	-2	2								
P	1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6							
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4						
R	-2	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6					
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2				
T	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3			
V	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4		
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10

PAM 250

Point Accepted Mutations

Matrici 20x20

es. 1 Matrice di Sostituzione

Ala

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2																			
C	-2	12																		
D	0	-5	4																	
E	0	-5	3	4																
F	-4	-4	-6	-5	9															
G	1	-3	1	0	-5	5														
H	-1	-3	1	1	-2	-2	6													
I	-1	-2	-2	-2	1	-3	-2	5												
K	-1	-5	0	0	-5	-2	0	-2	5											
L	-2	-6	-4	-3	2	-4	-2	2	-3	6										
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6									
N	0	-4	2	1	-4	0	2	-2	1	-3	-2	2								
P	1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6							
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4						
R	-2	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6					
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2				
T	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3			
V	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4		
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10

Le Ala sono facilmente rimpiazzabili

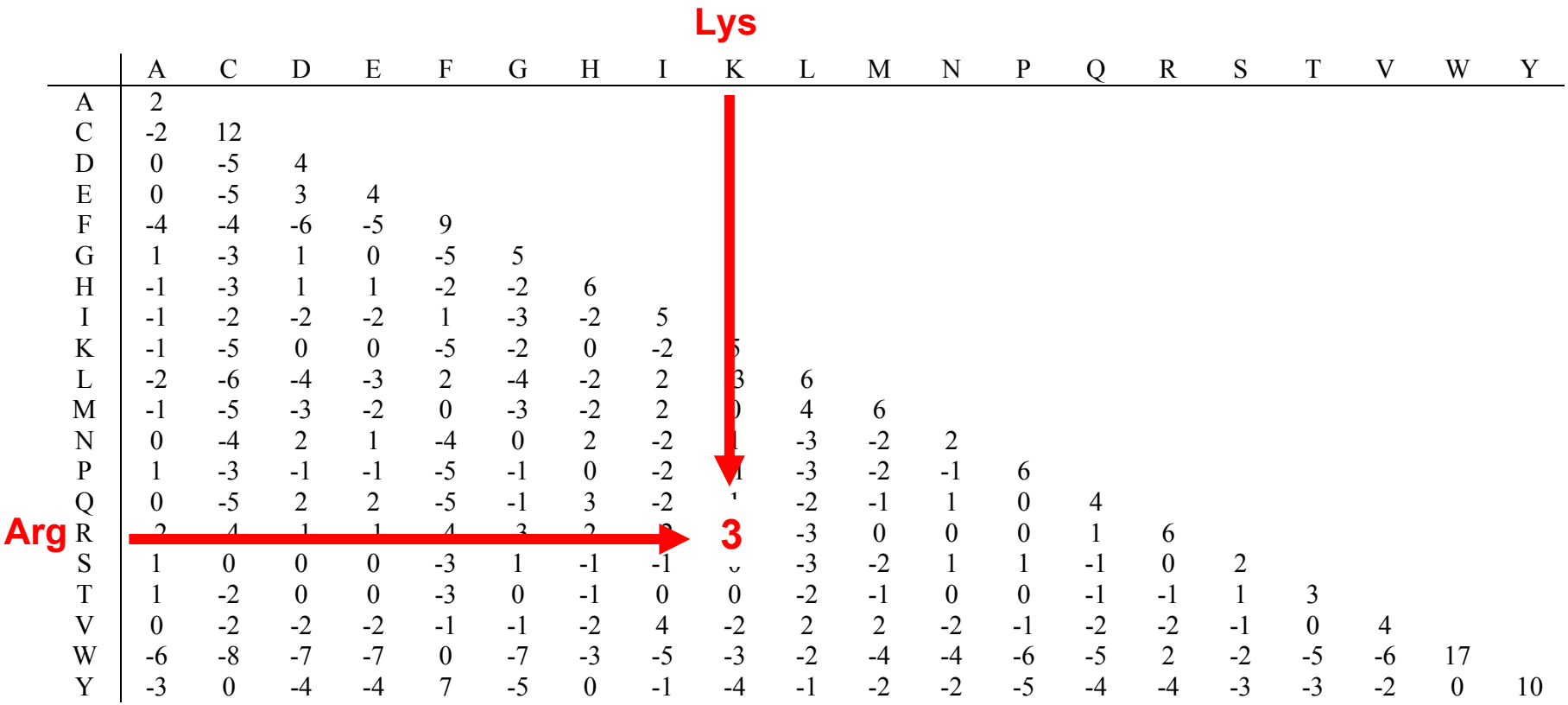
es. 1 Matrice di Sostituzione

Cys

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2																			
C	-2	12																		
D	0	-5	4																	
E	0	-5	3	4																
F	-4	-4	-6	-5	9															
G	1	-3	1	0	-5	5														
H	-1	-3	1	1	-2	-2	6													
I	-1	-2	-2	-2	1	-3	-2	5												
K	-1	-5	0	0	-5	-2	0	-2	5											
L	-2	-6	-4	-3	2	-4	-2	2	-3	6										
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6									
N	0	-4	2	1	-4	0	2	-2	1	-3	-2	2								
P	1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6							
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4						
R	-2	-4	-1	-1	-4	-3	2	-2	3	-3	0	0	0	1	6					
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2				
T	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3			
V	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4		
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10

La Cys non è facilmente rimpiazzabile

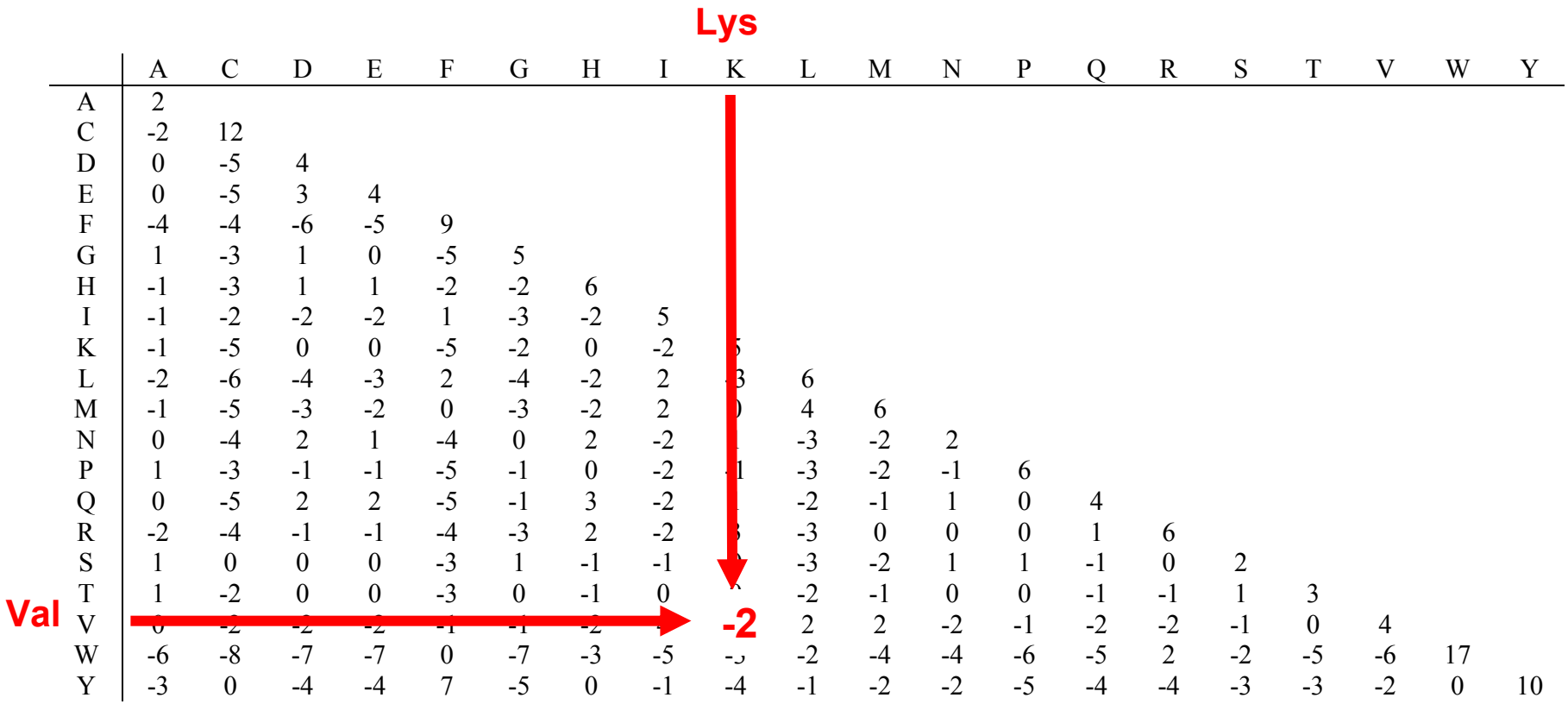
es. 1 Matrice di Sostituzione



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2																			
C	-2	12																		
D	0	-5	4																	
E	0	-5	3	4																
F	-4	-4	-6	-5	9															
G	1	-3	1	0	-5	5														
H	-1	-3	1	1	-2	-2	6													
I	-1	-2	-2	-2	1	-3	-2	5												
K	-1	-5	0	0	-5	-2	0	-2	5											
L	-2	-6	-4	-3	2	-4	-2	2	3	6										
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6									
N	0	-4	2	1	-4	0	2	-2	1	-3	-2	2								
P	1	-3	-1	-1	-5	-1	0	-2	1	-3	-2	-1	6							
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4						
R	-2	-4	-1	-1	-4	-3	-2	-3	3	-3	0	0	0	1	6					
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2				
T	1	-2	0	0	-3	0	-1	0	0	-2	-1	0	0	-1	-1	1	3			
V	0	-2	-2	-2	-1	-1	-2	4	-2	2	2	-2	-1	-2	-2	-1	0	4		
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10

Arg e Lys tendono a sostituirsi

es. 1 Matrice di Sostituzione



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2																			
C	-2	12																		
D	0	-5	4																	
E	0	-5	3	4																
F	-4	-4	-6	-5	9															
G	1	-3	1	0	-5	5														
H	-1	-3	1	1	-2	-2	6													
I	-1	-2	-2	-2	1	-3	-2	5												
K	-1	-5	0	0	-5	-2	0	-2	5											
L	-2	-6	-4	-3	2	-4	-2	2	-3	6										
M	-1	-5	-3	-2	0	-3	-2	2	0	4	6									
N	0	-4	2	1	-4	0	2	-2	1	-3	-2	2								
P	1	-3	-1	-1	-5	-1	0	-2	-1	-3	-2	-1	6							
Q	0	-5	2	2	-5	-1	3	-2	1	-2	-1	1	0	4						
R	-2	-4	-1	-1	-4	-3	2	-2	-3	-3	0	0	0	1	6					
S	1	0	0	0	-3	1	-1	-1	0	-3	-2	1	1	-1	0	2				
T	1	-2	0	0	-3	0	-1	0	-1	-2	-1	0	0	-1	-1	1	3			
V	-0	-2	-2	-2	-1	-1	-2	-1	-3	2	2	-2	-1	-2	-2	-1	0	4		
W	-6	-8	-7	-7	0	-7	-3	-5	-3	-2	-4	-4	-6	-5	2	-2	-5	-6	17	
Y	-3	0	-4	-4	7	-5	0	-1	-4	-1	-2	-2	-5	-4	-4	-3	-3	-2	0	10

Polari e non polari tendono a non sostituirsi

Allineamento

1) Punteggio per la corrispondenza di aa/basi
2) Penalizzazione di inserzioni / delezioni
3) Algoritmo che effettui l'allineamento
4) Misura della significativita' dell'allineamento

2. Penalizzazioni per inserzioni/delezioni

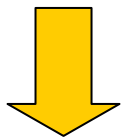
Le matrici di sostituzione sono state derivate da allineamenti che non presentavano inserzioni/delezioni. Queste vanno dunque trattate separatamente, su base empirica.

Nell'allineare due sequenze un algoritmo tenderebbe a massimizzare il punteggio (corrispondenza amminoacidi identici o simili) inserendo un gran numero di gap.

Questo modo di procedere rispecchia l'evoluzione?

2. Penalizzazioni per inserzioni/delezioni

In natura inserzioni e delezioni sono spesso letali



Dobbiamo quindi penalizzare inserzioni e delezioni. Cioè associare ad esse un punteggio negativo che si sottrae al punteggio totale per un allineamento

2. Penalizzazioni per inserzioni/delezioni

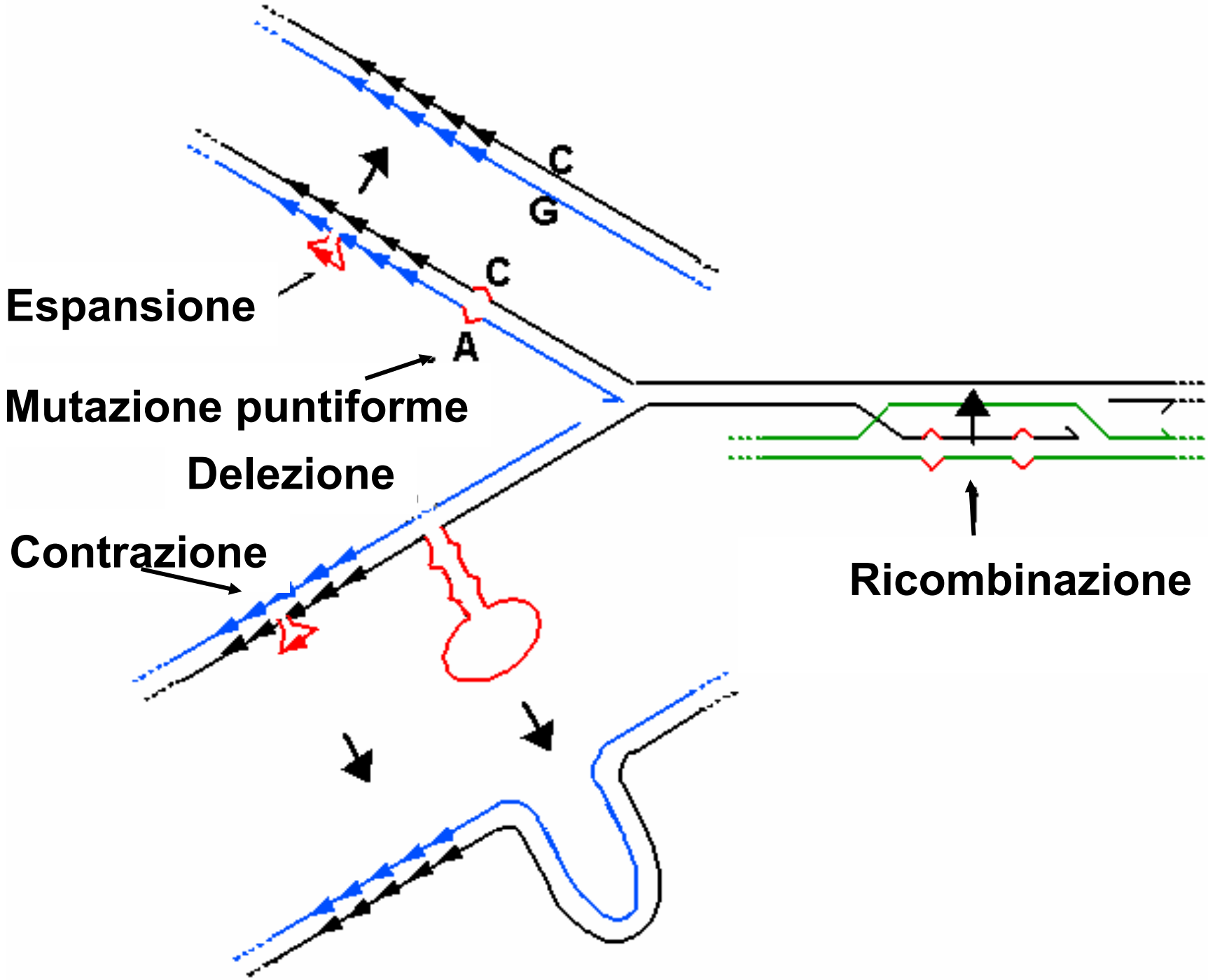
In natura la delezione di una serie di basi/amminoacidi contigui è un evento più probabile rispetto alla delezione indipendente dello stesso numero di basi/amminoacidi in posizioni non contigue

Distinguiamo l'inizio di un gap:

```
EGQTCA
AG-TCL
```

dall'estensione di un gap:

```
EGQQQTCA
AG---TCL
```


C
G
C
A
Espansione
Mutazione puntiforme
Delezione
Contrazione
Ricombinazione

2. Penalizzazioni per inserzioni/delezioni

In natura la delezione di una serie di basi/amminoacidi contigui è un evento più probabile rispetto alla delezione indipendente dello stesso numero di basi/amminoacidi in posizioni non contigue

Distinguiamo l'inizio di un gap:

```
EGQTCA
AG-TCL
```

dall'estensione di un gap:

```
EGQQQTCA
AG---TCL
```

2. Penalizzazioni per inserzioni/delezioni

In natura la delezione di una serie di basi/amminoacidi contigui è un evento più probabile rispetto alla delezione indipendente dello stesso numero di basi/amminoacidi in posizioni non contigue

Distinguiamo l'inizio di un gap:

```
EGQTCA
AG-TCL
```

dall'estensione di un gap:

```
EGQQQTCA
AG---TCL
```

Penalizziamo maggiormente l'inizio di un gap rispetto alla sua estensione

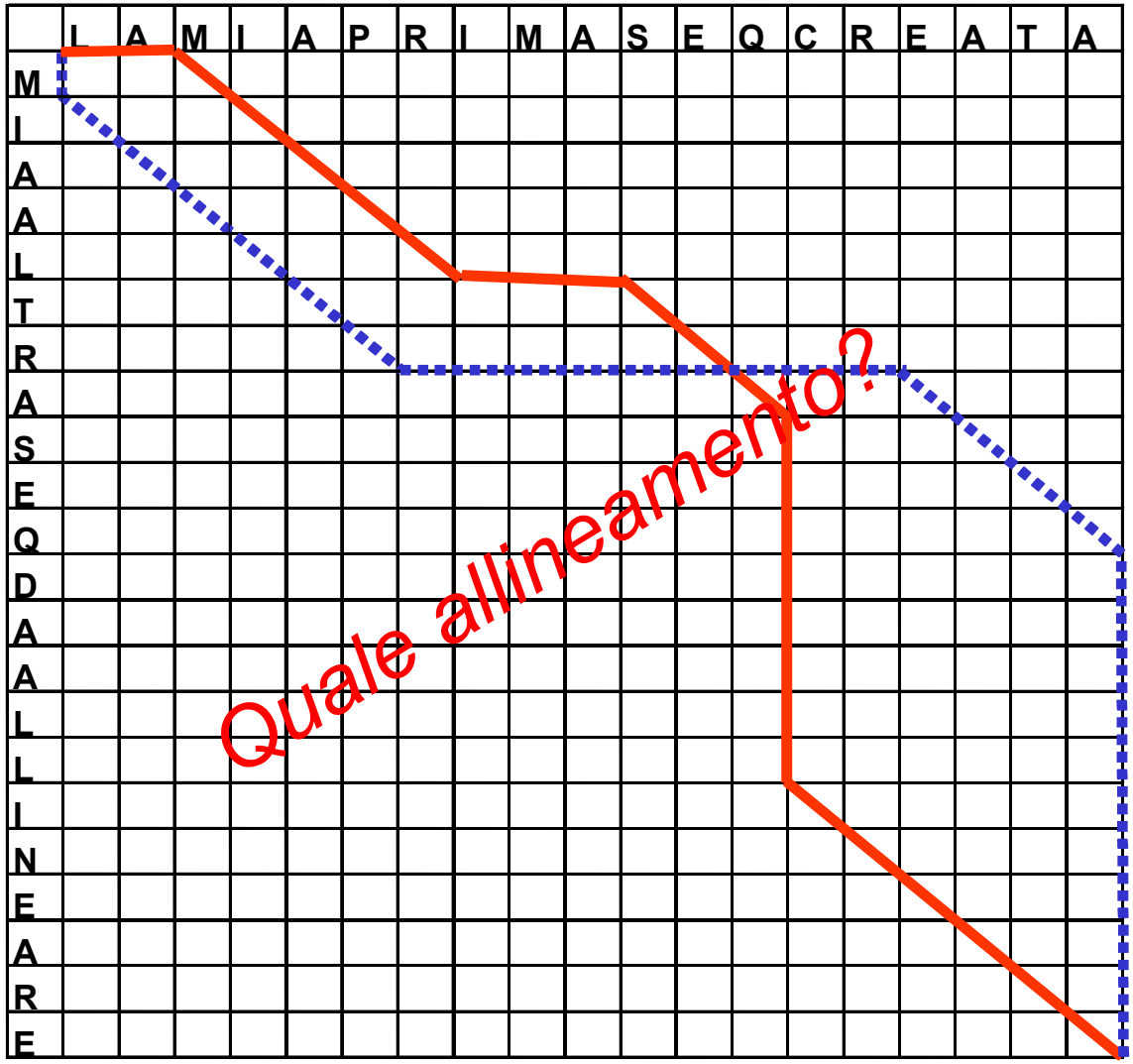
es.:

-11 inizio

-1 estensione

Allineamento

1) Punteggio per la corrispondenza di aa/basi
2) Penalizzazione di inserzioni / delezioni
3) Algoritmo che effettui l'allineamento
4) Misura della significativita' dell'allineamento



LAMIAPRIMASEQCREATA----------

-MIAALTR-------ASEQDAALLINEARE

LAMIAPRIMASEQ--------CREATA

--MIAAL---TRASEQDAALLINEARE

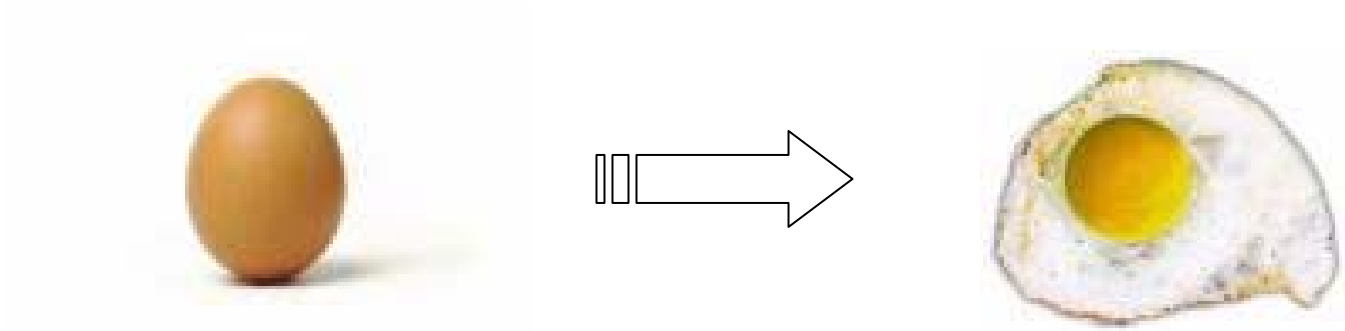
Qualche definizione…

dal nome di un matematico persiano al-Kharezmi (al-Khawarizmi) vissuto nel IX secolo

Algoritmo: Sequenza di istruzioni semplici (univocamente interpretabili) che permette di giungere ad un risultato in un numero determinato di passi.

Un esempio di algoritmo

1. *Prendere una padella antiaderente*
2. *versare un cucchiaio di olio extra vergine d'oliva*
3. *farlo scaldare per qualche secondo.*
4. *rompere l'uovo e adagiarlo al centro della padella*
5. *dare una spolverata di sale solo sul tuorlo*
6. *spegnere quando attorno all'albume si è formata una crosticina*



Qualche definizione…

dal nome di un matematico persiano al-Kharezmi (al-Khawarizmi) vissuto nel IX secolo

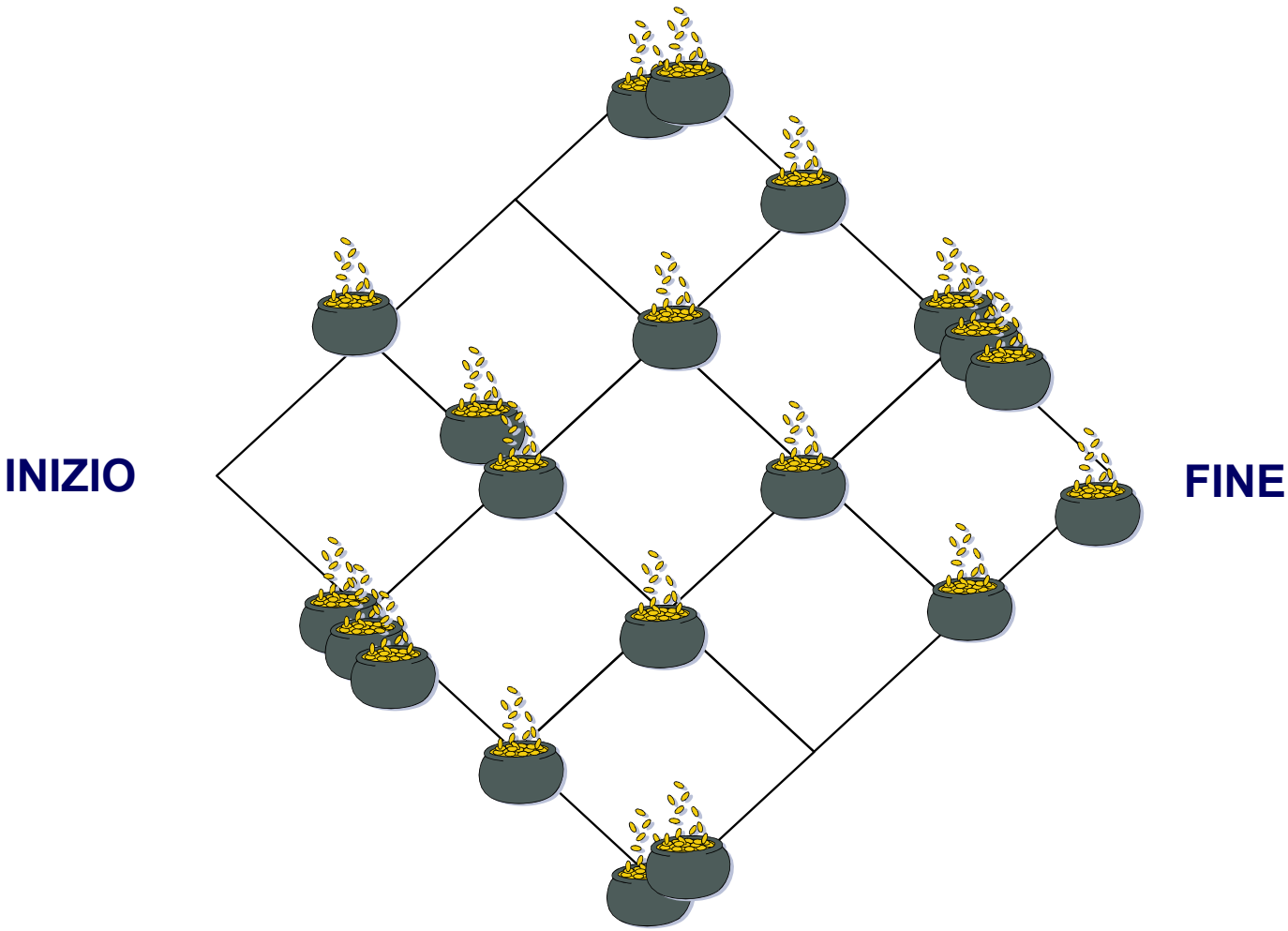
Algoritmo: Sequenza di istruzioni semplici (univocamente interpretabili) che permette di giungere ad un risultato in un numero determinato di passi.

Programma: descrizione di un algoritmo in uno specifico linguaggio di programmazione (C, fortran, perl).

Gli algoritmi di allineamenti esatti sono esempi di Programmazione Dinamica

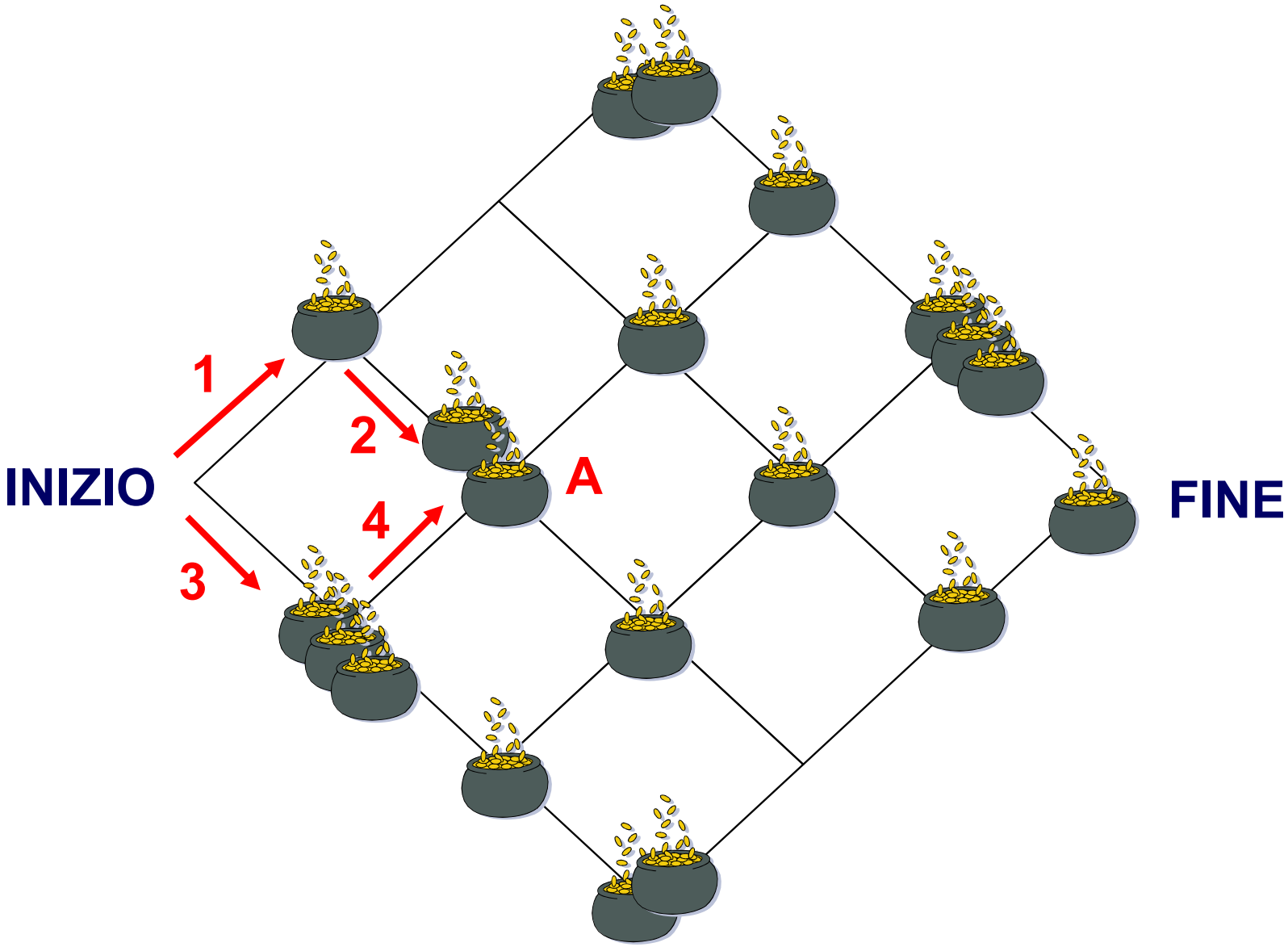
Dividere il problema in tanti sotto-problemi. Utilizzare la soluzione di ciascun sottoproblema per risolvere i successivi.

Il gioco delle pentole d'oro: regole



Andare da INIZIO a FINE senza mai passare per lo stesso punto e senza tornare indietro, raccogliendo il max numero di pentole d'oro

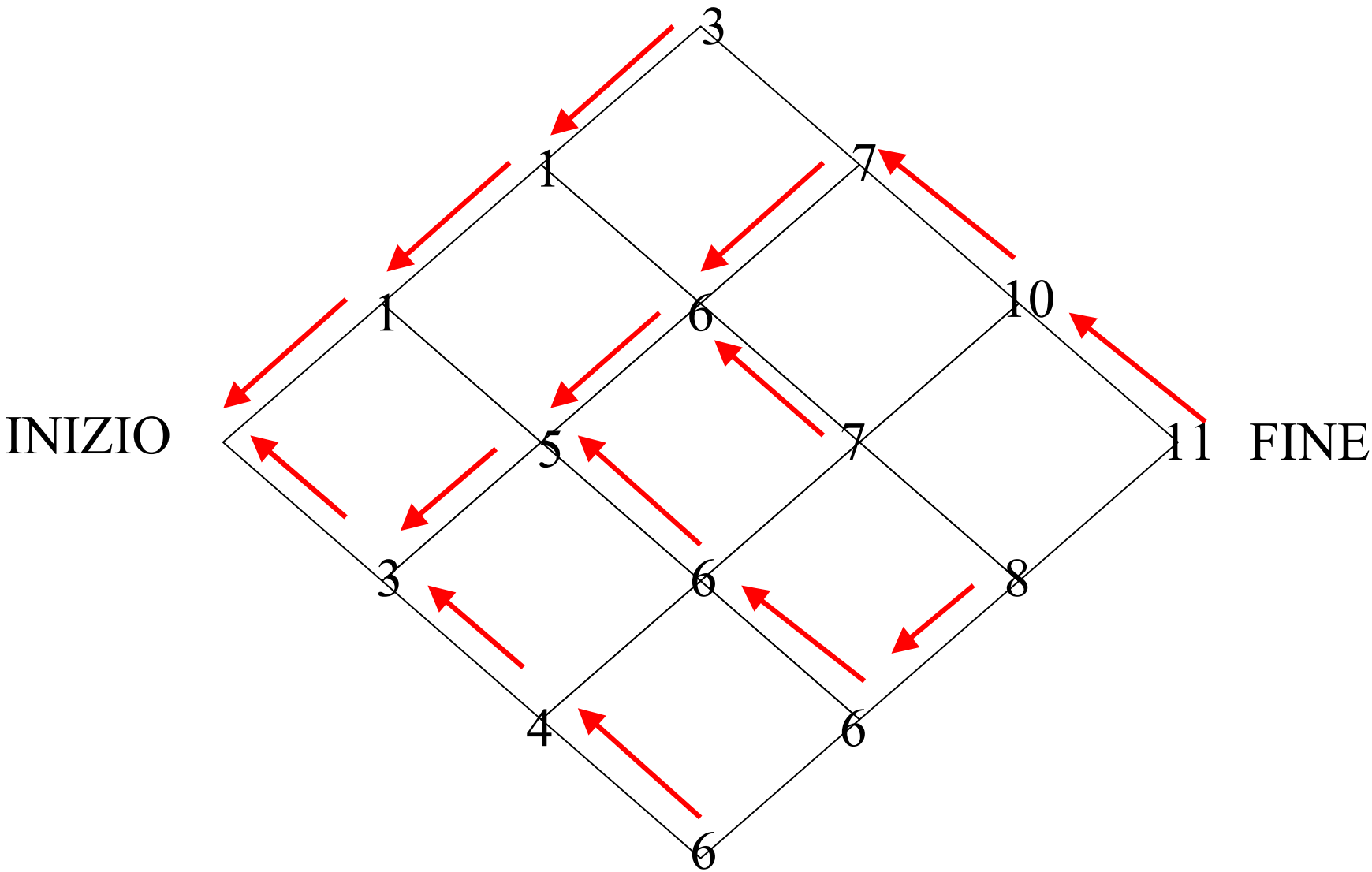
Il gioco delle pentole d'oro



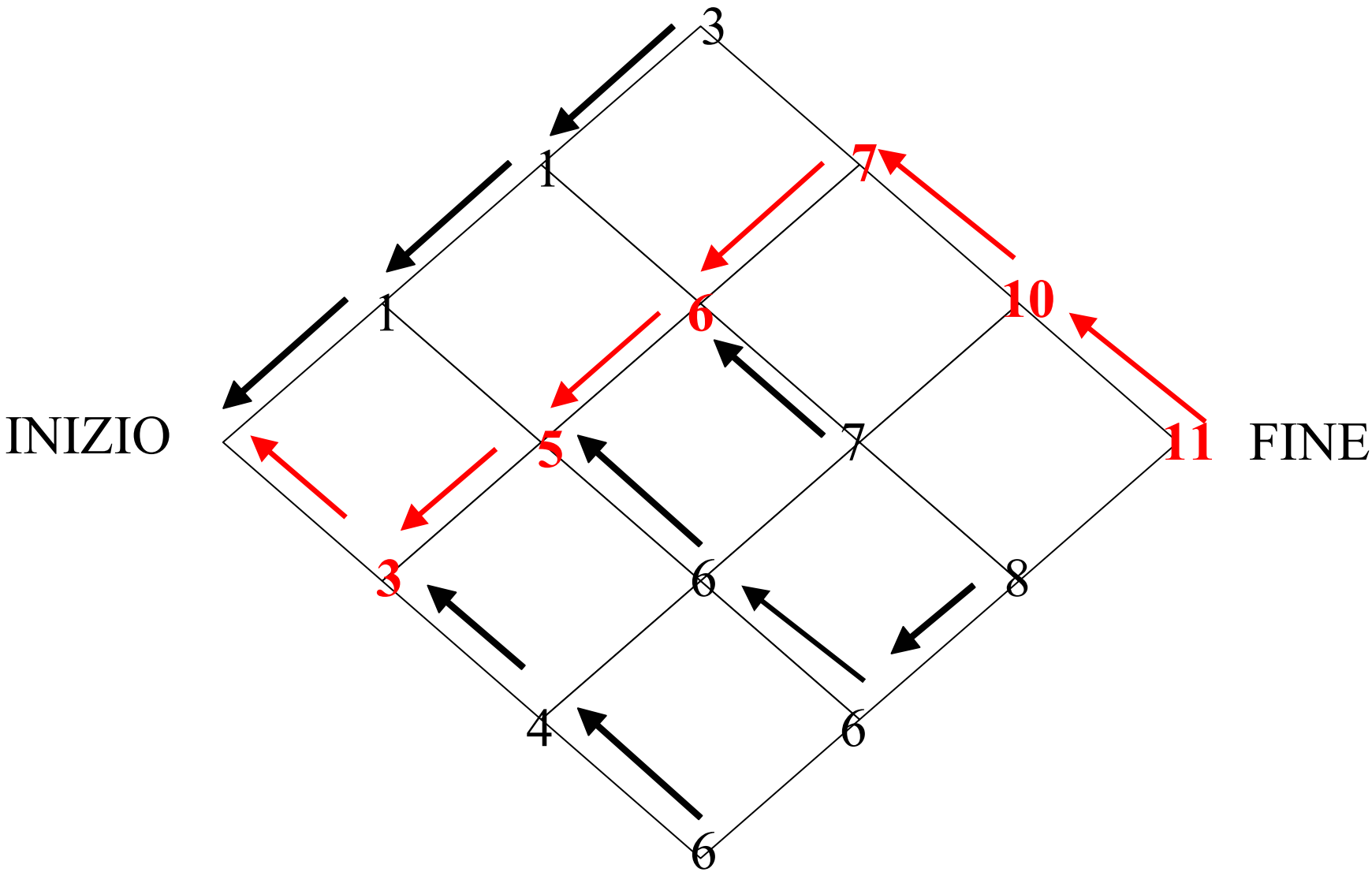
Il gioco delle pentole d'oro

La scelta del miglior percorso da A a FINE non dipende da come sono arrivata da INIZIO ad A.

Il gioco delle pentole d'oro: soluzione



Il gioco delle pentole d'oro: soluzione



Rispetto alle pentole d'oro, le matrici di similarita' contengono anche valori negativi

Un algoritmo di allineamento tendera' a saltarli e ad inserire un numero eccessivo di inserzioni e delezioni (*GAP*)

Sono necessarie le penalizzazioni di inserimento dei GAP (sia *di apertura*, che *di continuazione*)

Es. Come allineiamo le due sequenze:

HEAGAWGHEE

PAWHEAE

?

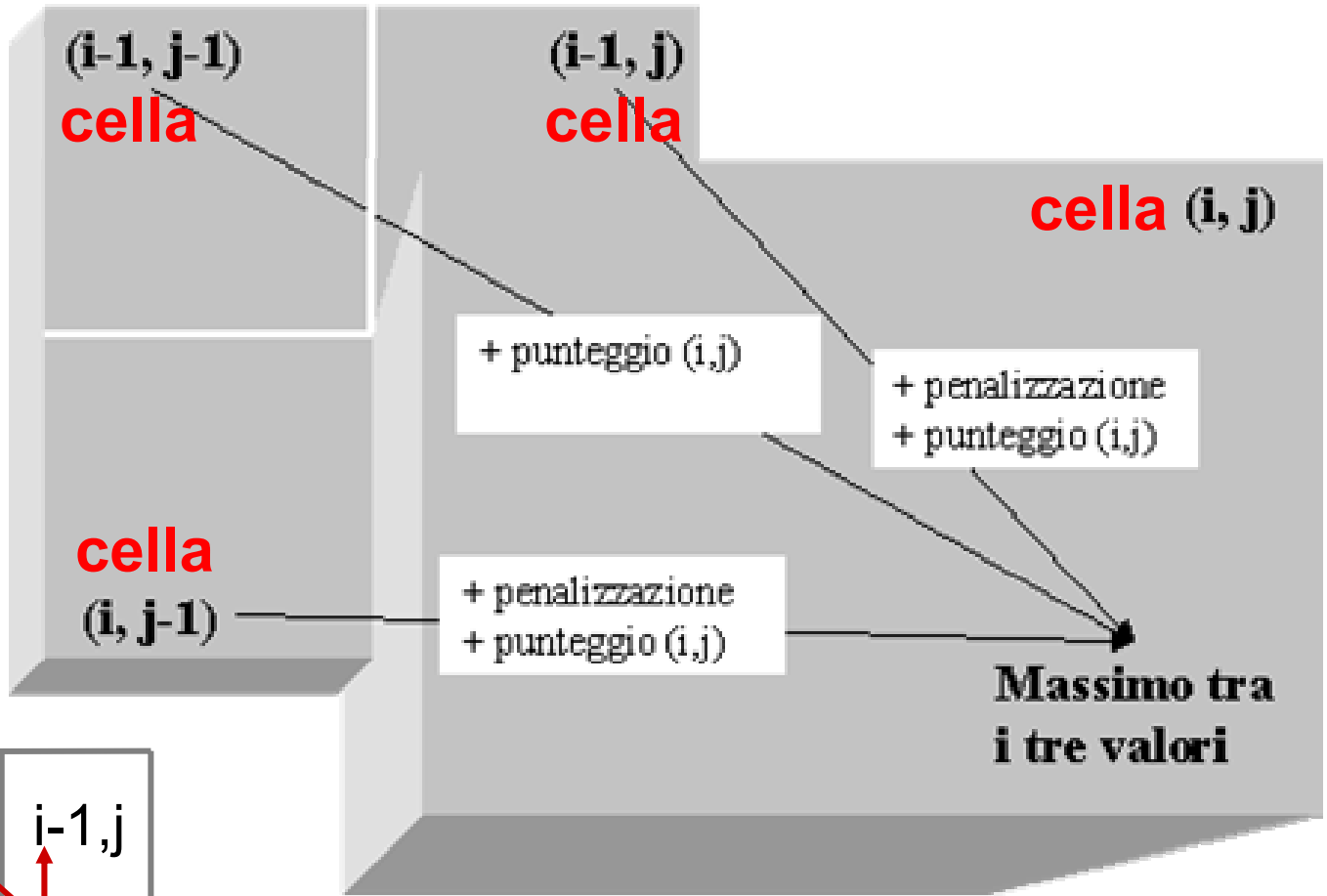
Un esempio di algoritmi di allineamento: matrici cumulative

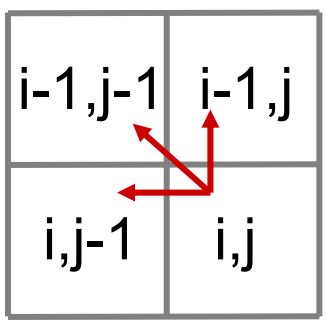
HEAGAWGHEE *vs* PAWHEAE

Step1: costruzione della matrice sito-specifica (valori da BLOSUM45)

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Algoritmo per l'allineamento: matrici cumulative

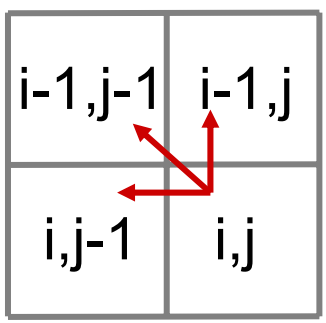


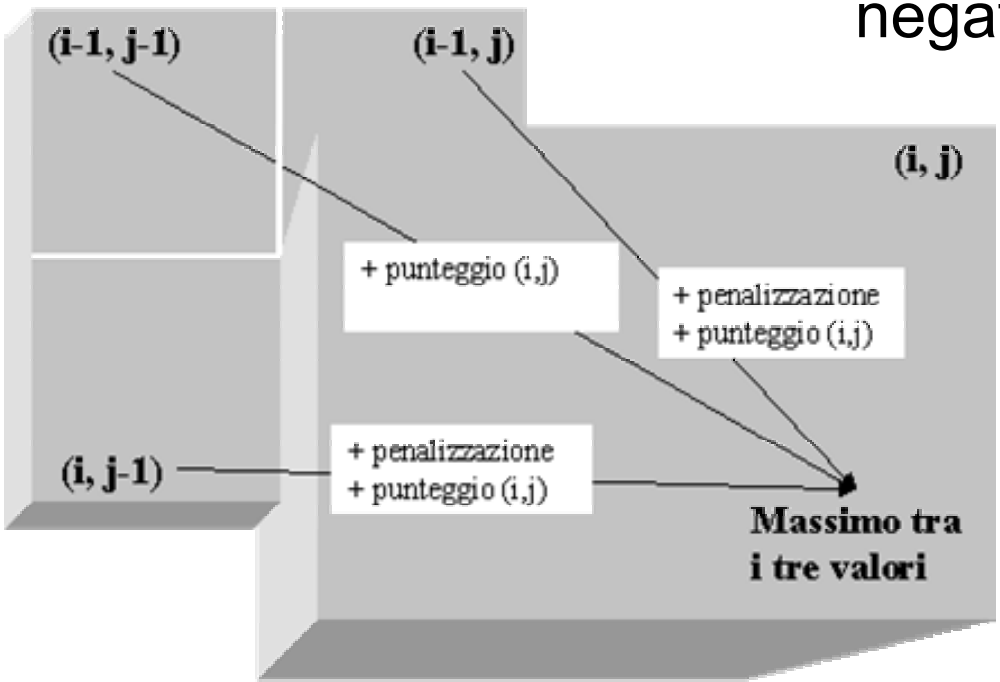


Step2: costruzione di una matrice *cumulativa*, in cui ogni elemento rappresenta il punteggio massimo ottenibile per arrivare dall'inizio fino a quel punto

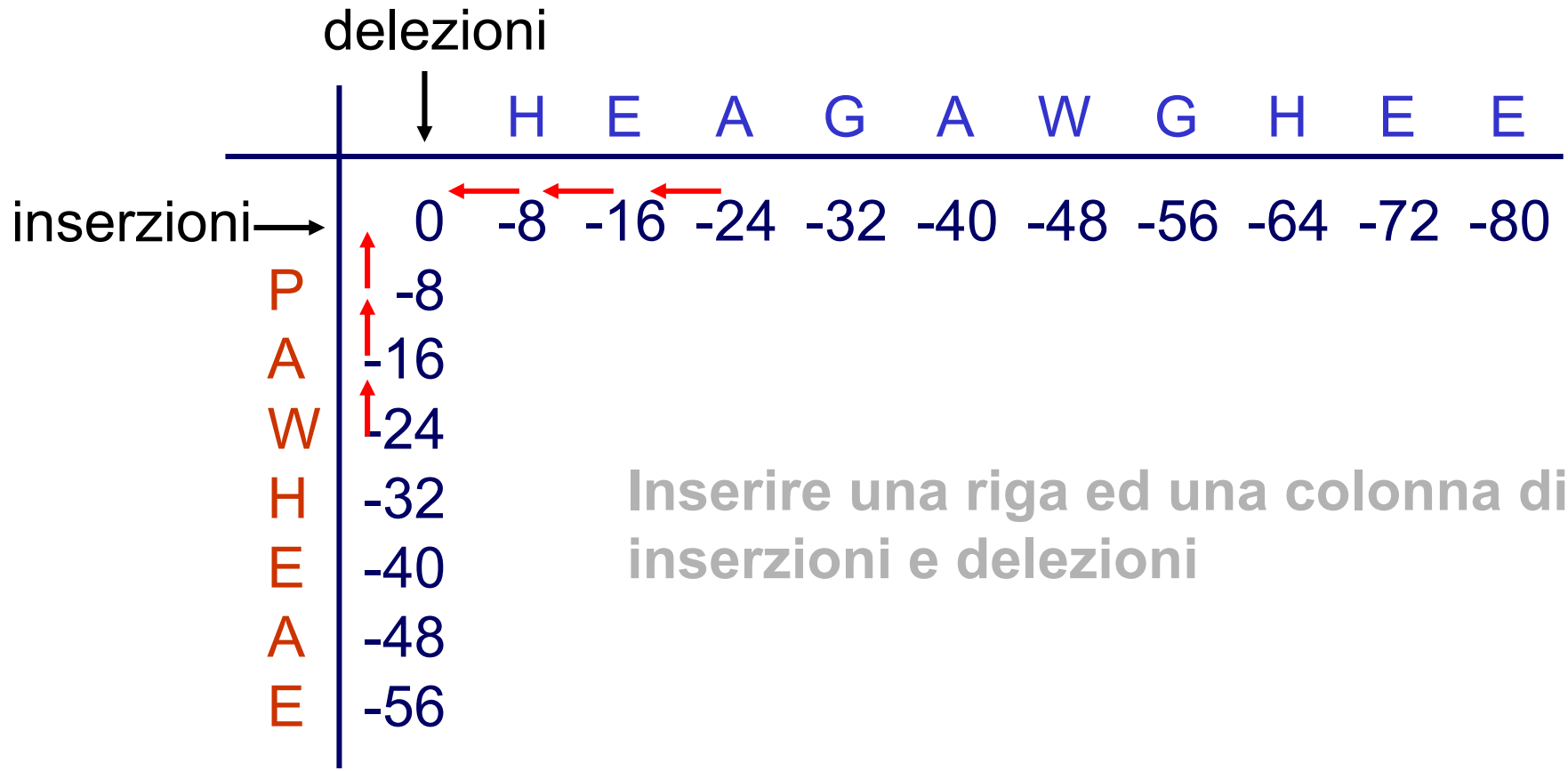
$$S_{i,j} = S_{i,j}\ (\text{sito-specifica}) + \max \begin{bmatrix} S_{i-1,j-1} \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{bmatrix}$$

dove g è la penalizzazione per in./del. e assume valori negativi





Matrice cumulativa



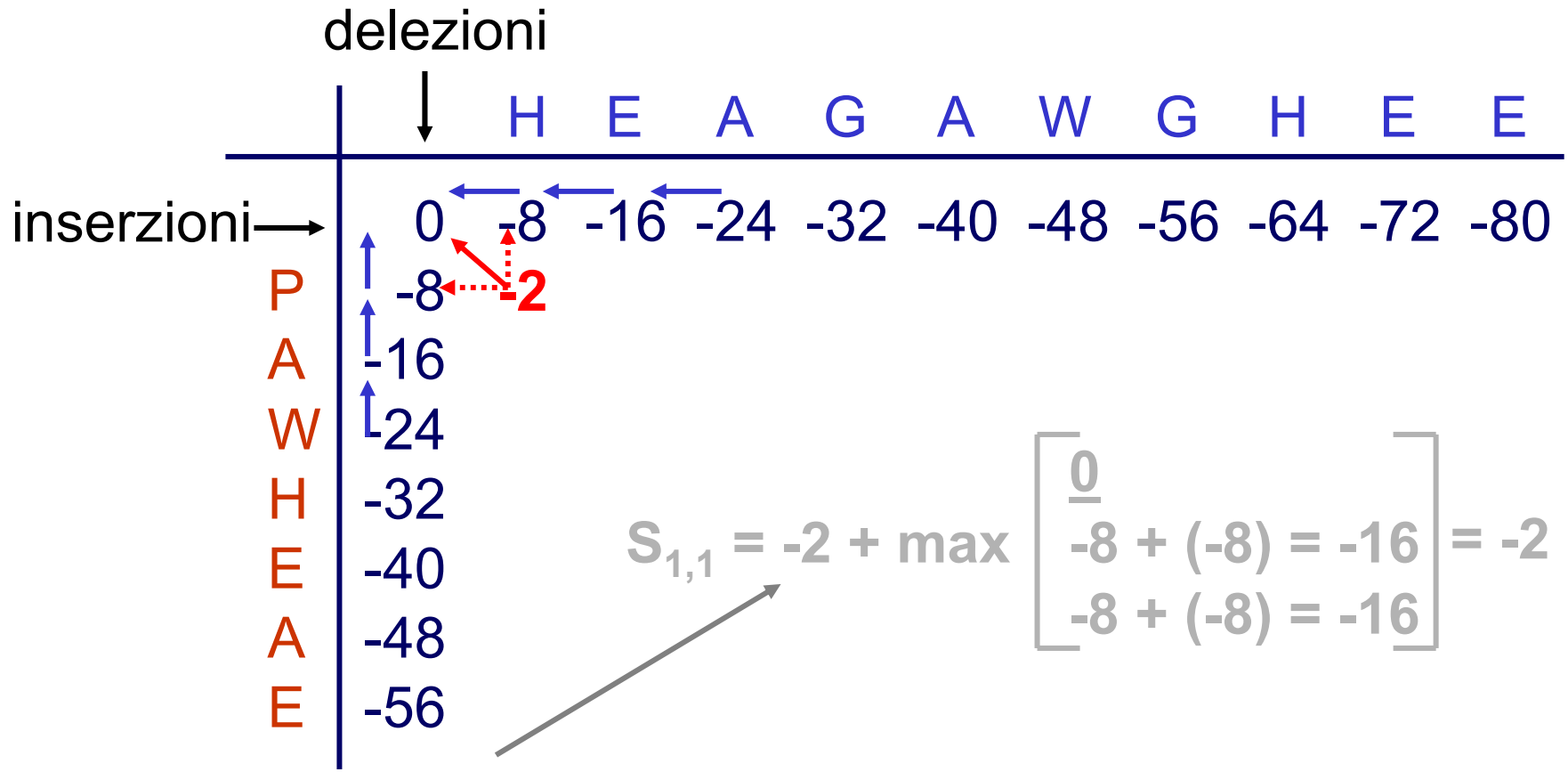
Inserire una riga ed una colonna di inserzioni e delezioni

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Penalizzazione GAP = -8

Matrice sito-specifica originaria

Matrice cumulativa



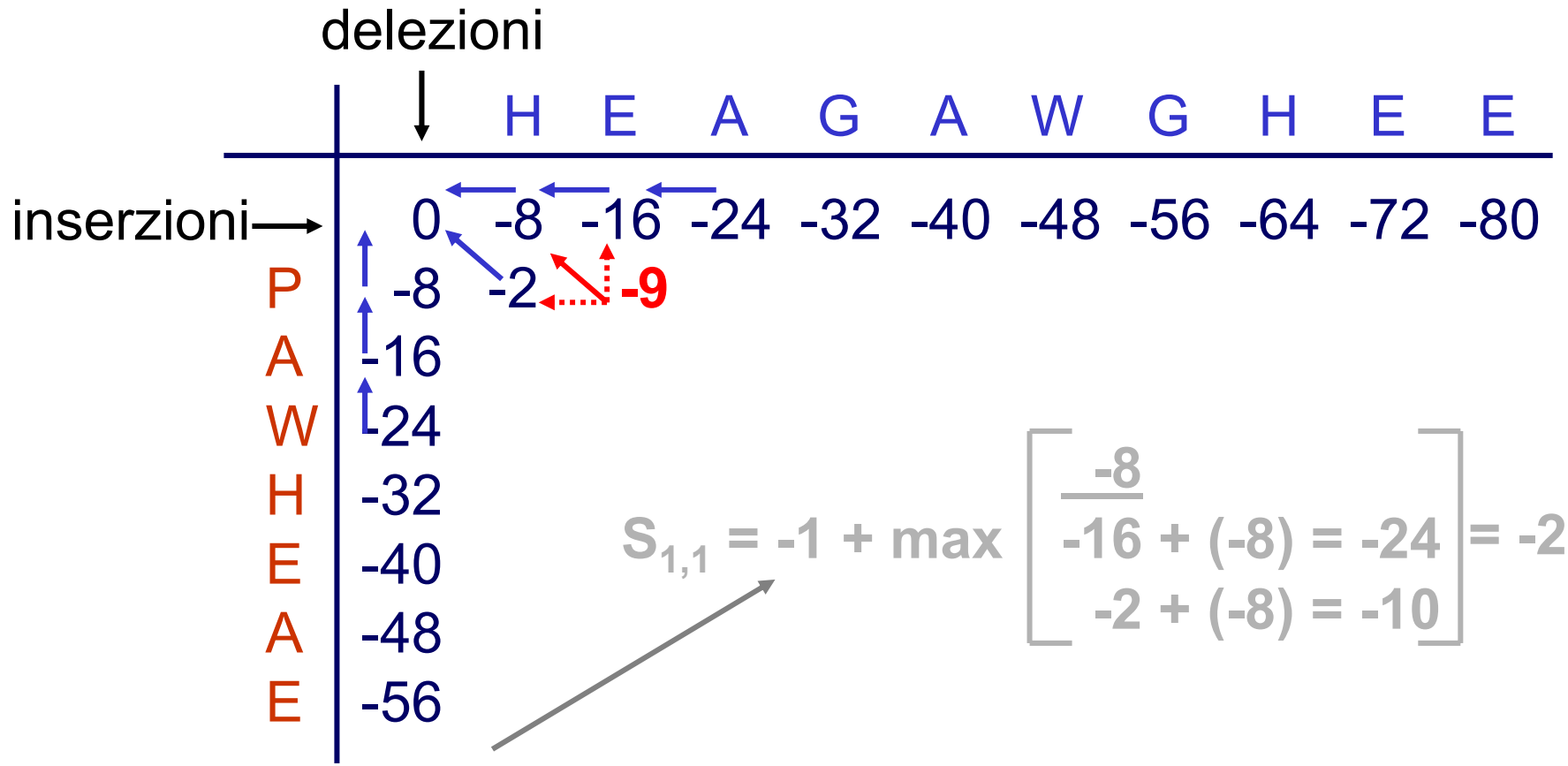
$$S_{1,1} = -2 + \max \begin{bmatrix} \underline{0} \\ -8 + (-8) = -16 \\ -8 + (-8) = -16 \end{bmatrix} = -2$$

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Penalizzazione GAP = -8

Matrice sito-specifica originaria

Matrice cumulativa



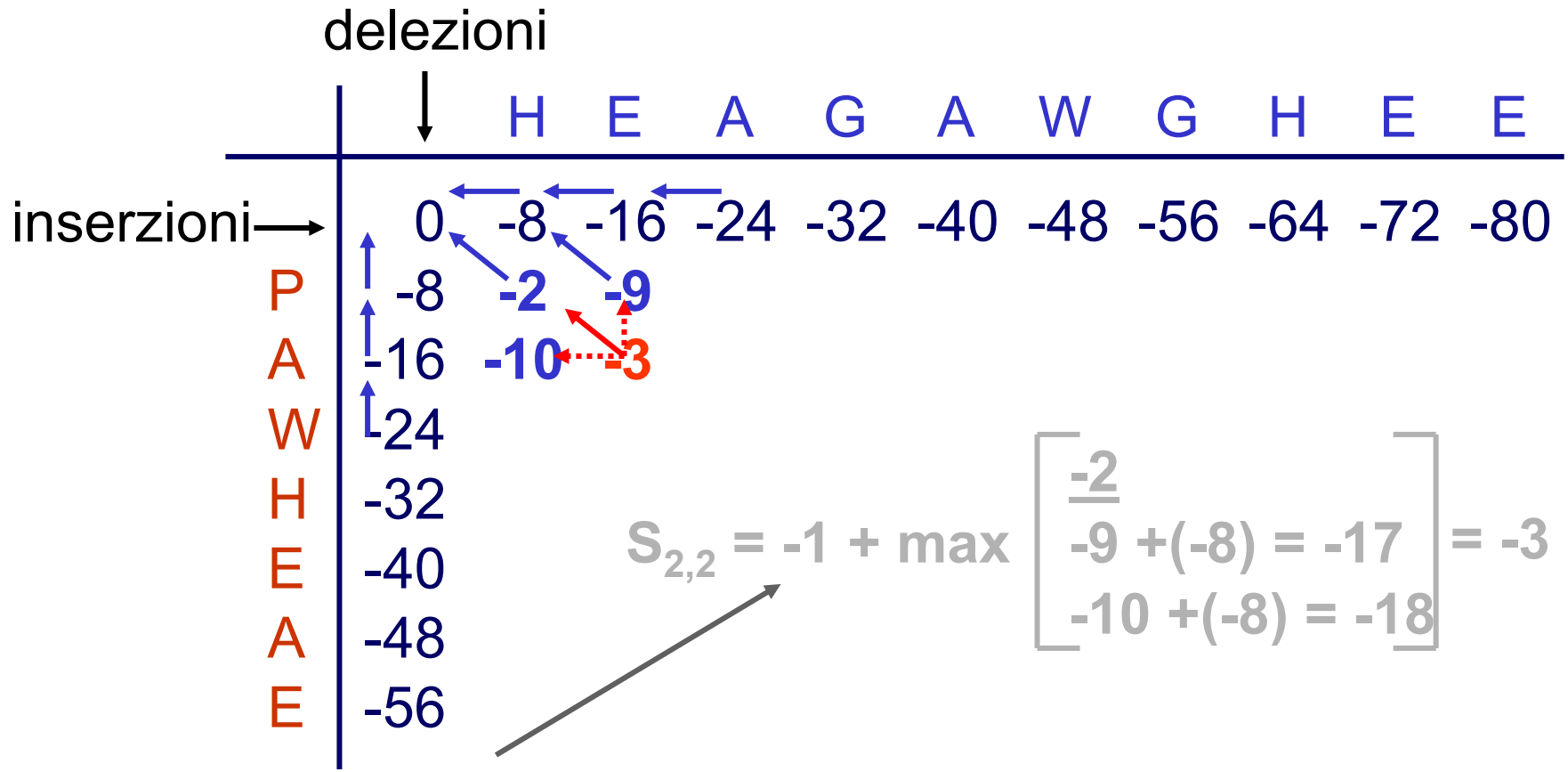
$$S_{1,1} = -1 + \max \begin{bmatrix} -8 \\ -16 + (-8) = -24 \\ -2 + (-8) = -10 \end{bmatrix} = -2$$

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Penalizzazione GAP = -8

Matrice sito-specifica originaria

Matrice cumulativa



$$S_{2,2} = -1 + \max \begin{bmatrix} \underline{-2} \\ -9 + (-8) = -17 \\ -10 + (-8) = -18 \end{bmatrix} = -3$$

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Penalizzazione GAP = -8

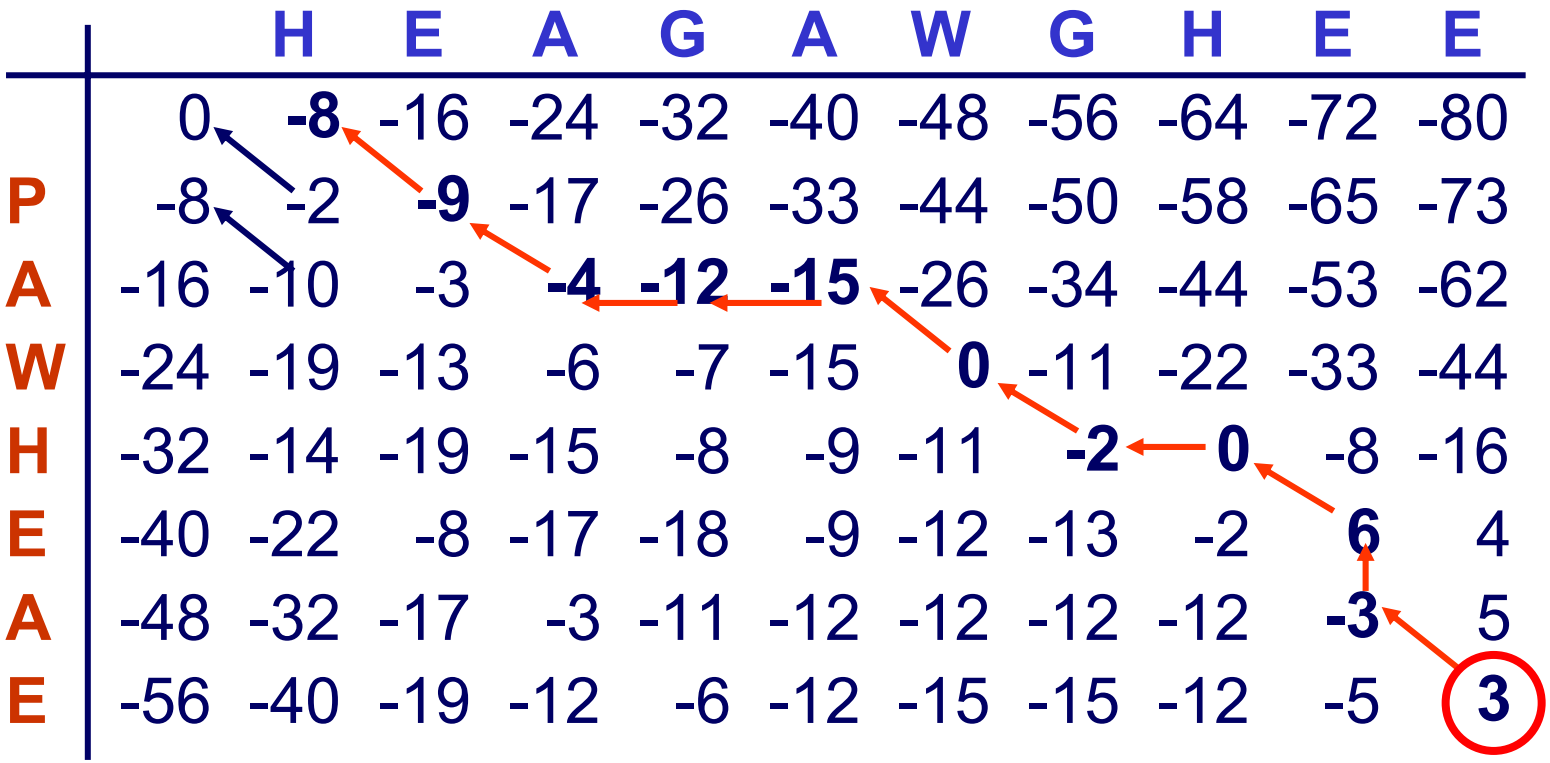
Matrice sito-specifica originaria

Matrice cumulativa

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-26	-33	-44	-50	-58	-65	-73
A	-16	-10	-3	-4	-12	-15	-26	-34	-44	-53	-62
W	-24	-19	-13	-6	-7	-15	0	-11	-22	-33	-44
H	-32	-14	-19	-15	-8	-9	-11	-2	0	-8	-16
E	-40	-22	-8	-17	-18	-9	-12	-13	-2	6	4
A	-48	-32	-17	-3	-11	-12	-12	-12	-12	-3	5
E	-56	-40	-19	-12	-6	-12	-15	-15	-12	-5	3

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-2	10	0	0
E	0	6	-1	-3	-1	-3	-3	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	-1	-3	-1	-3	-3	0	6	6

Matrice sito-specifica originaria



massimo punteggio ottenibile

HEAGAWGH–EE

–P––AW–HEAE

Step3: percorso a ritroso attraverso le celle che hanno permesso di ottenere i punteggi migliori ⇨ allineamento

		H	E	A	G	A	W	G	H	E	E
	0	**-8**	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	**-9**	-17	-26	-33	-44	-50	-58	-65	-73
A	-16	-10	-3	**-4**	**-12**	**-15**	-26	-34	-44	-53	-62
W	-24	-19	-13	-6	-7	-15	**0**	-11	-22	-33	-44
H	-32	-14	-19	-15	-8	-9	-11	**-2**	**0**	-8	-16
E	-40	-22	-8	-17	-18	-9	-12	-13	-2	**6**	4
A	-48	-32	-17	-3	-11	-12	-12	-12	-12	**-3**	5
E	-56	-40	-19	-12	-6	-12	-15	-15	-12	-5	**3**

HEAGAWGH–EE

–P––AW–HEAE

delezione

inserzione

Terza lezione.

1. Similarità e omologia.

2. Allineamenti di sequenze.

3. Sostituzioni e gap.

4. Algoritmo di allineamento