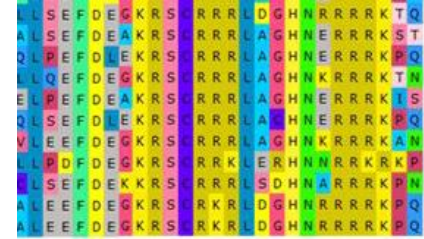


ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

BLAST

- ricerche per similarità in banca dati
- Definizione dell'algoritmo euristico
- Come funziona il processo di allineamento
- Suite di programmi blast: BlastN, BlastP, BlastX



Implementazione del BLAST

- Utilizzo di euristiche per accelerare la ricerca di allineamenti significativi
- Descrizione delle euristiche più comuni (ad esempio, algoritmi seed-and-extend)
- Esempio di applicazione del BLAST

Concetti chiave dell'algoritmo BLAST (Basic Local Alignment Search Tool)

1. Database e Sequenza di Interrogazione:

BLAST confronta una sequenza (query) con un database di sequenze. La query è tipicamente una sequenza di interesse che un ricercatore desidera confrontare con sequenze note per trovare somiglianze.

2. Identificazione di Parole:

BLAST inizia identificando brevi parole di lunghezza fissa (k-mers) nella sequenza query. Queste parole hanno lunghezza 'k' (solitamente 3 per le proteine e 11 per i nucleotidi).

2. Creazione del dizionario:

Dopo l'identificazione delle brevi parole (k-mer) nella sequenza query, BLAST procede con la creazione di un dizionario di parole affini. Questo passaggio coinvolge l'uso di una matrice di punteggio, come BLOSUM62 per le proteine, per identificare e includere nella ricerca quelle parole del database che hanno un grado di somiglianza al di sopra di un punteggio soglia, con i k-mer trovati nella sequenza query.

Questo processo aiuta a identificare possibili corrispondenze significative che non sono esatte, ma sufficientemente simili da meritare ulteriori indagini.

I passaggi 2 e 3 sono importanti per l'efficienza di BLAST, poiché evita di confrontare l'intera sequenza di interrogazione con tutte le sequenze nel database.

4. Estensione:

Una volta identificato un seed, BLAST tenta di estendere questa corrispondenza in entrambe le direzioni per creare un allineamento, fermandosi quando il punteggio dell'allineamento scende sotto una certa soglia. Questo passaggio è cruciale poiché aiuta a identificare somiglianze biologiche significative oltre la corrispondenza iniziale del seed.

5. Punteggio e Valutazione:

Ogni allineamento viene valutato in base al numero di corrispondenze, disallineamenti e spaziature (gaps). Il punteggio viene calcolato usando una matrice di sostituzione che assegna punteggi per la sostituzione degli amminoacidi (nelle sequenze proteiche) o delle sostituzioni dei nucleotidi (nelle sequenze di DNA/RNA).

L'algoritmo calcola anche parametri statistici come il valore E, che aiuta a stimare il numero di corrispondenze che ci si può aspettare di trovare per caso quando si cerca in un database di una determinata dimensione.

6. Filtraggio e Rapporto:

Infine, BLAST filtra i risultati, mantenendo solo quegli allineamenti che soddisfano criteri specifici come punteggio minimo

L'E-value esprime il numero di volte che ci si aspetta di trovare un allineamento con un determinato punteggio (o superiore) per puro caso, quando si effettua una ricerca contro un database di dimensioni specifiche.

Un E-value basso indica che l'allineamento trovato è altamente significativo, cioè è improbabile che sia avvenuto per caso.

Al contrario, un E-value alto suggerisce che l'allineamento potrebbe non essere significativo e potrebbe essere risultato dal caso.

Ad esempio un E-value di 1 indica che ci si aspetta di trovare una corrispondenza con un punteggio simile per caso una volta in un database delle dimensioni date. Un E-value di 0,01 indica che ci si aspetta una corrispondenza casuale con quel punteggio o superiore solo una volta su 100 ricerche.

6. Filtraggio e Rapporto:

nel filtraggio e nel report dei risultati di BLAST, si mantengono gli allineamenti che hanno un E-value basso, il che indica che sono meno probabili essere il risultato di casualità e quindi più probabili essere biologicamente significativi. Gli allineamenti con un E-value alto, al contrario, sono più probabili essere coincidenze casuali e quindi meno probabili rappresentare una corrispondenza biologicamente rilevante.

6. Filtraggio e Rapporto:

nel filtraggio e nel report dei risultati di BLAST, si mantengono gli allineamenti che hanno un E-value basso, il che indica che sono meno probabili essere il risultato di casualità e quindi più probabili essere biologicamente significativi. Gli allineamenti con un E-value alto, al contrario, sono più probabili essere coincidenze casuali e quindi meno probabili rappresentare una corrispondenza biologicamente rilevante.

Il BLAST è quindi progettato per bilanciare velocità e sensibilità. Ciò viene realizzato concentrandosi sui segmenti piccoli più promettenti (seeds) e estendendoli per trovare allineamenti significativi, piuttosto che eseguire ricerche esaustive, che sarebbero computazionalmente intensive.

ALGORITMI DI ALLINEAMENTO

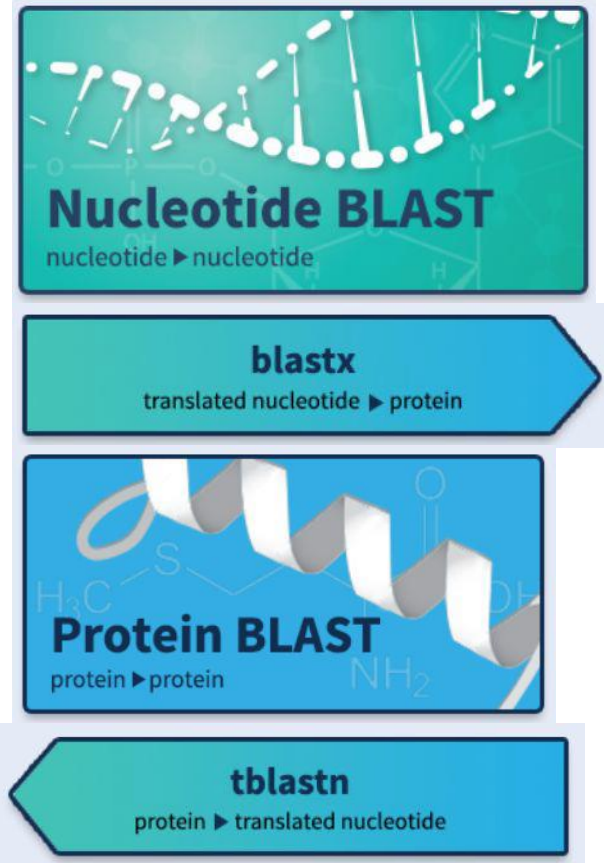
- Suite di programmi blast

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.



ALGORITMI DI ALLINEAMENTO



Gli algoritmi seed-and-extend sono tra le euristiche più comuni utilizzate nei programmi di allineamento di sequenze, tra cui anche Blast.

L'idea alla base dell'algoritmo seed-and-extend è di trovare queste sottosequenze conservate, chiamate "semi-omologhe" o "semi-corrispondenze", e utilizzarle per guidare la ricerca degli allineamenti completi.

In particolare, l'algoritmo ricerca le semi-corrispondenze tra la query e il database di riferimento, utilizzando una funzione di scoring per assegnare un punteggio a ciascuna semi-corrispondenza in base alla sua somiglianza con la sequenza di query.

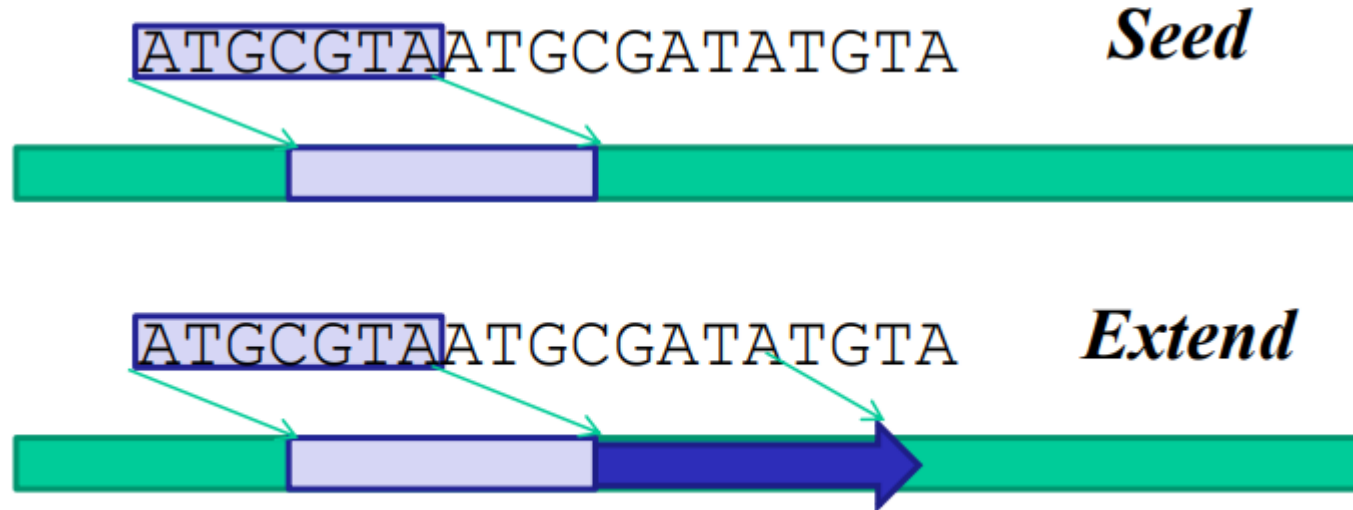
Una volta identificata una semi-corrispondenza, l'algoritmo utilizza la tecnica dell'estensione, ovvero cerca di estendere la semi-corrispondenza per ottenere un allineamento completo. In genere, l'estensione viene eseguita utilizzando un'altra funzione di scoring che tiene conto anche delle lacune (gap) presenti nelle sequenze.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Algoritmi seed-and-extend



Seed-and-Extend



ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



Le quattro fasi di una ricerca BLAST

1. Scegli la sequenza (query)
2. Selezionare il tipo di programma BLAST
3. Selezionare il database per la ricerca
4. Scegliere i parametri opzionali Quindi fare clic su "BLAST"

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



1. Scegli la sequenza (query)
2. Selezionare il tipo di programma BLAST
3. Selezionare il database per la ricerca
4. Scegliere i parametri opzionali

The screenshot shows the NCBI BLAST search interface. A blue box on the left contains four numbered steps. Arrows point from these steps to the corresponding fields in the interface:

- Step 1 points to the "Enter accession number, gi, or FASTA sequence" field, which contains "NP_000509".
- Step 2 points to the "Choose Search Set" section, specifically the "Database" dropdown menu, which is set to "Non-redundant protein sequences (nr)".
- Step 3 points to the "Program Selection" section, specifically the "Algorithm" dropdown menu, which is set to "blastp (protein-protein BLAST)".
- Step 4 points to the "BLAST" button and the "Algorithm parameters" section.

The interface also includes a "Job Title" field with the text "[NP_000509:beta globin [Homo sapiens]]", an "Entrez Query" field, and a "Show results in a new window" checkbox.

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



Passo 1: Scelta della sequenza La sequenza può essere inserita in formato FASTA o come accession number (RefSeq)

Esempio di formato FASTA per una query BLAST

NCBI Resources How To My NCBI Sign

Protein
Translations of Life

Search: Protein Limits Advanced search Help

Search Clear

Display Settings FASTA Send to: Change region shown

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

GenPept Graphics

>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDVGGGALGRLLVVPWTQRFESFGDLSTPDVVMGNPKVKAHGKKVLG
AFSDGLAHLDDLNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAAYQKVAVGAN
ALAHKYY

Analyze this sequence
Run BLAST
Identify Conserved Domains
Find in this Sequence

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



Passo 2:

Passo 2: Scegli il programma

- Scegli il programma BLAST
- blastn (nucleotide BLAST)
 - blastp (protein BLAST)
 - blastx (BLAST tradotto n ► P)
 - tblastn (BLAST tradotto p ► N...)
 - tblastx (BLAST tradotto n ► P... P ► N)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#)

- | | | |
|---|--|--|
| ☐ Human | ☐ Oryza sativa | ☐ Gallus gallus |
| ☐ Mouse | ☐ Bos taurus | ☐ Pan troglodytes |
| ☐ Rat | ☐ Danio rerio | ☐ Microbes |
| ☐ Arabidopsis thaliana | ☐ Drosophila melanogaster | ☐ Apis mellifera |

Basic BLAST

Choose a BLAST program to run.

[nucleotide blast](#)

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast

[protein blast](#)

Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast

[blastx](#)

Search **protein** database using a **translated nucleotide** query

[tblastn](#)

Search **translated nucleotide** database using a **protein** query

[tblastx](#)

Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- ☐ Search [trace archives](#)
- ☐ Find [conserved domains](#) in your sequence (cds)
- ☐ Find sequences with similar [conserved domain architecture](#) (cdart)
- ☐ Search sequences that have [gene expression profiles](#) (GEO)
- ☐ Search [immunoglobulins](#) (IgBLAST)
- ☐ Search for [SNPs](#) (snp)
- ☐ Screen sequence for [vector contamination](#) (vecscreen)
- ☐ [Align](#) two sequences using BLAST (b2seq)

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST

Passo 3: scegliere il database



Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTn programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#) [Query subrange](#) [?](#)

NM_002049 From 76 To 1317

Or, upload file [Scogli file](#) Nessun file selezionato [?](#)

Job Title NM_002049-Homo sapiens GATA binding protein... [?](#)

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database ☒ Standard databases (nr etc.): ☐ rRNA/ITS databases ☐ Genomic + transcript databases ☐ Betacoronavirus

Organism [Optional](#)

Exclude [Optional](#)

Limit to [Optional](#)

Entrez Query [Optional](#)

Program Selection

Optimize for

Nucleotide collection (nr/nt) [?](#)

Nucleotide collection (nr/nt) [?](#)

Reference RNA sequences (refseq_rna) [?](#)

RefSeq Representative genomes (refseq_representative_genomes) [?](#)

RefSeq Genome Database (refseq_genomes) [?](#)

Whole-genome shotgun contigs (wgs) [?](#)

Expressed sequence tags (est) [?](#)

Sequence Read Archive (SRA) [?](#)

Transcriptome Shotgun Assembly (TSA) [?](#)

High throughput genomic sequences (HTGS) [?](#)

Patent sequences(pat) [?](#)

PDB nucleotide database (pdb) [?](#)

Human RefSeqGene sequences(RefSeq_Gene) [?](#)

Genomic survey sequences (gss) [?](#)

Sequence tagged sites (dbsts) [?](#)

more dissimilar sequences (discontiguous megablast) [?](#)

Somewhat similar sequences (blastn) [?](#)

Choose a BLAST algorithm [?](#)

- nr = non ridondante (database più generale, ritorna una sequenza e diversi riferimenti in database in cui la stessa è presente)
- refseq = solo sequenze validate
- refseq dbest = database di EST
- dbsts = database di sequenze localizzate gss = genome sequence surveys (BAC, Yac, ecc)
- Altri... pdb, genomi completi, solo sequenze oggetto di brevetti (pat) ecc. ecc.

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST

Fase 4: parametri opzionali

Si può ...

- scegliere l'organismo di ricerca
- attivare filtri
- modificare la matrice di punteggio
- cambiare il valore minimo di affidabilità dei risultati
- modificare la dimensione della parola W

organismo



Entrez!



algoritmo



BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTp programs search protein databases using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#) [To](#)

NP_000509.1

Or, upload file [Browse...](#)

Job Title [Enter a descriptive title for your BLAST search](#)

☐ Align two or more sequences

Choose Search Set

Database [Non-redundant protein sequences \(nr\)](#)

Organism [Optional](#) [Enter organism name or id--completions will be suggested](#) ☐ Exclude [+](#)

[Optional](#) ☐ Models (X/M/X/P) ☐ Uncultured/environmental sample sequences

Entrez Query [Optional](#) [Enter an Entrez query to limit search](#)

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

[Choose a BLAST algorithm](#)

BLAST Search database **Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)** **4**

☐ Show results in a new window

[Algorithm parameters](#)

E' possibile limitare la ricerca con sintassi Entrez, come per esempio Smith[auth]

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST

Fase 4a: parametri opzionali di blastp



Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☒ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: ☐ Low complexity regions

Mask: ☐ Mask for lookup table only
☐ Mask lower case letters

BLAST Search database Non-redundant protein sequences (nr) using Blastp
☐ Show results in a new window

Soglia attesa

Valore massimo di incertezza richiesto ai risultati

W

Dimensione della parola usata dall'algoritmo di BLAST

Matrice di punteggio

Filtri & Maschere

Esclude parti della query perché non informative (i.e. ripetute) o non significative

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST

Fase 4a: parametri opzionali di blastn



Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☒ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 28

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

Filters and Masking

Filter: ☐ Low complexity regions
☐ Species-specific repeats for: Human

Mask: ☒ Mask for lookup table only
☐ Mask lower case letters

BLAST Search database Human G+T using Megablast (Optimize for highly similar sequences)
☐ Show results in a new window

Soglia attesa

Valore massimo di incertezza richiesto ai risultati

W

Dimensione della parola usata dall'algoritmo di BLAST

Punteggio

Determina il punteggio per match o mismatch

Filtri & Maschere

Esclude parti della query perché non informative (i.e. ripetute) o non significative

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



BLAST output: parte superiore

BLAST Basic Local Alignment Search Tool

My NCBI [Sign In] [Register]

Home Recent Results Saved Strategies Help

NCBI/BLAST/blastp suite/Formatting Results - GS1F74BK011

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#)

NP_000509:beta globin [Homo sapiens]

Query ID	gi 4504349 ref NP_000509.1	Database Name	nr
Description	beta globin [Homo sapiens] >gi 55635219 ref XP_508242.1 PREDICTED: hypothetical protein [Pan troglodytes] >gi 56749856 sp P68871.2 HBB_HUMAN RecName: Full=Hemoglobin subunit beta; AltName: Full=Hemoglobin beta chain; AltName: Full=Beta-	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
		Program	BLASTP 2.2.22+ Citation

hemoglobin, beta [synthetic construct]
>gi|189053145|dbj|BAG34767.1| unnamed protein
product [Homo sapiens]

Molecule type
Query Length 147

tassonomia

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#) [Multiple alignment](#) **NEW**

Graphic Summary

[Show Conserved Domains](#)

Putative conserved domains have been detected, click on the image below for detailed results.

Query seq. 1 25 50 75 100 125 147

Specific hits

Superfamilies

heme-binding site

globin

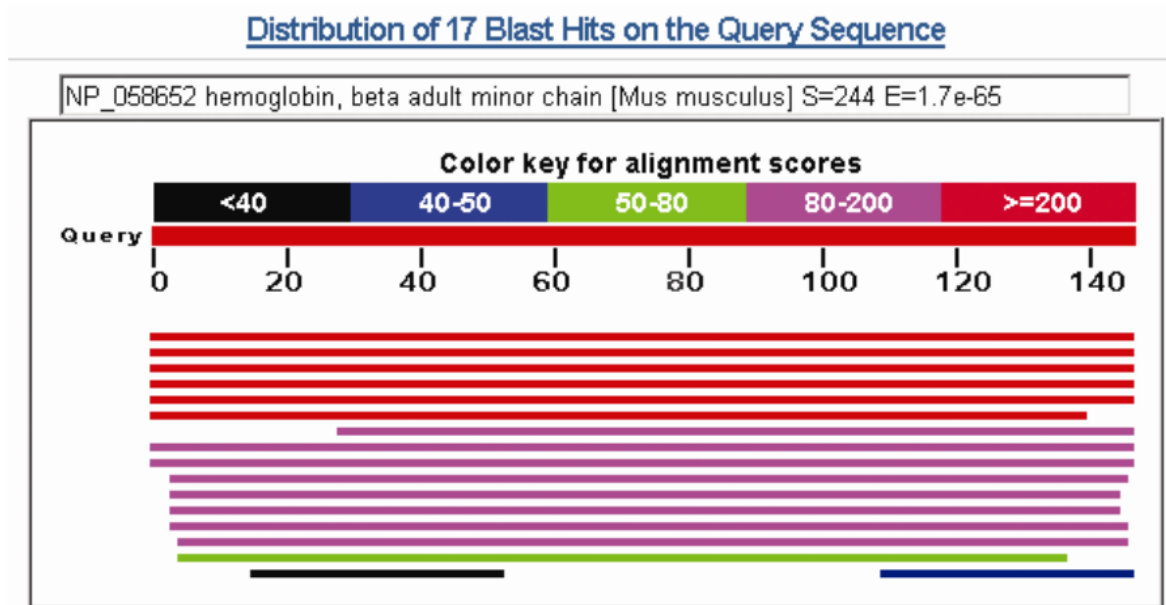
globin_like superfamily

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



BLAST output: output grafico



ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



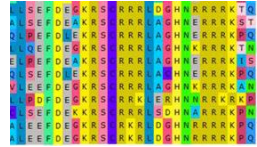
I risultati del Blast conterranno una classifica delle sequenze del database ordinate secondo valori crescenti di **E-value** rispetto alla sequenza query. Il numero di sequenze riportate sarà uguale al numero massimo N impostato nella pagina di input del programma se ci sono almeno N sequenze nella banca dati il cui allineamento con la sequenza query a E-value minore di 1.

Per ognuna delle sequenze della banca dati inclusi nella classifica vengono riportate le seguenti informazioni:

1. descrizione della sequenza e relativo numero di accesso della banca dati
2. punteggio dell'allineamento con la query
3. query coverage, ovvero, trattandosi di allineamenti locali, quale percentuale di residui della query è stata inclusa nell' allineamento locale con la sequenza della banca dati
4. E-value associato al punteggio dell'allineamento
5. percentuale di identità tra la sequenza e la query

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NP_000184.1	sonic hedgehog protein preproprotein [Homo sapiens] >gi 6094283 sp Q15465	941	941	100%	0.0	100%	UGM
NP_002122.2	indian hedgehog protein precursor [Homo sapiens]	459	459	94%	3e-160	59%	UGM
NP_066382.1	desert hedgehog protein preproprotein [Homo sapiens]	442	442	91%	2e-153	56%	UGM
NP_284941.2	mitofusin-1 [Homo sapiens]	31.2	31.2	17%	0.044	29%	UGM
NP_002572.2	pappalysin-1 preproprotein [Homo sapiens]	31.2	31.2	9%	0.048	34%	UGM
NP_001093861.1	RNA-binding Raly-like protein isoform 1 [Homo sapiens]	29.6	29.6	7%	0.096	37%	UGM
NP_001093862.1	RNA-binding Raly-like protein isoform 2 [Homo sapiens] >ref NP_001093863.1	29.6	29.6	7%	0.10	37%	UGM
NP_031293.2	RNA-binding protein Raly isoform 2 [Homo sapiens]	29.6	29.6	7%	0.11	38%	UGM
NP_057951.1	RNA-binding protein Raly isoform 1 [Homo sapiens]	29.6	29.6	7%	0.11	38%	UGM
NP_001013653.1	heterogeneous nuclear ribonucleoprotein C-like 1 [Homo sapiens]	27.2	27.2	7%	0.42	35%	GM
NP_001139653.1	heterogeneous nuclear ribonucleoprotein C-like [Homo sapiens]	27.2	27.2	7%	0.43	35%	UGM
NP_000278.3	peroxisome biogenesis factor 6 [Homo sapiens]	27.2	27.2	12%	0.47	38%	UGM
NP_001130033.2	heterogeneous nuclear ribonucleoprotein C-like [Homo sapiens]	27.2	27.2	7%	0.55	35%	UGM
NP_002580.2	protocadherin-7 isoform a precursor [Homo sapiens]	26.6	26.6	15%	1.2	29%	UGM
NP_115832.1	protocadherin-7 isoform b precursor [Homo sapiens]	26.6	26.6	15%	1.3	29%	GM
NP_001166994.1	protocadherin-7 isoform d precursor [Homo sapiens]	26.6	26.6	15%	1.4	29%	UGM
NP_115833.2	protocadherin-7 isoform c precursor [Homo sapiens]	26.2	26.2	15%	1.5	29%	UGM
NP_002128.2	interleukin-12 subunit beta precursor [Homo sapiens]	25.8	25.8	20%	1.5	22%	UGM
NP_057715.2	GC-rich sequence DNA-binding factor 1 isoform 1 [Homo sapiens]	26.2	26.2	6%	1.5	50%	UGM
NP_037461.2	GC-rich sequence DNA-binding factor 1 isoform 2 [Homo sapiens]	25.8	25.8	6%	1.8	50%	UGM
NP_275859.3	immunoglobulin superfamily member 22 [Homo sapiens]	25.8	25.8	11%	1.9	28%	UGM
NP_079060.1	zinc finger and BTB domain-containing protein 3 [Homo sapiens]	25.4	25.4	6%	2.8	44%	UGM
NP_570969.2	protein FAM71B [Homo sapiens]	25.0	25.0	9%	3.4	35%	UGM
NP_005215.1	AT-rich interactive domain-containing protein 3A [Homo sapiens]	24.3	24.3	5%	6.3	46%	UGM
NP_682820.2	uncharacterized protein C1orf177 isoform 1 [Homo sapiens]	23.9	23.9	11%	7.0	36%	UGM

Figura 5.6

Nel cerchio con linea continua gli allineamenti che possiamo considerare significativi perché hanno un E-value sufficientemente basso, nel rettangolo con linea tratteggiata quelli non significativi.

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST

BLAST output: valori in dettaglio



Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☐	hemoglobin subunit beta-2 [Mus musculus] >gi 122526 sp P02089.2 HBB2_MOUSE RecName: Full=Hemoglobin subunit beta-2; 303	303	303	100%	3e-103	100%	NP_058652.1
☐	beta-globin [Mus musculus castaneus] >gb ADD52624.1 beta-globin [Mus musculus castaneus] 302	302	302	100%	5e-103	99%	ADD52623.1
☐	beta-globin [Mus macedonicus] >gb ACY03396.1 beta-globin [Mus spicilegus] 301	301	301	100%	1e-102	99%	ACY03387.1
☐	beta-globin [Mus musculus] >gb ABU63284.1 beta-globin [Mus musculus] 301	301	301	100%	1e-102	99%	ABU63263.1
☐	beta-globin [Mus musculus] 300	300	300	100%	2e-102	99%	ABU63273.1
☐	beta-globin [Mus musculus castaneus] 300	300	300	100%	2e-102	99%	ADD52616.1
☐	Hbb-b2 protein [Mus musculus] >gb ACY03382.1 beta-globin [Mus musculus castaneus] >gb ACY03391.1 beta-globin [Mus mus 300	300	300	100%	3e-102	99%	AAH32264.1
☐	beta-2-globin [Mus musculus] 300	300	300	99%	4e-102	100%	BAA77357.1
☐	beta-globin [Mus musculus castaneus] 298	298	298	100%	2e-101	98%	ADD52606.1
☐	beta-2-globin [Mus musculus] >db BAA77356.1 beta-2-globin [Mus musculus] 297	297	297	99%	3e-101	99%	BAA77354.1
☐	haemoglobin beta-2 chain [Mus musculus] 297	297	297	100%	4e-101	97%	CAA32225.1
☐	beta-globin [Mus caroli] 290	290	290	100%	3e-98	95%	ACY03381.1
☐	beta-globin [Mus musculus castaneus] 289	289	289	100%	7e-98	95%	ADD52678.1
☐	hemoglobin beta 2 [Mus caroli] 287	287	287	100%	3e-97	94%	BAG31393.1

Punteggi più alti
E-value più bass

ALGORITMI DI ALLINEAMENTO

- Esempio di applicazione del BLAST



Il valore atteso E: è il numero di allineamenti (high scoring segment pairs o HSP) con punteggio maggiore o uguale a un punteggio S che dovrebbero verificarsi per caso in quella ricerca sul database.

Esso rappresenta la stima di probabilità che l'allineamento osservato non sia dovuto al caso. In altre parole si ha una stima del fatto che l'allineamento osservato sia frutto di una storia evolutiva comune tra le sequenze o se sia semplicemente dovuto al caso.

Si può pensarlo come un **indice di incertezza**.

Un valore E è correlato a un valore di probabilità p.

L'equazione fondamentale che descrive un valore E è:

$$E = KMN e^{-\lambda S}$$

Con: M: lunghezza della query

N: lunghezza della sequenza nel database

S: score

λ, K : parametri che dipendono dallo scoring system (λ) e dal database usato (K)

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



- Il valore di **E** **decrece esponenzialmente con l'aumentare S**. Valori più elevati di S corrispondono a migliori allineamenti e infatti hanno **E values** più bassi.
- Ottenere un allineamento con **E = 1** significa che esiste un altro allineamento con lo stesso score S che è risultato per caso.
- La stessa ricerca, su un **database più piccolo o più grande**, anche se restituisce lo stesso allineamento deve avere un valore di E diverso (ciò dipende da K)

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



Quindi...

L'E-value è un valore che va da 0 al numero totale di sequenze nel database, ed è ancora una volta tanto più piccolo di 1 quanto migliore è l'allineamento.

L'E-value associato a un allineamento locale di punteggio S ottenuto da ricerca per similarità in banca dati deve essere letto come “il numero atteso di sequenze della banca dati che – per caso– producono un allineamento con la query con punteggio maggiore uguale a S ”.

L'E-value è inversamente proporzionale al punteggio dell'allineamento: maggiore è il punteggio, più piccolo è l'E-value

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



Se ritroviamo nell'output una sequenza per cui l'E-value associato al suo allineamento con la query è ≥ 1 , significa che il risultato che stiamo osservando è quello che ci aspetteremmo per caso.

Quindi, possiamo ipotizzare che l'allineamento che osserviamo non sia indicativo di una storia evolutiva comune ma viceversa sia effetto proprio del caso.

Esaminando i risultati di un allineamento blast possiamo quindi fermarci quando nella classifica incontriamo la prima sequenza il cui E-value è > 1 (E-value $> e^{-5}$) in quanto le relative sequenze, almeno dal punto di vista statistico, difficilmente avranno un legame evolutivo o funzionale con la query nonostante Blast abbia determinato degli HSP nel loro allineamento con la query.

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



BLAST output: confrontiamo gli allineamenti: seconda hit

[Download](#) [GenPept](#) [Graphics](#)

beta-globin [Mus musculus castaneus]

Sequence ID: [gb|ADD52623.1|](#) Length: 147 Number of Matches: 1

[▶ See 1 more title\(s\)](#)

Range 1: 1 to 147 [GenPept](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
302 bits(773)	5e-103	Compositional matrix adjust	146/147(99%)	147/147(100%)	0/147(0%)

Query	1	MVHLTDAEKSAVSCWLAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPK	60
Sbjct	1	MVHLTDAEKSAVSCWLAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPK	60
Query	61	VKAHGKKVITAFNEGLKNLDNLKGTTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLG	120
Sbjct	61	VKAHGKKVITAFNEGLRNLNLDNLKGTTFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLG	120
Query	121	KDFTPAQAQAFQKVVGAVATALAHKYH	147
Sbjct	121	KDFTPAQAQAFQKVVGAVATALAHKYH	147

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



BLAST output: confrontiamo gli allineamenti: ultima hit (su 100)

Download GenPept Graphics						
RecName: Full=Hemoglobin subunit beta						
Sequence ID: sp B3EWD2.1 HBB SCICA Length: 146 Number of Matches: 1						
Range 1: 1 to 146 GenPept Graphics ▼ Next Match ▲ Previous Match						
Score	Expect	Method	Identities	Positives	Gaps	
253 bits(647)	7e-84	Compositional matrix adjusted	119/146(82%)	135/146(92%)	0/146(0%)	
Query 2	VHLTDAEKSAVSLWAKVNPDEVGGEALGRLLVVYPWTQRYFDSFGDLSSASAIMGNPKV					61
Sbjct 1	VHL+ EK+A++ LW KVNPD+GGEALGRLLVVYPWTQR+FDSEFGDLSSA+A+MGNPKV					60
Query 62	VHLSADEKNALATLWGKVNPDDELGGEALGRLLVVYPWTQRFDSFGDLSSATAVMGNPKV					60
Query 62	KAHGKKVITAFNEGLKNLDNLKGTFFASLSELHCDKLHVDPENFRLLGNAIVIVLGHHLGK					121
Sbjct 61	KAHGKKV+ +F++GLK+LDNLKGTFFSLSELHCDKLHVDPENFRLLGN +V+VL HLK					120
Query 122	KAHGKKVLDSEFSDGLKHLNLDNLKGTFFSSSELHCDKLHVDPENFRLLGNELVLVLAHLGK					120
Query 122	DFTPAQAQAFQKVVGAVATALAHKYH					147
Sbjct 121	DFTP QAAFQKVVGVA ALAHKYH					146
Sbjct 121	DFTPQVQAQFQKVVGAVANALAHKYH					146

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



Translated BLAST: blastx

❑ Andate al sito nell'NCBI:
<http://www.ncbi.nlm.nih.gov/>

❑ Selezionate la banca dati
dei nucleotidi
e cercate la entry
NM_000546

Allineiamo la sequenza
NM_000546, contro il database
Reference Proteins
(refseq_protein), escludendo
le sequenze proteiche predette
(Models XM/XP) ed i campioni
ambientali (Uncultured
environmental sequences).

blastn blastp **blastx** tblastn tblastx

Enter Query Sequence

BLASTX search protein databases using a translated nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#) [To](#)

NM_000546

Or, upload file [Scegli file](#) Nessun file selezionato [Genetic code](#) Standard (1) [Job Title](#) NM_000546:Homo sapiens tumor protein p53 (TP53),... Enter a descriptive title for your BLAST search [Align two or more sequences](#)

Choose Search Set

Database [Reference proteins \(refseq_protein\)](#) [Organism Optional](#) [exclude](#) [+](#) Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [Exclude Optional](#) ☒ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☒ Uncultured/environmental sample sequences

BLAST Search [database refseq_protein](#) using [Blastx \(search protein databases using a translated nucleotide query\)](#) ☐ Show results in a new window

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▾

Manage Columns ▾

Show

100 ▾

?

☒ select all

100 sequences selected

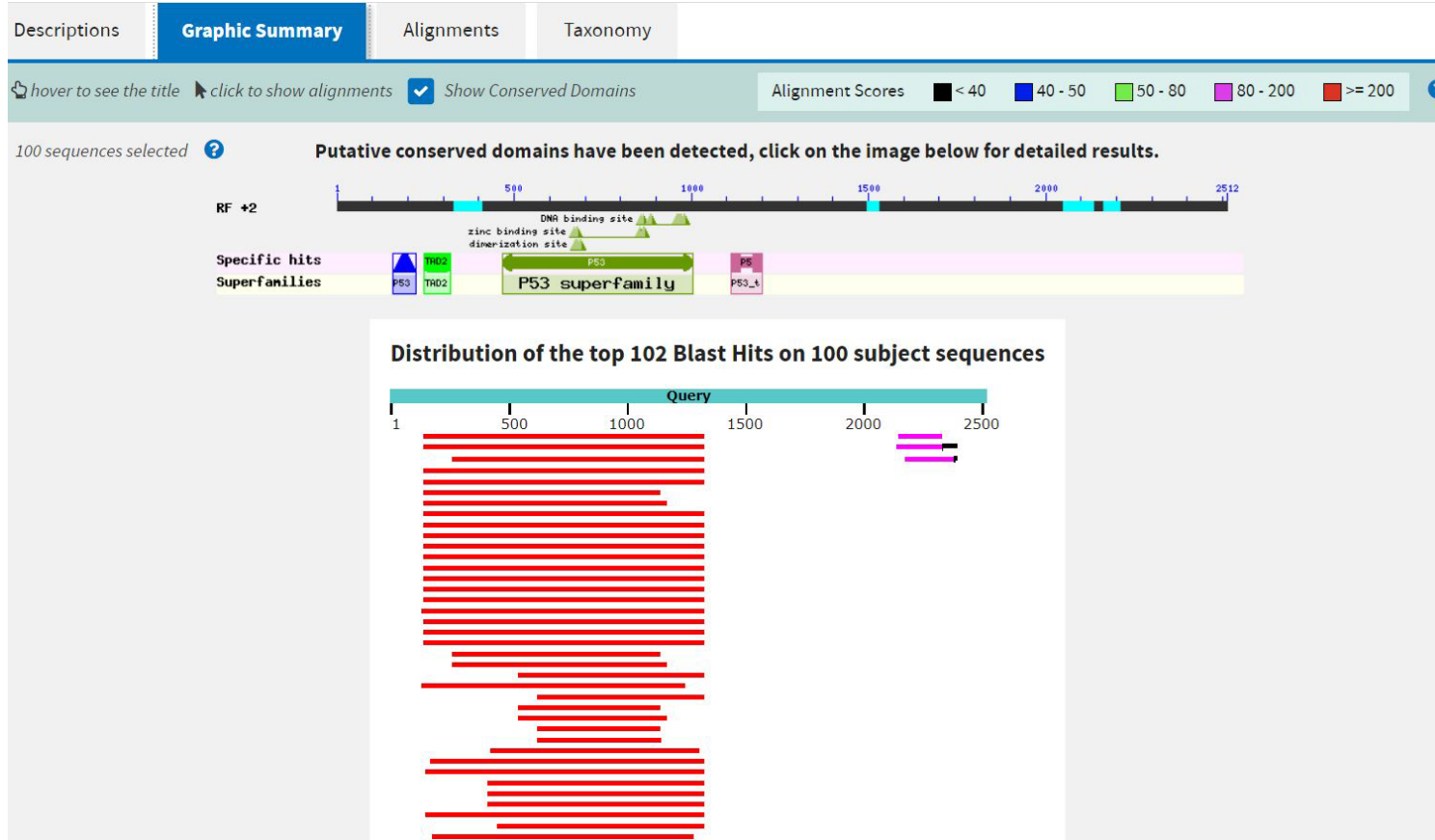
GenPept

Graphics

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☒	cellular tumor antigen p53 [Macaca mulatta]	725	725	46%	0.0	95.67%	NP_001040616.1
☒	cellular tumor antigen p53 [Tupaia chinensis]	695	695	46%	0.0	93.13%	NP_001274298.1
☒	cellular tumor antigen p53 isoform g [Homo sapiens]	674	674	42%	0.0	100.00%	NP_001119590.1
☒	cellular tumor antigen p53 [Nannospalax galii]	649	649	46%	0.0	84.52%	NP_001306217.1
☒	cellular tumor antigen p53 [Oryctolagus cuniculus]	645	645	46%	0.0	86.04%	NP_001075873.1
☒	cellular tumor antigen p53 isoform b [Homo sapiens]	625	625	39%	0.0	100.00%	NP_001119586.1
☒	cellular tumor antigen p53 isoform c [Homo sapiens]	625	625	40%	0.0	97.65%	NP_001119585.1
☒	cellular tumor antigen p53 [Heterocephalus glaber]	625	625	46%	0.0	83.21%	NP_001297199.1
☒	cellular tumor antigen p53 [Equus caballus]	608	608	46%	0.0	82.74%	NP_001189334.1
☒	cellular tumor antigen p53 [Sus scrofa]	606	606	46%	0.0	81.01%	NP_998989.3
☒	cellular tumor antigen p53 [Mesocricetus auratus]	597	597	46%	0.0	78.14%	NP_001268590.1
☒	cellular tumor antigen p53 [Cavia porcellus]	596	596	46%	0.0	78.12%	NP_001166211.1
☒	cellular tumor antigen p53 [Cricetulus griseus]	594	594	46%	0.0	77.16%	NP_001230905.1
☒	cellular tumor antigen p53 [Ovis aries]	581	581	46%	0.0	77.66%	NP_001009403.1
☒	cellular tumor antigen p53 [Felis catus]	581	581	46%	0.0	79.70%	NP_001009294.1
☒	cellular tumor antigen p53 [Bos taurus]	580	580	46%	0.0	78.93%	NP_776626.1
☒	cellular tumor antigen p53 isoform a [Mus musculus]	579	579	47%	0.0	77.27%	NP_035770.2
☒	cellular tumor antigen p53 [Rattus norvegicus]	577	577	46%	0.0	75.51%	NP_112251.2

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST

[Download](#) ▼ [GenPept](#) [Graphics](#)

cellular tumor antigen p53 [Macaca mulatta]

Sequence ID: [NP_001040616.1](#) Length: **393** Number of Matches: **1**

[See 3 more title\(s\)](#) ▼

Range 1: 1 to 393 [GenPept](#) [Graphics](#) [▼ Next Match](#) [▲ Previous Match](#)

	Score	Expect	Method	Identities	Positives	Gaps	Frame
	725 bits(1872)	0.0	Compositional matrix adjust.	376/393(96%)	382/393(97%)	0/393(0%)	+2
Query	143	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPPLSQAMDDLMLSPDDIEQWFTEDPGP				322	
Sbjct	1	MEEPQSDPS+EPPLSQETFSDLWKLLPENNVLSPPLSQ+DDLMLSPDD+ QW TEDPGP				60	
Query	323	DEaprmpaappvapapaaptapaapapSWPLSSSVPSQKTYQGSYGFRGLGFLHSGTAK				502	
Sbjct	61	DEAPRM EAAPP+AP PAAPTPAAPAPAPSWPLSSSVPSQKTY GSYGFRGLGFLHSGTAK				120	
Query	503	SVTCTYSPALNMFQCLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTVEVRRCPHHE				682	
Sbjct	121	SVTCTYSP LNKMFQCLAKTCPVQLWVDSTPPPG+RVRAMAIYKQSQHMTVEVRRCPHHE				180	
Query	683	RCSDSDGLAPPQHLIRVEGNLRVEYLDNRNTRFRHSVVVPYEPPEVGSDCCTIIHNYMCNS				862	
Sbjct	181	RCSDSDGLAPPQHLIRVEGNLRVEY DDRNTRFRHSVVVPYEPPEVGSDCCTIIHNYMCNS				240	
Query	863	SCMGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGDRDRTEENLRKKGEPHHELP				1042	
Sbjct	241	SCMGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGDRDRTEEN RKKGEP H+LP				300	
Query	1043	PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG				1222	
Sbjct	301	PGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEP				360	
Query	1223	GSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD		1321			
Sbjct	361	GSRAHSSHLKSKKGQSTSRHKK MFKTEGPDSD		393			



ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



Reports	Lineage	Organism	Taxonomy			
100 sequences selected ?						
Organism		Blast Name	Score	Number of Hits	Description	
Bilateria		animals		148		
. Deuterostomia		animals		146		
. . Chordata		chordates		145		
. . . Gnathostomata		vertebrates		143		
. . . . Euteleostomi		vertebrates		141		
. Tetrapoda		vertebrates		125		
. Amniota		vertebrates		122		
. Boreoeutheria		placentals		120		
. Euarchontoglires		placentals		105		
. Catarrhini		primates		61		
. Macaca		primates		4		
. Macaca mulatta		primates	725	2	Macaca mulatta hits	
. Macaca fascicularis		primates	725	2	Macaca fascicularis hits	
. Homo sapiens		primates	674	41	Homo sapiens hits	
. Pan paniscus		primates	270	11	Pan paniscus hits	
. Gorilla gorilla gorilla		primates	269	2	Gorilla gorilla gorilla hits	
. Pan troglodytes		primates	268	3	Pan troglodytes hits	
. Tupaia chinensis		placentals	695	1	Tupaia chinensis hits	
. Nannospalax galili		rodents	649	2	Nannospalax galili hits	
. Oryctolagus cuniculus		rabbits & hares	645	1	Oryctolagus cuniculus hits	
. Heterocephalus glaber		rodents	625	1	Heterocephalus glaber hits	
. Mesocricetus auratus		rodents	597	1	Mesocricetus auratus hits	
. Cavia porcellus		rodents	596	1	Cavia porcellus hits	

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



<https://www.youtube.com/watch?v=hTTvq8KJthA>

<https://www.youtube.com/watch?v=HXEpBnUbAMo>

<https://www.youtube.com/watch?v=RzC-V67z5LA&t=64s>

<https://www.youtube.com/watch?v=JKD5laNtwSc&t=76s>

E per finire la lista dei tutorial di ncbi:

<https://www.youtube.com/playlist?list=PLH-TjWpFfWrtjzMCivUe-YbrlleFQIKMq>

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



Rispondete alle seguenti domande relative alla entry **NM_000546** :

- Di che tipo di molecola si tratta e quanto è lunga?
- Quando è stata aggiornata l'ultima volta la entry?
- Cosa dice la sua definizione?
- Da che organismo proviene?
- Quale gene produce questa molecola? Questo gene ha nomi alternativi?
- E' un trascritto codificante? Se sì, dove inizia la 5' UTR? E la 3' UTR? Quanto sono lunghe? E la CDS invece quanto è lunga? E la proteina codificata?
- Riesci a trovare un modo per recuperare la sequenza di NM_000546 in formato FASTA?
- Su che cromosoma si trova il gene che produce questo trascritto?

ALGORITMI DI ALLINEAMENTO A COPPIE DI SEQUENZE

- Esempio di applicazione del BLAST



- Cerca di capire se esistono malattie umane note associate a mutazioni di questo gene.
- Procedi alla entry della proteina prodotta da questo trascritto.
- Qual è identificativo (Accession, o ID) della proteina? La sua definizione?
- Esistono dei siti noti di fosforilazione annotati su questa proteina? In che posizione?
- Ottieni la sequenza in formato FASTA della proteina.
- Vai alla entry del gene che codifica questa proteina.
- Su che filamento è annotato? Ci sono altri geni nelle vicinanze?
- Cerca di capire se il gene produce altri trascritti oltre a quello che codifica per questa proteina.

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.