ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

1. Introduzione al confronto tra sequenze
- Spiegazione del concetto di omologia tra sequenze
- Similarità e funzione
- Motivazione del confronto tra sequenze

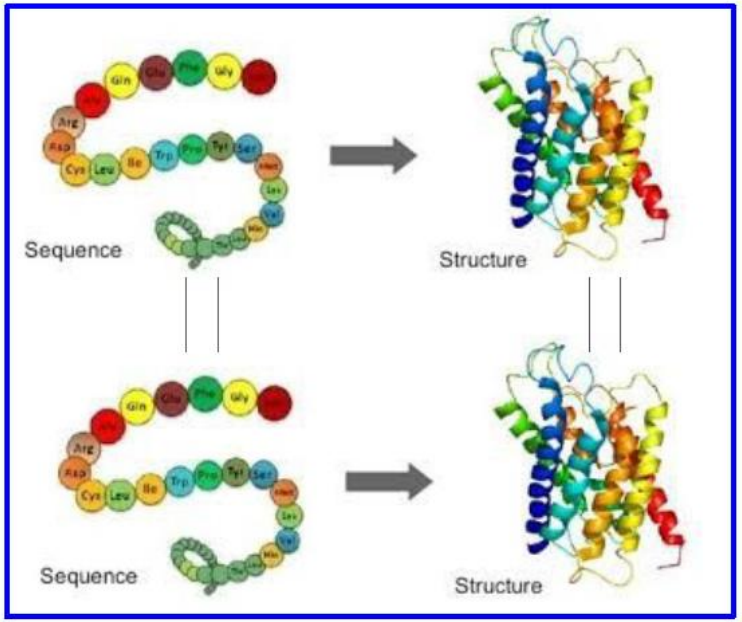
2. Introduzione all'algoritmo di allineamento a coppie di sequenza
- Matrice di sostituzione
- Descrizione delle matrici di sostituzione più comuni (ad esempio BLOSUM, PAM)

ALLINEAMENTO A COPPIE DI SEQUENZE

1. Introduzione al confronto tra sequenze

Uno dei metodi fondamentali dell'analisi bioinformatica di sequenze biologiche consiste nel confronto tra **due** o **più sequenze** diverse, limitandosi ad esaminare gli elementi che le compongono.

<u>Assunzione</u>: due sequenze identiche, ovvero formate dagli stessi elementi disposti nella stessa successione, avranno comportamento identico
nelle cellule viventi.



ALLINEAMENTO A COPPIE DI SEQUENZE

Il principio riportato nella slide precedente può essere esteso anche a sequenze non identiche:

● tanto più 2 o più sequenze sono simili tra loro quanto più sarà probabile che la loro funzione sia simile.

Limitandosi allo studio della sequenza si ignorano altri fattori spesso fondamentali: per esempio la funzione

❑ di una sequenza di DNA potrà dipendere anche dal fatto che sia metilata un meno
❑ o di una proteina potrà dipendere dalla presenza o meno di modificazioni post traduzionali come acetilazione o fosforilazione di determinati residui.

ALLINEAMENTO A COPPIE DI SEQUENZE

Tuttavia i risultati di fondamentale importanza possono essere ottenuti semplicemente dall'analisi di sequenza come testimoniano da migliaia di articoli scientifici che si basano su di essa.

Il confronto tra sequenze biologiche e quindi oggi pratica comune nella ricerca biomolecolare moderna, dalla biologia molecolare, alla genetica alla biochimica.

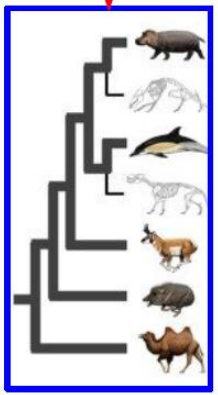
Tutte le sequenze che vengono analizzate oggi sono il risultato di centinaia di milioni di anni di evoluzione molecolare. quindi nella quasi totalità dei casi di similarità significativa tra due o più sequenze non sono frutto del caso, bensì di legami evolutivi tra le sequenze stesse.

ALLINEAMENTO A COPPIE DI SEQUENZE

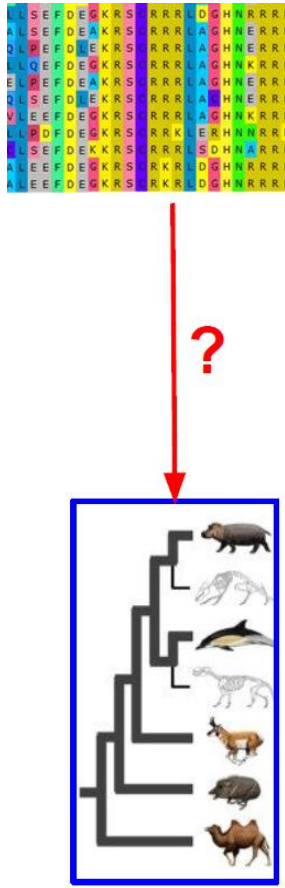
Due sequenze proteiche identiche per l'80% dei residui che le compongono evidentemente discendono da una sequenza antenata comune.

Nel corso del tempo ciascuna di esse ha accumulato variazioni, mantenendo però inalterata l'80% della sequenza antenata. In questo caso, con elevata probabilità, le funzioni fondamentali della sequenza antenata sono rimaste inalterate nelle sequenze che da essa discendono.

Viceversa, confrontando due sequenze, possiamo cercare di stabilire se la similarità che riscontriamo può essere indicativa di una storia evolutiva comune, oppure no.



SIMILARITÀ e OMOLOGIA tra SEQUENZE

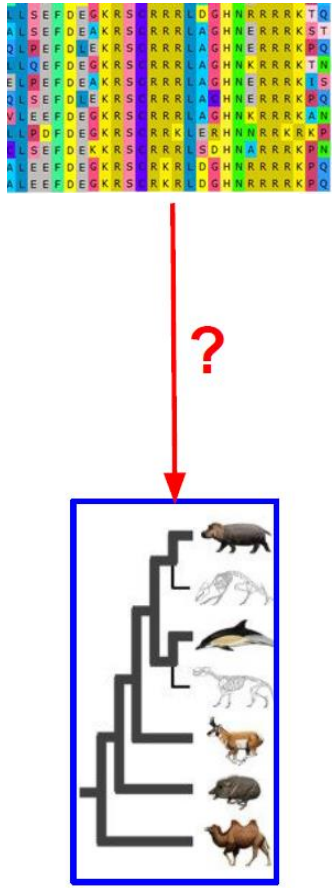


Il concetto di similarità e omologia tra sequenze biologiche sono termini chiave in bioinformatica e biologia molecolare, ma spesso possono causare confusione. Ecco le differenze principali tra i due:

Similarità: La similarità tra due sequenze biologiche (che possono essere di DNA, RNA o proteine) si riferisce al grado in cui le sequenze sono identiche o simili tra loro quando allineate. Questa similarità può essere quantificata in percentuale o con altri punteggi che considerano identità (base o aminoacido identico nelle due sequenze), conservazioni (sostituzioni accettabili che non cambiano la funzione della proteina, ad esempio) e differenze.

La similarità non implica necessariamente una relazione evolutiva comune; può semplicemente indicare che due sequenze hanno tratti comuni.

SIMILARITÀ e OMOLOGIA tra SEQUENZE



Omologia: L'omologia tra sequenze biologiche indica che le sequenze hanno un antenato comune. In altre parole, due sequenze omologhe sono derivate dalla stessa sequenza ancestrale attraverso il processo di evoluzione. Le omologie possono essere di due tipi principali: ortologia e paralogia.

Le sequenze **ortologhe** sono quelle che si sono divise a seguito di un evento di speciazione (e quindi si trovano in specie diverse ma svolgono funzioni simili), mentre le sequenze **paraloghe** sono il risultato di duplicazioni di geni all'interno dello stesso organismo (e possono avere sviluppato funzioni diverse).

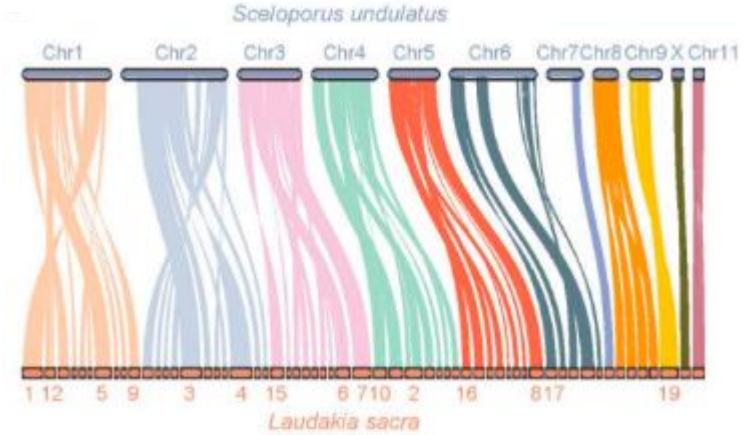
L'identificazione dell'omologia è fondamentale per comprendere la funzione dei geni e delle proteine e per studiare l'evoluzione dei genomi.

SIMILARITÀ e OMOLOGIA tra SEQUENZE

In pratica, quando si confrontano due sequenze, si può iniziare valutando la loro similarità attraverso vari metodi di allineamento di sequenze.

Se le sequenze sono significativamente simili, si può ipotizzare che siano omologhe, ma la similarità da sola non prova l'omologia.

Per confermare l'omologia, sono necessarie ulteriori analisi, come la costruzione di alberi filogenetici o l'esame del contesto genetico (ad esempio, la conservazione dell'ordine dei geni), per stabilire l'effettiva relazione evolutiva.



ALLINEAMENTO A COPPIE DI SEQUENZE

Esempio di output dell'allineamento tra proteina SSH di uomo e di topo

```
SSH_UOMO        -MLLLARCLLLVLVSSLLVCSGLACGPGRGFGKRRHPKKLTPLAYKQFIPNVAEKTLGAS
SSH_TOPO        MLLLLARCFLVILASSLLVCPGLACGPGRGFGKRRHPKKLTPLAYKQFIPNVAEKTLGAS
                 :******:*::*.******.****************************************

SSH_UOMO        GRYEGKISRNSERFKELTPNYNPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPG
SSH_TOPO        GRYEGKITRNSERFKELTPNYNPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPG
                *******:****************************************************

SSH_UOMO        VKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRSKYGMLARLAVEAGFDWVYYESKA
SSH_TOPO        VKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRSKYGMLARLAVEAGFDWVYYESKA
                ************************************************************

SSH_UOMO        HIHCSVKAENSVAAKSGGCFPGSATVHLEQGGTKLVKDLSPGDRVLAADDQGRLLYSDFL
SSH_TOPO        HIHCSVKAENSVAAKSGGCFPGSATVHLEQGGTKLVKDLRPGDRVLAADDQGRLLYSDFL
                ************************************ ********************

SSH_UOMO        TFLDRDDGAKKVFYVIETREPRERLLLTAAHLLFVAPHNDSATGEPEASSGSGPPSGGAL
SSH_TOPO        TFLDRDEGAKKVFYVIETLEPRERLLLTAAHLLFVAPHND-----------SGPTPG---
                ******:*********** ********************          ***..*

SSH_UOMO        GPRALFASRVRPGQRVYVVAERDGDRRLLPAAVHSVTLSEEAAGAYAPLTAQGTILINRV
SSH_TOPO        -PSALFASRVRPGQRVYVVAERGGDRRLLPAAVHSVTLREEEAGAYAPLTAHGTILINRV
                 * ******************.*************** ** *********:********

SSH_UOMO        LASCYAVIEEHSWAHRAFAPFRLAHALLAALAPARTDRGGDSGGGDRGGGGGRVALTAPG
SSH_TOPO        LASCYAVIEEHSWAHRAFAPFRLAHALLAALAPARTD----------GGGGGSIP-AAQS
                ************************************          ***** :. :* .

SSH_UOMO        AADAPGAGATAGIHWYSQLLYQIGTWLLDSEALHPLGMAVKSS
SSH_TOPO        ATEARGAEPTAGIHWYSQLLYHIGTWLLDSETMHPLGMAVKSS
                *::* ** .************:*********::**********
```

* = identità

: e . = sostituzioni conservative

\- = gap (inserzione/delezione)

ALLINEAMENTO A COPPIE DI SEQUENZE

** = identità*

: e . = sostituzioni conservative

- = gap (inserzione/delezione)

Esempio di output dell'allineamento tra proteina SSH di uomo e di drosofila

```
CLUSTAL W (1.81) multiple sequence alignment


NP_000184       ------------------------------------------------------MLLLARC
CG4637-RA       MDNHSSVPWASAASVTCLSLDAKCHSSSSSSSSKSAASSISAIPQEETQTMRHIAHTQRC
                                                                     :    **

NP_000184       L--------LLVLVSSLLVCSGLACGPGRGFGKRRHPKKLTPLAYKQFIPNVAEKTLGAS
CG4637-RA       LSRLTSLVALLLIVLPMVFSPAHSCGPGRGLG-RHRARNLYPLVLKQTIPNLSEYTNSAS
                *        **::* .::.... :******:* *:::.::* **. ** ***::* * .**

NP_000184       GRYEGKISRNSERFKELTPNYNPDIIFKDEENTGADRLMTQRCKDKLNALAISVMNQWPG
CG4637-RA       GPLEGVIRRDSPKFKDLVPNYNRDILFRDEEGTGADRLMSKRCKEKLNVLAYSVMNEWPG
                *  ** * *:* :**:*.**** **:*:***.*******::***:***.** ****:***

NP_000184       VKLRVTEGWDEDGHHSEESLHYEGRAVDITTSDRDRSKYGMLARLAVEAGFDWVYYESKA
CG4637-RA       IRLLVTESWDEDYHHGQESLHYEGRAVTIATSDRDQSKYGMLARLAVEAGFDWVSYVSRR
                ::* ***.**** **.:********** *:*****:****************** * *:

NP_000184       HIHCSVKAENSVAAKSGGCFPGSATVHLEQGGTKLVKDLSPGDRVLAADDQGRLLYSDFL
CG4637-RA       HIYCSVKSDSSISSHVHGCFTPESTALLESGVRKPLGELSIGDRVLSMTANGQAVYSEVI
                **:****::.*::::  ***. .:*. **.*  * : :** *****:   :*: :**:.:

NP_000184       TFLDRDDGAKKVFYVIETREPRERLLLTAAHLLFVAPHNDSATGEPEASSGSGPPSGGAL
CG4637-RA       LFMDRN---------LEQMQNFVQLHTDGGAVLTVTPAHLVSVWQPESQK----------
                 *:**:         :*  :   :*   .. :* *:* :  :. :**:..
```

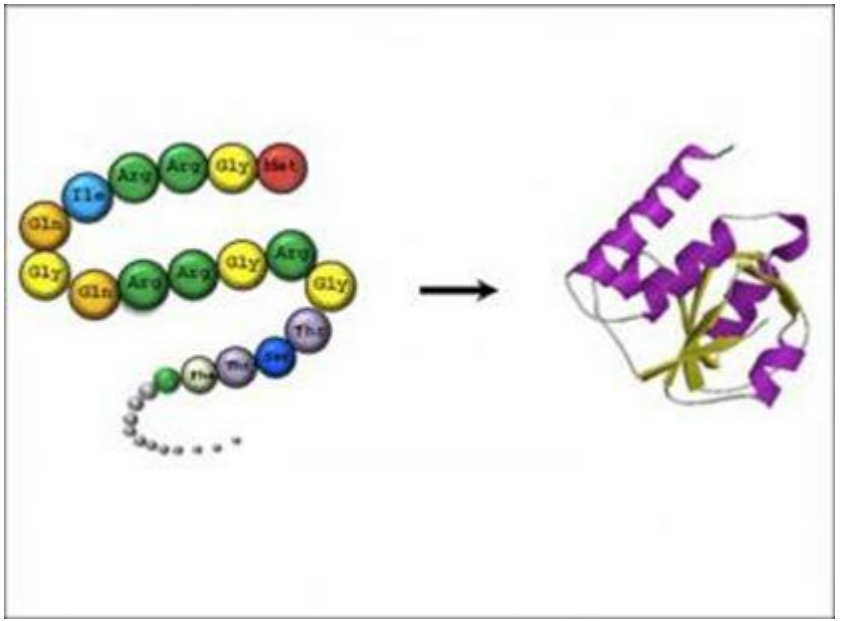
NP_00184 = proteina SHH di H. sapiens
CG4637-RA = proteina HH di D. melanogaster

Qui notiamo maggiori differenze rispetto al confronto con la proteina SSH di topo

ALLINEAMENTO A COPPIE DI SEQUENZE

- Similarità e funzione

- In altre parole, sequenze identiche hanno funzione identica, più simili sono due sequenze tra loro e più alta la probabilità che svolgano la stessa funzione.
- Geni con sequenza simile codificano per proteine simili.
- Questo ci permette di bypassare il problema della struttura delle proteine e continuare a lavorare a livello di sequenza.

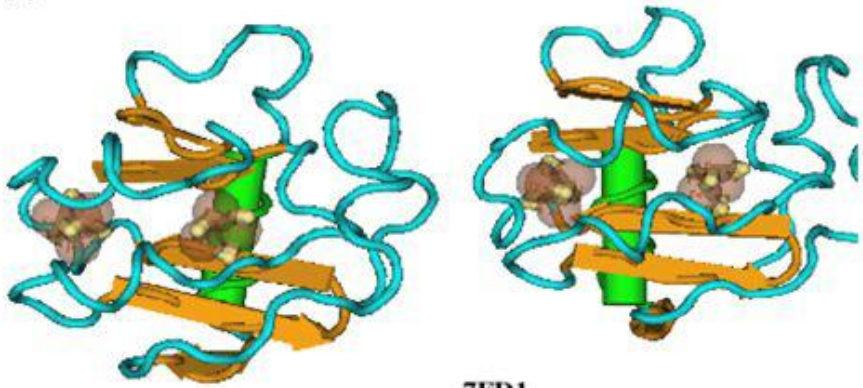


ALLINEAMENTO A COPPIE DI SEQUENZE

- Similarità e funzione

```
  1 .AFVVTDNCIKCKYTDCVEV.CPVDCFYEGPNFLVIHPDECID...CALC 45
    |: ||      .. |: ||. || |.:      :|    .:|.||    :. .
  1 XAYKVT....LVTPTGNVEFQCPDDVY.....ILDAAEEEGIDLPYSCRA 41
             .          .          .          .          .
 46 EPECPAQAIFSEDEVPEDMQEFIQLNAELAEVWPNITEKKDPLPDAEDWD 95
    :. :...: :..:.: :| |.|:: :.::.|.|. :|     |:.|..   .
 42 GSCSSCAGKLKTGSLNQDDQSFLD.DDQIDEGWV.LTCAAYPVSDVTIET 89
             .
 96 GVKGKLQHLER 106
      |:.|
 90 HKKEELTA... 97
```

(A)



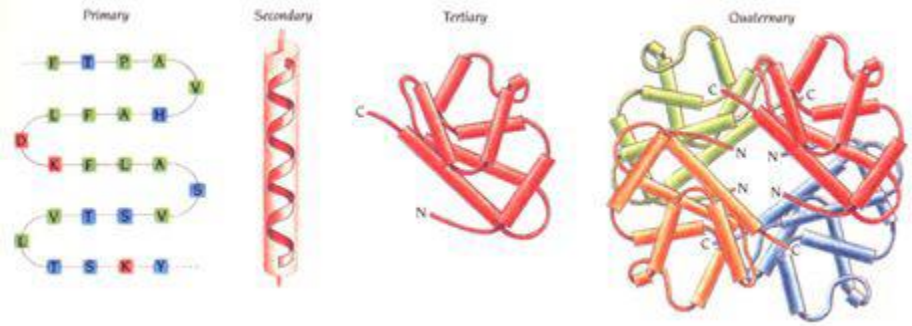
1A70 7FD1

Ferrodoxina di spinacio vs Ferrodoxina di batterio. La struttura terziaria appare più simile di quanto l'allineamento di sequenza farebbe supporre.

ALLINEAMENTO A COPPIE DI SEQUENZE

- Similarità e funzione

- Quanto possono essere diverse due proteine che mantengono la stessa struttura?
- Si è visto che proteine con un'identità di solo il 25% continuano ad avere la stessa struttura. In generale si è visto che una similarità >40% implica funzione identica o molto simile.
- Esistono eccezioni. Se le differenze riguardano posizioni critiche (ad es. sito catalitico di un enzima) la funzione può differire notevolmente.
- I domini funzionali tendono ad essere più conservati, mentre le regioni strutturali meno.
- Ad esempio, il core idrofobico tende ad essere meno conservato degli amminoacidi che stanno sulla superficie di una proteine globulare.
- Questo ci consente, quando abbiamo una proteina di funzione ignota, di confrontarla con le sequenze di proteine con funzione nota e di dedurre quindi la sua funzione in base alla similarità di sequenza.



ALLINEAMENTO A COPPIE DI SEQUENZE

- Similarità e funzione

- Anche se esistono milioni di sequenze di proteine diverse, i fold possibili sono in numero molto minore.
- Mentre è molto difficile "predire" il fold di una proteina a partire dalla sequenza, è molto più semplice "riconoscerlo" in base alla sua similarità con proteine di cui già conosciamo la struttura.
- O almeno possiamo vedere se esistono all'interno della proteina domini conservati in altre proteine di cui già conosciamo la struttura e la funzione.
- Se riusciamo ad individuare uno o più domini noti, possiamo assegnare una funzione alla proteina.

ALLINEAMENTO A COPPIE DI SEQUENZE

- Motivazione del confronto tra sequenze

Il confronto tra sequenze è un'attività importante nella ricerca biologica e nella biologia molecolare. Perché permette di:

1. Identificazione di **omologie** tra sequenze: il confronto tra sequenze consente di identificare sequenze simili o omologhe. Questo può aiutare a comprendere la funzione delle proteine e dei geni, nonché a stabilire relazioni tra specie diverse.

1. Ricerca di **nuove** proteine: il confronto tra sequenze può essere utilizzato per identificare proteine simili a quelle già conosciute. Ciò può aiutare a scoprire nuove proteine e nuove funzioni biologiche.

1. Analisi **filogenetica**: il confronto tra sequenze può essere utilizzato per costruire alberi filogenetici, che sono diagrammi che mostrano le relazioni evolutive tra le specie.

1. Scoperta di nuovi **farmaci**: il confronto tra sequenze può essere utilizzato per identificare proteine bersaglio che possono essere utilizzate per lo sviluppo di nuovi farmaci.

COSA E' UNA MATRICE

Una matrice è un concetto matematico che rappresenta una tabella di numeri, simboli o espressioni, disposti in righe e colonne. Per esemplificare, ecco come appare una matrice:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

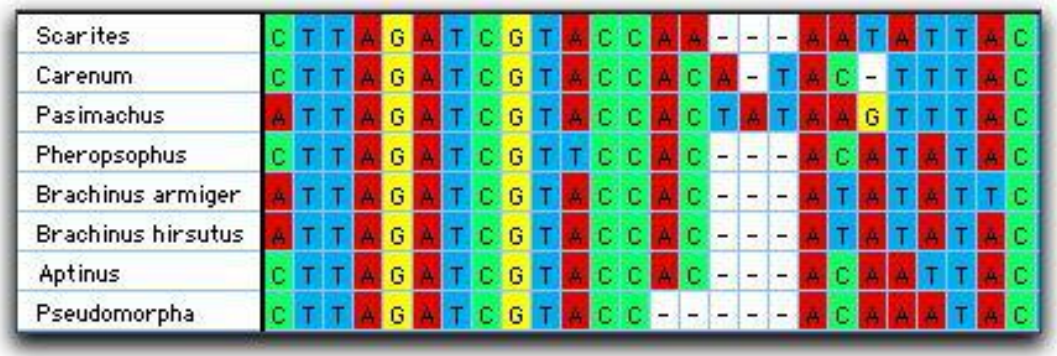
In questa rappresentazione, **aij** rappresenta l'elemento situato all'incrocio della i-esima riga e della j-esima colonna. Le dimensioni di una matrice sono date dal numero di righe (m) e colonne (n) e si indicano come m × n.

Le matrici sono strumenti fondamentali in vari campi della matematica e delle scienze applicate. Esse possono essere utilizzate per rappresentare sistemi di dati.

ALLINEAMENTO DI SEQUENZE

Il metodo più utilizzato oggi per confrontare sequenze di nucleotidi o aminoacidi consiste nel loro allineamento.

Allineare le sequenze consiste nel rappresentarle, una sotto l'altra (una riga per sequenza), in colonna andando a vedere i residui che le compongono in modo da formare una matrice. Ciascuna colonna della matrice conterrà un residuo di ognuna delle sequenze da allineare oppure uno spazio rappresentato da un trattino (gap).

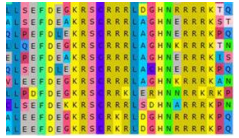


ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

	1	2	3	4	5	6	7	8	9	10	11	12	13
sequenza 1	A	C	T	T	G	T	C	T	T	A	T	G	C
sequenza 2	A	C	T	_	G	A	_	T	T	A	_	_	C

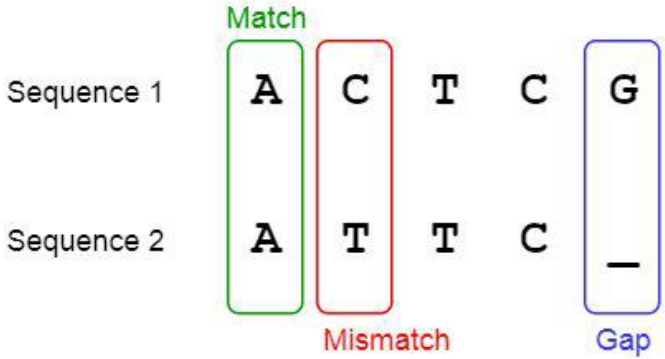
Nell'esempio sopra le due sequenze sono di lunghezza diversa: è possibile infatti allineare sequenze di lunghezza qualsiasi.
In 8 posizioni (1-3, 5, 8-10, 13) la sequenza è la stessa. Per questi 8 nucleotidi si ipotizza che rispetto alla sequenza antenata non siano avvenute variazioni, ossia i nucleotidi sono conservati.
In posizione 6 vi è un mismatch, quindi potrebbe esserci stata almeno una sostituzione rispetto alla sequenza antenata se non addirittura due.
In posizione (4, 7, 11-12) vi sono delle gap. Possiamo quindi ipotizzare che vi siano state inserzioni o delezioni rispetto alla sequenza antenata.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE



Le mutazioni delle sequenze nucleotidiche o proteiche sono alla base dell'evoluzione.
Esse sono dovute principalmente a:

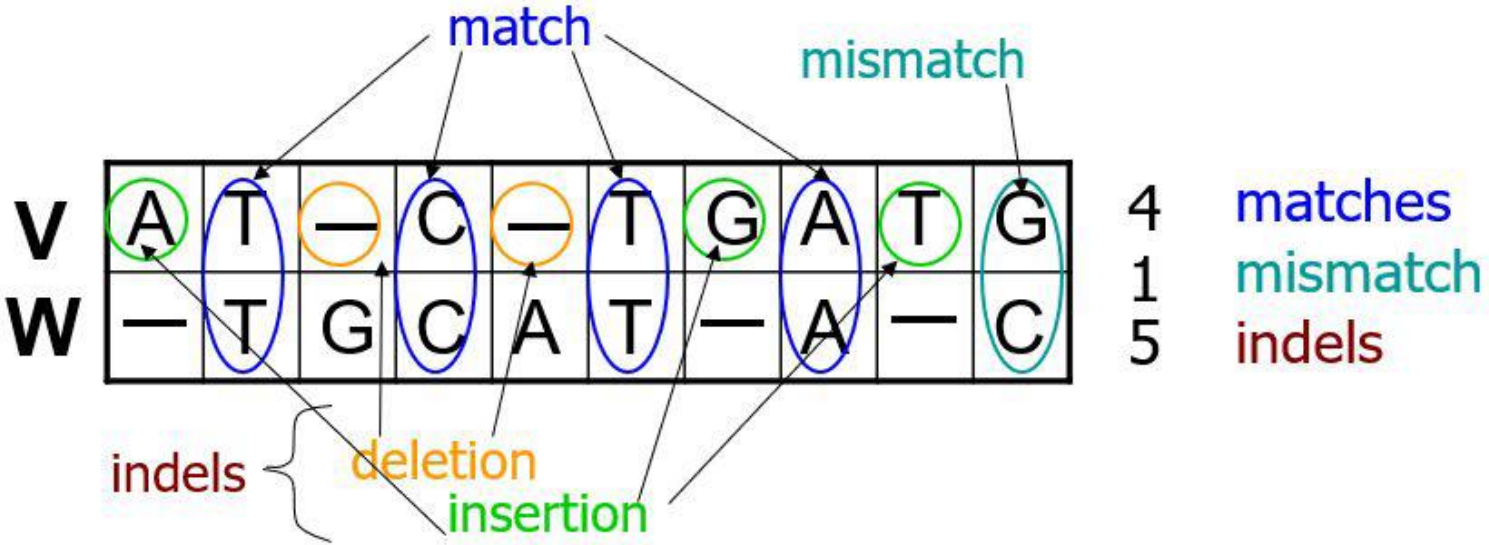
1. sostituzioni: ACA ⇒ AGA
2. cancellazioni: CTTG ⇒ CTG
3. inserzioni: AAA ⇒ AATA
4. inversioni: AAGAG ⇒ AAAGG
5. ricombinazioni: dovuti a scambi di porzioni di DNA



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

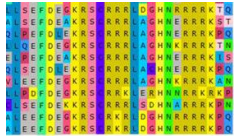
V = ATCTGATG n = 8

W = TGCATAC m = 7

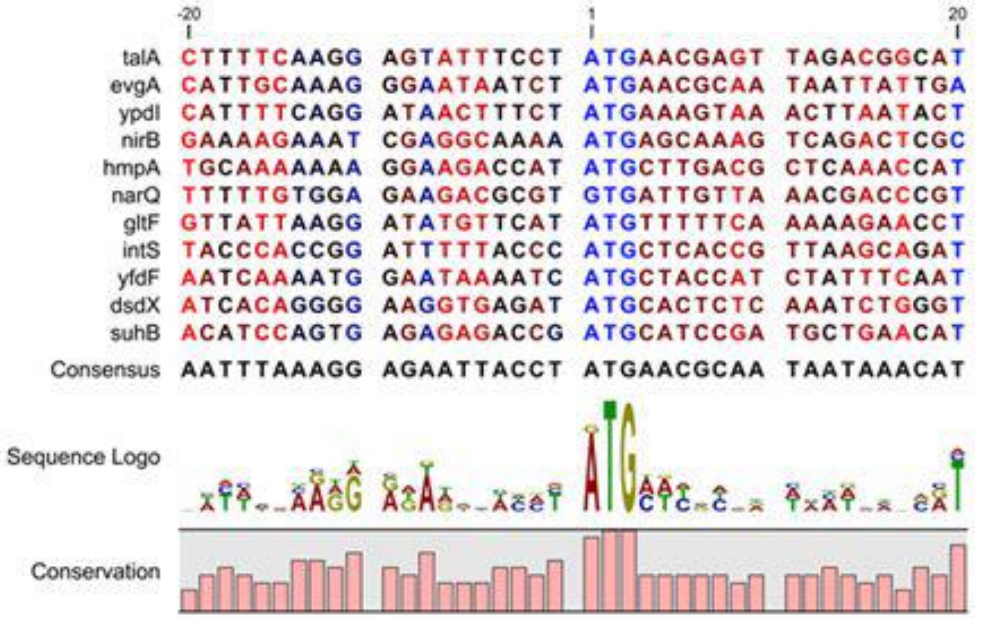


4 matches
1 mismatch
5 indels

ALGORITMI DI ALLINEAMENTO DI SEQUENZE



- L'allineamento di sequenze è il metodo bioinformatico per misurare la "similarità di sequenza".
- Come abbiamo visto, oltre che per stabilire la relazione evolutiva tra due sequenze può essere utilizzato anche per assegnare la funzione a partire dalla sequenza, ammesso che ci siano altre sequenze simili di cui conosciamo già la funzione.
- Gli allineamenti possono essere fatti tra coppie di sequenze ("pairwise") oppure tra molte sequenze contemporanemaente ("multipli").
- Gli allineamenti possono essere di due tipi: "globale" e "locale". Gli allineamenti multipli sono sempre globali.



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

A COPPIE

```
AGTTTGAATGTTTTGTGTGAAAGGAGTATACCATGAGATGAGATGACCACCAATCATTTC
|||||||||||||||||||  |||||||| ||| | |||||| ||||||||||||||||||
AGTTTGAATGTTTTGTGTGTGAGGAGTATTCCAAGGGATGAGTTGACCACCAATCATTTC
```

MULTIPLO

```
KFKHHLKEHLRIHSGEKPFECPNCKKRFSHSGSYSSHMSSKKCISLILVNGRNRALLKTl
KYKHHLKEHLRIHSGEKPYECPNCKKRFSHSGSYSSHISSKKCIGLISVNGRMRNNIKT-
KFKHHLKEHVRIHSGEKPFGCDNCGKRFSHSGSFSSHMTSKKCISMGLKLNNNRALLKRl
KFKHHLKEHIRIHSGEKPFECQQCHKRFSHSGSYSSHMSSKKCV----------------
KYKHHLKEHLRIHSGEKPYECPNCKKRFSHSGSYSSHISSKKCISLIPVNGRPRTGLKTs
```

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

2. Introduzione all'algoritmo

Dobbiamo definire un criterio che permetta di stabilire quale sia l'allineamento più corretto tra diverse possibilità.
Questo criterio dovrà essere implementato in un algoritmo di allineamento, ovvero, un programma che, date due o più sequenze, dovrà cercare in modo automatico l'allineamento migliore.

- Definizione dell'algoritmo

In bioinformatica, **un algoritmo di allineamento è un metodo per confrontare due o più sequenze di DNA, RNA o proteine e trovare le corrispondenze tra di esse.**

Date due sequenze vogliamo:
- misurare la loro similarità
- determinare la corrispondenza nucleotide-nucleotide o residuo-residuo
- osservare i pattern di conservazione e variabilità
- inferire relazioni evoluzionistiche

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Definizione dell'algoritmo

L'algoritmo di allineamento utilizza una serie di **regole** per assegnare **punteggi** alle corrispondenze tra le sequenze. L'obiettivo dell'algoritmo è trovare l'allineamento che **massimizza** il punteggio totale delle corrispondenze tra le sequenze.

Ad esempio, potremmo definire un punteggio pari ad uno se le due basi sono uguali il pari a zero altrimenti.

```
---IPLMTRWDQEQESDFGHKLPIYTREWCTRG
   |||||||||              ||||||
CHKIPLMTRWDQQESDFGHKLPVIYTREW----
```

Punteggio totale =
somma dei punteggi = 15

```
---IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
   |||||||||  ||||||||||  ||||||
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW----
```

Punteggio totale =
somma dei punteggi = 25

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Definizione dell'algoritmo

Altro esempio, potremmo porre un punteggio pari a 0 se le due basi sono uguali altrimenti (in caso di mismatch e gap) un punteggio negativo, pari a “-1”.

```
---IPLMTRWDQEQESDFGHKLPIYTREWCTRG
   |||||||||              ||||||
CHKIPLMTRWDQQESDFGHKLPVIYTREW----
```

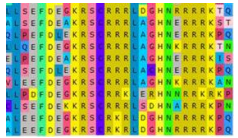
Punteggio totale =
somma dei punteggi = - 18

```
---IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
   |||||||||| |||||||||| ||||||
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW----
```

Punteggio totale =
somma dei punteggi = - 9
(punteggio più alto poiché ci sono molte meno sostituzioni)

Il calcolo dell’esempio 1 può essere visto come una misura di similarità, laddove invece quello dell’esempio 2 come una misura di divergenza.

ALGORITMI DI ALLINEAMENTO GLOBALE O LOCALE



In bioinformatica, gli algoritmi di allineamento vengono spesso classificati in base alla loro modalità di allineamento: allineamento **globale** o allineamento **locale**.

- L'allineamento **locale** è una modalità di allineamento in cui si cerca di identificare le regioni di **maggiore omologia** tra due sequenze, senza preoccuparsi della somiglianza complessiva tra le sequenze.

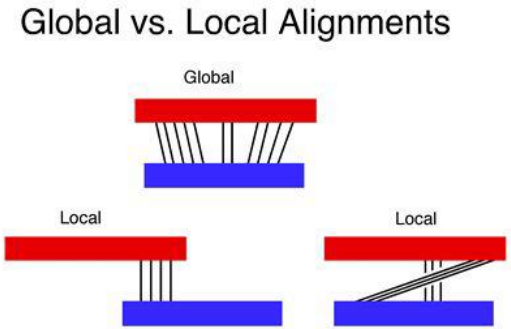
- Questa modalità di allineamento è utile quando si cerca di identificare regioni di somiglianza tra sequenze che possono avere molte differenze in altre parti.

ALGORITMI DI ALLINEAMENTO GLOBALE O LOCALE

L'allineamento **globale**, invece, cerca di allineare l'intera lunghezza delle due sequenze.

- Questa modalità di allineamento è utile quando si confrontano due sequenze che sono simili in gran parte della loro lunghezza, come le sequenze di DNA o proteine di specie simili o strettamente imparentate.

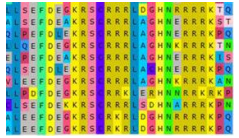
- In questo caso, l'allineamento globale consente di evidenziare le differenze e le somiglianze tra le sequenze in modo più completo.




```
Global  FTFTALILLAVAV
        F--TAL-LLA-AV

Local   FTFTALILL-AVAV
        --FTAL-LLAAV--
```

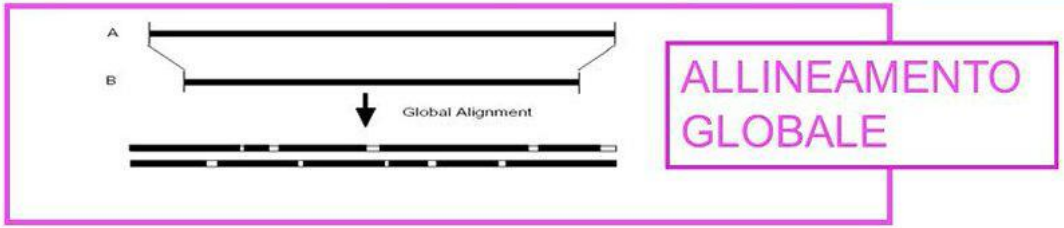
ALGORITMI DI ALLINEAMENTO GLOBALE O LOCALE

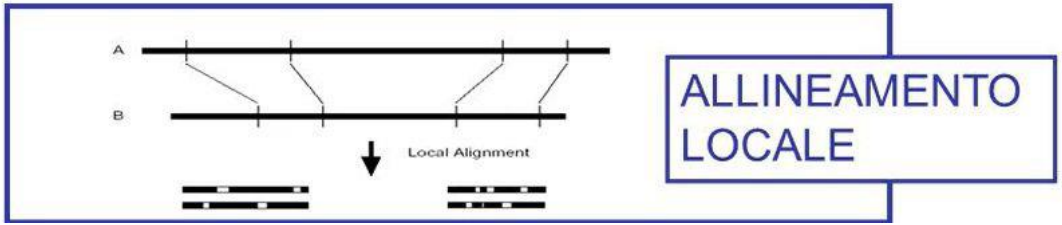


GLOBALE: considera la similarita' tra due sequenze in tutta la loro lunghezza

LOCALE: considera solo specifiche REGIONI simili tra alcune parti delle sequenze in analisi

```
Globale     LTGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK
             ||.  | |  |  .|     .|  ||  || | ||
            TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHKAG
Locale    LTGARDWEDIPLWTDWDIEQESDFKTRAFGTANCHK
                 ||||||||.||||
               TGIPLWTDWDLEQESDNSCNTDHYTREWGTMNAHK
```





ALGORITMO DI NEEDLEMAN-WUNSCH (ALLINEAMENTO GLOBALE)

L'algoritmo di **Needleman-Wunsch** è un algoritmo di allineamento utilizzato per allineare coppie di sequenze.

- L'algoritmo utilizza una **matrice di sostituzione** (o matrice di punteggio) per valutare la somiglianza tra le coppie di nucleotidi o aminoacidi nelle sequenze e produce un allineamento globale tra le due sequenze.

- è definito **algoritmo esatto** in quanto produce un risultato esatto e ottimale in termini di allineamento tra due sequenze, ovvero un allineamento che massimizza il punteggio di somiglianza tra le sequenze in base alla matrice di punteggio definita.

MATRICI DI SOSTITUZIONE

Una matrice di sostituzione è uno strumento utilizzato in bioinformatica per assegnare punteggi alle sostituzioni tra i vari amminoacidi o basi nucleotidiche in un allineamento di sequenze. Questi punteggi riflettono le probabilità che un amminoacido (o nucleotide) sia stato sostituito con un altro nel corso dell'evoluzione.

Le matrici di sostituzione sono fondamentali per gli allineamenti di sequenze poiché aiutano a identificare regioni di somiglianza che potrebbero indicare relazioni funzionali, strutturali o evolutive tra le sequenze studiate.

MATRICI DI SOSTITUZIONE

Una matrice di sostituzione si presenta come una tabella quadrata dove le righe e le colonne rappresentano tutti i possibili amminoacidi (nel caso delle proteine) o le basi nucleotidiche (nel caso del DNA o RNA).

Ogni cella della matrice contiene un punteggio che rappresenta la penalità o il premio per la sostituzione di un amminoacido (o nucleotide) con un altro.

I punteggi delle matrici di sostituzione sono calcolati sulla base di studi empirici e statistici che analizzano le frequenze delle sostituzioni in sequenze strettamente correlate dal punto di vista evolutivo.

I punteggi positivi indicano sostituzioni favorevoli o tollerate (che si verificano più frequentemente di quanto ci si aspetterebbe per caso), mentre i punteggi negativi indicano sostituzioni sfavorevoli o rare.

MATRICI DI SOSTITUZIONE

Durante l'allineamento di sequenze, i punteggi della matrice vengono utilizzati per calcolare il punteggio totale dell'allineamento.
Un allineamento con un punteggio complessivo più alto è considerato migliore, poiché suggerisce una maggiore somiglianza evolutiva o funzionale tra le sequenze.

Una matrice di sostituzione fornisce quindi uno schema quantitativo per valutare quanto sia probabile che una data sostituzione tra amminoacidi o basi si sia verificata, facilitando l'interpretazione degli allineamenti di sequenze e contribuendo alla comprensione delle relazioni evolutive tra le sequenze.

Tra le matrici di sostituzione più utilizzate ci sono le matrici **PAM (Point Accepted Mutation)** e **BLOSUM (BLOcks SUbstitution Matrix)**.

MATRICI DI SOSTITUZIONE

Abbiamo detto che la matrice di punteggio è utilizzata per valutare la somiglianza tra due sequenze di DNA o di proteine e viene utilizzata dagli algoritmi di allineamento per assegnare un punteggio a ogni possibile allineamento tra le sequenze.

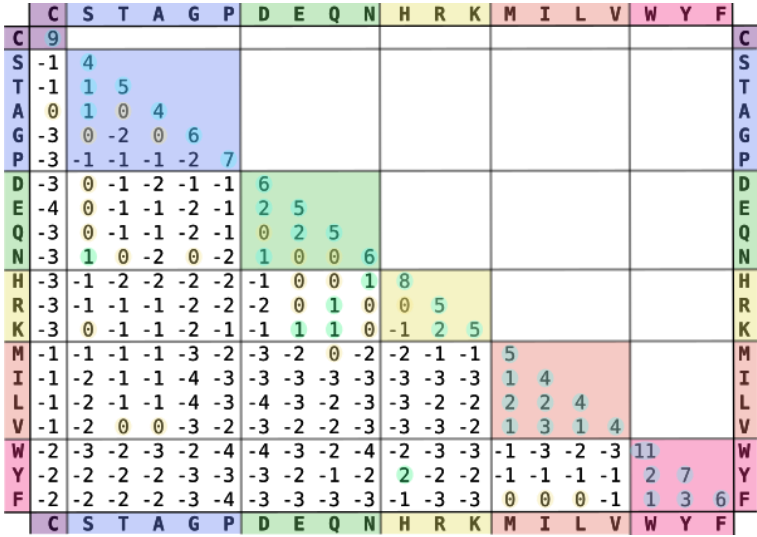
Se sto allineando sequenze nucleotidiche la matrice quadrata sarà molto piccola: dimensione 4*4 (abbiamo solo 4 basi)

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

MATRICI DI SOSTITUZIONE

Se invece sto allineando proteine, la matrice di sostituzione è più grande perché ho 20 simboli e non solo 4!

❏ Inoltre, mentre con le sequenze nucleotidiche va bene avere solo due possibili punteggi (uno per il “match” ed uno per il “mismatch”), per le proteine questo non va più bene. Infatti ci sono amminoacidi più o meno simili tra loro. Sostituire un amminoacido idrofobico con un altro idrofobico non è la stessa cosa che sostituirlo con uno polare.



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

MATRICI DI SOSTITUZIONE

Le matrici BLOSUM sono state sviluppate a partire da blocchi di sequenze allineate di proteine omologhe (cioè che hanno un'origine evolutiva comune).

Questi blocchi sono insiemi di sequenze proteiche altamente conservate che si ritiene mantengano una struttura e una funzione simili attraverso diverse specie.

I numeri all'interno della matrice BLOSUM rappresentano i punteggi di sostituzione tra due **amminoacidi**, dove i valori maggiori indicano che una sostituzione è meno probabile che si verifichi casualmente rispetto ad un valore minore.

MATRICI DI SOSTITUZIONE

Nella matrice BLOSUM62, come in altre matrici di sostituzione, i valori (o punteggi) riflettono la probabilità relativa di che due amminoacidi si sostituiscano l'uno con l'altro durante l'evoluzione. Questi valori sono calcolati in base alla frequenza delle sostituzioni osservate in confronto a quelle attese, e sono poi trasformati in logaritmi per essere utilizzati negli allineamenti delle sequenze di proteine.

Nella matrice BLOSUM62, un **valore positivo** indica che la sostituzione tra due amminoacidi è avvenuta più frequentemente di quanto ci si aspetterebbe per caso, basandosi sulle frequenze generali di questi amminoacidi nelle sequenze proteiche.
In pratica, un punteggio positivo suggerisce che la sostituzione è conservativa, ovvero che tende a conservare la funzione e la struttura della proteina. Questo perché aminoacidi con proprietà chimiche o fisiche simili sono più propensi a essere sostituiti tra loro senza alterare significativamente la funzione della proteina. Ad esempio, la sostituzione di un amminoacido idrofobico con un altro amminoacido idrofobico tende ad avere un valore positivo.

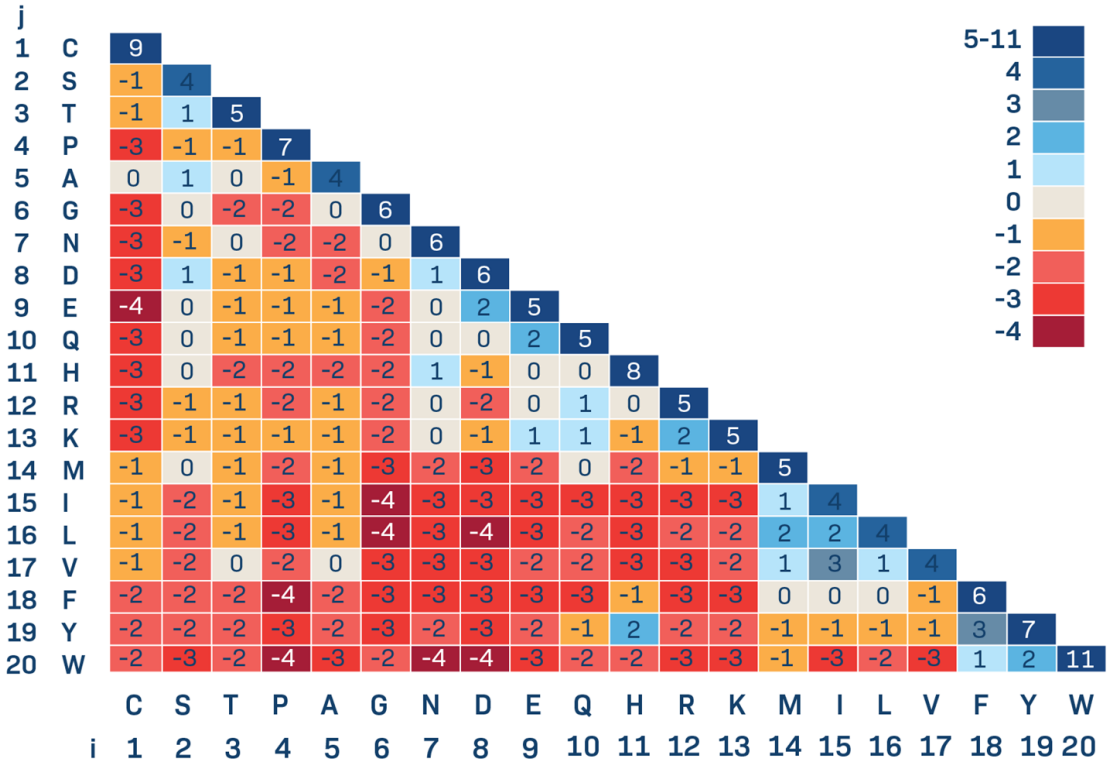
MATRICI DI SOSTITUZIONE

Al contrario, un **valore negativo** nella matrice BLOSUM62 indica che la sostituzione tra due amminoacidi è meno frequente di quanto ci si aspetterebbe per caso. Questo suggerisce che la sostituzione è sfavorevole dal punto di vista evolutivo, probabilmente perché porta a un cambiamento significativo nella struttura o nella funzione della proteina. Le sostituzioni che coinvolgono aminoacidi con proprietà molto diverse, come da idrofobico a idrofilico, tendono ad avere valori negativi perché tali cambiamenti possono alterare drasticamente la conformazione della proteina o il suo comportamento in un ambiente acquoso.

Quindi, durante l'allineamento delle sequenze, l'utilizzo di una matrice come BLOSUM62 aiuta a identificare le sostituzioni che sono più probabili dal punto di vista evolutivo, fornendo punteggi più alti per allineamenti che mantengono le proprietà biologiche delle proteine e punteggi più bassi per allineamenti che probabilmente alterano la struttura o la funzione.

MATRICI DI SOSTITUZIONE

La matrice di sostituzione **BLOSUM62**



MATRICI DI SOSTITUZIONE

Perché nella matrice Blosum62, come in altre matrici di sostituzione, i punteggi di sostituzione di un aminoacido con se stesso sono diversi?

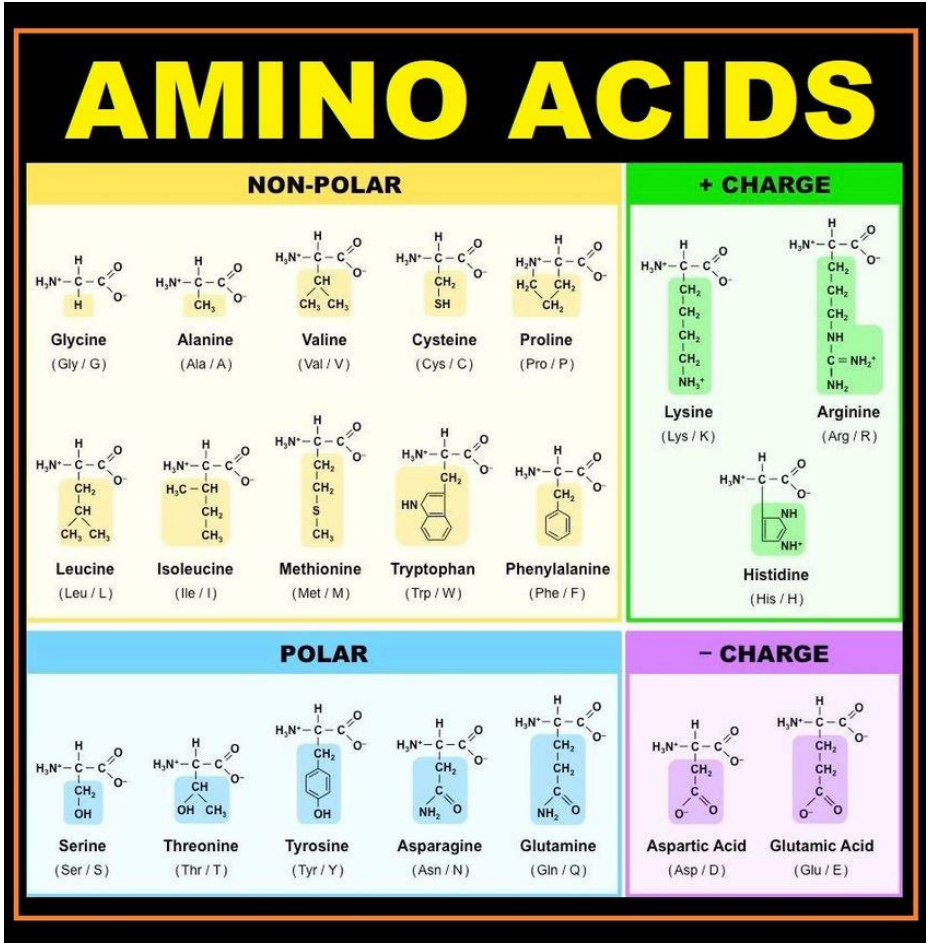
Questi punteggi riflettono la "conservazione" di ciascun amminoacido durante l'evoluzione: un punteggio più alto indica che l'amminoacido è generalmente più conservato (cioè meno probabile che cambi durante l'evoluzione), mentre un punteggio più basso indica che l'amminoacido è meno conservato.

Il punteggio di auto-sostituzione per ciascun amminoacido riflette quindi la sua conservazione generale in questo vasto contesto evolutivo.

MATRICI DI SOSTITUZIONE

- La matrice BLOSUM62 assegna un punteggio maggiore ad amminoacidi che si **sostituiscono** l'un l'altro **frequentemente** nelle sequenze di proteine omologhe, e un punteggio minore a quelli che si sostituiscono meno spesso.

- Ad esempio, una sostituzione di un residuo di leucina (L) con uno di valina (V) viene valutata con un punteggio di +1, mentre una sostituzione di un residuo di leucina (L) con uno di asparagina (N) viene valutata con un punteggio di -4, poiché tale sostituzione è considerata molto meno probabile nella natura.

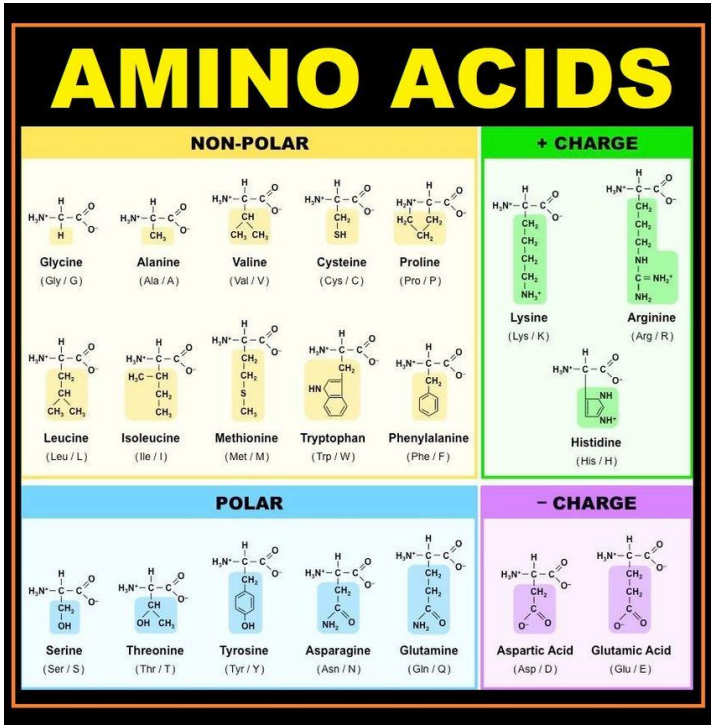


La matrice di sostituzione **BLOSUM30** è una delle matrici BLOSUM di bassa conservazione, che viene utilizzata per l'allineamento di sequenze molto divergenti.

- Essa è stata calcolata a partire da un insieme di sequenze di proteine aventi al massimo il 30% di identità tra loro.

- In questa matrice, gli amminoacidi sostituiti tra loro meno frequentemente nelle sequenze di proteine omologhe ricevono punteggi più alti, mentre gli amminoacidi sostituiti più frequentemente ricevono punteggi più bassi.

- Ad esempio, una sostituzione di un residuo di leucina (L) con uno di valina (V) viene valutata con un punteggio di +2, mentre una sostituzione di un residuo di leucina (L) con uno di asparagina (N) viene valutata con un punteggio di -3.



MATRICI DI SOSTITUZIONE

Le matrici **PAM** sono invece state costruite a partire dalla teoria dell'evoluzione molecolare, che suggerisce che le sostituzioni di amminoacidi in sequenze di proteine seguono una distribuzione di Poisson.

Le matrici PAM descrivono quindi la probabilità di osservare una specifica sostituzione di amminoacido in una sequenza di proteine a seguito di un dato numero di eventi di sostituzione accettati, assumendo che la sequenza sia stata sottoposta ad una specifica quantità di evoluzione.

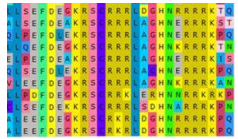
MATRICI DI SOSTITUZIONE

Le matrici PAM sono utilizzate quindi per quantificare la distanza evolutiva tra le sequenze di proteine.

Sviluppate originariamente da Margaret Dayhoff negli anni '70, queste matrici sono state tra le prime a essere utilizzate per studiare le relazioni evolutive tra le proteine.

La denominazione "PAM" si riferisce alla quantità di evoluzione che si verifica attraverso mutazioni accettate, ovvero mutazioni che sono state fissate (mantenute) nella popolazione e non causano un cambiamento deleterio nella funzione della proteina.

MATRICI DI SOSTITUZIONE



Il numero (N) dopo “PAM” indica che quella è la matrice derivata da sostituzioni in proteine omologhe a N “passaggi evolutivi” di distanza.

- La matrice PAM1 rappresenta il livello di sostituzione atteso per 1 punto accettato per 100 residui amminoacidici. È la matrice più fine e riflette le mutazioni più immediate e minori.

- Ogni unità PAM misura le differenze tra le sequenze con circa l'1% di sostituzioni amminoacidiche per sito. Matrici con numeri più elevati (ad esempio, PAM250) descrivono le sostituzioni attese su periodi evolutivi più lunghi e sono usate per allineare sequenze meno correlate.

- Le matrici PAM sono particolarmente utili nell'allineamento di sequenze che sono vicine dal punto di vista evolutivo (per esempio, con una grande quantità di omologia). Tuttavia, per sequenze molto diverse, vengono spesso preferite matrici come quelle BLOSUM, poiché le PAM ad alto numero possono essere meno precise nel prevedere le sostituzioni tra sequenze distanti.

MATRICI DI SOSTITUZIONE

➢ Un passaggio evolutivo è un “PAM” e equivale ad una sostituzione ogni 100 aa.

➢ Ogni indice dice il numero di passi evolutivi che si prevede per le proteine in analisi.

Quindi:
PAM 100 non significa che il 100% degli aminoacidi vengono sostituiti, ma che si prevedono 100 passi evolutivi, ognuno con le sue probabilità.

➢ Più è alta la distanza evolutiva tra le due sequenza da allinare, più alto l’N della matrice da usare.

➢ Le matrici PAM comunemente usate sono PAM30, PAM70, PAM120, PAM250

MATRICI DI SOSTITUZIONE

PAM Versus BLOSUM

- PAM is based on an evolutionary model.
- BLOSUM is based on protein families.
- PAM is based on global alignment.
- BLOSUM is based on local alignment.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	← →	*More divergent*

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- matrice di sostituzione

La matrice **PAM120** è stata calcolata a partire da un insieme di sequenze di proteine aventi al massimo il 120% di divergenza.

In questa matrice, gli amminoacidi sostituiti tra loro meno frequentemente nelle sequenze di proteine omologhe ricevono punteggi più alti, mentre gli amminoacidi sostituiti più frequentemente ricevono punteggi più bassi.

Ad esempio, una sostituzione di un residuo di leucina (L) con uno di valina (V) viene valutata con un punteggio di +1, mentre una sostituzione di un residuo di leucina (L) con uno di asparagina (N) viene valutata con un punteggio di -3.

In generale, la matrice PAM120 è più conservativa rispetto a matrici come la BLOSUM62, ed è quindi più adatta per l'allineamento di sequenze molto simili tra loro.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- matrice di sostituzione

PAM 120

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

small, polar

small, nonpolar

polar or acidic

basic

large, hydrophobic

aromatic

Positive score – frequency of substitutions is greater than would have occurred by random chance.

Zero score – frequency is equal to that expected by chance.

Negative score – frequency is less than would have occurred by random chance.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- matrice di sostituzione

La matrice **PAM250** è stata calcolata a partire da un insieme di sequenze di proteine aventi al massimo il 250% di divergenza.

In questa matrice, gli amminoacidi sostituiti tra loro meno frequentemente nelle sequenze di proteine omologhe ricevono punteggi più alti, mentre gli amminoacidi sostituiti più frequentemente ricevono punteggi più bassi.

Ad esempio, una sostituzione di un residuo di leucina (L) con uno di valina (V) viene valutata con un punteggio di +2, mentre una sostituzione di un residuo di leucina (L) con uno di asparagina (N) viene valutata con un punteggio di -2.

In generale, la matrice PAM250 è meno conservativa rispetto a matrici come la BLOSUM62, ed è quindi più adatta per l'allineamento di sequenze meno simili tra loro.

La matrice **PAM250**

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

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

Riepilogando, mentre il numero dopo PAM indica divergenza, quello dopo BLOSUM indica similarità.