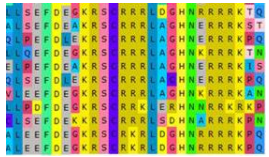
ALLINEAMENTI TRA SEQUENZE

7. Ancora su BLAST
- Limiti e sfide di Blast

7. Allineamento di sequenze contro genomi
- Blat e Ssaha
- Algoritmi di read mappers: Bowtie e BWA
- Formato SAM

7. Allineamento multiplo di sequenze
- Clustal
- SAGA e RAGA

ALLINEAMENTI TRA SEQUENZE

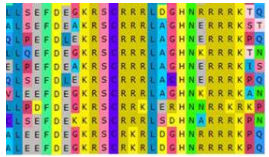


- Limiti e sfide di Blast:

Pur essendo uno degli strumenti di allineamento di sequenze più utilizzati in bioinformatica ci sono alcune limitazioni e sfide nell'uso di Blast che possono influenzare i risultati dell'allineamento.

1. Sequenze altamente **simili**: BLAST può avere difficoltà a distinguere tra sequenze altamente simili, poiché le sequenze condividono molte regioni identiche. In tali casi, BLAST potrebbe non essere in grado di distinguere tra le due sequenze o di identificare le regioni di divergenza.
2. **Lunghezza** della sequenza: la lunghezza della sequenza può influenzare la precisione di BLAST. Sequenze troppo corte possono non fornire informazioni sufficienti per l'allineamento corretto, mentre sequenze troppo lunghe possono richiedere troppo tempo per l'allineamento.

ALLINEAMENTI TRA SEQUENZE



- Limiti e sfide di Blast:

3. **Sovrapposizione** di domini proteici: BLAST può avere difficoltà a rilevare le sovrapposizioni tra i domini proteici all'interno di una singola sequenza. In tali casi, BLAST può non essere in grado di identificare la corretta struttura proteica o di riconoscere le relazioni di omologia tra sequenze.

4. **Bias** di composizione: la composizione nucleotidica può influenzare la precisione di BLAST. Sequenze con una forte composizione di un particolare nucleotide possono produrre falsi positivi o falsi negativi nell'allineamento.

ALLINEAMENTI TRA SEQUENZE

- Limiti e sfide di Blast:

5. Selezione del **database** di riferimento: la scelta del database di riferimento può influenzare la capacità di BLAST di identificare relazioni di omologia tra le sequenze. Un database di riferimento troppo ampio o troppo piccolo può compromettere la precisione dell'allineamento.

6. Scalabilità: BLAST può essere limitato nella sua capacità di gestire grandi quantità di sequenze, soprattutto quando si tratta di sequenze di dimensioni diverse.

Tuttavia, nonostante queste sfide, BLAST rimane uno strumento utile per l'analisi di sequenze e l'identificazione di relazioni di omologia tra le sequenze.

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

L'allineamento di sequenze contro genomi è una tecnica utilizzata in bioinformatica per comparare le sequenze di DNA o RNA di un organismo o con i genomi di altri organismi correlati o con il suo stesso genoma (ad esempio partendo da un RNA posso localizzare il gene corrispondente sul genoma)

L'obiettivo principale dell'allineamento di sequenze contro genomi è identificare le regioni di conservazione tra i genomi, che possono essere utilizzate per identificare geni, regioni funzionali e differenze genetiche tra gli organismi.

L'allineamento di sequenze contro genomi è diverso dall'allineamento di sequenze tra loro. Mentre l'allineamento di sequenze tra loro è utilizzato per comparare sequenze simili, l'allineamento di sequenze contro genomi confronta sequenze che possono avere **molte differenze**. Inoltre, l'allineamento di sequenze contro genomi richiede l'utilizzo di tecniche avanzate per gestire la complessità dei genomi, come la presenza di regioni non codificanti, sequenze ripetitive e introni.

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

Se la sequenza da allineare è una sequenza di RNA, in questo caso l'allineamento verrà calcolato tra l'RNA ed entrambi i filamenti di DNA della sequenza del genoma.

Ovviamente l'allineamento contro il genoma è un allineamento locale ovvero soltanto una piccola regione del genoma sarà allineata con l'RNA, anche nel caso in cui tutti i nucleotidi dell'RNA corrispondano esattamente alla sequenza del DNA.

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

Quando un RNA viene allineato con il rispettivo genoma di riferimento, è normale aspettarsi riscontrare differenze a livello dei singoli nucleotidi, corrispondenti per esempio ai polimorfismi dell'individuo da cui proviene, o da siti in cui l'RNA è soggetto a **editing**.

Inoltre poiché vengono solitamente sequenziati gli RNA maturi, negli eucarioti superiori il loro allineamento con il genoma comporterà l'introduzione di **gap** in corrispondenza degli **introni** del gene corrispondente.
Infine per gli RNA poliadenilati non verrà trovata una corrispondenza tra i nucleotidi della coda di **poly-A** e la sequenza genomica.

Anche gli algoritmi di allineamento sul genoma sono basati su metodi euristici simili al Blast. Viene prima individuata una corrispondenza esatta di almeno n nucleotidi, che viene estesa prevedendo sostituzioni.

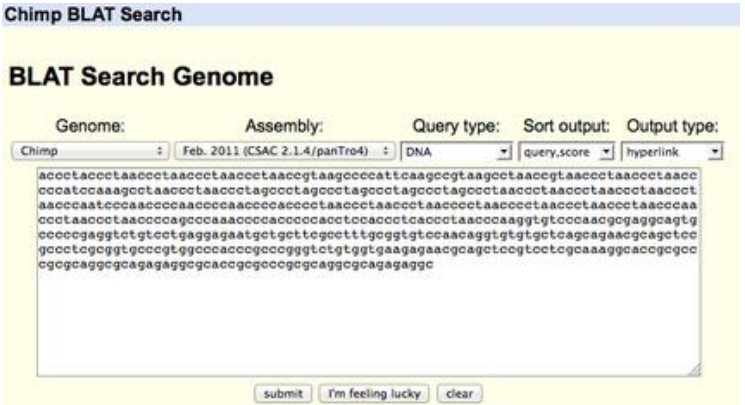
ALLINEAMENTI TRA SEQUENZE

- Blat e Ssaha

Ci sono diversi strumenti di allineamento disponibili per allineare le sequenze contro i genomi, come BLAT e SSAHA.

- BLAT (BLAST-like Alignment Tool) è un algoritmo di allineamento di sequenze simile a BLAST, ma progettato per l'allineamento di sequenze lunghe come i genomi. BLAT utilizza una strategia di "allineamento rapido" basata sulla ricerca di occorrenze di k-mer (sottostringhe di lunghezza k) nella sequenza di query e nel genoma di riferimento. BLAT è molto veloce e può allineare sequenze di dimensioni diverse con alta precisione. Tuttavia, BLAT non è raccomandato per l'allineamento di sequenze altamente simili o per l'identificazione di sequenze ripetitive.

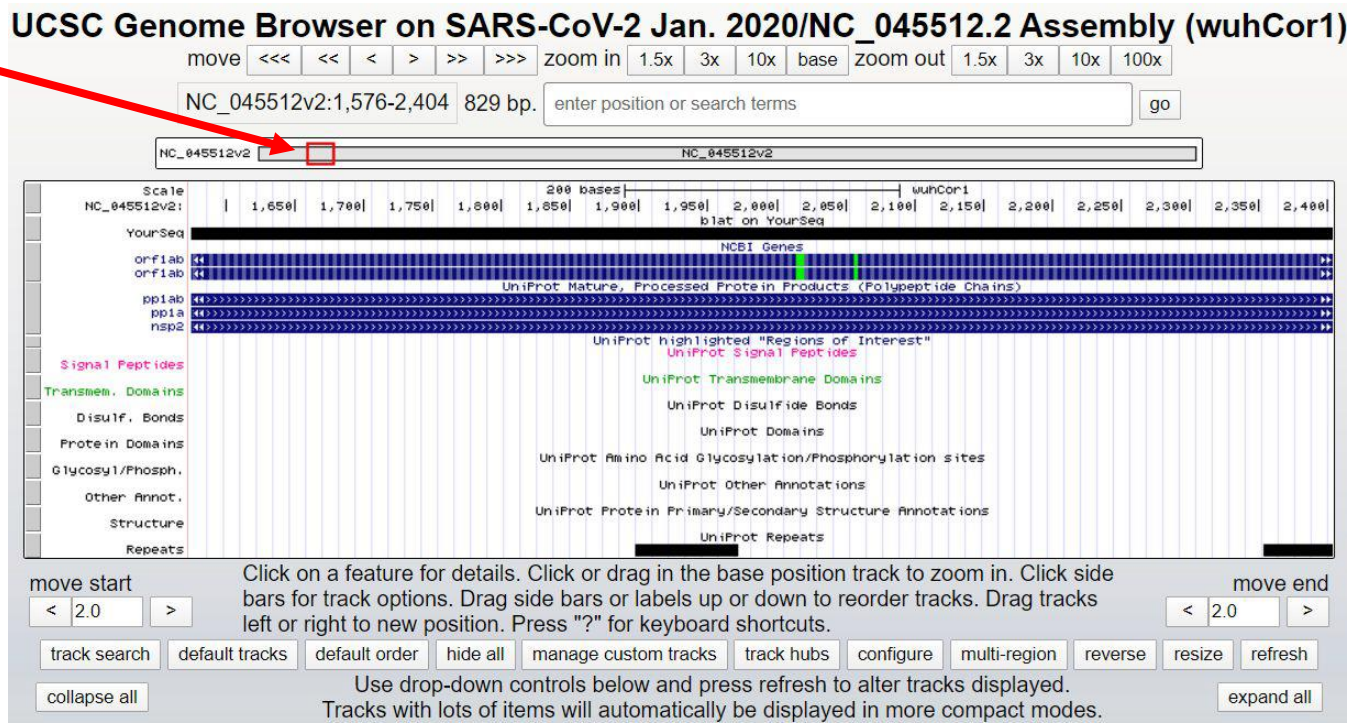
Il browser genomico UCSC (https://genome.ucsc.edu/cgi-bin/hgBlat) include il form per l'allineamento di sequenze a genomi attraverso l'utilizzo dell'algoritmo Blat.



ALLINEAMENTI TRA SEQUENZE

- Blat e Ssaha

Output di Blat:
La sequenza sottomessa nel form mappa nella regione evidenziata con il rettangolo rosso nel genome browser



ALLINEAMENTI TRA SEQUENZE

- Blat e Ssaha

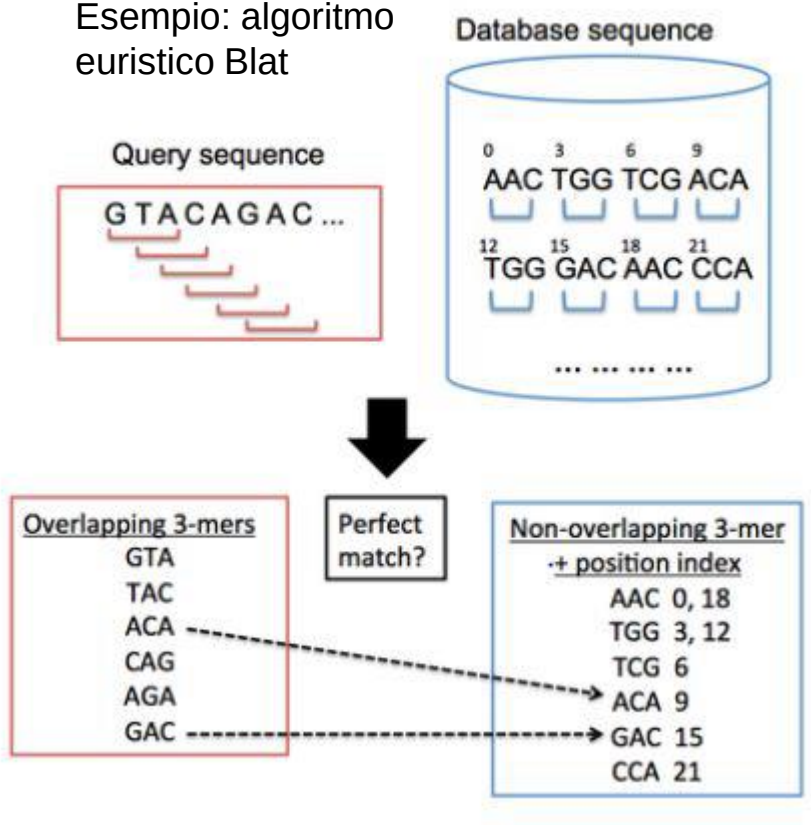
- SSAHA (Sequence Search and Alignment by Hashing Algorithm) è un algoritmo di allineamento di sequenze che utilizza una struttura di dati chiamata "tabella di hash" per effettuare una ricerca efficiente delle sequenze di query nel genoma di riferimento. SSAHA è particolarmente utile per l'allineamento di sequenze di lunghezza inferiore ai 20.000 nucleotidi e per l'identificazione di sequenze altamente simili. SSAHA è stato utilizzato in diversi progetti di sequenziamento del genoma, tra cui il Progetto del Genoma Umano.

In generale, BLAT e SSAHA sono due importanti algoritmi di allineamento di sequenze utilizzati in bioinformatica per allineare sequenze di DNA o RNA di un organismo con i genomi di altri organismi correlati. L'uso di questi algoritmi è importante per identificare le regioni di conservazione tra i genomi, identificare nuovi geni e varianti di geni, e studiare l'evoluzione dei genomi e la biologia molecolare dei caratteri fenotipici.

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

Esempio: algoritmo euristico Blat



Gli algoritmi di allineamento contro genomi e quelli di allineamento locale presentano anche differenze significative dal punto di vista dell'implementazione, in particolare per quanto riguarda l'uso di **indici** per l'accelerazione dell'allineamento.

Questi indici sono basati sulla costruzione di strutture dati che rappresentano il genoma di riferimento, che possono essere interrogate rapidamente per trovare le corrispondenze tra la sequenza di query e il genoma. Gli indici permettono di **ridurre il tempo** di ricerca delle regioni di interesse all'interno del genoma e quindi velocizzano l'allineamento.

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

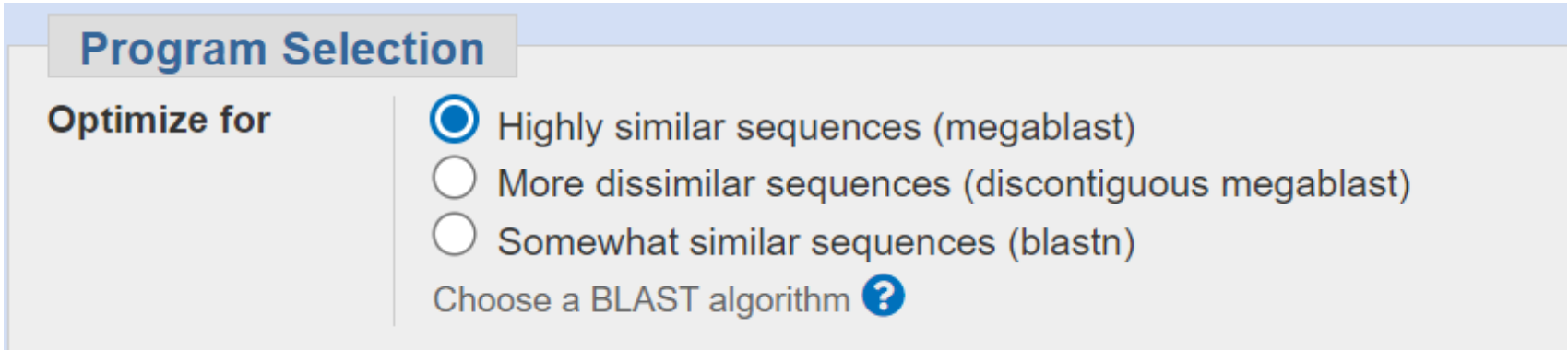
Un'analogia che può servire a chiarire il concetto è quella dell'indice analitico di un libro. Se per esempio volessimo localizzare in quali pagine di un libro compare un dato numero o una data parola invece di leggere dall'inizio alla fine il libro stesso potremmo consultare l'indice analitico in cui le parole più importanti sono riportate, in ordine alfabetico , insieme alle pagine in cui compaiono.

Analogamente un indice per un genoma permette di recuperare, per ogni oligonucleotide di una data lunghezza, tutte le posizioni del genoma in cui esso può essere trovato. Gli algoritmi di allineamento, quindi, consultano innanzitutto l'indice del genoma per tutte le parole di una data lunghezza della sequenza da allineare e per tutte le posizioni ritrovate cercano di estendere l'allineamento

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

Un algoritmo di questo tipo detto **MegaBlast** è una variante di Blast che fa parte dei programmi messi a disposizione sul sito dell'NCBI.
Mega Blast è stato sviluppato per gestire grandi database di sequenze e per trovare corrispondenze di sequenze con alta similarità.
Può essere utilizzato per identificare rapidamente sequenze di interesse all'interno di un vasto database di sequenze.
A differenza di BLAT, MegaBlast è specifico per le sequenze nucleotidiche



ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

SPLING (https://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi) è un programma di allineamento di trascritti al genoma che tiene conto della presenza di **siti di splicing** canonici durante l'introduzione di gap nell'allineamento.

SPLING tiene conto della presenza di siti di splicing canonici durante l'allineamento di trascritti al genoma, al fine di produrre una suddivisione in esoni e introni più accurata e precisa. Ciò consente di identificare con maggiore precisione i limiti dei geni e di generare trascritti completi e funzionali. Inoltre, SPLING è in grado di allineare sequenze di trascritti di diverse lunghezze e di diverse origini, facilitando la comparazione tra i trascritti e il genoma di riferimento.

ALLINEAMENTI TRA SEQUENZE

- Allineamento di sequenze contro genomi

Un'ulteriore importante applicazione gli algoritmi di alliineamento al genoma è derivata dalle tecnologie di sequenziamento di nuova generazione (NGS). Queste sono molto spesso utilizzate nei progetti di **risequenziamento**, il cui scopo è di ottenere la sequenza del genoma (o parte di esso) da un campione proveniente da una specie della quale il genoma è stato già sequenziato e ricostruito.

Lo scopo è quello di identificare differenze tra individui, popolazioni, ceppi, cercare mutazioni somatiche in condizioni patologiche, individuare regioni arricchite in un campione (per es. Copy Number Variation, CNV), ecc.
In questo caso non è necessario assemblare un nuovo genoma, ma si può usare il genoma noto della specie come riferimento.

ALLINEAMENTI TRA SEQUENZE

- Algoritmi di read mappers

Negli ultimi anni parallelamente alla diffusione dei metodi di sequenziamento *next-generation*, è stata sviluppata una serie di algoritmi specifici detti **read mapper** (mappatori) allo scopo di risolvere il problema di dover allineare (mappare) centinaia di milioni di sequenze di poche decine di basi su un genoma in maniera rapida e accurata.

Quindi gli algoritmi di read mappers sono una classe di algoritmi di bioinformatica utilizzati per allineare le sequenze di lettura generate dalle tecnologie di sequenziamento ad alta velocità (ad es. Illumina) ai genomi di riferimento.

Inoltre, questi algoritmi devono anche essere in grado di gestire le regioni ripetitive o altamente variabili presenti nel genoma, che possono causare difficoltà nell'allineamento. Esistono molti algoritmi di read mappers disponibili, tra cui Bowtie (Bowtie2), BWA e molti altri. Ciascun algoritmo ha i propri punti di forza e debolezze, e la scelta dell'algoritmo dipende spesso dalle specifiche esigenze dell'applicazione.

ALLINEAMENTI TRA SEQUENZE

- Algoritmi di read mappers

Come discusso in precedenza, gli algoritmi di allineamento contro genomi spesso utilizzano degli indici specifici, come l'Index di Burrows-Wheeler (BWT) che consente di velocizzare il processo di allineamento.

Bowtie è un algoritmo di allineamento read mapper contro genomi basato sull'Index di Burrows-Wheeler (BWT). Bowtie è stato progettato per allineare sequenze di DNA di breve lunghezza (tipicamente meno di 50 nucleotidi) contro grandi genomi di riferimento, come il genoma umano. Bowtie è stato progettato per essere veloce ed efficiente, e per questo utilizza diverse tecniche di ottimizzazione, come la ricerca bidirezionale e il pruning delle regioni non promettenti del genoma di riferimento.

ALLINEAMENTI TRA SEQUENZE

- Algoritmi di read mappers

BWA (Burrows-Wheeler Aligner) è un altro algoritmo di allineamento, read mapper, contro genomi basato sull'Index di Burrows-Wheeler (BWT).

BWA è stato progettato per allineare sequenze di DNA o RNA di breve e lunga lunghezza contro i genomi di riferimento.

BWA utilizza un approccio di allineamento a due fasi, in cui la prima fase utilizza un algoritmo basato sull'Index di Burrows-Wheeler per trovare le regioni di corrispondenza tra la sequenza di query e il genoma di riferimento, e la seconda fase utilizza un algoritmo di allineamento di Smith-Waterman per allineare accuratamente le regioni trovate.

lh3/**bwa**

Burrow-Wheeler Aligner for short-read alignment
(see minimap2 for long-read alignment)

ALLINEAMENTI TRA SEQUENZE

- Formato SAM

L'output di tutti i software di read mapping ha un formato tabulare standard per descrivere in maniera compatta ma completa allineamenti delle read a un genoma, detto *Sequence Aligment* o **SAM** (BAM nlla sua versione binaria compressa).
A ciascuna read allineata corrisponde una riga nel file, come in questo esempio:

```
r001 163 ref  7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
```

Il formato specifica una serie di colonne, ognuna delle quali contiene una precisa informazione, tra cui:
-identificati della reads
-un codice numerico che indica se la read mappa sul DNA e in caso positivo su quale dei due filamenti
-le coordinate genomiche a cui esso è allineato
-la sequenza della reads stessa

ALLINEAMENTI TRA SEQUENZE

- Formato SAM

Il risultato dell'allineamento vero è proprio è descritto con un formato compatto, detto stringa CIGAR.

```
r001 163 ref   7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
```

In questo formato una serie di caratteri indica quanti nucleotidi della read sono allineati al genoma.

Per esempio la stringa CIGAR dell'esempio sopra indica l'allineamento di una read di 17 nucleotidi del genoma, in cui i primi 8 nucleotidi della read sono allineati a nucleotidi del genoma (8M), seguiti da una inserzione di due nucleotidi (2I), poi altri quattro nucleotidi allineati (4M), poi una delezione di un nucleotide (1D) e infine altri tre allineamenti (3M).

ALLINEAMENTI TRA SEQUENZE

- Formato SAM

Il risultato dell'allineamento vero è proprio è descritto con un formato compatto, detto stringa CIGAR.

M: match/mismatch
I: insertion
D: deletion
P: padding
N: skip
S: soft-clip
H: hard-clip

```
Ref:  GCATTCAGATGCAGTACGC
Read:   ccTCAG--GCAGTAgtg

CIGAR   2S4M2D6M3S
POS     5
```

ALLINEAMENTI TRA SEQUENZE

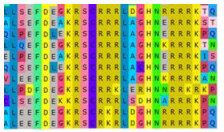
- Formato SAM

Complessivamente questo formato può sembrare di difficile lettura, ma contiene tutte le informazioni necessarie per svolgere analisi successive sul risultato dell'allineamento e viene solitamente letto non dall'occhio umano ma da appositi programmi.

Per esempio si utilizzano software che partendo dal risultato dell'allineamento delle read al genoma interpretano le informazioni associate a ciascuna delle read per identificare tutte le posizioni in cui questi contengono sostituzioni rispetto al genoma, posizioni candidate a contenere degli SNP.

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze



L'allineamento multiplo di sequenze è un'operazione di bioinformatica che consiste nell'allineare contemporaneamente **più sequenze** di DNA o di proteine, anziché confrontare solo due sequenze alla volta, come avviene nell'allineamento globale e locale che abbiamo discusso finora.

Il vantaggio dell'allineamento multiplo deriva dal fatto che consente di analizzare un insieme più ampio di sequenze e di estrarre informazioni più dettagliate sui **rapporti evolutivi** tra le sequenze.

In particolare, l'allineamento multiplo può essere utilizzato per identificare **regioni conservate** tra le sequenze (ad esempio, regioni di un gene che sono presenti in tutte le specie considerate), per studiare la **diversità genetica** tra diversi organismi, o per ricostruire la **filogenesi** di un gruppo di organismi.
Tuttavia, l'allineamento multiplo può essere computazionalmente più oneroso rispetto all'allineamento di due sole sequenze e richiede l'uso di risorse informatiche più potenti.

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

Se analizziamo una sola sequenza, ogni residuo ha lo stesso peso degli altri. I due triptofani (W nel codice a una lettera) presenti nella sequenza in alto non possono essere associati a una maggiore o minore importanza. Se invece abbiamo un allineamento multiplo di sequenze omologhe, ogni residuo viene immediatamente caratterizzato dalla sua maggiore o minore conservazione nelle altre sequenze omologhe. Per esempio, uno dei due W può essere molto conservato mentre l'altro può non essere conservato per nulla.

Una sola sequenza non contiene informazioni sull'importanza relativa dei vari residui

```
VLSAADWTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSA
```

Molte sequenze possono dare **MOLTE** informazioni

```
-VLSAADWTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-----HGSA
-VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNP
VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
-GLSDGEKQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE
  *:     *   * **       * *.* *::  .* *   *  * .*.     ..
```

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

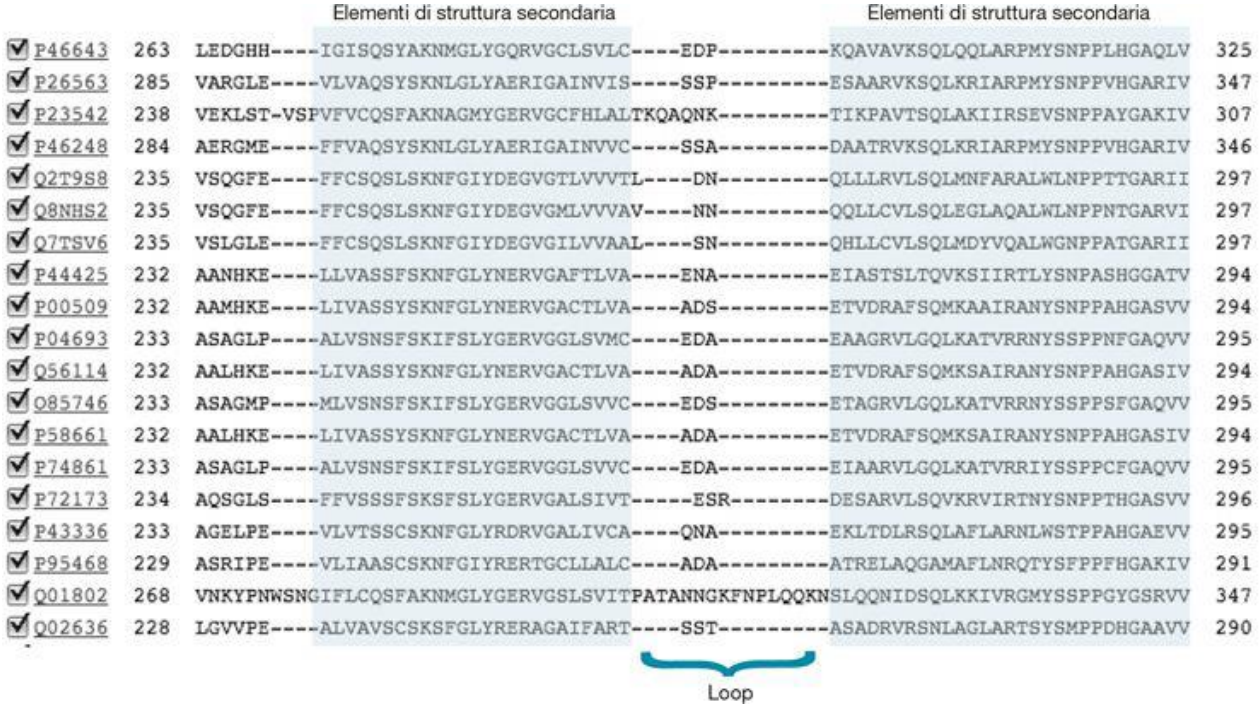


Figura 5.9
Un allineamento multiplo di proteine omologhe contiene importanti informazioni sulla struttura secondaria dei residui che lo compongono. Risulta evidente la possibile identificazione di elementi di struttura secondaria separati da un numero di residui variabile nelle diverse sequenze, e quindi probabilmente appartenenti a un loop. Ulteriori analisi a carico dei residui nei putativi elementi di struttura secondaria possono rendere facilmente identificabile il tipo di struttura secondaria (α-elica o filamento β).

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

L'ipotesi indotta da un allineamento multiplo è che tutti i residui nella stessa colonna siano **evolutivamente correlati.**
L'allineamento multiplo è un allineamento di tipo **globale** di un numero arbitrario di k sequenze.
Supponendo che le seguente da allineare siano tutte della stessa lunghezza n, l'algoritmo

dovrebbe riempire una tabella costituita da n^k celle.
In questo caso la funzione utilizzata per stabilire l'allineamento ottimale tra tutti quelli possibili è una generalizzazione di quella dell'allineamento a coppie ed è data dalla somma dei punteggi degli allineamenti A di tutte le coppie di sequenze i e j indotti dall'allenamento multiplo:

$$S = \sum_{i=1}^{k} \sum_{j \neq i}^{k} A_{i,j}$$

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

Se dovessimo applicare l' algoritmo di Needleman e Wunsch per fare l'allineamento a coppie sulle K sequenze i tempi sarebbero ragionevoli solo per 3-4 sequenze da allineare.
Per ottimizzare le prestazioni anche in questo caso si utilizzano algoritmi euristici per ottenere i risultati in tempi ragionevoli allineando un gran numero di sequenze.
Pur non garantendo di calcolare l'allineamento ottimale l'algoritmo di allineamento multiplo euristico produce comunque allineamenti corretti dal punto di vista biologico ed evolutivo.
In un allineamento globale tra K sequenze di lunghezza fissata n vengono calcolati

$k(k-1)/2$ allineamenti a coppie.

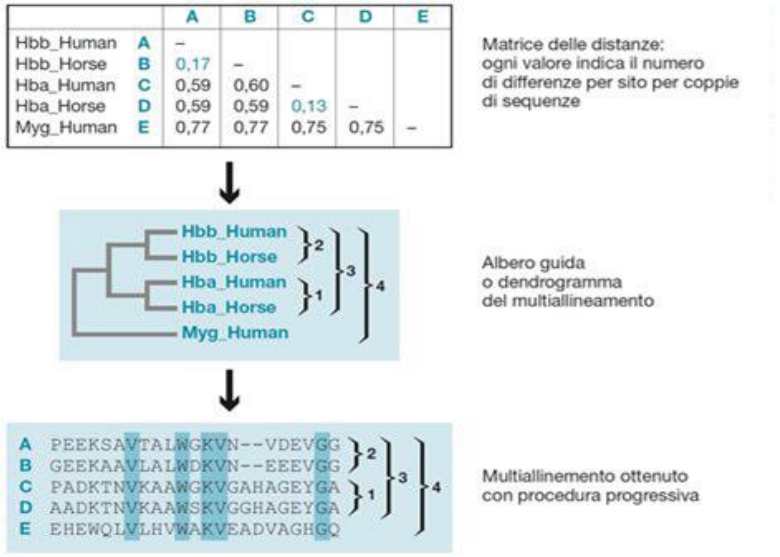
I punteggi risultanti vengono quindi utilizzati per costruire un albero, o dendogramma, che guiderà tutti i passaggi successivi, e quindi l'ordine con cui le sequenze verranno aggiunte all'allineamento multiplo.

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

In figura è mostrato un esempio di allineamento multiplo in cui vengono allineate cinque proteine omologhe appartenenti alla famiglia delle globine.

Dapprima viene calcolato il punteggio dell'allineamento di ogni possibile coppia di sequenze, espresso in figura come numero di sostituzioni per sito. Sulla base di questi punteggi viene costruito l'albero Guida.



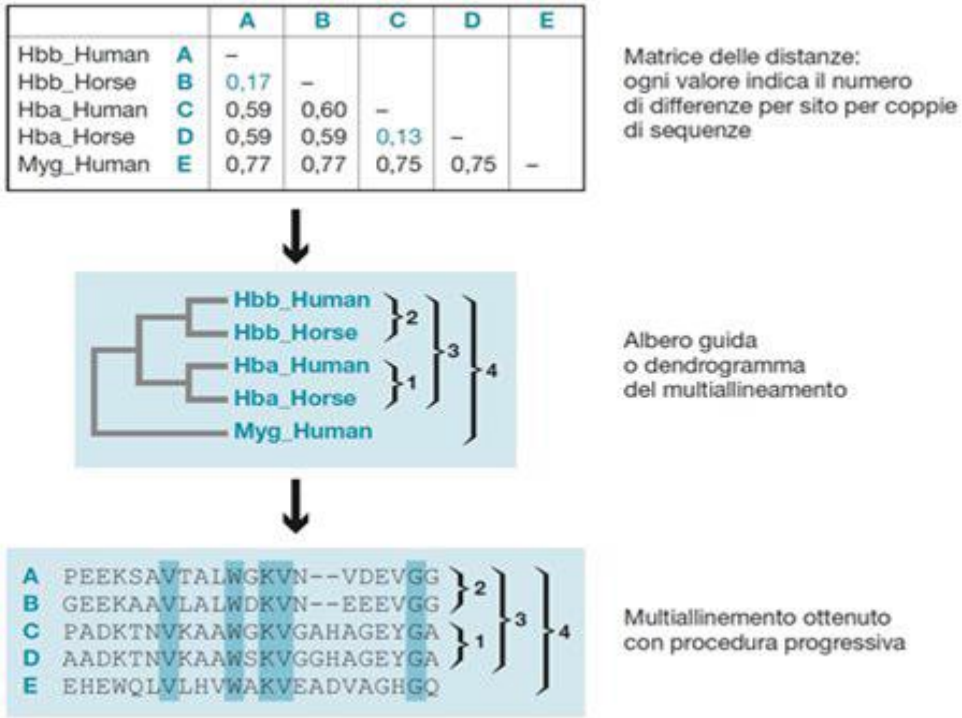
ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

L'allineamento multiplo verrà costruito con i seguenti passi:

1. vengono allineate le sequenze 3 e 4 che costituiscono la coppia più simile. Le due sequenze formano un cluster.
2. vengono allineate le sequenze 1 2, che tra tutte le scelte rimanenti formano la coppia più simile. Queste due sequenze formano un secondo cluster.
3. il cluster prodotto al passo uno viene allineato con quello prodotto al passo due.
4. la sequenza cinque viene allineato al cluster prodotto al passo tre.

Il risultato sarà l'allineamento multiplo di tutte e cinque le Sequenze.



ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

In ciascuno dei passi viene calcolato quindi l'allineamento di una sequenza a un allineamento, oppure l' allineamento di due allineamenti.
Il tempo stimato dell'algoritmo di allineamento multiplo per l'allineamento a coppie di sequenze progressivo è pari a: $O(kn^2)$

In pratica, algoritmi di questo tipo richiedono pochi minuti anche per allineamenti di decine di sequenze.

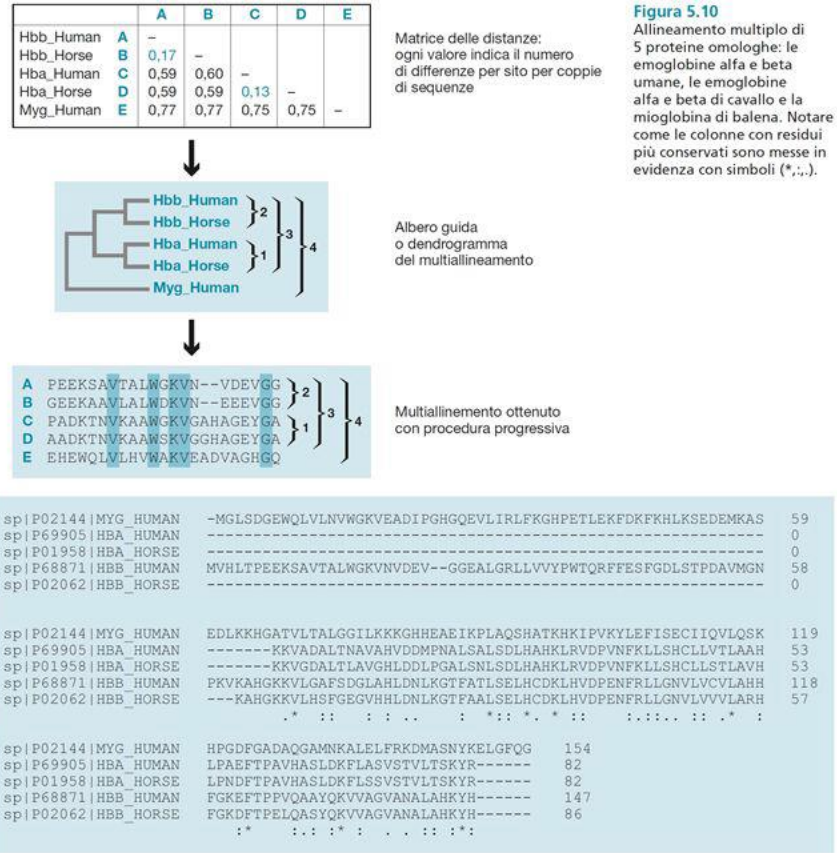


Figura 5.10
Allineamento multiplo di 5 proteine omologhe: le emoglobine alfa e beta umane, le emoglobine alfa e beta di cavallo e la mioglobina di balena. Notare come le colonne con residui più conservati sono messe in evidenza con simboli (*,:,.).

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

La grande diffusione degli algoritmi progressivi di allenamento multiplo ha mostrato come l'euristica di scegliere ad ogni passo la coppia migliore, basata sull'ipotesi di una storia evolutiva, comune tra le due sequenze sia in effetti affidabile e produca allineamenti corretti dal punto di vista biologico.
Poiché alla base di ogni allineamento c'è l'ipotesi che le sequenze studiate abbiano una storia evolutiva comune, se per errore si includesse una sequenza non omologa alle altre, l'allineamento risulterebbe drasticamente alterato è privo di significato biologico in quanto cercherebbe legami evolutivi tra sequenze che non ne hanno.

L'allineamento multiplo può risultare inoltre problematico quando le sequenze che si vogliono allineare sono di lunghezza molto diversa per la presenza di grosse inserzione e delezioni.

ALLINEAMENTI TRA SEQUENZE

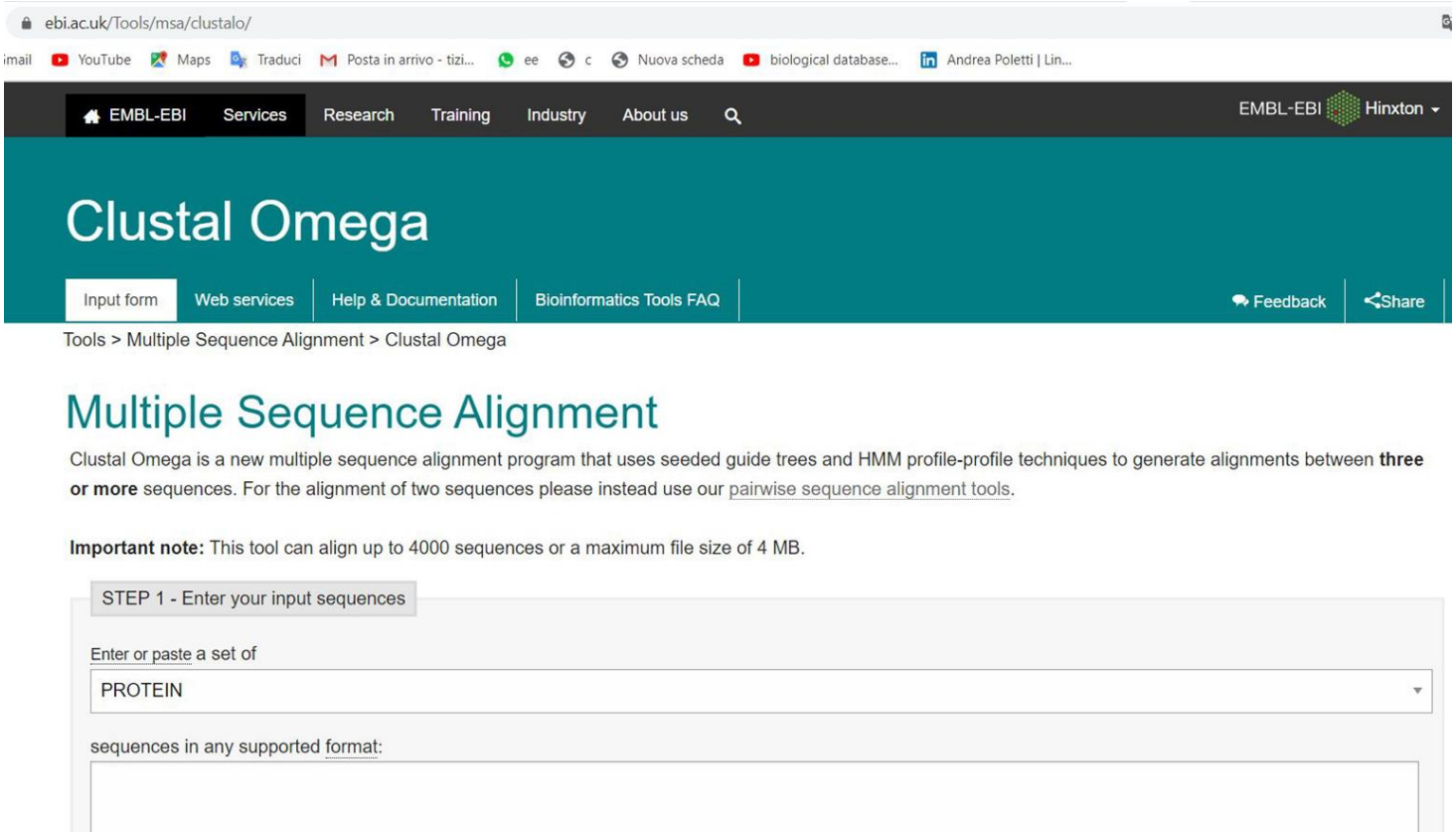
- Allineamento multiplo di sequenze

Esistono ad oggi diversi programmi per la costruzione di allineamenti multipli: CLUSTAL è uno dei più utilizzati in campo bioinformatico.
Tra le sue caratteristiche, CLUSTAL permette anche di aggiungere una o più sequenze a un allineamento precedente determinato e di generare un albero filogenetico basato sull'allineamento multiplo.

Tuttavia, quando le sequenze sono molto divergenti, CLUSTAL potrebbe non fornire sempre la soluzione ottimale per l'allineamento multiplo e possono essere necessari approcci alternativi.

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze



ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

Output di Clustal

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Alignment File | Show Colors

```
CLUSTAL O(1.2.4) multiple sequence alignment


AAI64788.1                --------MENSSEPSRWVSPGLLSPDPLVNYTTDSGLLPPVDVDEPFYS--SSETDLLP      50
NP_002040.1               MEFPGLGSLGTSEPLPQFVDPALVSSTP------ESGVFFPSGPEGLDAAASSTAPSTAT      54
sp|P15976.1|GATA1_HUMAN   MEFPGLGSLGTSEPLPQFVDPALVSSTP------ESGVFFPSGPEGLDAAASSTAPSTAT      54
                                 : .*.   ::*.*.*:*  *      :**:: * . :    :  *:  .


AAI64788.1                SYYSTSVQSRASSSYRHSPVRQVYSSQSILGNIQWLDNSAGHSLNSPYNPTSTVWSSSPF      110
NP_002040.1               AAAAALAYYRDAEAYRHSPVFQVYPLLNCMEGIPG---------GSPYA----GWAY---      98
sp|P15976.1|GATA1_HUMAN   AAAAALAYYRDAEAYRHSPVFQVYPLLNCMEGIPG---------GSPYA----GWAY---      98
                          :  :: .  * :.:****** ***   . : .*           .***      *:


AAI64788.1                AKTPLHSHTSTSIYQNTATPSFTSPKEGFPSPSRDGKDSPRLQEVLKAERLSPMSGSGSS      170
NP_002040.1               GKTGL--------YPASTVCPTREDSPPQAVEDLDGKGSTSFLETLKTERLSPDLLT---      147
sp|P15976.1|GATA1_HUMAN   GKTGL--------YPASTVCPTREDSPPQAVEDLDGKGSTSFLETLKTERLSPDLLT---      147
                          .** *        *  ::.    . .      . ***.*  : *.**:*****   :


AAI64788.1                FLNLSSATGGVYGPSPHMLSPYGSYMSTSQDYSSAALYSTGGPWMSPSSYSPKLRNKMRL      230
NP_002040.1               ---L--------GPALPSSLPVPNSAYGGPDFSSTFFSPTGSPLNSAAYSSPKLRGTLPL      196
sp|P15976.1|GATA1_HUMAN   ---L--------GPALPSSLPVPNSAYGGPDFSSTFFSPTGSPLNSAAYSSPKLRGTLPL      196
                             *        **:     *  .    . *:**: :  **.*  * :  *****..: *


AAI64788.1                SPPEARECVNCGATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIRPKKRLIVSKRAGT      290
NP_002040.1               PPCEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGQNRPLIRPKKRLIVSKRAGT      256
sp|P15976.1|GATA1_HUMAN   PPCEARECVNCGATATPLWRRDRTGHYLCNACGLYHKMNGQNRPLIRPKKRLIVSKRAGT      256
                           * ******************* ****************************************


AAI64788.1                QCANCHTSTTTLWRRNASGEPVCNACGLYFKLHNVNRPLTMKKDGIQTRNRKVSNRNKKG      350
NP_002040.1               QCTNCQTTTTTLWRRNASGDPVCNACGLYYKLHQVNRPLTMRKDGIQTRNRKASGKGKKK      316
sp|P15976.1|GATA1_HUMAN   QCTNCQTTTTTLWRRNASGDPVCNACGLYYKLHQVNRPLTMRKDGIQTRNRKASGKGKKK      316
                          **:**:*:***********:*********:***:*******:**********.*.:.**


AAI64788.1                KKNAASEVY----PDMSH--------------------MAPPDEHVGAYSLNPGPLLSYS      386
NP_002040.1               RGSSLGGTGAAEGPAGGFMVVAGGSGSGNCGEVASGLTLGPPGTAH--LYQGLGPVVLSG      374
sp|P15976.1|GATA1_HUMAN   RGSSLGGTGAAEGPAGGFMVVAGGSGSGNCGEVASGLTLGPPGTAH--LYQGLGPVVLSG      374
                          : .: . .    *  ..                   :.**.       . **::  .
```

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

Output di Clustal

Results for job clustalo-I20200420-205637-0878-27645928-p1m

Alignments | Result Summary | Guide Tree | Phylogenetic Tree | Results Viewers | Submission Details

Download Guide Tree Data

Phylogram

Branch length: Cladogram Real

Guide Tree

```
(
AAI64788.1:0.319613
,
(
NP_002040.1:0
,
sp|P15976.1|GATA1_HUMAN:0
):0.319613
)
;
```

https://www.youtube.com/watch?v=tq-bWYVnQCM

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

EMBL-EBI Services Research Training Industry About us EMBL-EBI Hinxton

MUSCLE

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback | Share

Tools > Multiple Sequence Alignment > MUSCLE

Multiple Sequence Alignment

MUSCLE stands for **MU**ltiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of sequences in any supported format:

ALLINEAMENTI TRA SEQUENZE

- Allineamento multiplo di sequenze

Esistono anche metodi per il calcolo di allineamenti multipli non basati sull'allineamento progressivo.
Ad esempio i programmi SAGA (sequence alignment by genetic algorithm) e RAGA (RNA alignment by genetic algorithm) che esplorano lo spazio di tutti i possibili allineamenti valutando la funzione obiettivo definito in precedenza. Questi programmi possono essere utili per allineamenti di sequenze molto divergenti.

I programmi specializzati per l'allineamento multiplo di sequenze di RNA, come RAGA, tiene conto della conservazione della struttura secondaria delle sequenze. Questi programmi possono essere particolarmente utili per analisi di RNA, dove la struttura tridimensionale delle molecole gioca un ruolo importante nella loro funzione biologica.