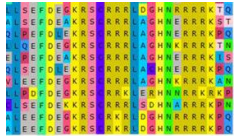
ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE



Per fortuna esiste un metodo più veloce che non calcolarsi i punteggi di tutti i possibili allineamenti per poi fare la classifica!

❏ Gli “algoritmi” di allineamento sfruttano il concetto di “programmazione dinamica”.

-La programmazione dinamica è una t**ecnica algoritmica** utilizzata per risolvere problemi che possono essere suddivisi in **sotto-problemi** più piccoli.

-Gli algoritmi di allineamento utilizzano spesso la tecnica di programmazione dinamica per risolvere il problema dell'allineamento ottimale tra due sequenze. Questa tecnica aiuta a trovare l'allineamento ottimale tra due sequenze minimizzando un punteggio di penalità per mismatch, inserzioni e delezioni o massimizzando un punteggio di somiglianza.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

La programmazione dinamica è una metodica computazionale usata per allineare 2 o più sequenze (aminoacidiche o nucleotidiche).

Tale tecnica algoritmica permette di combinare gli appaiamenti (**matches**), i non appaiamenti(**mismatches**) e le **gaps** nelle sequenze in modo da trovare il massimo numero possibile degli appaiamenti dei residui tra loro correlati.

Attenzione!! Bisogna sapere che gli allineamenti ottenuti dipendono dalla scelta del sistema di punteggio per gli appaiamenti (matrici di sostituzione) e le penalità delle inserzioni e delezioni.

ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE) per l'ALLINEAMENTO A COPPIE DI SEQUENZE

L'algoritmo di Needleman-Wunsch è un classico metodo di programmazione dinamica utilizzato per l'allineamento globale di sequenze, che può includere sequenze di DNA, RNA o proteine. Sviluppato da Saul B. Needleman e Christian D. Wunsch nel 1970, questo algoritmo è stato uno dei primi approcci computazionali progettati per identificare l'allineamento ottimale tra due sequenze biologiche.

L'obiettivo dell'algoritmo di Needleman-Wunsch è trovare l'allineamento che massimizza un punteggio di somiglianza totale, considerando corrispondenze, disallineamenti (mismatch) e gap (inserzioni o delezioni).

ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE) per l'ALLINEAMENTO A COPPIE DI SEQUENZE

Ecco come funziona l'algoritmo:

1. Inizializzazione:

- Viene creata una matrice di punteggi dove le righe rappresentano i caratteri (ad esempio, nucleotidi o amminoacidi) della prima sequenza, e le colonne rappresentano i caratteri della seconda sequenza.

- La prima riga e la prima colonna vengono inizializzate con valori che riflettono le penalità cumulative per l'inserimento di gap. Per esempio, se la penalità per un gap è -2, allora la matrice sarà inizializzata a 0, -2, -4, -6, e così via lungo la prima riga e la prima colonna.

$$A(0,j) = j * gap \qquad A(i,0) = i * gap$$

ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE) per l'ALLINEAMENTO A COPPIE DI SEQUENZE

2. Riempimento della Matrice:

- Ogni cella (i,j) della matrice viene riempita calcolando il massimo di tre possibili scenari:
  - Il punteggio nella cella immediatamente sopra $(i-1,j)$, più la penalità per un gap (allineamento di un carattere della sequenza verticale con un gap).
  - Il punteggio nella cella immediatamente a sinistra $(i, j-1)$, più la penalità per un gap (allineamento di un carattere della sequenza orizzontale con un gap).
  - Il punteggio nella cella in diagonale superiore sinistra $(i-1,j-1)$, più il punteggio per l'allineamento dei caratteri correnti (che può essere positivo per una corrispondenza o negativo per un mismatch).

- Questo processo continua fino a che tutte le celle sono state riempite.

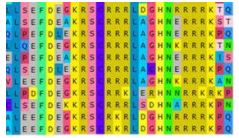
$$A(i,j) = max \begin{cases} A(i-1,j) + gap \\ A(i,j-1) + gap \\ A(i-1,j-1) + \sigma(i,j) \end{cases}$$

ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE) per l'ALLINEAMENTO A COPPIE DI SEQUENZE

3. Backtracking:

- Partendo dalla cella in basso a destra della matrice, si segue il percorso che ha portato al punteggio più alto per arrivare a quella cella. Questo percorso determina l'allineamento ottimale.
- Si procede all'indietro (backtracking) fino a raggiungere la cella in alto a sinistra, registrando l'allineamento. Ogni passo può essere una corrispondenza/mismatch (muovendosi in diagonale), un gap nella sequenza superiore (muovendosi verticalmente) o un gap nella sequenza inferiore (muovendosi orizzontalmente).

ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE)



Esempio: Allineamento delle sequenze
seq1 = GCATGC
seq2 = GATTAC

j->		0	1	2	3	4	5	6
i->			G	C	A	T	G	C
0		0	−1	−2	−3	−4	−5	−6
1	G	−1						
2	A	−2						
3	T	−3						
4	T	−4						
5	A	−5						
6	C	−6						

Inizializzazione della matrice:

$$gap = -1$$

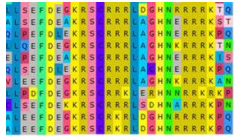
$$A(i,0) = i * gap$$

$$A(0,j) = j * gap$$

	A	G	C	T
A	1	-1	-1	-1
G	-1	1	-1	-1
C	-1	-1	1	-1
T	-1	-1	-1	1

Matrice dei punteggi di similarità

ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE)



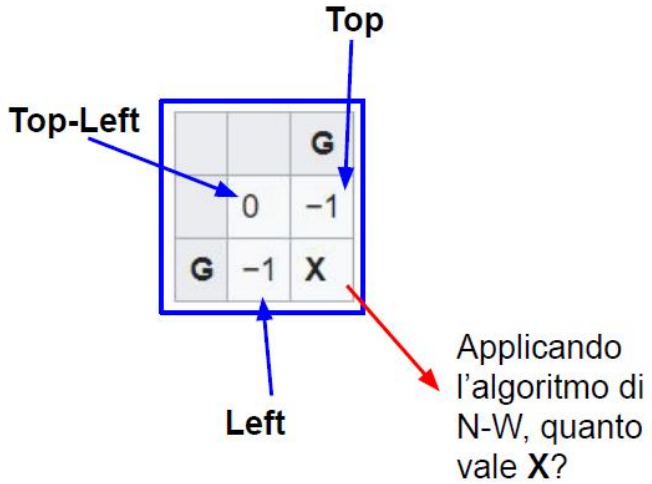
- Programmazione dinamica

Consideriamo la prima casella da popolare corrispondente alle G (primo nucleotide di ognuna delle sequenze).

I punteggi delle celle vicine sono (0,-1,-1).
Applichiamo l'algoritmo di N-W:

- Top: (−1) + (−1) = (−2)
- Left: (−1) + (−1) = (−2)
- Top-Left: 0+1 = 1

X = 1



ALGORITMO DI NEEDLEMAN & WUNSCH (GLOBALE)

- Programmazione dinamica

Andando avanti nel popolamento della matrice calcoliamo altri due valori corrispondenti all' appaiamento di C con A e G con C.

Applichiamo l'algoritmo di N-W:

X:

- Top: (−2)+(−1) = (−3)
- Left: (+1)+(−1) = (0)
- Top-Left: (−1)+(−1) = (−2)

→ **X = 0**

Y:

- Top: (1)+(−1) = (0)
- Left: (−2)+(−1) = (−3)
- Top-Left: (−1)+(−1) = (−2)

→ **Y = 0**

		G	C
	0	−1	−2
G	−1	1	X
A	−2	Y	

↓

		G	C
	0	−1	−2
G	−1	1	0
A	−2	0	

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

	–	A	G	A	T	T	C	C	A	T
–	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	0	1	0
G	0	1	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4
C	0	1	2	2	3	3	4	5	5	5
C	0	1	2	2	3	3	4	5	5	5
A	0	1	2	3	3	3	3	3	6	6
T	0	1	2	3	4	4	4	4	6	7

Avendo popolato l'intera matrice l'algoritmo si sposterà dall'ultima casella andando a ritroso fino a tornare alla prima casalla A(0,0) ripercorrendo i valori che hanno portato al valore massimo in fondo a destra della matrice.
Questo processo viene chiamato ***traceback***, indicato con il tratto azzurro nella slide successiva.

Qui la gap e' stata posta pari a zero

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Programmazione dinamica

	–	A	G	A	T	T	C	C	A	T
–	0	0	0	0	0	0	0	0	0	0
A	0	1	0	1	0	0	0	0	1	0
G	0	1	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4
C	0	1	2	2	3	3	4	5	5	5
C	0	1	2	2	3	3	4	5	5	5
A	0	1	2	3	3	3	3	3	6	6
T	0	1	2	3	4	4	4	4	6	7

L'algoritmo appena visto è un'**algoritmo di allineamento globale** (vengono allineate le due sequenze dalla prima all'ultima base), sviluppato nel 1970 da ***Needleman e Wunsch.***

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

Quindi possiamo riassumere dicendo che l'algoritmo di Needleman-Wunsch funziona in tre fasi principali:

1. **Scelta della matrice di sostituzione e valoredi gap**: viene costruita una matrice in cui le righe e le colonne rappresentano i singoli nucleotidi o amminoacidi delle due sequenze. I punteggi sono assegnati in base alla somiglianza tra i nucleotidi/amminoacidi.

1. **Riempimento della matrice**: viene riempita la matrice di punteggio a partire dal primo elemento in alto a sinistra fino all'ultimo elemento in basso a destra. In ogni posizione, vengono calcolati tre possibili punteggi, rappresentanti le tre possibilità di allineamento: un allineamento diagonale, un allineamento verticale e un allineamento orizzontale. Il punteggio massimo tra le tre possibilità viene quindi assegnato alla posizione corrente.

1. **Costruzione dell'allineamento:** partendo dall'ultimo elemento in basso a destra della matrice di punteggio, viene ricostruito l'allineamento a ritroso (traceback). In base al punteggio massimo assegnato alla posizione corrente, si decide in quale direzione andare (diagonale, verticale o orizzontale) e si inseriscono i nucleotidi/amminoacidi corrispondenti nell'allineamento finale.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

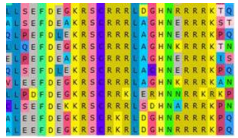
- Esempio pratico

..un esercizio pratico: confronto tra proteina codificata dal gene gata1 del gallo ed uomo!

```
>AAA49055.1 Eryf1 protein [Gallus gallus]
MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLGTPQWVPPATQ
MEPPHYLELLQPPRGSPPHPSSGPLLPLSSGPPPCEARECVNCGATATPLWRRDGTGHYLCNACGLYHRL
NGQNRPLIRPKKRLLVSKRAGTVCSNCQTSTTTLWRRSPMGDPVCNACGLYYKLHQVNRPLTMRKDGIQT
RNRKVSSKGKKRRPPGGGNPSATAGGGAPMGGGGDPSMPPPPPPPAAAPPQSDALYALGPVVLSGHFLPF
GNSGGFFGGGAGGYTAPPGLSPQI
```

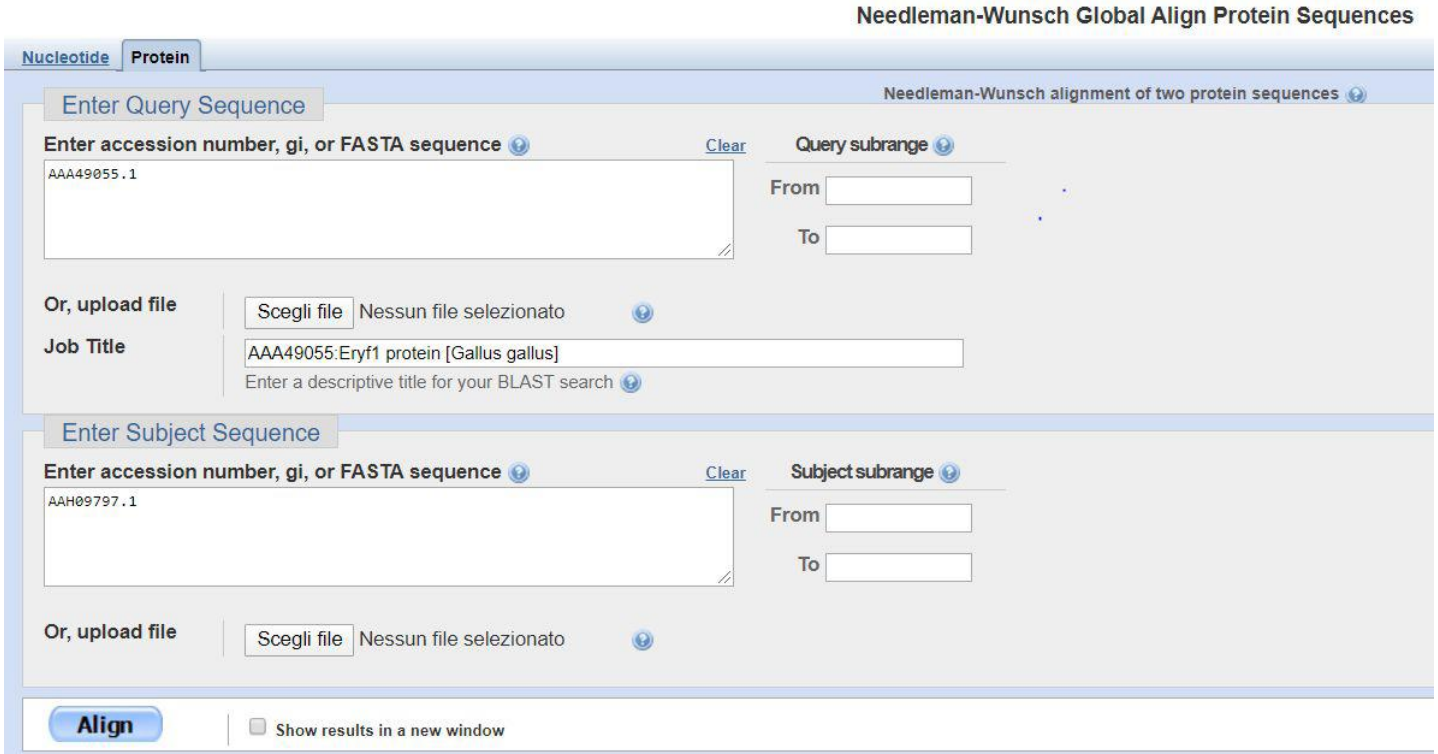
```
>AAH09797.1 GATA1 protein [Homo sapiens]
MEFPGLGSLGTSEPLPQFVDPALVSSTPESGVFFPSGPEGLDAAASSTAPSTATAAAAALAYYRDAEAYR
HSPVFQVYPLLNCMEGIPGGSPYAGWAYGKTGLYPASTVCPTREDSPPQAVEDLDGKGSTSFLETLKTER
LSPDLLTLGPALPSSLPVPNSAYGGPDFSSTFFSPTGSPLNSAAYSSPKLRGTLPLPPCEARECVNCGAT
ATPLWRRDRTGHYLCNACGLYHKMNGQNRPLIRPKKRLIVSKRAGTQCTNCQTTTTTLWRRNASGDPVCN
ACGLYYKLHHQHYCGGSAQLMRAQSMASRGGVVSFSSCSQNSGQPKSLGPRHPLA
```

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE



- Esempio pratico

utilizziamo l'algoritmo di N-W sul portale di ncbi:
(nel form di google scrivete "Needleman-Wunsch ncbi")



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio pratico

BLAST® » Global Alignment » results for RID-8W9MYW40114

< Edit Search | Save Search | Search Summary

Job Title	**gb|AAA49055.1|**
RID	8W9MYW40114 *Search expires on 04-10 03:31 am* Download All
Program	Needleman-Wunsch alignment of two sequences Citation
Query ID	AAA49055.1 (amino acid)
Query Descr	Eryf1 protein [Gallus gallus]
Query Length	304
Subject ID	AAH09797.1 (amino acid)
Subject Descr	GATA1 protein [Homo sapiens]
Subject Length	335

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio pratico

GATA1 protein [Homo sapiens]

Sequence ID: AAH09797.1 **Length:** 335 **Number of Matches:** 1

See 2 more title(s)

Alignments

Range 1: 1 to 335 GenPept Graphics

NW Score	Identities	Positives	Gaps
400	135/402(34%)	152/402(37%)	165/402(41%)

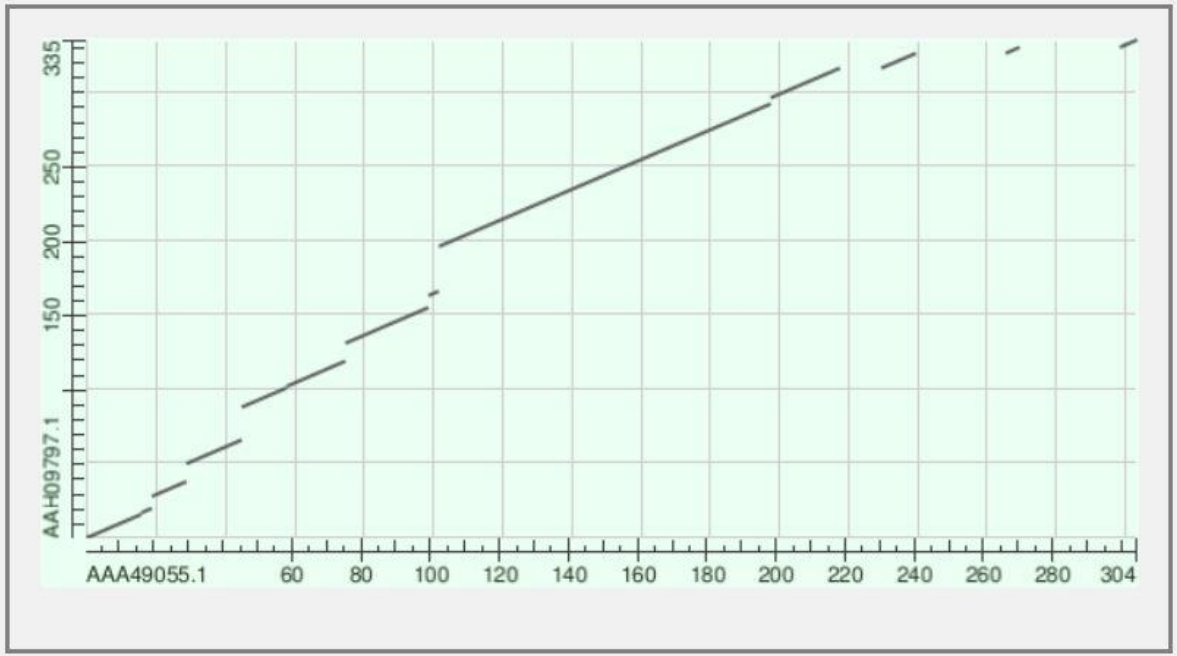
```
Query  1    MEFVALGGPDAGSPTP-FPD--------EAGAFLGLGG------------GERTEAGGLL  39
            MEF  LG      P P F D        E+G F   G                T A   L
Sbjct  1    MEFPGLGSLGTSEPLPQFVDPALVSSTPESGVFFPSGPEGLDAAASSTAPSTATAAAAAL  60

Query  40   ASYPPS----------------------GRVSLVPWADTGT-LGTPQWVPPATQMEPPH-  75
            A Y  +                      G      WA   T L     V P  +  PP
Sbjct  61   AYYRDAEAYRHSPVFQVYPLLNCMEGIPGGSPYAGWAYGKTGLYPASTVCPTREDSPPQA  120

Query  76   -----------YLELLQPPRGSPPHPSSGPLLPLS--------SGP--------------  102
                      +LE L+  R SP   + GP LP S          GP
Sbjct  121  VEDLDGKGSTSFLETLKTERLSPDLLTLGPALPSSLPVPNSAYGGPDFSSTFFSPTGSPL  180

Query  103  ----------------PPCEARECVNCGATATPLWRRDGTGHYLCNACGLYHRLNGQNRP  146
                            PPCEARECVNCGATATPLWRRD TGHYLCNACGLYH++NGQNRP
Sbjct  181  NSAAYSSPKLRGTLPLPPCEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGQNRP  240

Query  147  LIRPKKRLLVSKRAGTVCSNCQTSTTTLWRRSPMGDPVCNACGLYYKLHQVN----RPLT  202
            LIRPKKRL+VSKRAGT C+NCQT+TTTLWRR+  GDPVCNACGLYYKLH  +
Sbjct  241  LIRPKKRLIVSKRAGTQCTNCQTTTTTLWRRNASGDPVCNACGLYYKLHHQHYCGGSAQL  300

Query  203  MRKDGIQTRNRKVSSKGKKRRPPGGGNPSATAGGGAPMGGGGDPSMPPPPPPPAAAPPQS  262
            MR   + +R   VS               S +   G P
Sbjct  301  MRAQSMASRGGVVSFS------------SCSQNSGQPK----------------------  326

Query  263  DALYALGPVVLSGHFLPFGNSGGFFGGGAGGYTAPPGLSPQI  304
                +LGP                              P
Sbjct  327  ----SLGP----------------------------RHPLA  335
```

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio pratico

Plot of AAA49055.1 vs AAH09797.1

Dot Plot



Vi riporto il Dot Plot delle due sequenze allineate con l'algoritmo di N-W.

A cosa corrispondono nell'allineamento i tratti di interruzione delle linee rette?

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

Ricordiamo...

L'allineamento può essere globale o locale, a seconda del tipo di analisi che si vuole effettuare.

L'allineamento globale di sequenze cerca di trovare l'allineamento migliore tra due sequenze per l'intera lunghezza delle sequenze stesse. Ciò significa che ogni singolo carattere delle sequenze viene confrontato con ogni singolo carattere dell'altra sequenza. L'allineamento globale viene solitamente utilizzato per analizzare sequenze di DNA o proteine che hanno una forte somiglianza tra loro e sono di lunghezza simile.

D'altra parte, **l'allineamento locale** di sequenze si concentra su una regione specifica delle sequenze, ignorando le regioni che non sono rilevanti per l'analisi. In questo modo, l'allineamento locale è in grado di rilevare regioni di somiglianza tra sequenze che non sarebbero state rilevate dall'allineamento globale. L'allineamento locale viene solitamente utilizzato per identificare regioni di interesse all'interno di sequenze di lunghezza molto differente o con una bassa somiglianza globale.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Allineamento globale vs locale

An Introduction to Bioinformatics Algorithms www.bioalgorithms.info

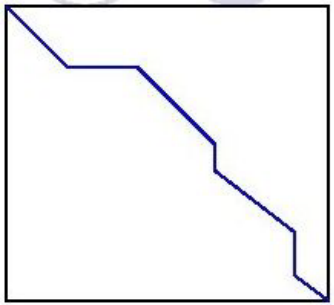
Local vs. Global Alignment: Example

- Global Alignment:

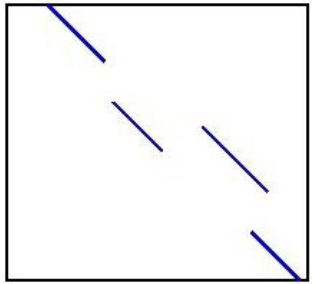
```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |   || |  ||   | | | |||     || | | |   | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

- Local Alignment—better alignment to find conserved segment:

```
                    tccCAGTTATGTCAGgggacacgagcatgcagagac
                      ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```



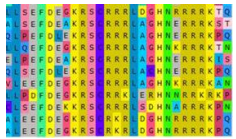
Allineamento globale



Allineamento locale

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Allineamento globale vs locale

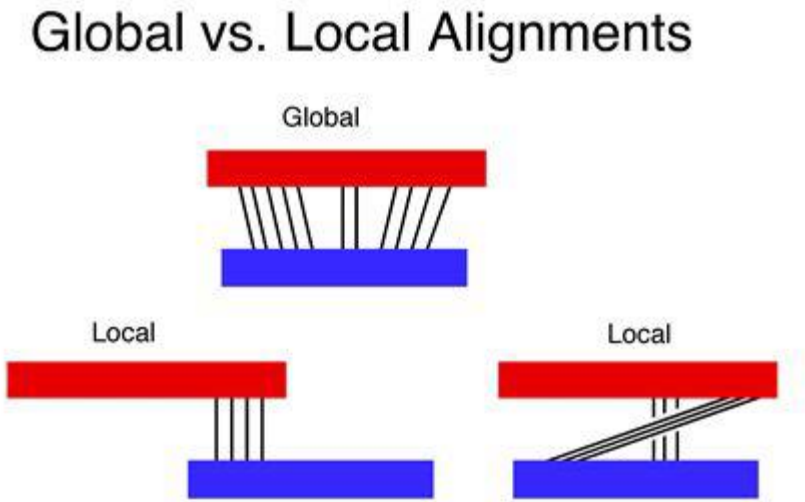


I **vantaggi** dell'allineamento locale sono molteplici.

- è più sensibile nell'individuare regioni di somiglianza tra sequenze, in quanto si concentra solo su queste regioni e non viene influenzato da regioni divergenti.

- può essere utilizzato per identificare regioni di somiglianza tra sequenze che sono di lunghezza molto differente o che hanno una bassa somiglianza globale.

- può essere utilizzato per l'analisi di sequenze di proteine, in cui i domini funzionali possono essere distribuiti in modo non uniforme all'interno della sequenza.



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

4. Algoritmi di allineamento locale: Smith-Waterman (allineamento **locale**)

L'algoritmo di Smith-Waterman è un algoritmo di allineamento locale di sequenze di DNA o proteine, che permette di individuare le regioni di somiglianza tra due sequenze.

-Si basa sulla creazione di una **matrice di punteggio**, che viene utilizzata per confrontare ogni possibile coppia di caratteri delle due sequenze. Questa matrice di punteggio può essere creata utilizzando una serie di parametri, come ad esempio la **somiglianza** tra i caratteri o la presenza di **gap**.

-A partire dalla matrice di punteggio, l'algoritmo di Smith-Waterman calcola il **punteggio** di allineamento **massimo** per ogni possibile posizione di allineamento nella matrice. Questo viene fatto utilizzando un metodo di programmazione dinamica, che permette di evitare il confronto di tutte le possibili coppie di sequenze.

-Una volta calcolati tutti i punteggi di allineamento, cerca la **posizione** con il **punteggio massimo** e crea l'allineamento tra le due sequenze partendo da quella posizione.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

4. Algoritmi di allineamento locale: Smith-Waterman

$$A(i,0) = A(0,j) = 0$$

$$A(i,j) = max \begin{cases} A(i-1,j) + gap \\ A(i,j-1) + gap \\ A(i-1,j-1) + \sigma(i,j) \\ 0 \end{cases}$$

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

3	6	9	7	10	13
G	T	T	-	A	C
I	I	I		I	I
G	T	T	G	A	C

Che valore di gap penalty è stato dato nell'esempio qui di fianco?

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

4. Algoritmi di allineamento locale: Smith-Waterman

L'allineamento locale di sequenze di Smith-Waterman è un'importante tecnica utilizzata in bioinformatica per diverse applicazioni. Alcune di queste applicazioni includono la ricerca di motivi, la classificazione di sequenze e la ricostruzione di filogenesi. Ad esempio:

- Ricerca di motivi: l'allineamento locale di sequenze può essere utilizzato per cercare regioni conservate tra sequenze. Queste regioni possono essere sequenze funzionali, come i siti di legame di una proteina o le regioni codificanti di un gene. L'allineamento locale di sequenze permette di individuare queste regioni anche in sequenze molto divergenti, in cui l'allineamento globale non sarebbe possibile.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

4. Algoritmi di allineamento locale: Smith-Waterman

- Classificazione di sequenze: l'allineamento locale di sequenze può essere utilizzato per la classificazione di sequenze. Ad esempio, l'allineamento di una sequenza sconosciuta con sequenze di riferimento può permettere di determinare la sua appartenenza a una determinata famiglia di proteine o di geni. Questa tecnica è utilizzata ampiamente in bioinformatica per la caratterizzazione di nuove sequenze.

- Ricostruzione di filogenesi: l'allineamento locale di sequenze può essere utilizzato anche per la ricostruzione di filogenesi, ovvero per la costruzione di alberi filogenetici che rappresentano la relazione evolutiva tra sequenze. L'allineamento locale di sequenze permette di identificare regioni conservate tra sequenze di diversi organismi, che possono essere utilizzate per la ricostruzione di alberi filogenetici.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Comparazione con l'algoritmo di Needleman-Wunsch

L'algoritmo di Smith-Waterman e l'algoritmo di Needleman-Wunsch sono entrambi utilizzati per l'allineamento di sequenze di nucleotidi o di proteine, ma differiscono per l'obiettivo che cercano di raggiungere.

- L'algoritmo di Needleman-Wunsch cerca di trovare l'allineamento **globale** ottimale tra due sequenze.

- L'algoritmo di Smith-Waterman, invece, cerca di trovare l'allineamento **locale** ottimale tra due sequenze.

In termini di complessità computazionale, entrambi gli algoritmi hanno una complessità di tempo O(NM), dove N e M sono le lunghezze delle due sequenze. Tuttavia, l'algoritmo di Smith-Waterman richiede **più tempo** per eseguire perché deve trovare tutte le sotto-sequenze di somiglianza tra le due sequenze.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

5. BLAST
- ricerche per similarità in banca dati

L'allineamento di sequenze è uno strumento estremamente utile è importante in tutti gli ambiti della ricerca biomolecolare moderna. L'allineamento locale o globale di due sequenze omologhe ci permette di studiare come si sono evolute nel corso del tempo e di quali siano i residui o i domini rimasti conservati . In questo modo è possibile ipotizzare quali funzioni possono essere ritenute comuni alle sequenze studiate.

E' però necessario a priori individuare quali siano le sequenze da studiare e sapere che esse sono omologhe. Questa informazione potrebbe non essere disponibile.

Ad esempio, se studiamo un gene **appena sequenziato**, o una proteina da esso codificata per cercarne la funzione, senza ulteriori informazioni dovremmo formulare delle ipotesi preliminari: dovremmo selezionare geni candidati omologhi e verificarne la similarità di sequenza con il gene in oggetto, in maniera da confermare l'ipotesi di una storia evolutiva e di funzioni comuni.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- ricerche per similarità in banca dati

Gli strumenti informatici oggi a disposizione dei ricercatori permettono di **automatizzare** eseguire nel giro di pochi minuti un'analisi come quella scritta nella slide precedente.

Possiamo allineare la sequenza che vogliamo studiare con tutte le sequenze note dello stesso tipo allo scopo di determinare, sulla base dei risultati, le sequenze che possano essere considerate omologhe ad essa, o comunque che presentino qualsiasi tipo di similarità.

Ad esempio per studiare una sequenza di amminoacidi l'analisi può essere effettuata allineando la sequenza con ciascuna delle seguenti contenute in una banca dati di proteine.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- ricerche per similarità in banca dati

Il processo è automatizzato, ovvero non è l'utente che deve scaricare le sequenze della banca dati, ma è il programma che si occupa di calcolare ciascun allineamento recuperando una per una le sequenze della banca dati.

Terminato l'allineamento i risultati vengono presentati in ordine decrescente di similarità ovvero dalle sequenze più simili e quindi con più elevata probabilità di essere omologhe, fino a quelle meno simili.

Questo procedimento è noto come **ricerca per similarità** in banca dati (**similarity search**). La sequenza di partenza è detta sequenza **query**.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- ricerche per similarità in banca dati

Un programma che esegue una ricerca in banca dati per similarità deve quindi calcolare in pochi minuti migliaia migliaia di allenamenti a coppie di sequenze, ovvero l'allineamento di ciascuna delle sequenze della banca dati con la sequenza query.

Gli algoritmi di allineamento di sequenze locali e globali visti finora sono estremamente veloci, richiedendo una frazione di secondo per l'allineamento di coppie di sequenze di centinaia di residui.

Tuttavia, nonostante la loro velocità, la loro applicazione diretta e sequenziale a centinaia di migliaia di allineamenti richiederebbe ugualmente un tempo di esecuzione dell'ordine delle ore o dei giorni a seconda della dimensione della banca dati.

Per poter velocizzare ulteriormente la procedura i moderni algoritmi utilizzati per le ricerche in banca dati basate sulla similarità usano dei **procedimenti euristici.**

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Definizione dell'algoritmo euristico

I procedimenti euristici sono una classe di algoritmi di ricerca che utilizzano delle euristiche, ovvero delle regole generali o "buone pratiche" per trovare una **soluzione approssimata** a un problema di ottimizzazione, quando non è possibile utilizzare un approccio esatto.

L'idea alla base dei procedimenti euristici è quella di identificare le soluzioni migliori, pur **non garantendo la soluzione ottimale.**

I procedimenti euristici sono spesso utilizzati in problemi di ottimizzazione complessi, in cui le soluzioni esatte sono difficili o impossibili da trovare in modo efficiente.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

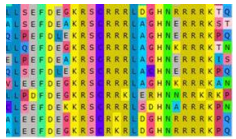
- Blast

Blast (Basic Local Alignment Search Tool) è lo strumento considerato lo standard di fatto per le ricerche di similarità in banche dati di sequenze.

È infatti incluso nelle principali banche dati mondiali di sequenze come quelle curate dalle e dall' NCBI (https://www.ncbi.nlm.nih.gov/) e dall'EBI (https://www.ebi.ac.uk/).

Non è necessario avere Blast in locale sul proprio computer ma si può utilizzare attraverso delle interfacce web collegate con le banche dati. il Blast è un algoritmo euristico basato sull'allineamento locale. Esso non esegue esaustivamente l'algoritmo tradizionale di Smith-Waterman per ognuna delle sequenze in banca dati, ma utilizza un metodo euristico che permette di selezionare a priori un sottoinsieme delle sequenze della banca dati che soddisfino certi criteri.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

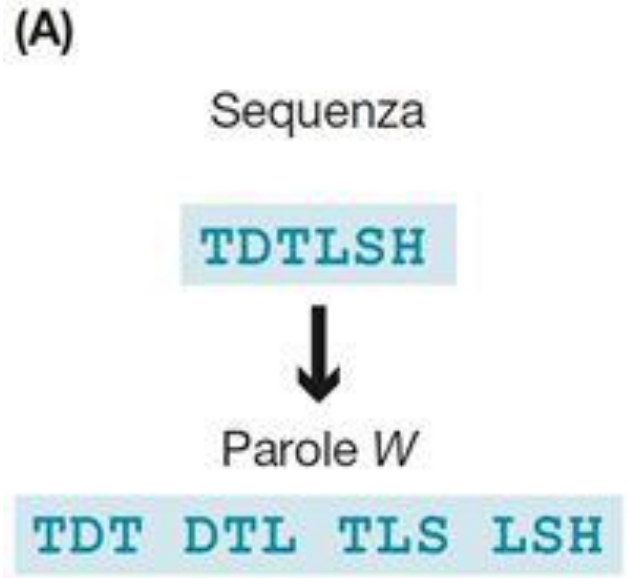


- Come funziona il processo di allineamento

Data una sequenza query da utilizzare per la ricerca e una banca dati di sequenze il primo passo che compie l'algoritmo di Blast è il seguente:

A. esegue la suddivisione della sequenza query in tutte le possibili parole o frammenti sovrapposte di una data lunghezza W.
Ogni sequenza viene suddivisa in parole di lunghezza W (per esempio W = 3) sovrapposte in quanto ottenute spostandosi di un residuo alla volta.

Le parole di lunghezza 3 che si formano sono quelle mostrate in figura.



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

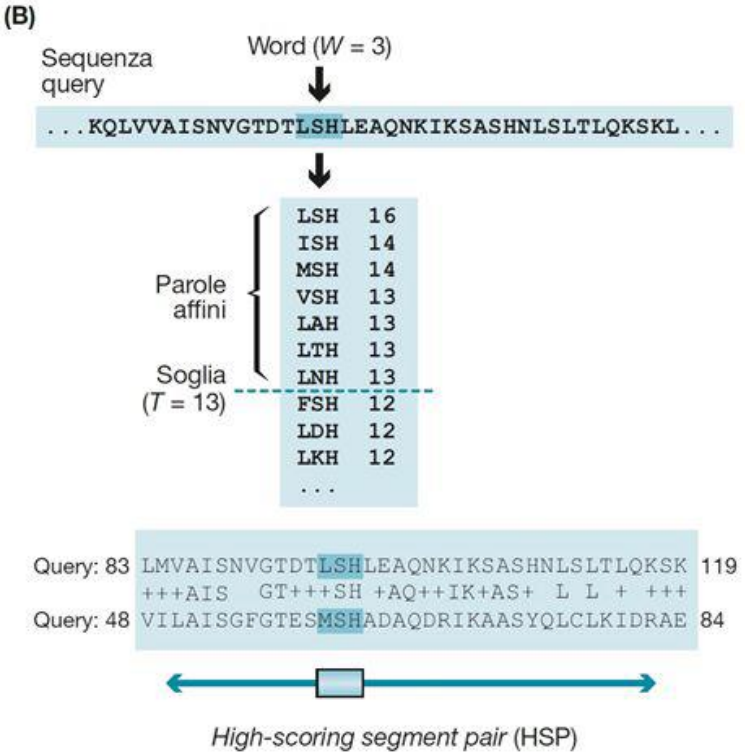
- Come funziona il processo di allineamento

Il secondo passo dell'algoritmo Blast e al seguente:

B. per ognuna delle parole generate viene creato un elenco di parole affini, dette **W-mer**, ottenute attraverso l'allineamento senza gap.
Nella lista vengono considerate le parole che hanno un valore di allineamento superiore a T con la parola di riferimento.

Ad esempio alla parola LSH corrisponde la lista di parole affini riportata in figura, ottenuta utilizzando la matrice Blosum62 è una soglia T = 13.

In corrispondenza della parola LSH ci sono altre 6 triplette (parole affini) che superano la soglia fissata per l'allineamento

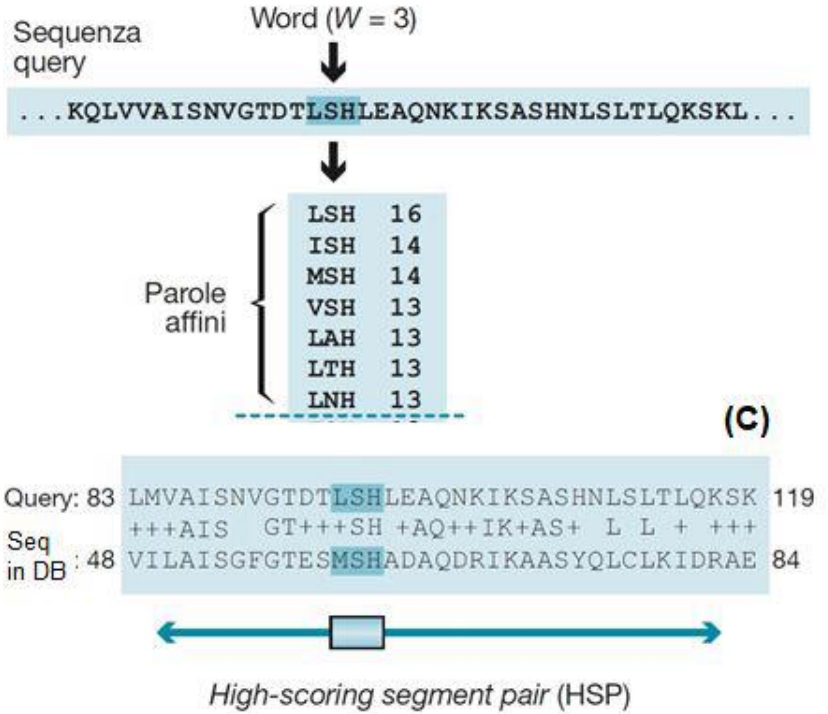


ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Come funziona il processo di allineamento

Completato il passo precedente, l'algoritmo selezionerà dalla banca dati, per allinearle con la query, soltanto quelle sequenze che contengono almeno un frammento di lunghezza W uguale ad uno dei W-mer prodotti esaminando tutta la sequenza query.

Le corrispondenze trovate tra W-mer della sequenza query e frammenti di lunghezza W delle sequenze della banca dati (dette hit) vengono utilizzate dall'algoritmo come punto di partenza per la costruzione degli allineamenti.
Le corrispondenze trovate dovranno essere incluse nell'allineamento finale. Questo verrà costruito partendo dagli hit ed estendendo opportunamente sia a monte sia valle.



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Come funziona il processo di allineamento

I 4 passi dell'algoritmo di Blast

1: Query Preprocessing

Break query into words

DTLVRAIP -> DTL, TLV, LVR, VRA ...

Make a table of similar words

TLV -> TLI, TIV, SLV

2: Search query words in indexed table of database words

Find exact match between table word & db.

TIV

SDTDGDKNADGWIETIVRALPTSD

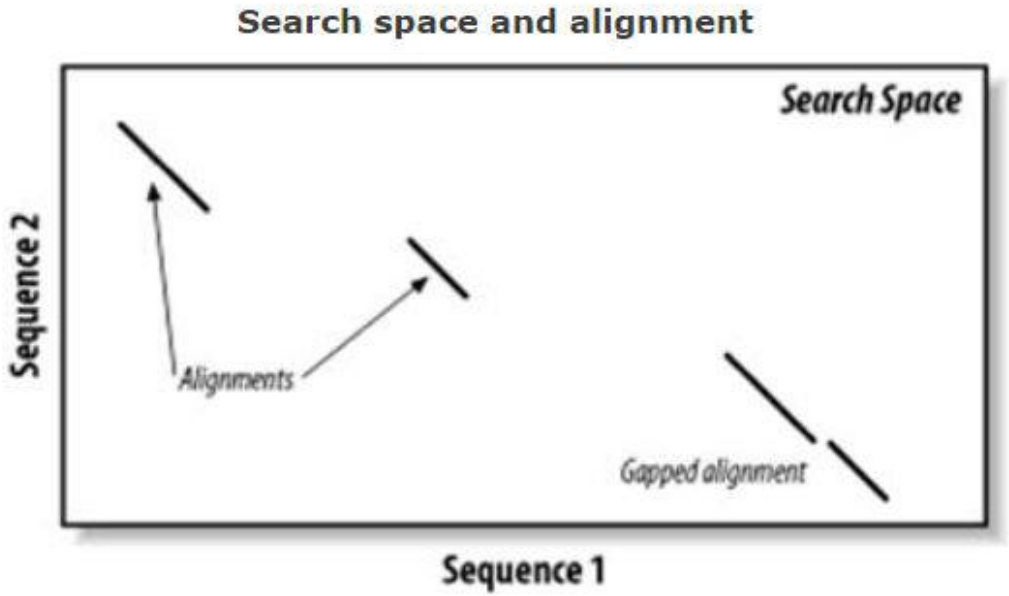
3: Extend query match to HSP

– keep HSPs of significant quality.

DTLVRAIP

SDTDGDKNADGWIETIVRALPTSD

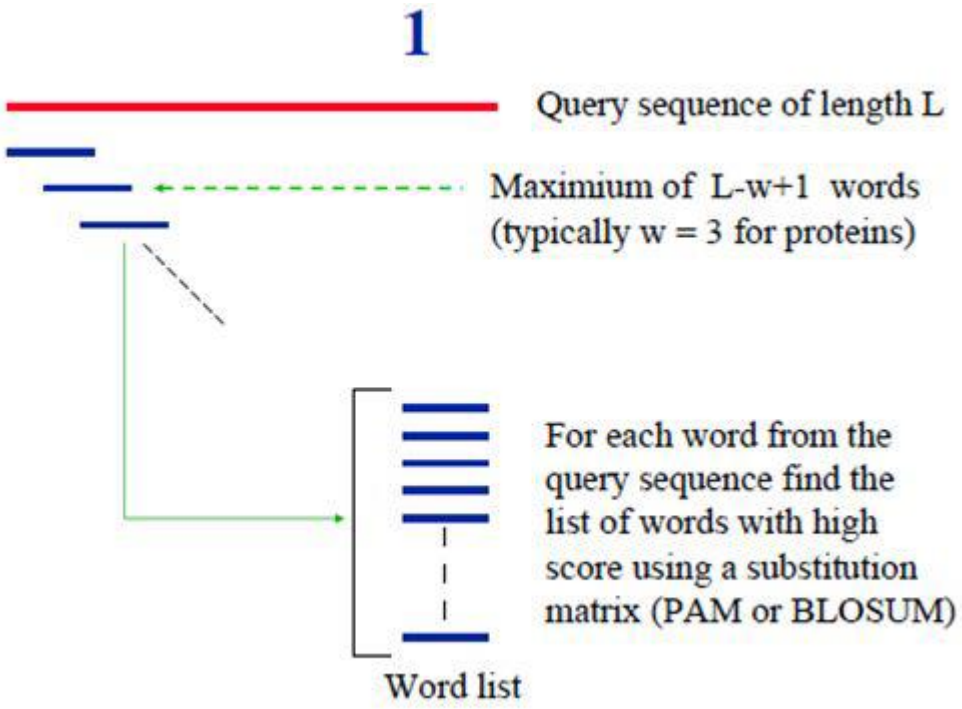
4: Assemble HSPs into gapped alignment



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Essendo un algoritmo "euristico" non abbiamo la certezza "matematica" di trovare le sequenze in banca dati che danno gli allineamenti migliori quando allineate alla Query. Le probabilità di sbagliare sono comunque abbastanza basse da essere accettabili, visto che in cambio riusciamo ad ottenere una risposta in tempi ragionevoli.
- La prima fase consiste nel suddividere la query in piccole "parole" (words) di lunghezza definita, ad es. per le proteine la lunghezza delle parole è tre.
- Per ogni parola "P" cerchiamo nella lista di tutte le parole di uguale lunghezza quelle che allineate senza gap a P diano un punteggio uguale o superiore ad una certa soglia T.
- Ad esempio per le proteine e con parole lunghe 3, dobbiamo cercare quale tra le 8000 (20x20x20) parole di lunghezza 3 abbiano un punteggio superiore a T.

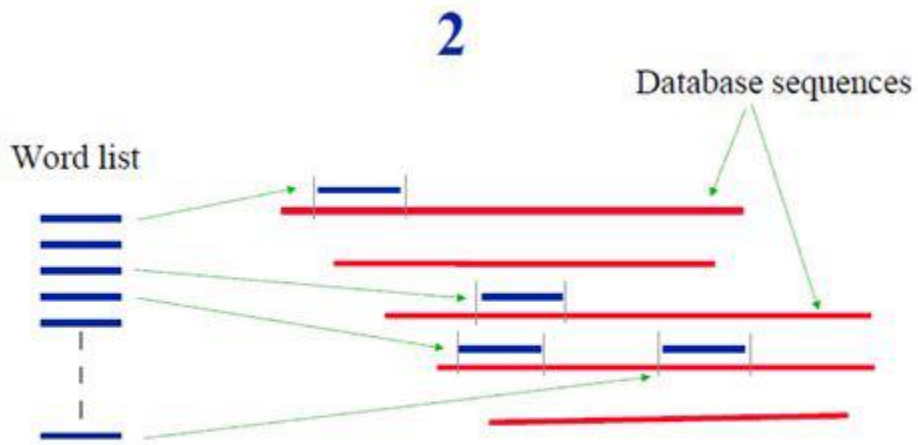
to



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Come funziona il processo di allineamento

- A questo punto abbiamo ottenuto una lista "L" di parole (lunghe 3 per le proteine) che si allineano con buoni punteggi ad almeno una delle parole della mia Query.
- Se io ho un "indice" che mi dice, per ciascuna parola, in quale delle sequenze in banca dati compare, posso già buttar via tutte quelle sequenze in cui non compare mai una delle parole di L.
- L'indice è già pronto, non viene calcolato ogni volta, e viene aggiornato solo quando una nuova sequenza è aggiunta alla bancadati.



Exact matches of words from the word list to the database sequences

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Come funziona il processo di allineamento
  - Per ogni sequenza in banca dati che abbia almeno una parola di L, provo ad "estendere" l'allineamento partendo dal match esatto con la parola e proseguendo su entrambi i lati finché il punteggio dell'allineamento non scende sotto una determinata soglia.
  - Una volta che ho fatto questo per tutte le sequenze in banca dati che abbiano almeno un match con le parole di L posso fare una classifica dei punteggi.
  - Gli allineamenti con punteggi maggiori saranno quelli con sequenze più simili a quella Query.

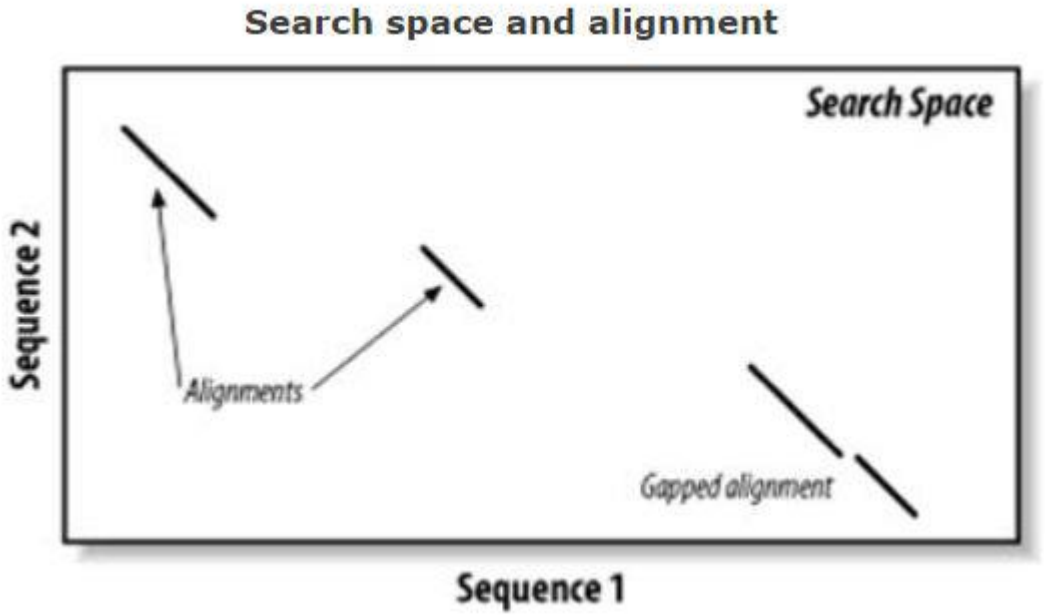
For each exact word match, alignment is extended in both directions to find high score segments

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Come funziona il processo di allineamento

4

Ottenuti gli HSPs come ultimo passo l'algoritmo cerca di estendere l'allineamento tra HSPs introducendo le gaps.



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Suite di programmi blast

Blast utilizza un algoritmo di allineamento di sequenze per cercare somiglianze tra una query (sequenza di interesse) e un database di riferimento (insieme di sequenze già note).

Esistono diverse varianti di Blast, ognuna delle quali è ottimizzata per l'analisi di specifiche tipologie di sequenze:

- **BlastN**: è progettato per l'allineamento di sequenze nucleotidiche. Questo programma confronta la query nucleotidica con un database di sequenze nucleotidiche e restituisce una lista di allineamenti potenziali.

- **BlastP**: è progettato per l'allineamento di sequenze proteiche. Questo programma confronta la query proteica con un database di sequenze proteiche e restituisce una lista di allineamenti potenziali.

- **BlastX**: è progettato per l'allineamento di sequenze nucleotidiche che vengono tradotte in sequenze proteiche. Questo programma confronta la query nucleotidica con un database di sequenze proteiche tradotte e restituisce una lista di allineamenti potenziali.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Suite di programmi blast

Esistono anche altre due varianti del programma Blast chiamate tblastx e tblastn, che sono progettate per l'allineamento di sequenze miste di nucleotidi e proteine.

- **tblastx**: è progettato per l'allineamento di sequenze nucleotidiche tradotte in tutte le possibili sequenze proteiche in sei frame di lettura. Questo programma confronta la query nucleotidica con un database di sequenze nucleotidiche tradotte e restituisce una lista di allineamenti potenziali.
- **tblastn**: è progettato per l'allineamento di sequenze proteiche con un database di sequenze nucleotidiche. Questo programma traduce la query proteica in tutte le possibili sequenze nucleotidiche e le confronta con un database di sequenze nucleotidiche, restituendo una lista di allineamenti potenziali.

Queste varianti di Blast sono utilizzate quando si hanno sequenze che non sono facilmente confrontabili con le altre tre varianti di Blast. Ad esempio, tblastx può essere utilizzato per confrontare sequenze nucleotidiche non tradotte con un database di sequenze proteiche, mentre tblastn può essere utilizzato per confrontare sequenze proteiche con un database di sequenze nucleotidiche.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Suite di programmi blast

BLASTN programs search nucleotide databases using a nucleotide query.

BLASTP programs search protein databases using a protein query.

BLASTX search protein databases using a translated nucleotide query.

TBLASTN search translated nucleotide databases using a protein query.

TBLASTX search translated nucleotide databases using a translated nucleotide query.