Lezione 8

Ricerche in banche dati (databases) attraverso l’uso di BLAST

BLAST: Basic Local Alignment Search Tool

Basic Local Alignment Search Tool. Altschul et al. 1990,1994,1997

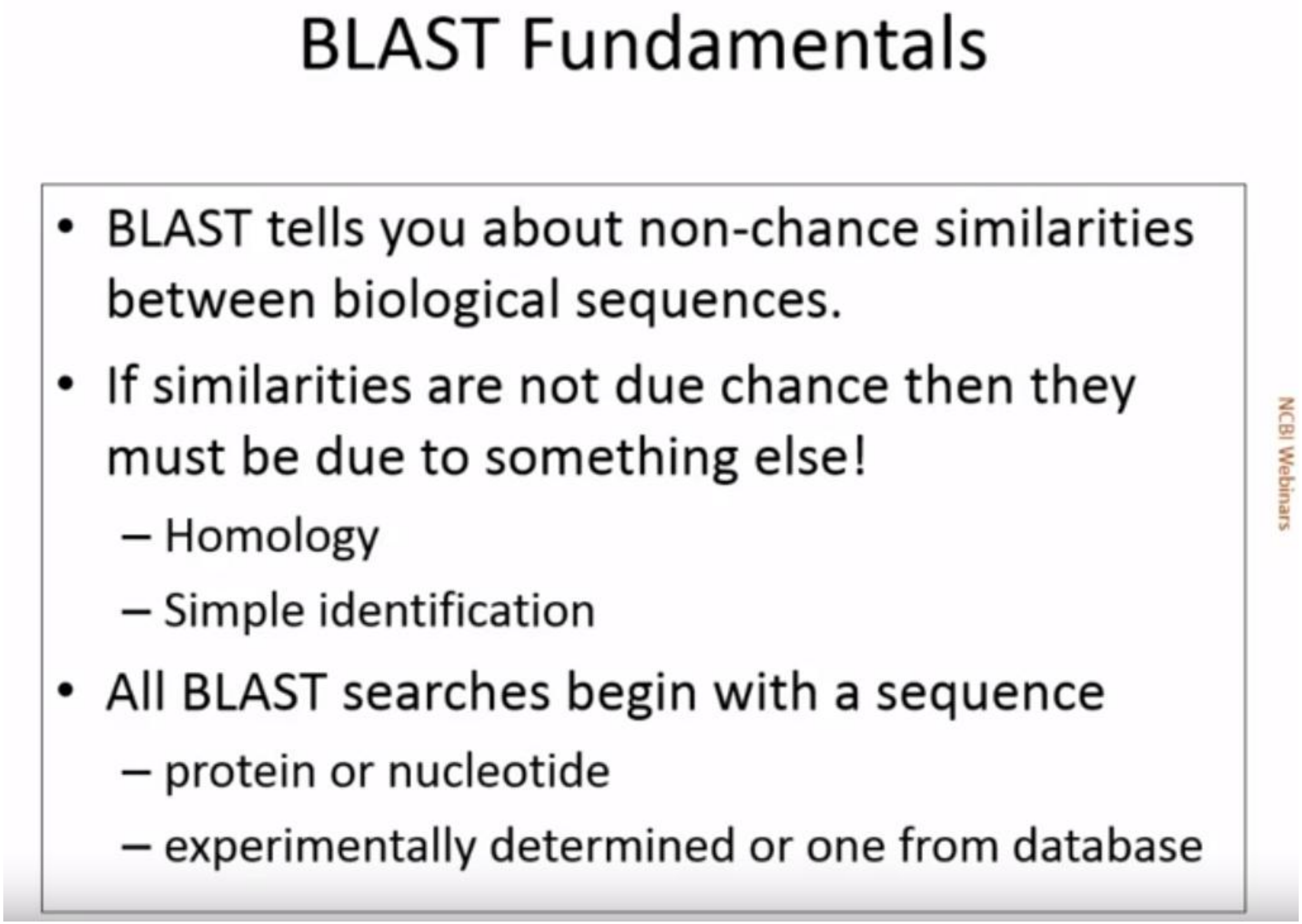
- Sviluppato per rendere ancora più veloci le ricerche nelle banche dati rispetto a FASTA, senza perdere in sensibilità e selettività
- Metodo euristico per allineamenti locali
- Pensato specificamente per ricerche in database
- Basato sulle stesse assunzioni di FASTA: un buon allineamento contiene corti frammenti di match esatti

BLAST: Basic Local Alignment Search Tool

Basic Local Alignment Search Tool. Altschul et al. 1990,1994,1997

- Input:
  - Query sequence Q (la vostra sequenza!)
  - Database of sequences DB
  - Minimal score S

- Output:
  - Sequenze presenti nel DB (Seq), per le quali Q e Seq abbiano uno score > S



BLAST Fundamentals

- BLAST tells you about non-chance similarities between biological sequences.
- If similarities are not due chance then they must be due to something else!
  - Homology
  - Simple identification
- All BLAST searches begin with a sequence
  - protein or nucleotide
  - experimentally determined or one from database

https://www.youtube.com/watch?v=mvjHYMgJDTQ

What BLAST tells you

Here's my sequence.

1. What is it related to? (What does it do?)
   - Homology
   - Function
2. Is it already in the database? (Identification)
   - find the matching sequence in the database
   - organism of origin
3. Where is it located or how is it organized?
   - in a genome
   - other annotation problems
     - comparing sequences
     - looking for frame shifts

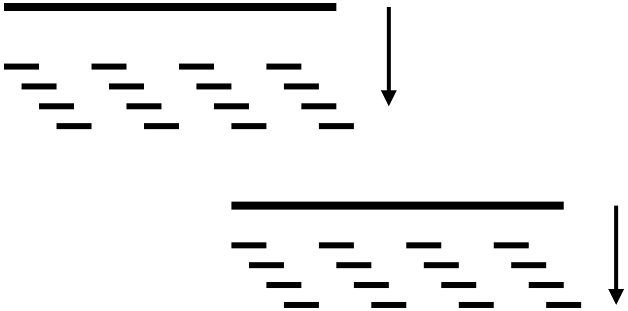
NCBI Webinars

BLAST Word Matching

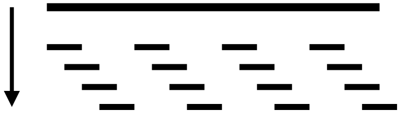
```
MEAAVKEEISVEDEAVDKNI

MEA
 EAA
  AAV
   AVK
    VKE
     KEE
      EEI
       EIS
        ISV
         ...
```

Break query into words:



Break database sequences into words:



Alignment starts with initial word of 11

```
ACACTGAGTGA
|||||||||||
ACACTGAGTGA
```

Extension to the left has no mismatches, no penalty points
Extension to the right has mismatches and penalty points

```
GCACCTTTGCCACACTGAGTGAGCTGCTCTATG
|||||||||||||||||||||| |||| || |
GCACCTTTGCCACACTGAGTGACCTGCACTGTA
```

Extension to the left has no penalty points and can continue to grow
Extension to the right accumulates too many mismatch penalty points; extension in this direction stops

```
CAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCTCTATGGTCCTTTGGGG
||||||||||||||||||||||||||||||||| |||| || |     ||||
CAACCTCAAGGGCACCTTTGCCACACTGAGTGACCTGCACTGTAAAGTTTTGCAT
```

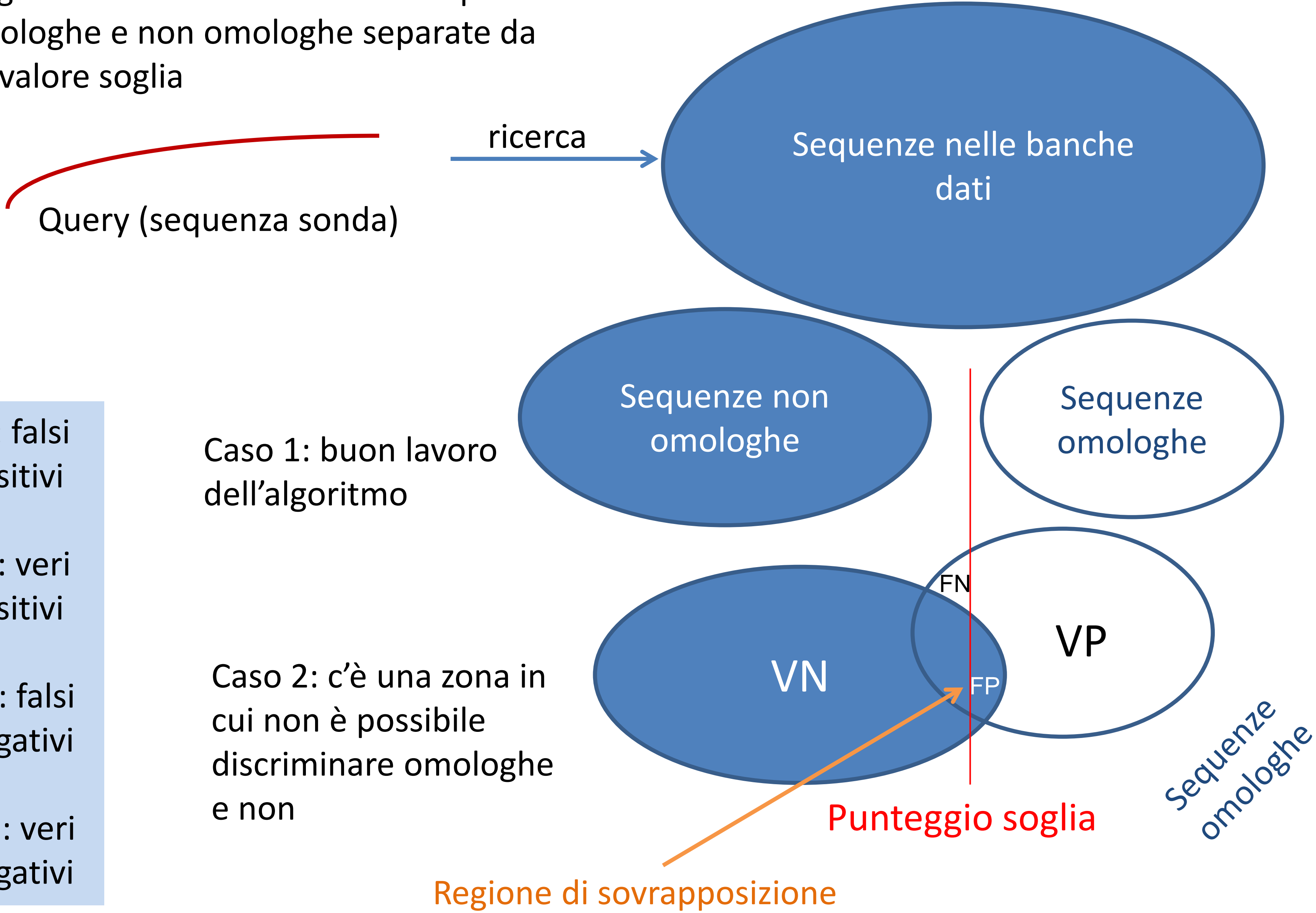
If left side cannot grow any more, the final alignment looks like this:

```
CAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCTCTATG
||||||||||||||||||||||||||||||||| |||| || |
CAACCTCAAGGGCACCTTTGCCACACTGAGTGACCTGCACTGTA
```

Figure 3.1 Simple extension example for BLASTN. Starting with an initial match of “words,” BLAST extends the alignment between query and hit, keeping track of penalty points against, and increasing significance for, extending the alignment.

Ricerche in database

L'algoritmo deve identificare le sequenze omologhe e non omologhe separate da un valore soglia



Caso 1: buon lavoro dell'algoritmo

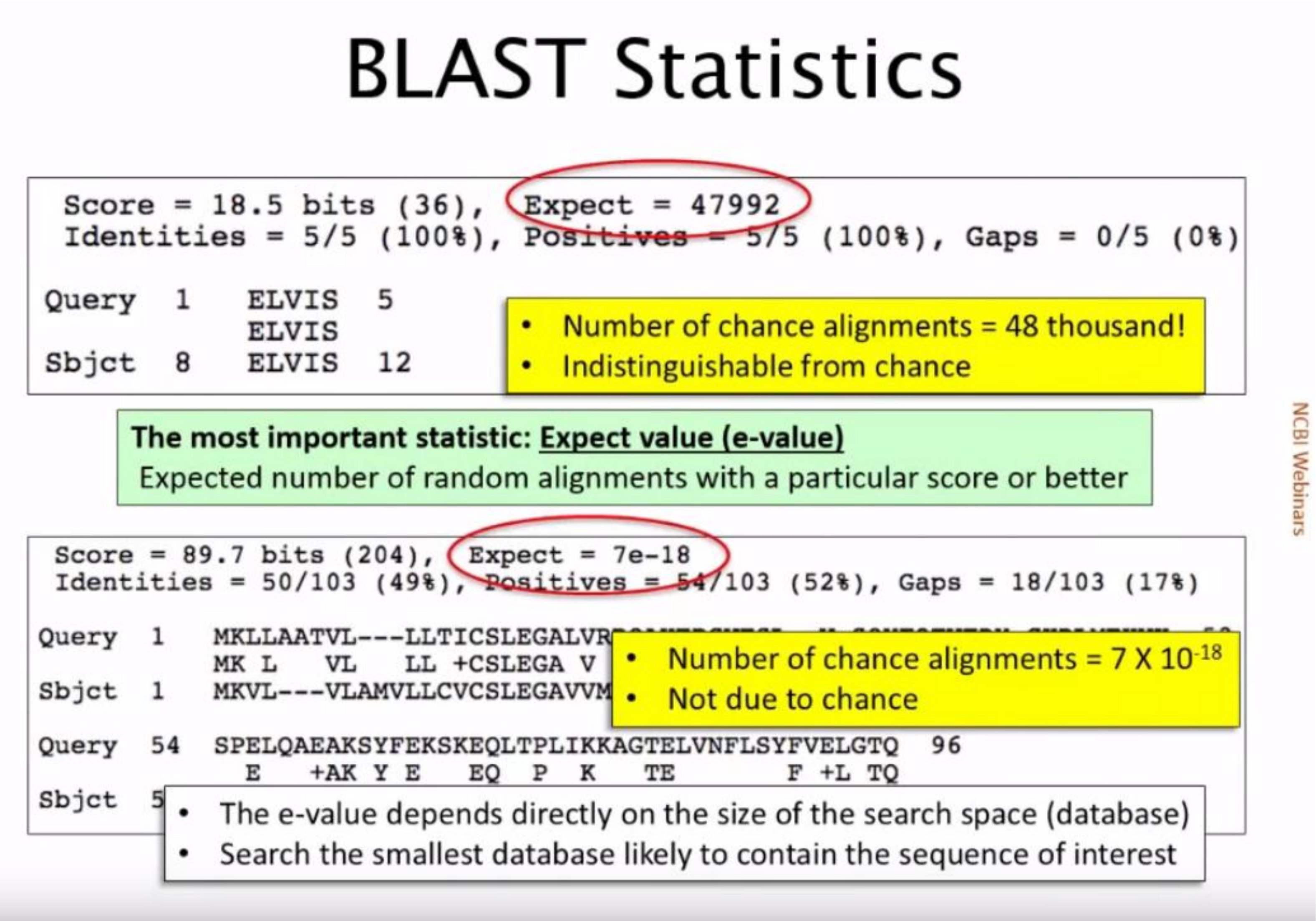
Caso 2: c'è una zona in cui non è possibile discriminare omologhe e non

FP: falsi positivi

VP: veri positivi

FN: falsi negativi

VN: veri negativi



BLAST Statistics

```
 Score = 18.5 bits (36),  Expect = 47992
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query  1    ELVIS  5
            ELVIS
Sbjct  8    ELVIS  12
```

- Number of chance alignments = 48 thousand!
- Indistinguishable from chance

The most important statistic: Expect value (e-value)

Expected number of random alignments with a particular score or better

```
 Score = 89.7 bits (204),  Expect = 7e-18
 Identities = 50/103 (49%), Positives = 54/103 (52%), Gaps = 18/103 (17%)

Query  1    MKLLAATVL---LLTICSLEGALVR
            MK L    VL    LL +CSLEGA V
Sbjct  1    MKVL---VLAMVLLCVCSLEGAVVM

Query  54   SPELQAEAKSYFEKSKEQLTPLIKKAGTELVNFLSYFVELGTQ  96
              E   +AK Y E    EQ   P   K    TE         F +L TQ
Sbjct  5
```

- Number of chance alignments = 7 X 10^{-18}
- Not due to chance

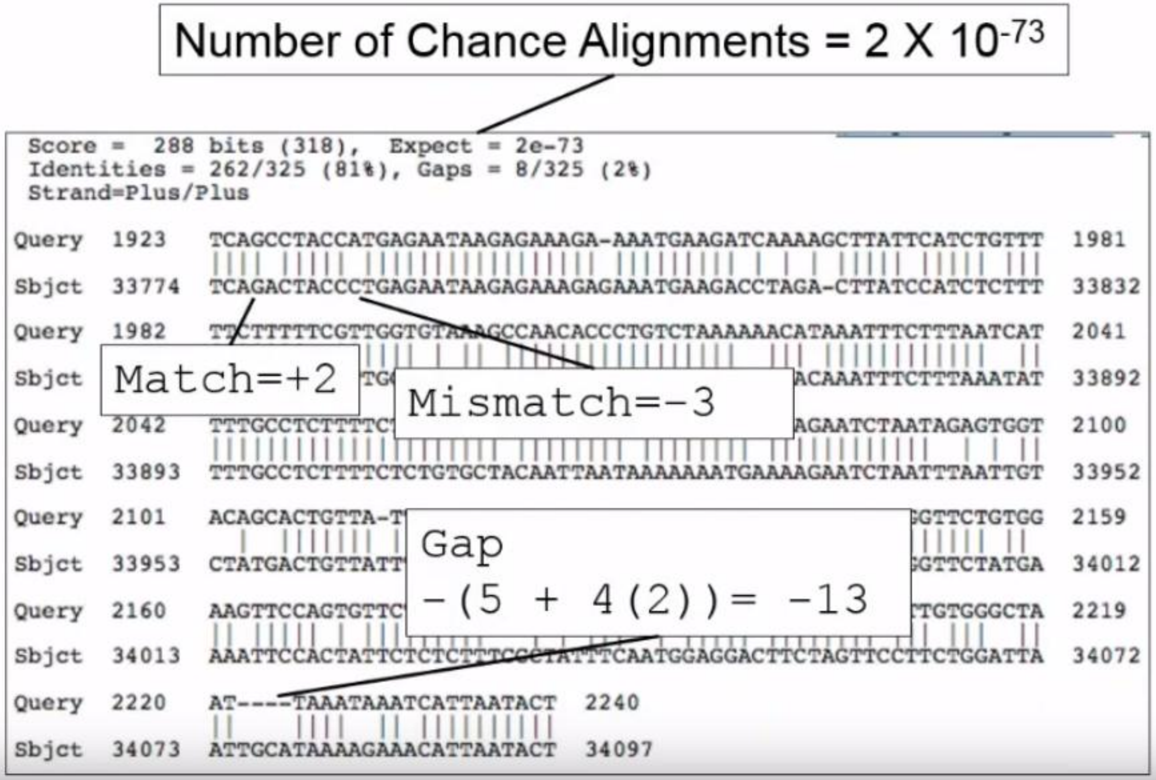
- The e-value depends directly on the size of the search space (database)
- Search the smallest database likely to contain the sequence of interest

NCBI Webinars

Attesa (**E**xpectation) di trovare PER CASO uno Score come quello osservato

Scoring: Nucleotide

Number of Chance Alignments = 2 X 10^{-73}



```
 Score =  288 bits (318),  Expect = 2e-73
 Identities = 262/325 (81%), Gaps = 8/325 (2%)
 Strand=Plus/Plus

Query  1923   TCAGCCTACCATGAGAATAAGAGAAAGA-AAATGAAGATCAAAAGCTTATTCATCTGTTT  1981
              |||| ||||| ||||||||||||||||| ||||||||| | | | ||||| ||||| |||
Sbjct  33774  TCAGACTACCCTGAGAATAAGAGAAAGAGAAATGAAGACCTAGA-CTTATCCATCTCTTT  33832

Query  1982   TTCTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCAT  2041
                        ||||| | ||  |||||||||||||||||  ||| |||||||||||   ||
Sbjct                   TG                              ACAAATTTCTTTAAATAT  33892

Query  2042   TTTGCCTCTTTTC                             AGAATCTAATAGAGTGGT  2100
              |||||||||||||||||||| |||||||| |||||||| |||||||||||||  | | ||
Sbjct  33893  TTTGCCTCTTTTCTCTGTGCTACAATTAATAAAAAAATGAAAAGAATCTAATTTAATTGT  33952

Query  2101   ACAGCACTGTTA-T                                    GGTTCTGTGG  2159
                |  ||||||| |                                    |||||| ||
Sbjct  33953  CTATGACTGTTATT                                    GGTTCTATGA  34012

Query  2160   AAGTTCCAGTGTTC                                     TGTGGGCTA  2219
              || ||||| | ||||                                    | |||  ||
Sbjct  34013  AAATTCCACTATTCTCTCTTTCGCTATTTCAATGGAGGACTTCTAGTTCCTTCTGGATTA  34072

Query  2220   AT----TAAATAAATCATTAATACT  2240
              ||    ||||  || ||||||||||
Sbjct  34073  ATTGCATAAAAGAAACATTAATACT  34097
```

Match=+2

Mismatch=-3

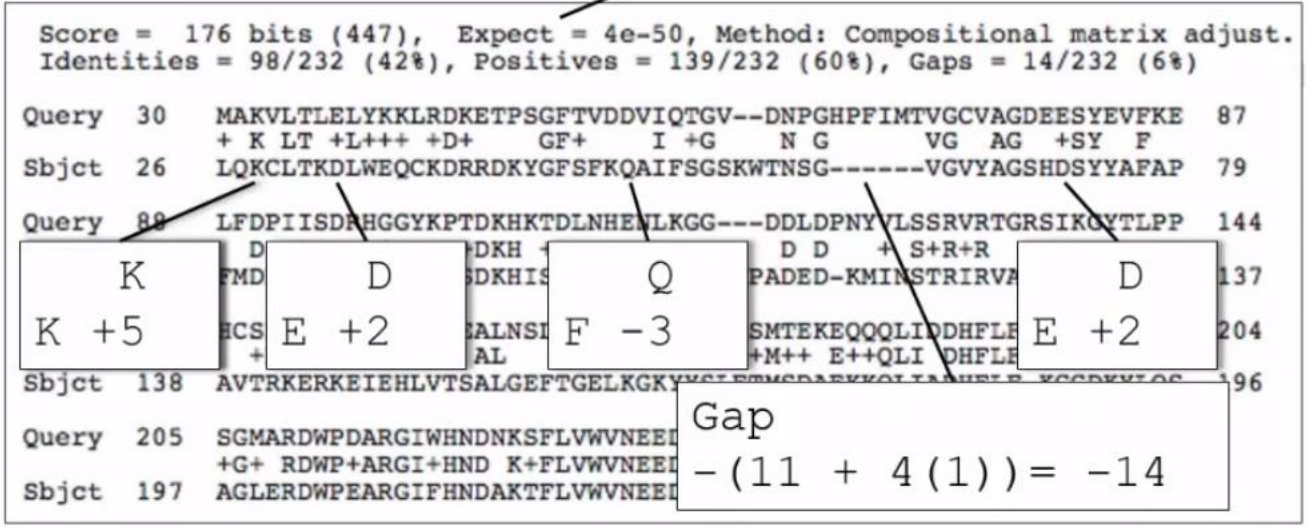
Gap

$-(5 + 4(2))= -13$

NCBI Webinars

Scoring: Protein

Number of Chance Alignments = 4 X 10^{-50}



```
 Score =  176 bits (447),  Expect = 4e-50, Method: Compositional matrix adjust.
 Identities = 98/232 (42%), Positives = 139/232 (60%), Gaps = 14/232 (6%)

Query  30    MAKVLTLELYKKLRDKETPSGFTVDDVIQTGV--DNPGHPFIMTVGCVAGDEESYEVFKE  87
             + K LT +L+++ +D+     GF+     I +G    N G        VG  AG  +SY   F
Sbjct  26    LQKCLTKDLWEQCKDRRDKYGFSFKQAIFSGSKWTNSG------VGVYAGSHDSYYAFAP  79

Query  88    LFDPIISDRHGGYKPTDKHKTDLNHENLKGG---DDLDPNYVLSSRVRTGRSIKGYTLPP  144
               D                  DKH +             D D   + S+R+R
             FMD              DKHIS              PADED-KMINSTRIRVA          137

             HCS              ALNSI             SMTEKEQQQLIDDHFLE          204
              +                AL               +M++ E++QLI DHFLE
Sbjct  138   AVTRKERKEIEHLVTSALGEFTGELKGK                                    196

Query  205   SGMARDWPDARGIWHNDNKSFLVWVNEED
             +G+ RDWP+ARGI+HND K+FLVWVNEED
Sbjct  197   AGLERDWPEARGIFHNDAKTFLVWVNEED
```

K
K +5

D
E +2

Q
F −3

D
E +2

Gap
−(11 + 4(1))= −14

Scores from BLOSUM62, a position independent matrix

- Same substitution gets the same score at all positions
- All positions equally likely to change

E value: significatività statistica

Non si interpretano come p values dove

$$\xcancel{p < 0.05}$$

sono generalmente considerati significativi

Regola generale

E values < 10^{-6} sono molto probabilmente significativi.

10^{-6} < E values < 10^{-3} meritano una seconda occhiata.

E values < 10^{-3} andrebbero scartati (ci aspettiamo di trovare 0.001 sequenze non correlate alla nostra-falsi positivi- che ottengono un punteggio superiore a quell'S).

BLAST Programs

BLAST has five programs

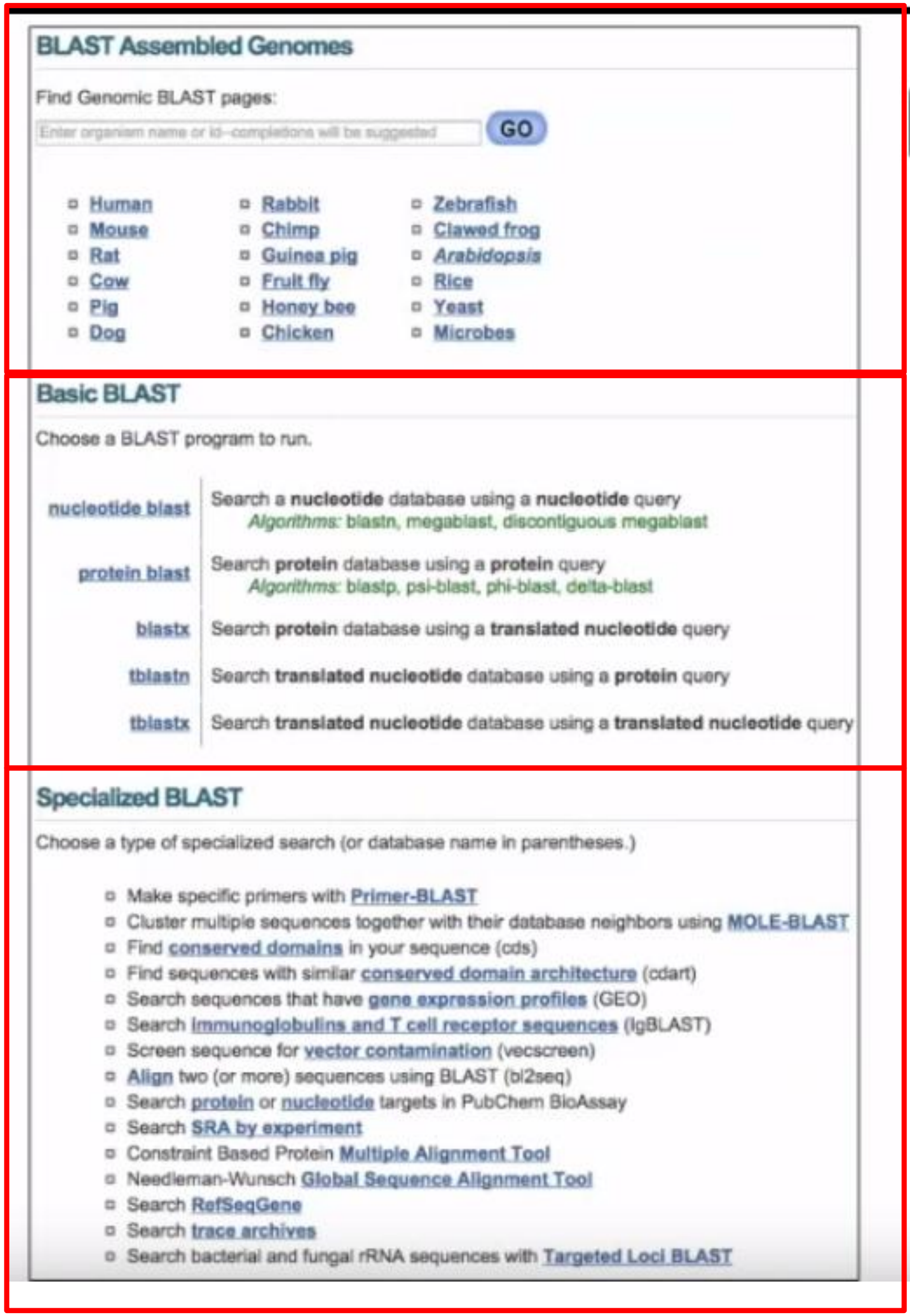
Differ in the types of sequences they align and at what level

Program	Query Seq. Type	Database Seq. Type	Alignment Level
blastn	nucleotide	nucleotide	nucleotide
blastp	protein	protein	protein
blastx	nucleotide	protein	protein
tblastn	protein	nucleotide	protein
tblastx	nucleotide	nucleotide	protein

Six-frame translation

BLAST Homepage

blast.ncbi.nlm.nih.gov



BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-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

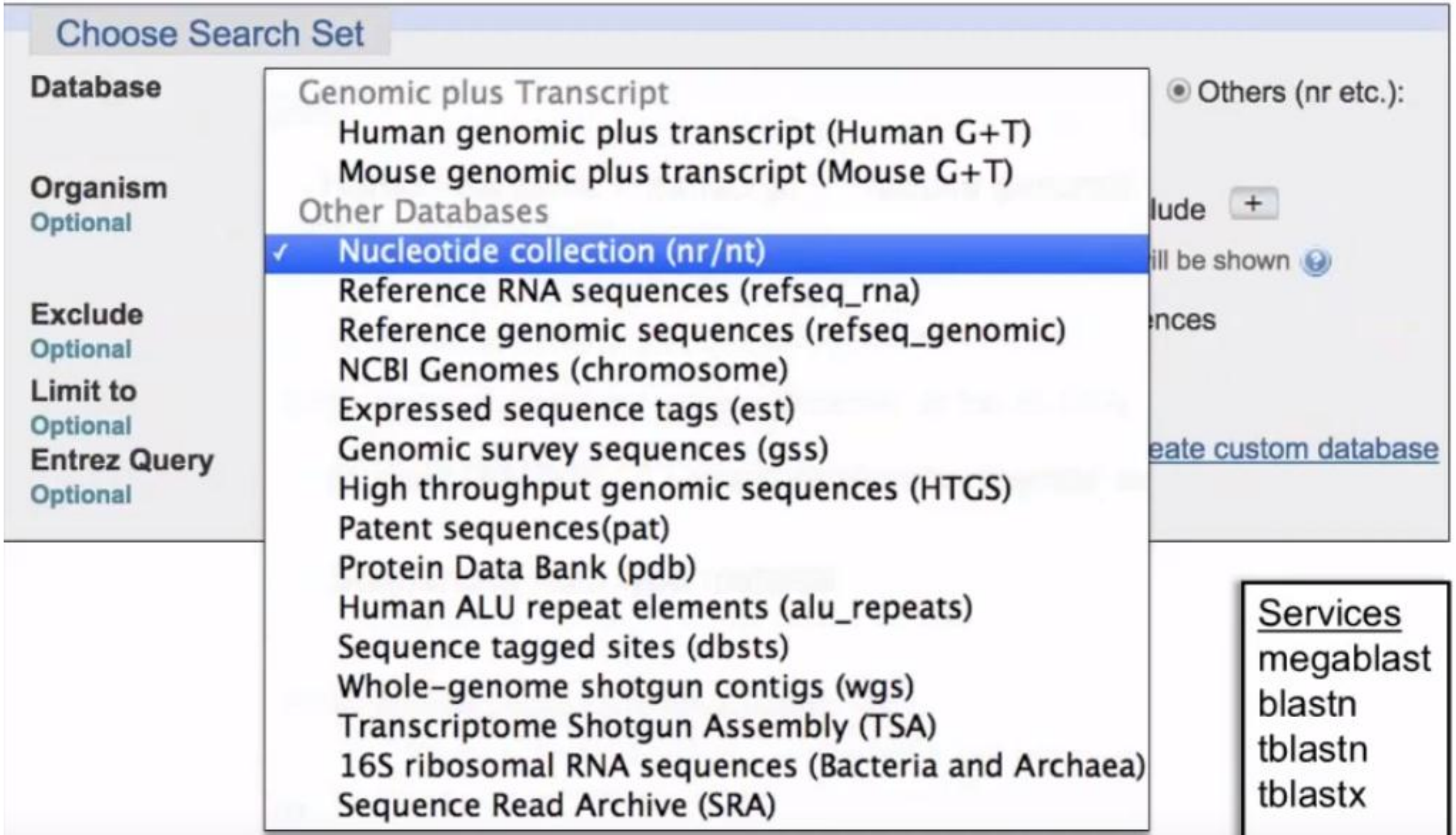
Specialized BLAST

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

- Make specific primers with Primer-BLAST
- Cluster multiple sequences together with their database neighbors using MOLE-BLAST
- 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 and T cell receptor sequences (IgBLAST)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA by experiment
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search trace archives
- Search bacterial and fungal rRNA sequences with Targeted Loci BLAST

NCBI Webinars

Nucleotide Databases



Choose Search Set

Database

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- ✓ Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)
- 16S ribosomal RNA sequences (Bacteria and Archaea)
- Sequence Read Archive (SRA)

● Others (nr etc.):

Organism
Optional

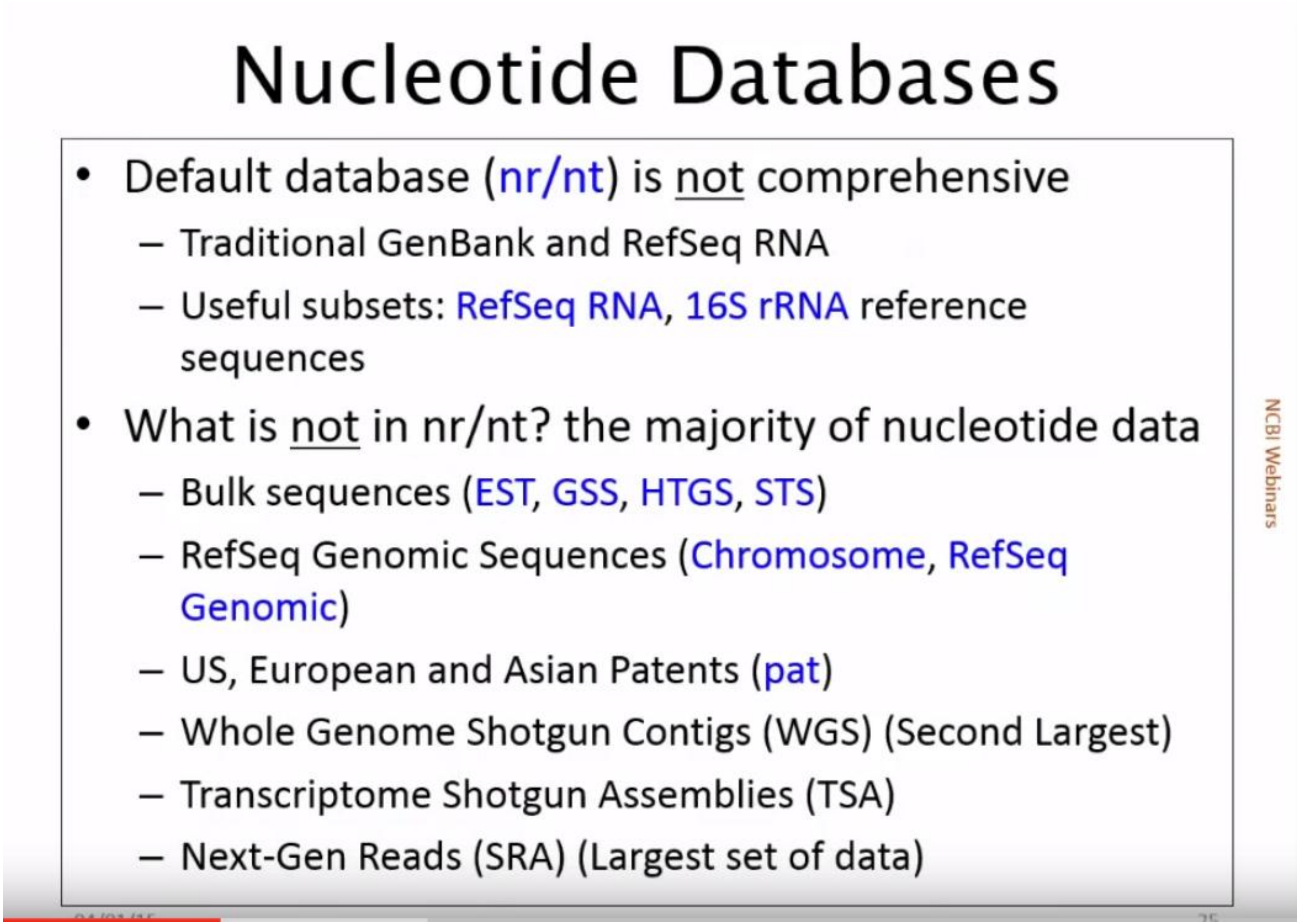
Exclude
Optional

Limit to
Optional

Entrez Query
Optional

Services
megablast
blastn
tblastn
tblastx

NCBI Webinars



Nucleotide Databases

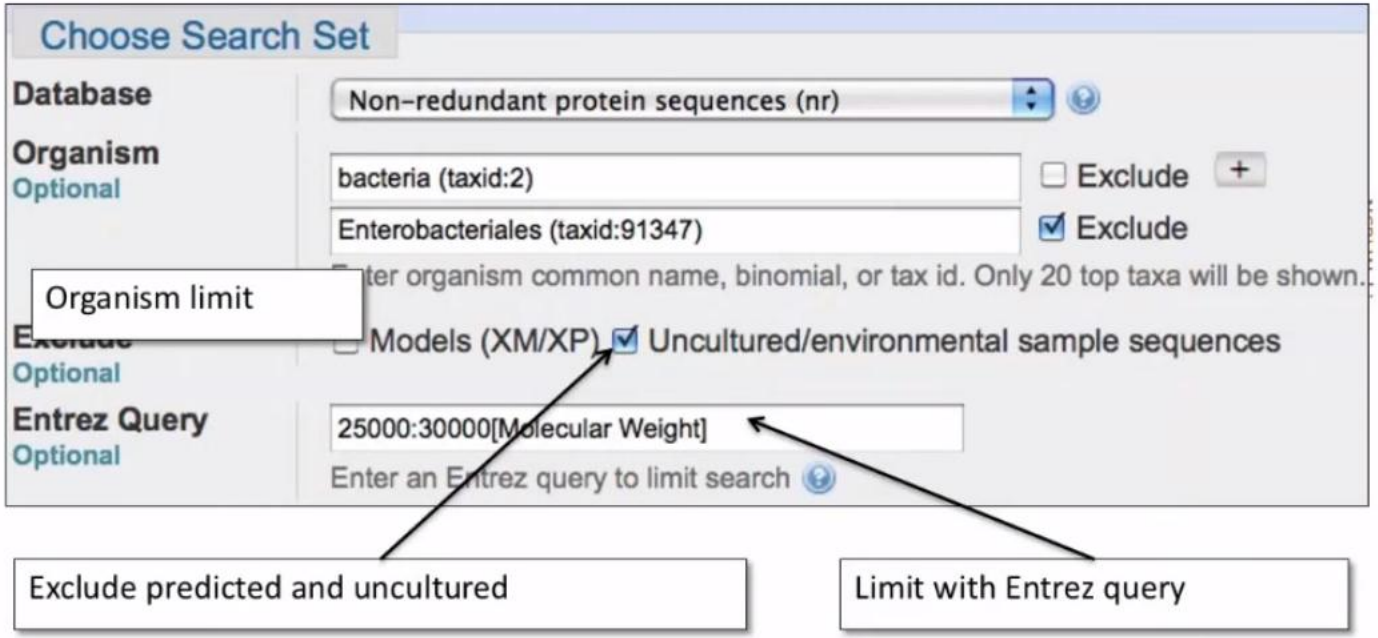
- Default database (nr/nt) is not comprehensive
  - Traditional GenBank and RefSeq RNA
  - Useful subsets: RefSeq RNA, 16S rRNA reference sequences
- What is not in nr/nt? the majority of nucleotide data
  - Bulk sequences (EST, GSS, HTGS, STS)
  - RefSeq Genomic Sequences (Chromosome, RefSeq Genomic)
  - US, European and Asian Patents (pat)
  - Whole Genome Shotgun Contigs (WGS) (Second Largest)
  - Transcriptome Shotgun Assemblies (TSA)
  - Next-Gen Reads (SRA) (Largest set of data)

NCBI Webinars

Ricordiamo che l’efficienza della ricerca aumenta se limitiamo il database che interroghiamo

Limiting Databases

Search the smallest database likely to contain the sequence of interest.



Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: bacteria (taxid:2) □ Exclude

Enterobacteriales (taxid:91347) ☑ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) ☑ Uncultured/environmental sample sequences

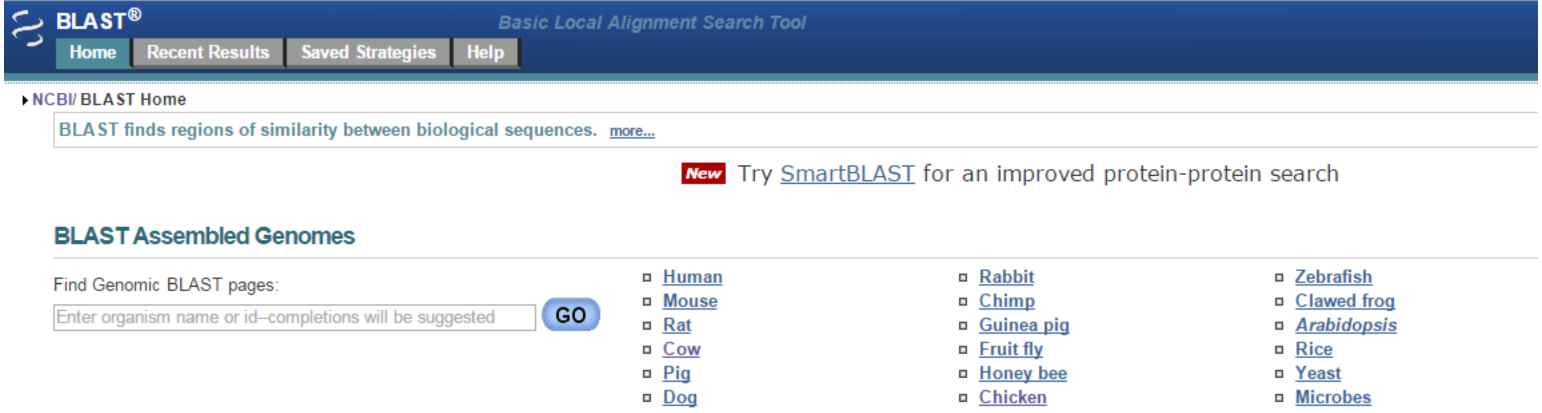
Entrez Query Optional: 25000:30000[Molecular Weight]

Enter an Entrez query to limit search

Organism limit

Exclude predicted and uncultured

Limit with Entrez query



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New Try SmartBLAST for an improved protein-protein search

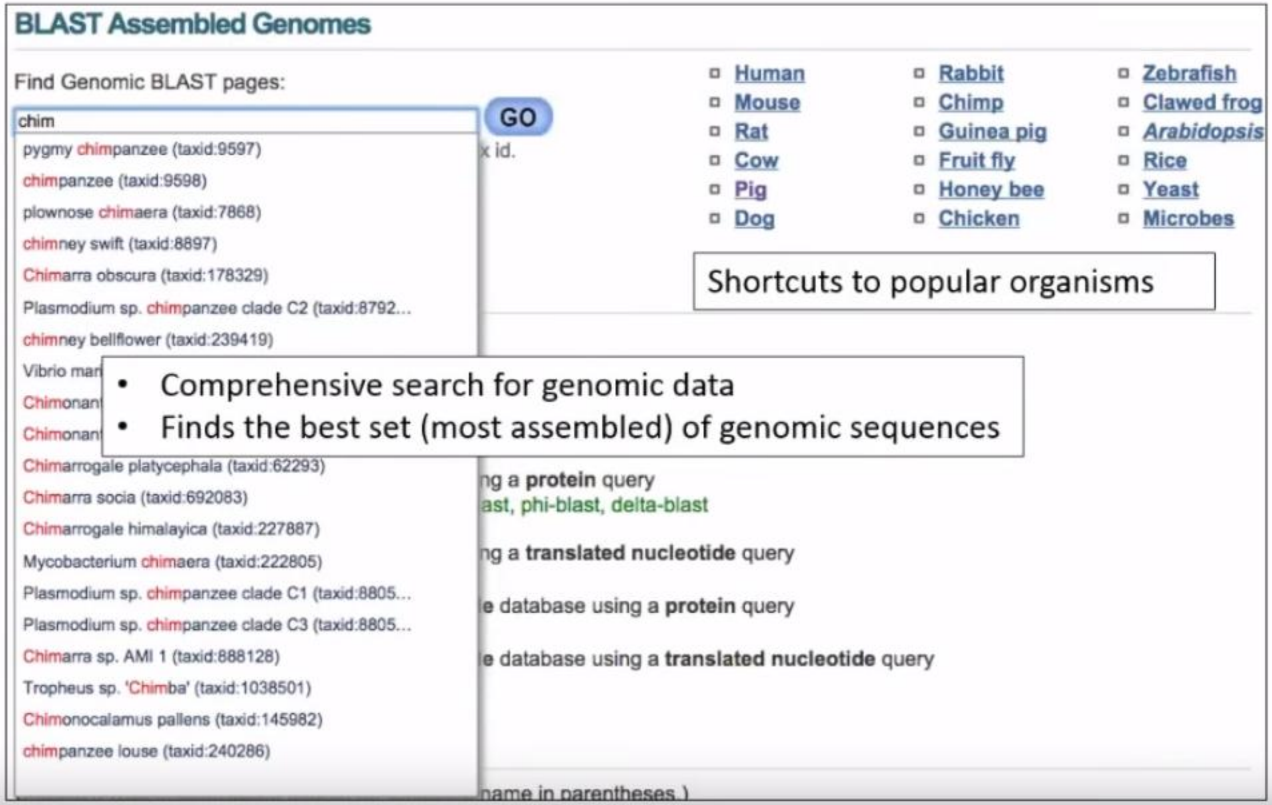
BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Genome Databases



BLAST Assembled Genomes

Find Genomic BLAST pages:

chim GO

pygmy chimpanzee (taxid:9597)
chimpanzee (taxid:9598)
plownose chimaera (taxid:7868)
chimney swift (taxid:8897)
Chimarra obscura (taxid:178329)
Plasmodium sp. chimpanzee clade C2 (taxid:8792...
chimney bellflower (taxid:239419)
Vibrio mar
Chimonan
Chimonan
Chimarrogale platycephala (taxid:62293)
Chimarra socia (taxid:692083)
Chimarrogale himalayica (taxid:227887)
Mycobacterium chimaera (taxid:222805)
Plasmodium sp. chimpanzee clade C1 (taxid:8805...
Plasmodium sp. chimpanzee clade C3 (taxid:8805...
Chimarra sp. AMI 1 (taxid:888128)
Tropheus sp. 'Chimba' (taxid:1038501)
Chimonocalamus pallens (taxid:145982)
chimpanzee louse (taxid:240286)

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

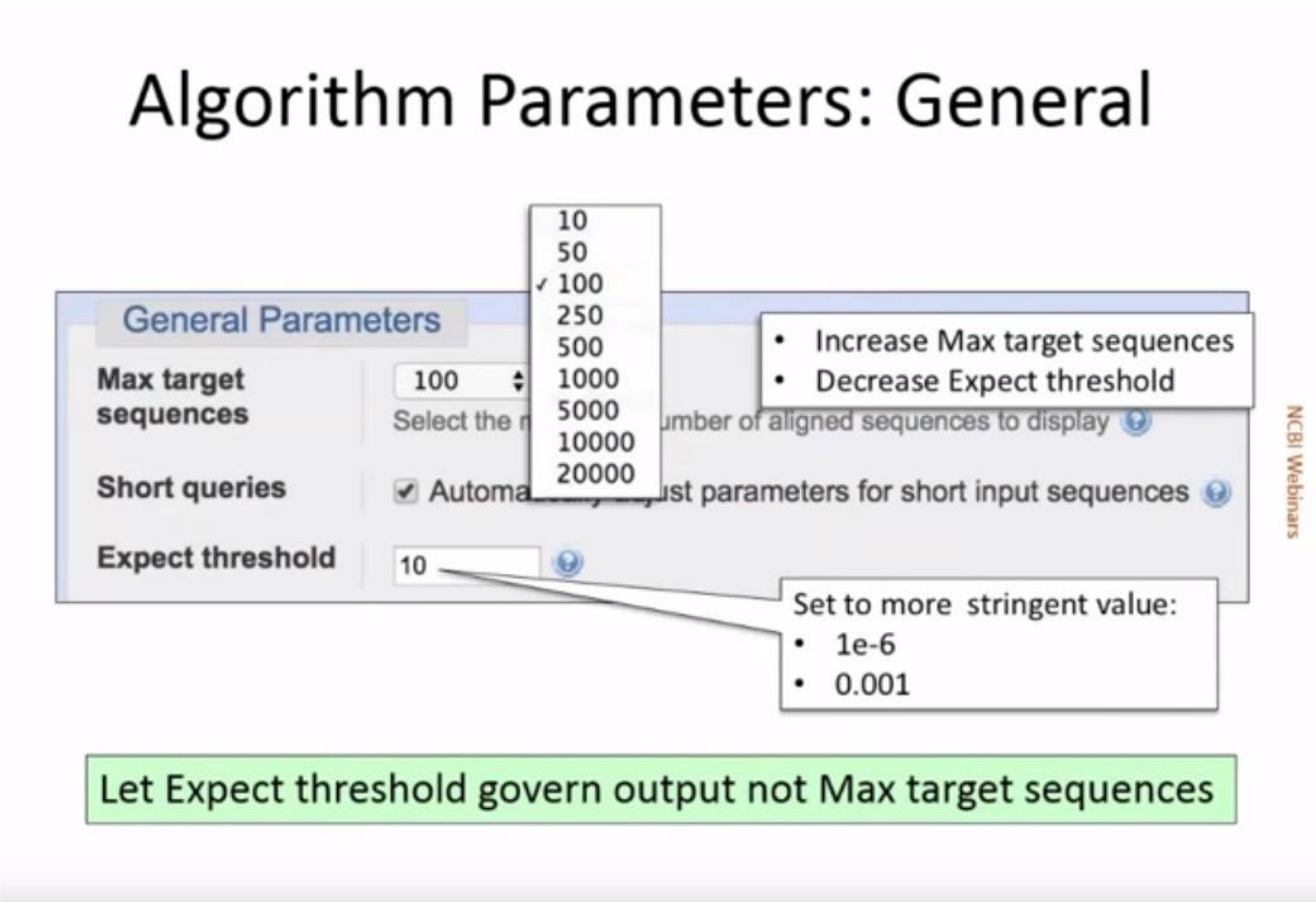
x id.

ng a **protein** query
ast, phi-blast, delta-blast
ng a **translated nucleotide** query
e database using a **protein** query
e database using a **translated nucleotide** query
name in parentheses.)

Shortcuts to popular organisms

- Comprehensive search for genomic data
- Finds the best set (most assembled) of genomic sequences

NCBI Webinars



Algorithm Parameters: General

General Parameters

Max target sequences: 100

Select the m... number of aligned sequences to display

Short queries: ☑ Automa... ust parameters for short input sequences

Expect threshold: 10

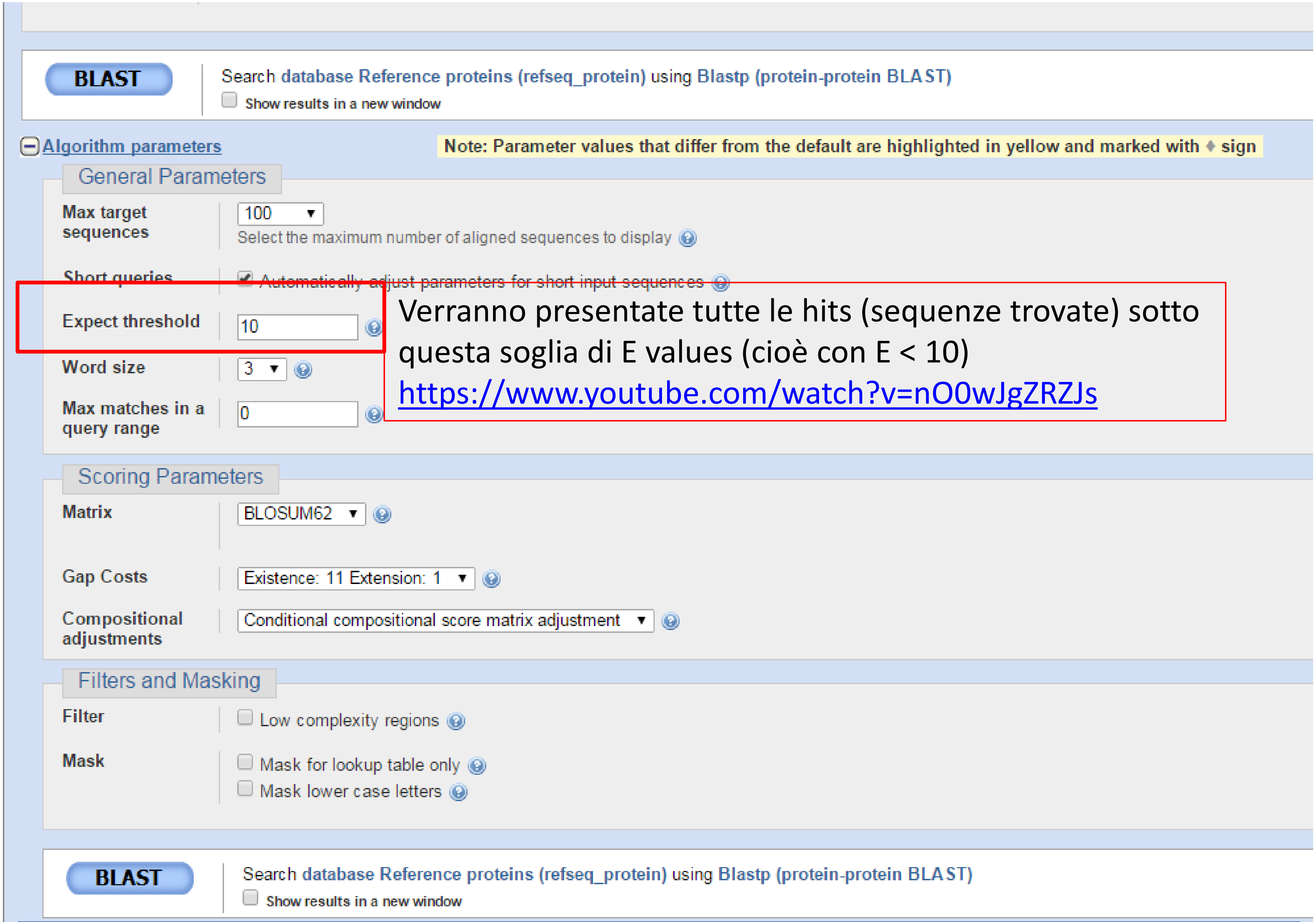
- Increase Max target sequences
- Decrease Expect threshold

Set to more stringent value:

- 1e-6
- 0.001

Let Expect threshold govern output not Max target sequences

Threshold = soglia (vedi diapositiva 5)



BLAST — Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

□ Show results in a new window

⊟ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

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

Verranno presentate tutte le hits (sequenze trovate) sotto questa soglia di E values (cioè con E < 10)
https://www.youtube.com/watch?v=nO0wJgZRZJs

Scoring Parameters

- Matrix: BLOSUM62
- Gap Costs: Existence: 11 Extension: 1
- Compositional adjustments: Conditional compositional score matrix adjustment

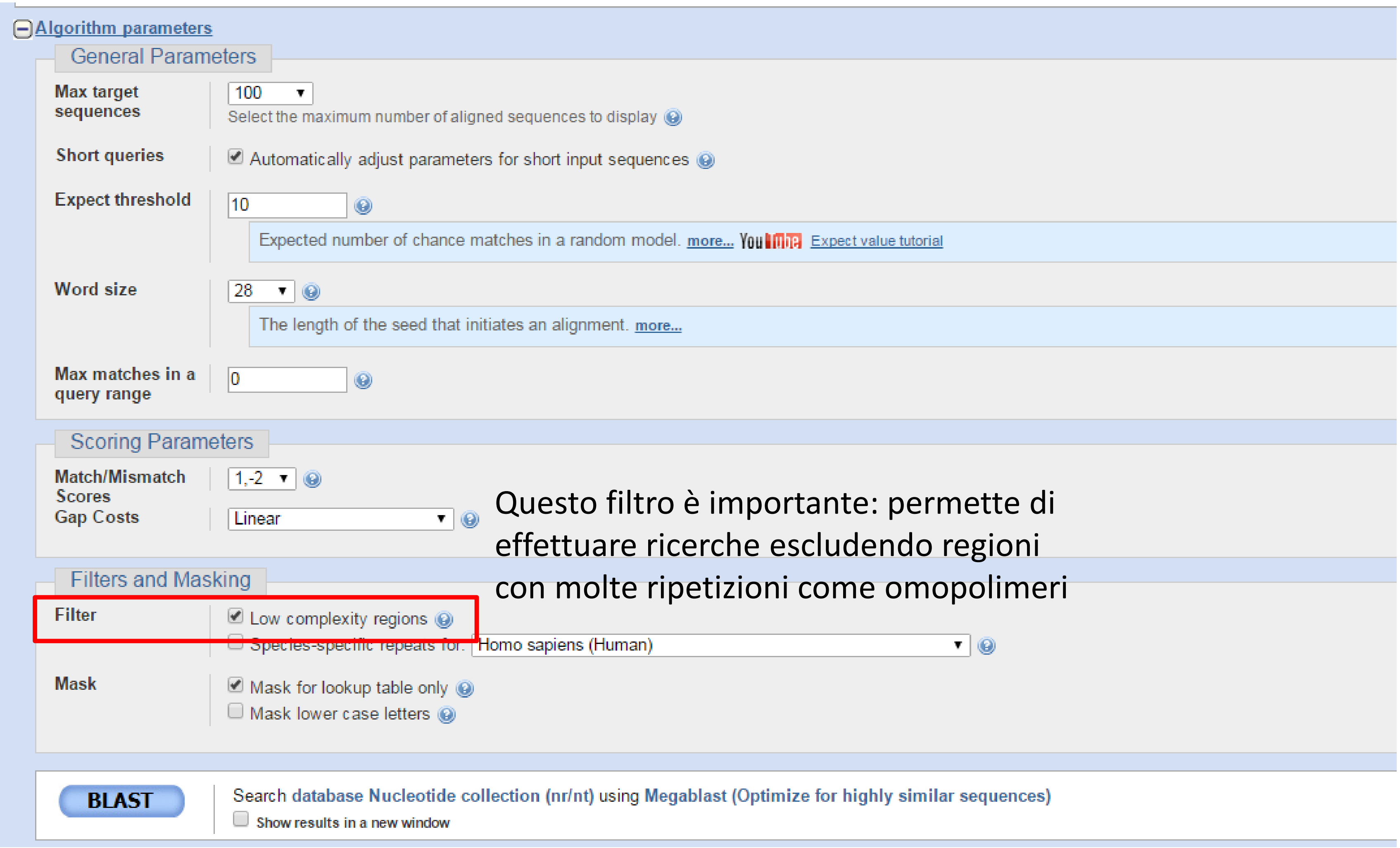
Filters and Masking

- Filter: □ Low complexity regions
- Mask: □ Mask for lookup table only
  □ Mask lower case letters

BLAST — Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

□ Show results in a new window

Ricordiamo che l'E risponde alla domanda: quante sequenze mi aspetto che abbiano **per caso** uno score maggiore o uguale a quello che ho osservato (falsi positivi!)



⊟ 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
Expected number of chance matches in a random model. more... YouTube Expect value tutorial

Word size: 28 ▾
The length of the seed that initiates an alignment. more...

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2 ▾

Gap Costs: Linear ▾

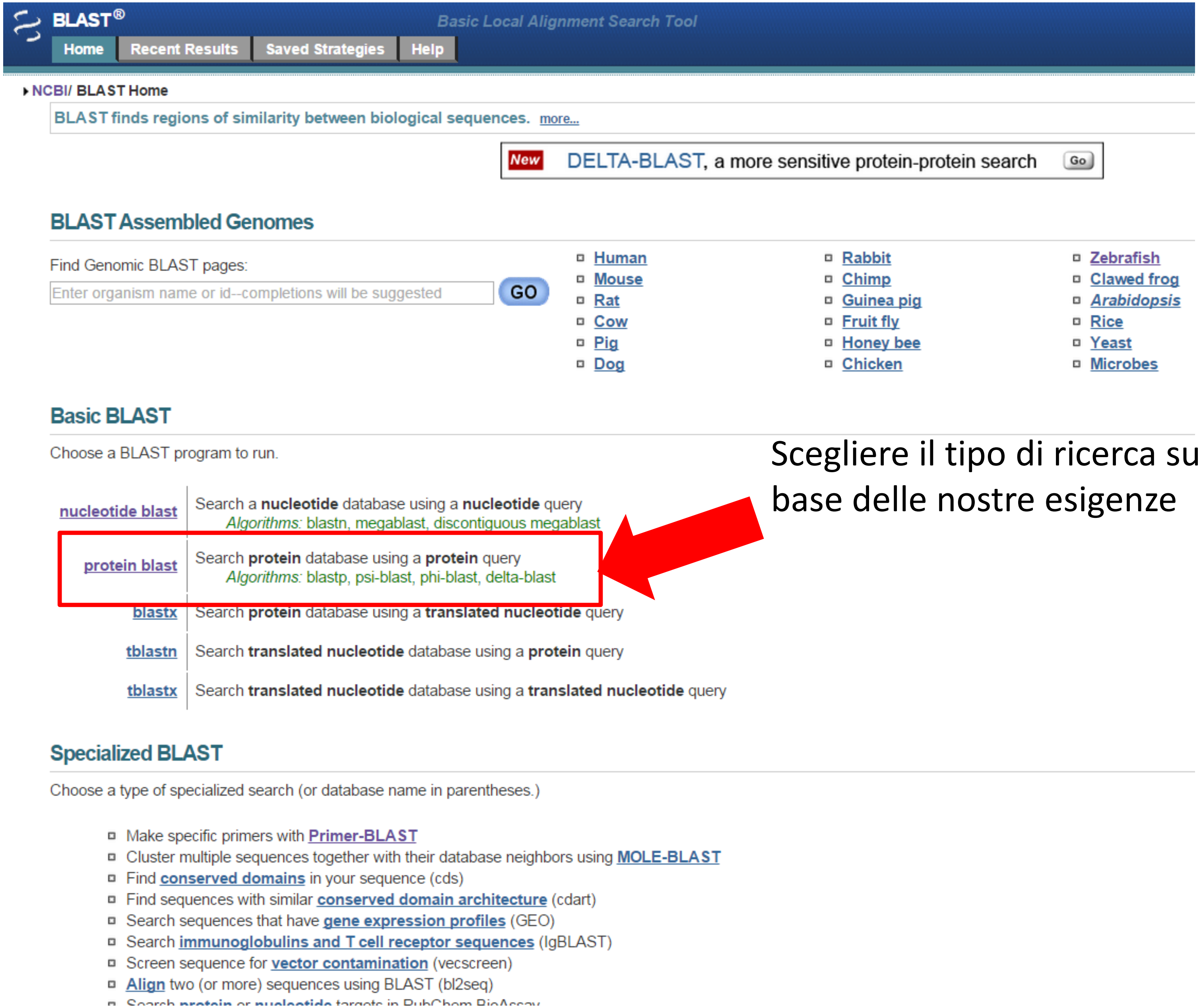
Filters and Masking

Filter: ☑ Low complexity regions
☐ Species-specific repeats for: Homo sapiens (Human) ▾

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

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
☐ Show results in a new window

Questo filtro è importante: permette di effettuare ricerche escludendo regioni con molte ripetizioni come omopolimeri



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

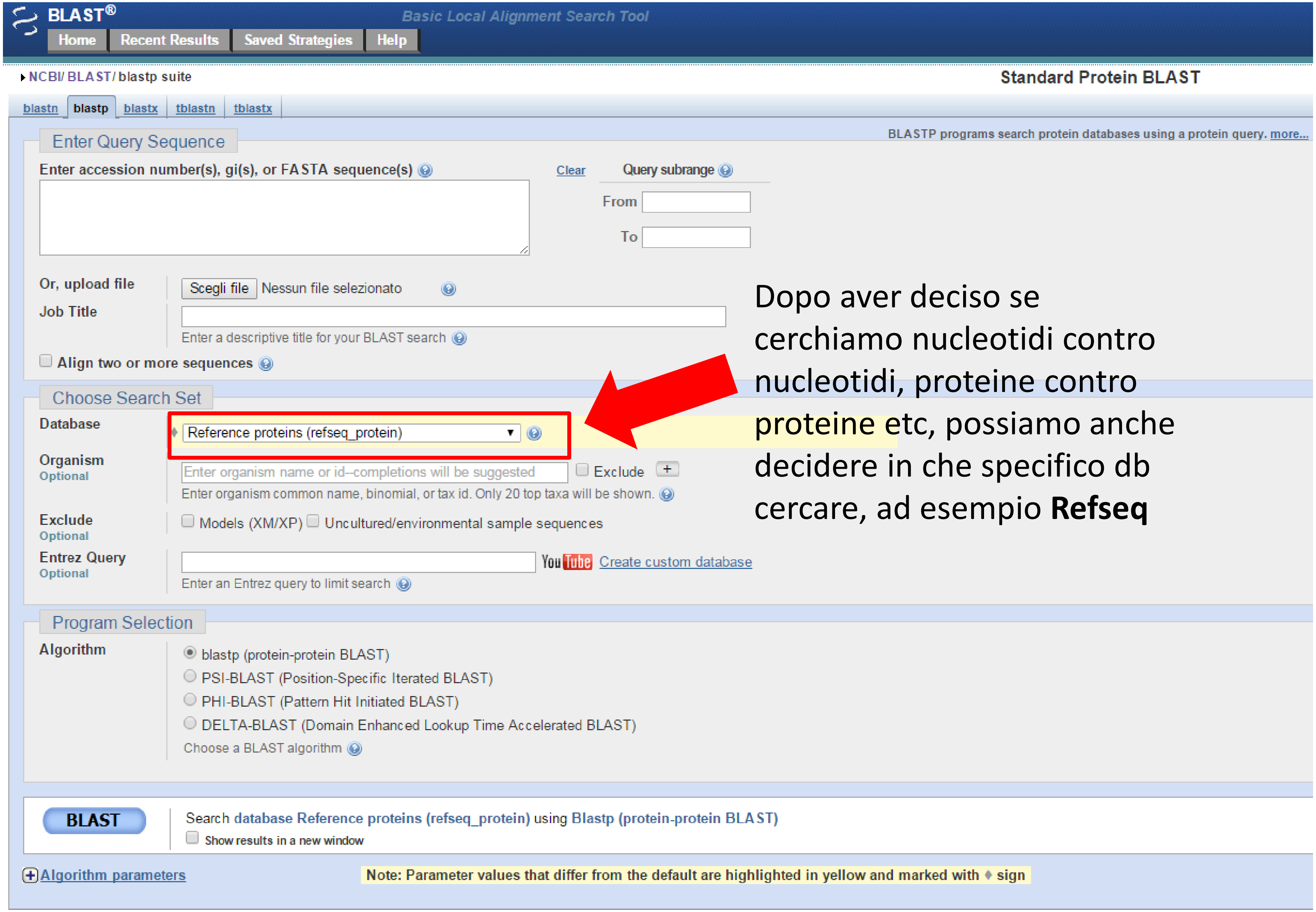
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using 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.)

- Make specific primers with Primer-BLAST
- Cluster multiple sequences together with their database neighbors using MOLE-BLAST
- 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 and T cell receptor sequences (IgBLAST)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay

Scegliere il tipo di ricerca sulla base delle nostre esigenze



BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange
From
To

Or, upload file: Scegli file Nessun file selezionato

Job Title: Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Reference proteins (refseq_protein)

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional): YouTube Create custom database
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)
- ○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST | Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)
□ Show results in a new window

+ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Dopo aver deciso se cerchiamo nucleotidi contro nucleotidi, proteine contro proteine etc, possiamo anche decidere in che specifico db cercare, ad esempio **Refseq**

NCBI RefSeq Database

- *Goal:* Provide a single reference sequence for each molecule of the central dogma (DNA, mRNA, and protein)

- Distinguishing features
  - Non-redundancy
  - Updates to reflect the current knowledge of sequence data and biology
  - Includes biological attributes of the gene, gene transcript, or protein
  - Encompasses a wide taxonomic range, with primary focus on mammalian and human species
  - Ongoing updates and curation (both automated and manual review), with review status indicated on each record

Pruitt et al., Nucleic Acids Res. 42: D756-D763, 2014

RefSeq Accession Number Prefixes

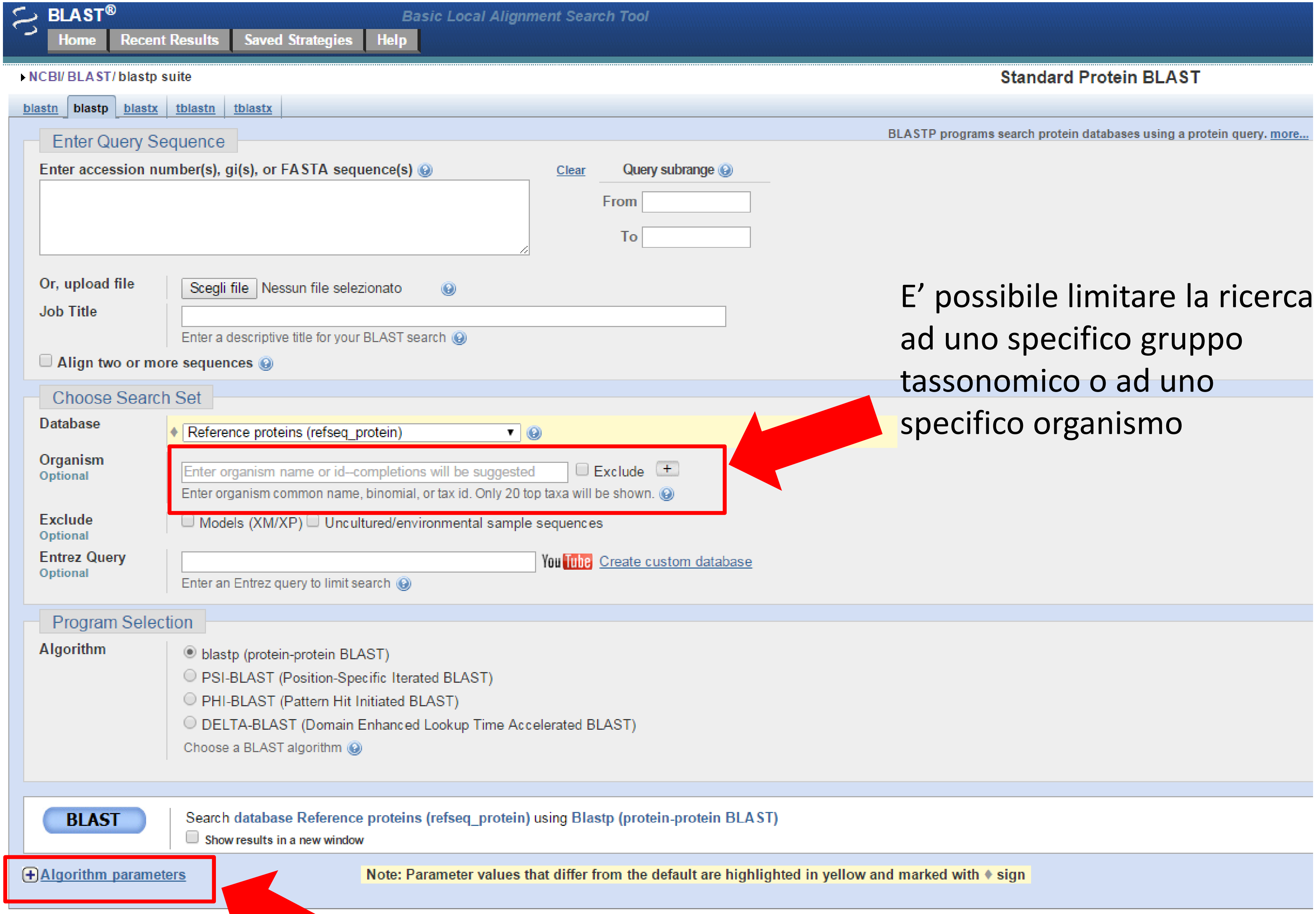
From curation of GenBank entries:

NT_	Genomic contigs
NM_	mRNAs
NP_	Proteins
NR_	Non-coding transcripts

From genome annotation:

XM_	Model mRNA
XP_	Model proteins

Complete list of molecule types in Chapter 18 of the NCBI Handbook
http://ncbi.nlm.nih.gov/books/NBK21091



BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

▸ NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) | Clear | Query subrange

From

To

Or, upload file: Scegli file Nessun file selezionato

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ♦ Reference proteins (refseq_protein) ▾

Organism (Optional): Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query (Optional): YouTube Create custom database

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)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST | Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

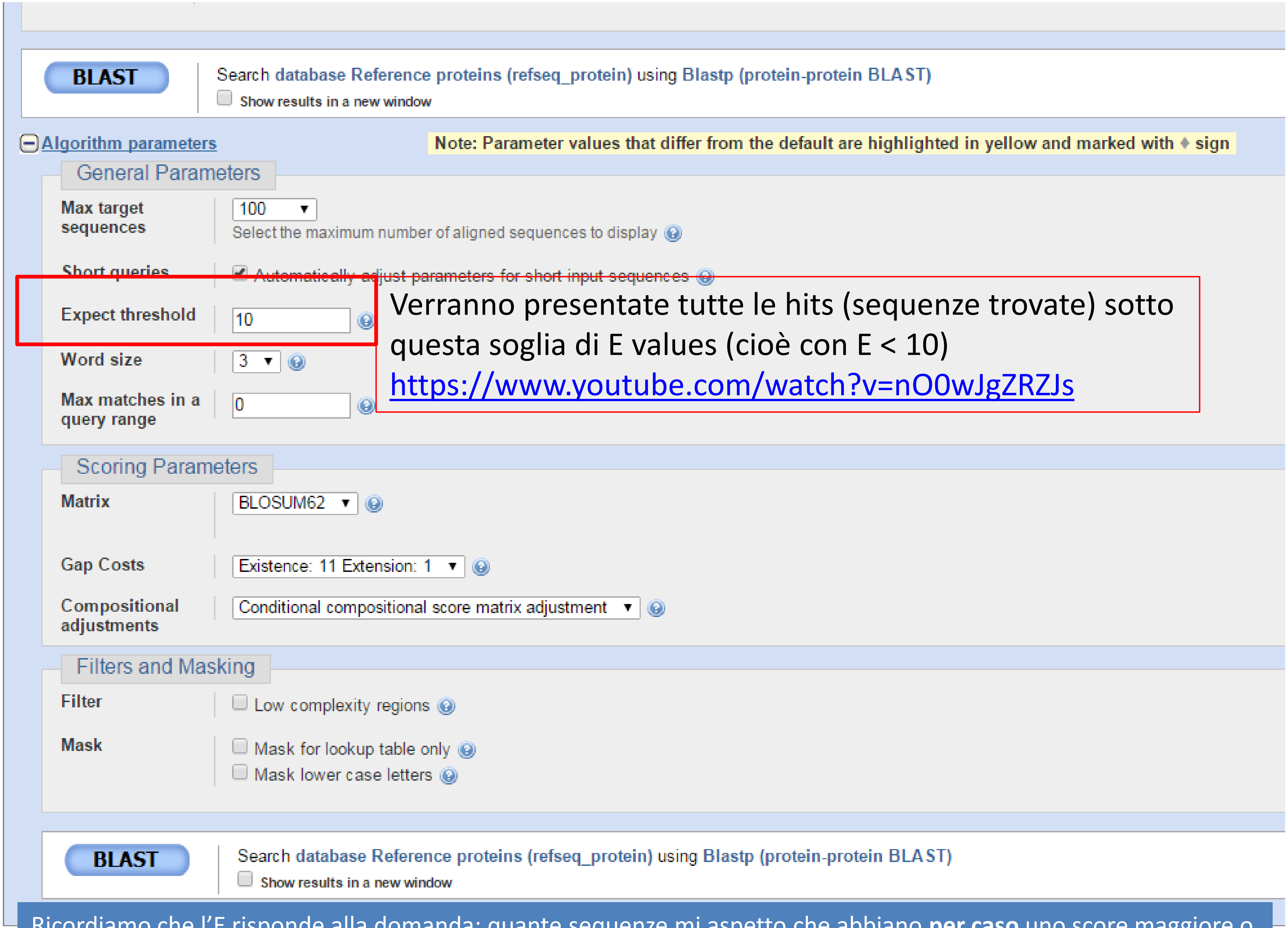
□ Show results in a new window

⊞ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

E’ possibile limitare la ricerca ad uno specifico gruppo tassonomico o ad uno specifico organismo

E’ possibile definire specifici parametri per la ricerca



BLAST | Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)
□ Show results in a new window

⊟ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

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

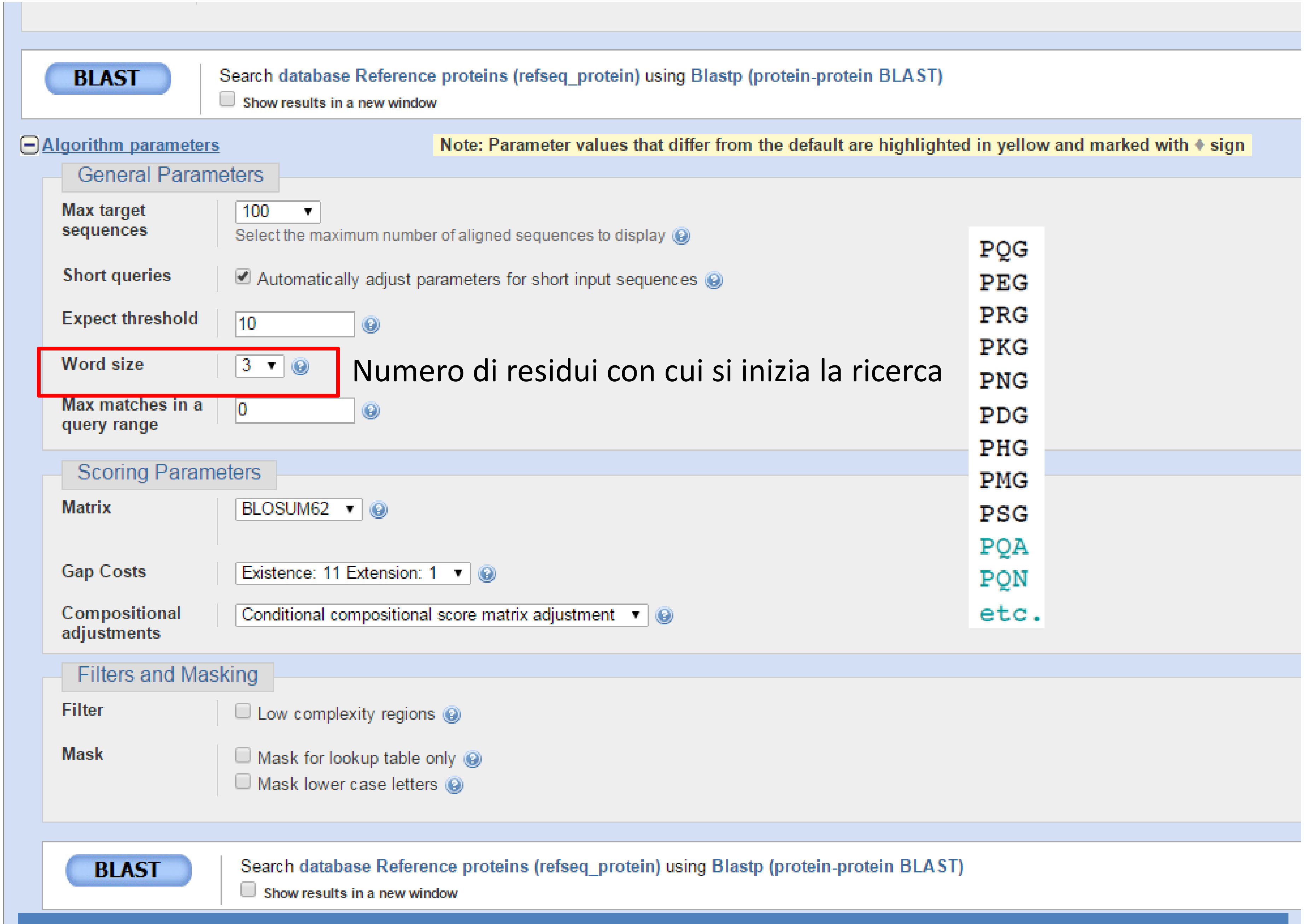
Filter: □ Low complexity regions

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

BLAST | Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)
□ Show results in a new window

Verranno presentate tutte le hits (sequenze trovate) sotto questa soglia di E values (cioè con E < 10)
https://www.youtube.com/watch?v=nO0wJgZRZJs

Ricordiamo che l’E risponde alla domanda: quante sequenze mi aspetto che abbiano **per caso** uno score maggiore o uguale a quello che ho osservato (falsi positivi!)



Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

□ Show results in a new window

⊟ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

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

Numero di residui con cui si inizia la ricerca

Scoring Parameters

Matrix	BLOSUM62 ▾
Gap Costs	Existence: 11 Extension: 1 ▾
Compositional adjustments	Conditional compositional score matrix adjustment ▾

Filters and Masking

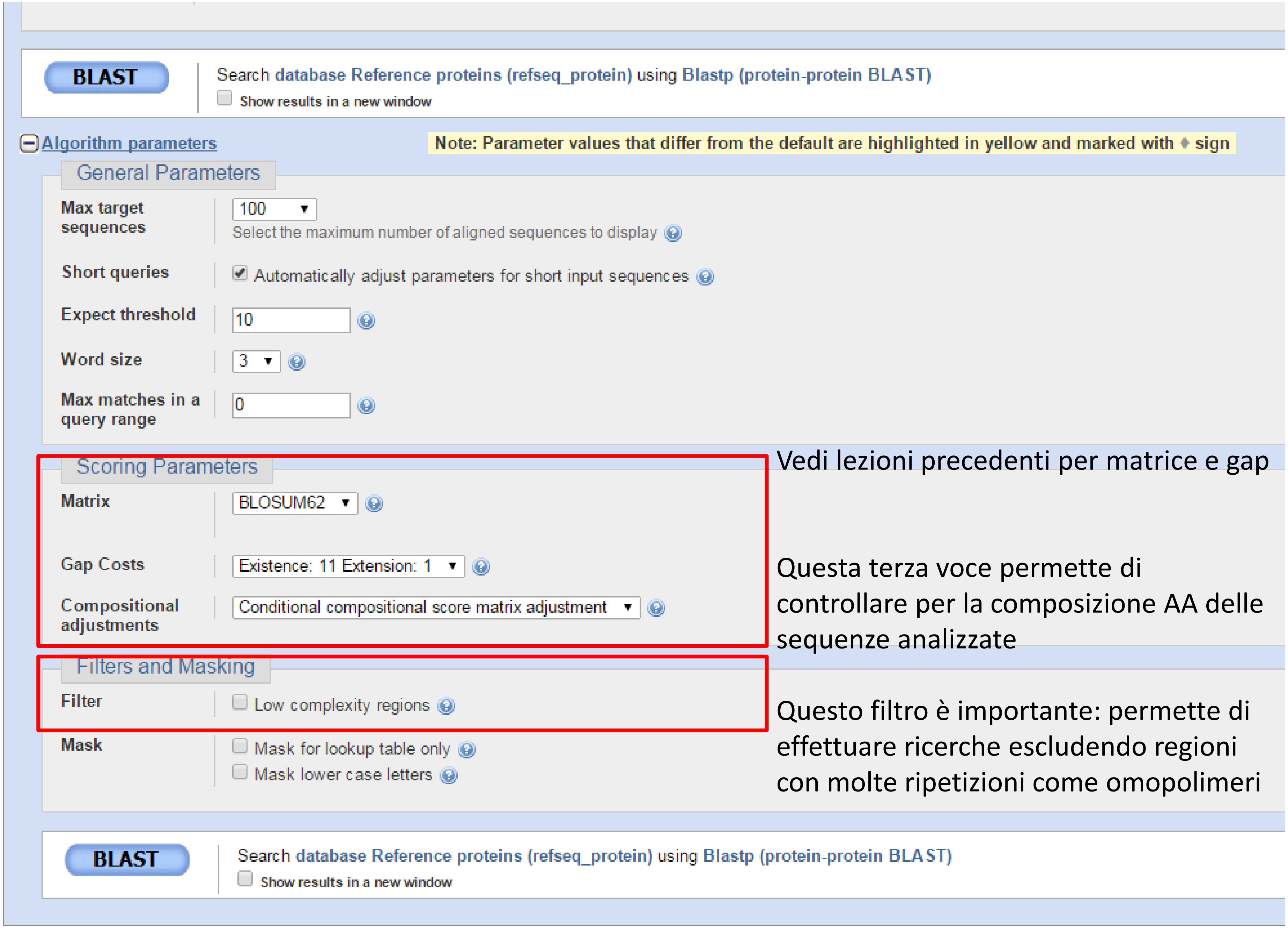
Filter	□ Low complexity regions
Mask	□ Mask for lookup table only □ Mask lower case letters

Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

□ Show results in a new window

PQG
PEG
PRG
PKG
PNG
PDG
PHG
PMG
PSG
PQA
PQN
etc.

Ricordiamo che l'E risponde alla domanda: quante sequenze mi aspetto che abbiano **per caso** uno score maggiore o uguale a quello che ho osservato (falsi positivi!)



Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

☐ Show results in a new window

⊟ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

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

Search database Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)

☐ Show results in a new window

Vedi lezioni precedenti per matrice e gap

Questa terza voce permette di controllare per la composizione AA delle sequenze analizzate

Questo filtro è importante: permette di effettuare ricerche escludendo regioni con molte ripetizioni come omopolimeri

Low-Complexity Regions

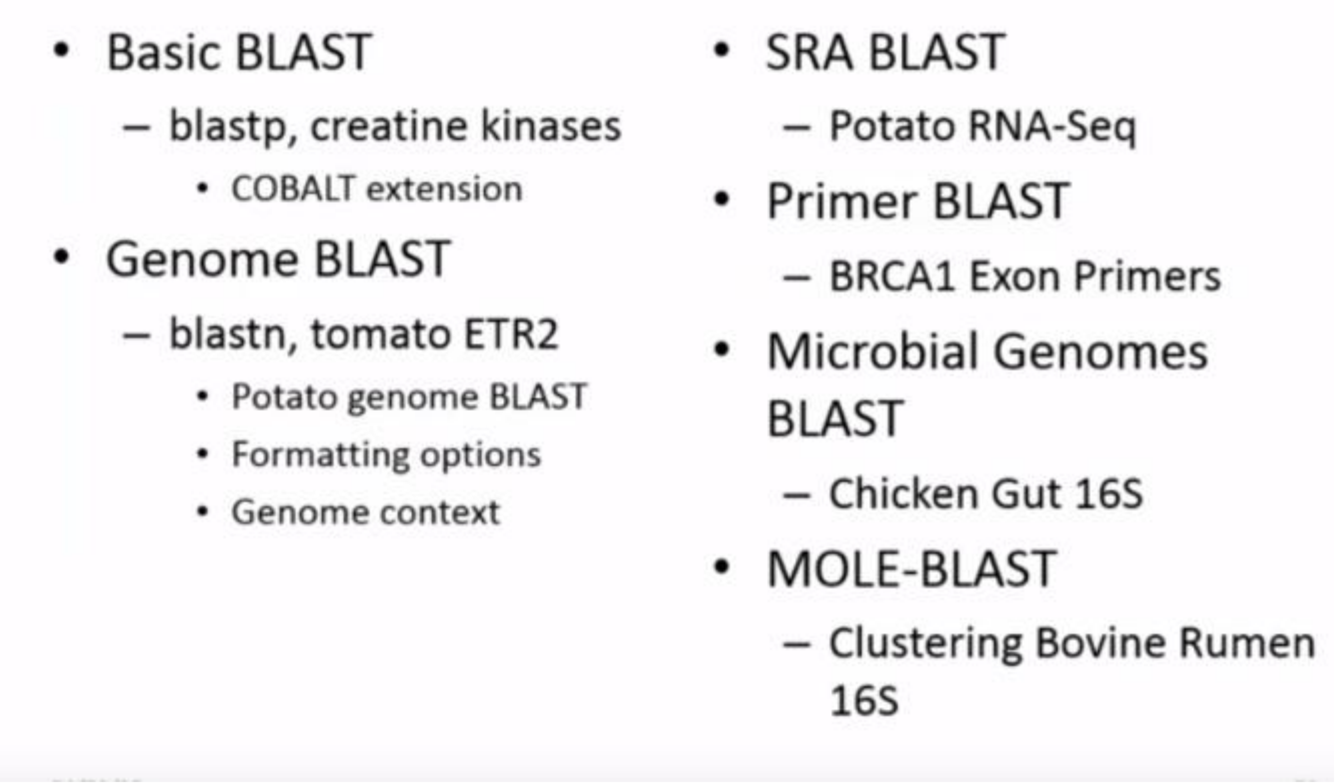
- Defined as regions of “biased composition”
  - Homopolymeric runs
  - Short-period repeats
  - Subtle over-representation of several residues

- May confound sequence analysis
  - BLAST relies on uniformly-distributed amino acid frequencies
  - Often lead to false positives

- Filtering is advised (but *not* enabled by default)

Esercizi con BLAST

- Proviamo ad effettuare una ricerca con le sequenze disponibili nel file
- BLAST



- Basic BLAST
  - blastp, creatine kinases
    - COBALT extension
- Genome BLAST
  - blastn, tomato ETR2
    - Potato genome BLAST
    - Formatting options
    - Genome context
- SRA BLAST
  - Potato RNA-Seq
- Primer BLAST
  - BRCA1 Exon Primers
- Microbial Genomes BLAST
  - Chicken Gut 16S
- MOLE-BLAST
  - Clustering Bovine Rumen 16S