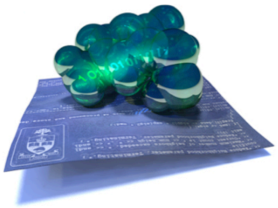
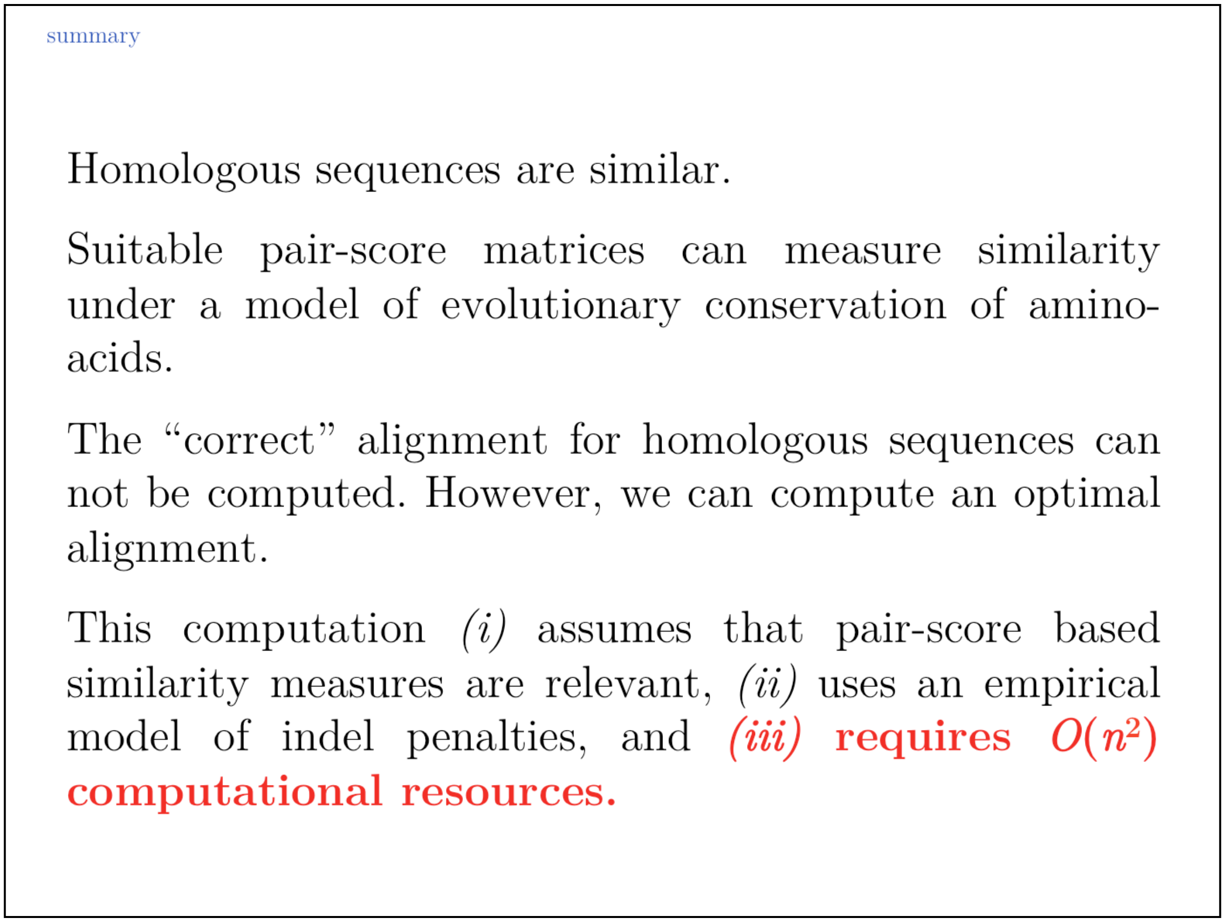
A
Bioinformatics
Course

BLAST



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



summary

Homologous sequences are similar.

Suitable pair-score matrices can measure similarity under a model of evolutionary conservation of amino-acids.

The “correct” alignment for homologous sequences can not be computed. However, we can compute an optimal alignment.

This computation *(i)* assumes that pair-score based similarity measures are relevant, *(ii)* uses an empirical model of indel penalties, and ***(iii)* requires $O(n^2)$ computational resources.**

The $O(n^2)$ resource requirement (where n is the alignment length) means the algorithm is too slow for searches on a database scale i.e. in very large search spaces.

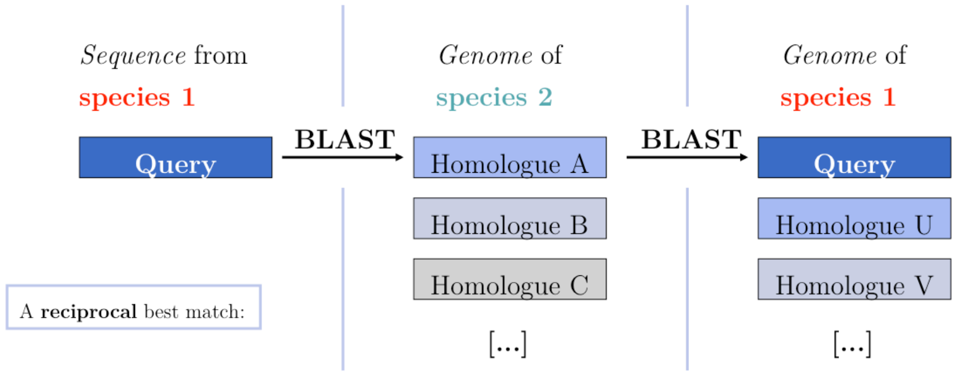
why large search spaces?

Some procedures require **genome-wide** or **database-wide** similarity searches. Such searches are not feasible with *optimal sequence alignment* algorithms.

Example: discovering orthologues

Computationally, an orthologue has been defined when the best match in another specie's genome has the original sequence as it's own best match. This is called the "**reciprocal best-match**" criterion (RBM).

(This is a useful procedure, but not exactly the same as the definition of an orthologue.)



Sequence from **species 1**

Genome of **species 2**

Genome of **species 1**

Query

BLAST

Homologue A

Homologue B

Homologue C

[...]

BLAST

Query

Homologue U

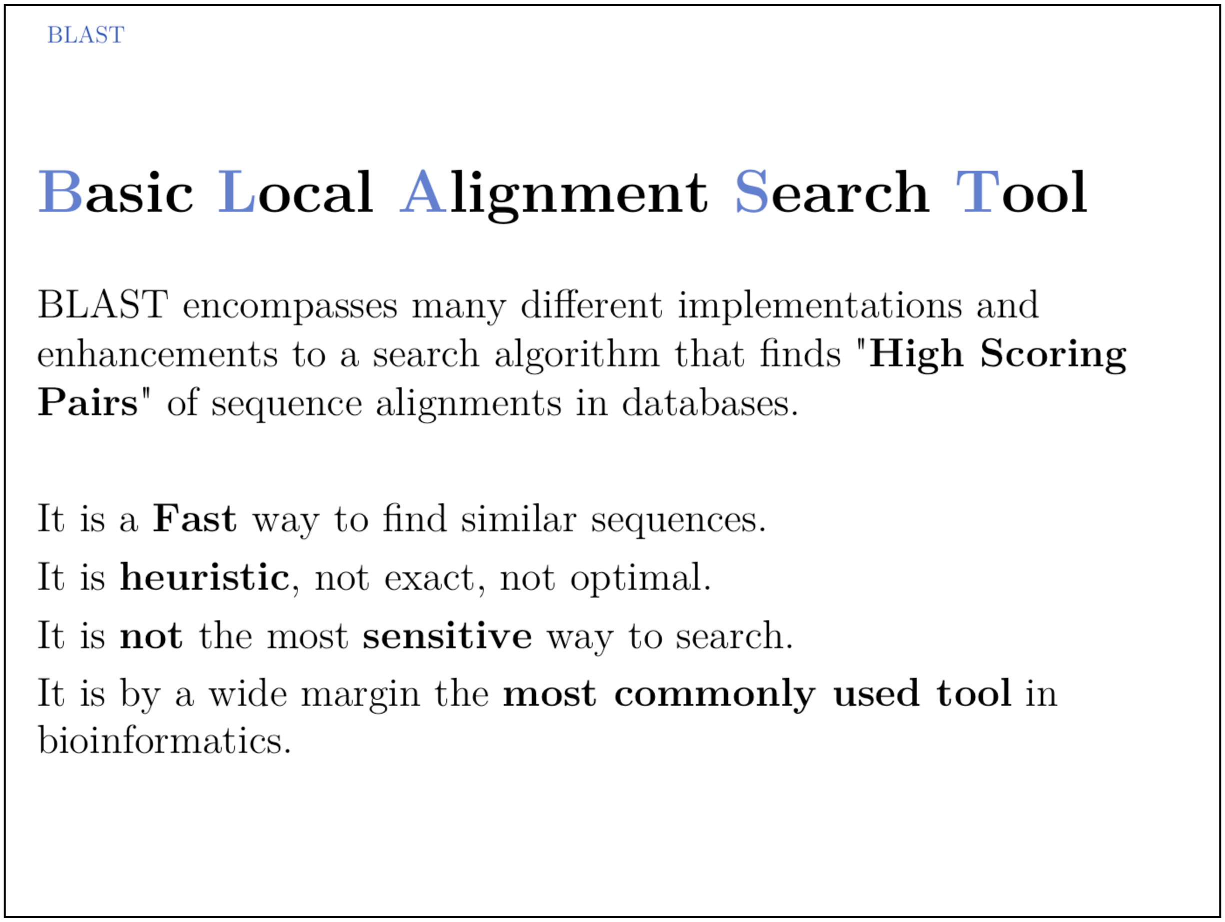
Homologue V

[...]

A **reciprocal** best match:

Some computational strategies absolutely require genome-wide searches: e.g. the computational definition of orthologues, or compiling evolutionary conservation patterns.

The tight integration of search capacity with database holdings is the key to the utility of the data. Investments in sequencing **only** pay off when the sequences are easily accessible!



BLAST

Basic Local Alignment Search Tool

BLAST encompasses many different implementations and enhancements to a search algorithm that finds "**High Scoring Pairs**" of sequence alignments in databases.

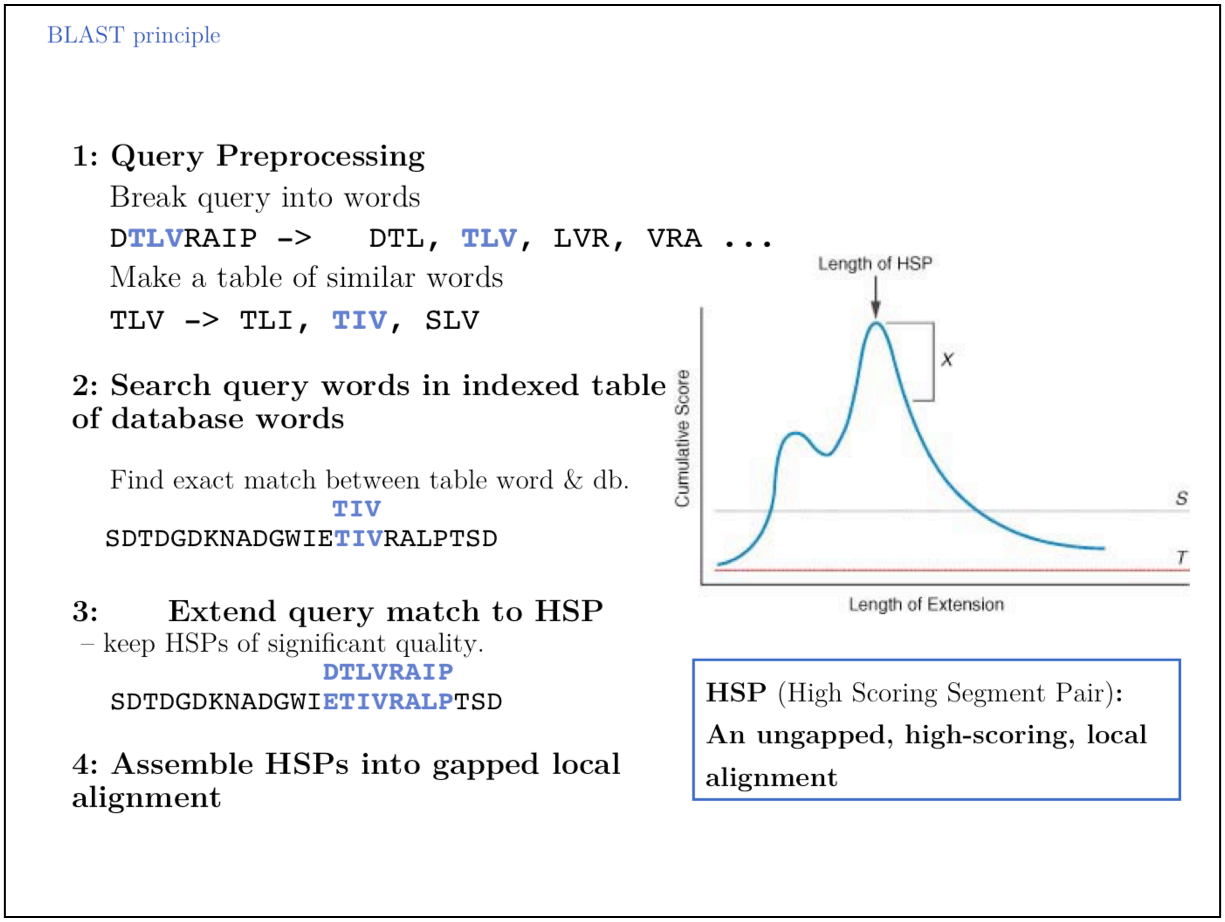
It is a **Fast** way to find similar sequences.

It is **heuristic**, not exact, not optimal.

It is **not** the most **sensitive** way to search.

It is by a wide margin the **most commonly used tool** in bioinformatics.

BLAST was developed as a heuristic alternative to exact alignment, looking for a way to compute fast, repeated searches in large search spaces much more efficiently than what is possible with optimal pairwise alignments. The strategy is to pre-compute similarities and then piece a match together from quickly retrieved partial matches.



BLAST principle

1: Query Preprocessing

Break query into words

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

Make a table of similar words

`TLV -> TLI, TIV, SLV`

2: Search query words in indexed table of database words

Find exact match between table word & db.

```
                  TIV
SDTDGDKNADGWIETIVRALPTSD
```

3: Extend query match to HSP

– keep HSPs of significant quality.

```
                  DTLVRAIP
SDTDGDKNADGWIETIVRALPTSD
```

4: Assemble HSPs into gapped local alignment

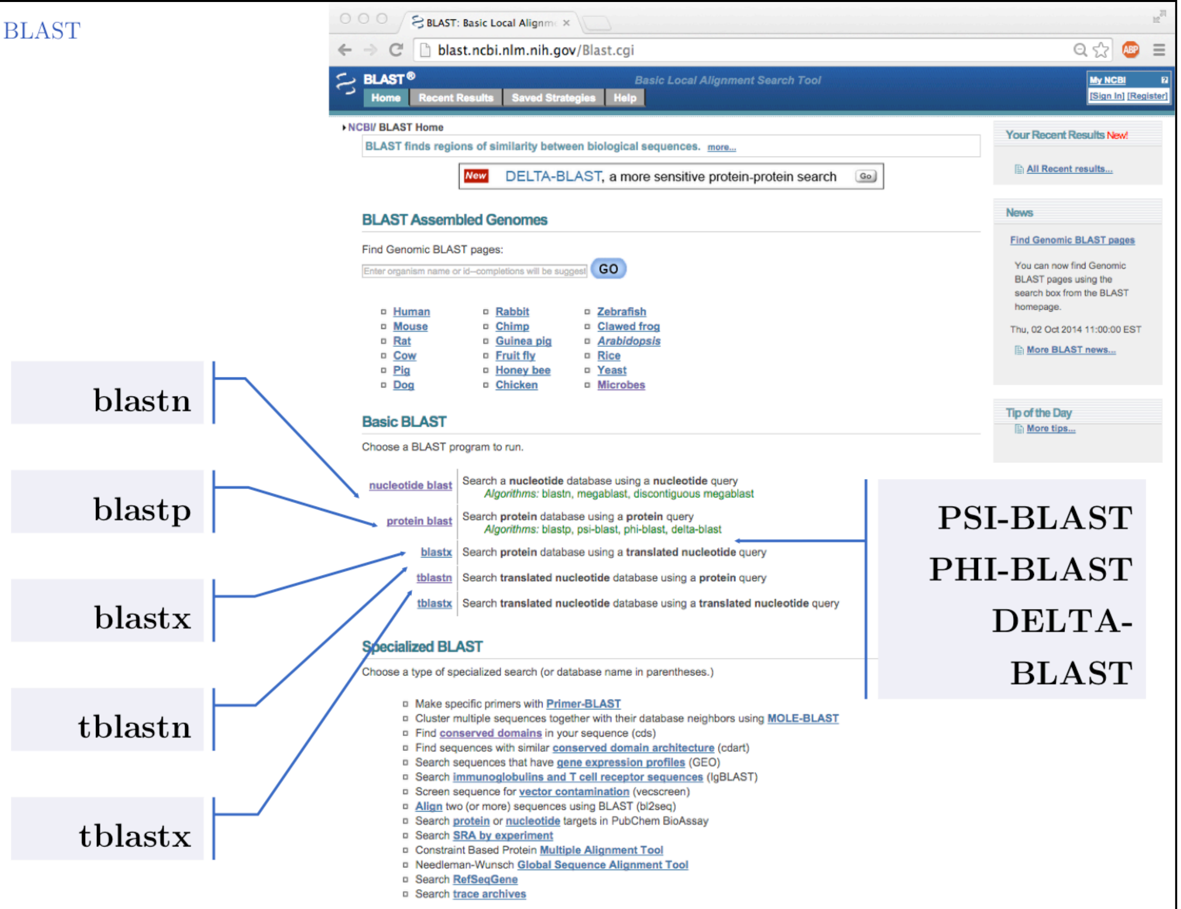
HSP (High Scoring Segment Pair): **An ungapped, high-scoring, local alignment**

The enormous speed-up of BLAST is due to its use of an **indexed table** of database "words". The index is a list of positions at which each word occurs in the database. Using an index, it is very easy to examine every occurrence of a word in the database and try to extend the word match on both sides with additional similar sequence. The extension does not introduce gaps, because this is faster, but also because the statistics of ungapped alignments are tractable! The final step is the assembly of significant hits into longer alignments.

Note that BLAST is **heuristic**, not **optimal** and that it is a **local**, not **global** alignment algorithm.

See also: Altschul *et al.* (1990): http://www.ncbi.nlm.nih.gov/pubmed/2231712

BLAST



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 suggest GO

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

Your Recent Results New!

All Recent results...

News

Find Genomic BLAST pages

You can now find Genomic BLAST pages using the search box from the BLAST homepage.

Thu, 02 Oct 2014 11:00:00 EST

More BLAST news...

Tip of the Day

More tips...

Basic BLAST

Choose a BLAST program to run.

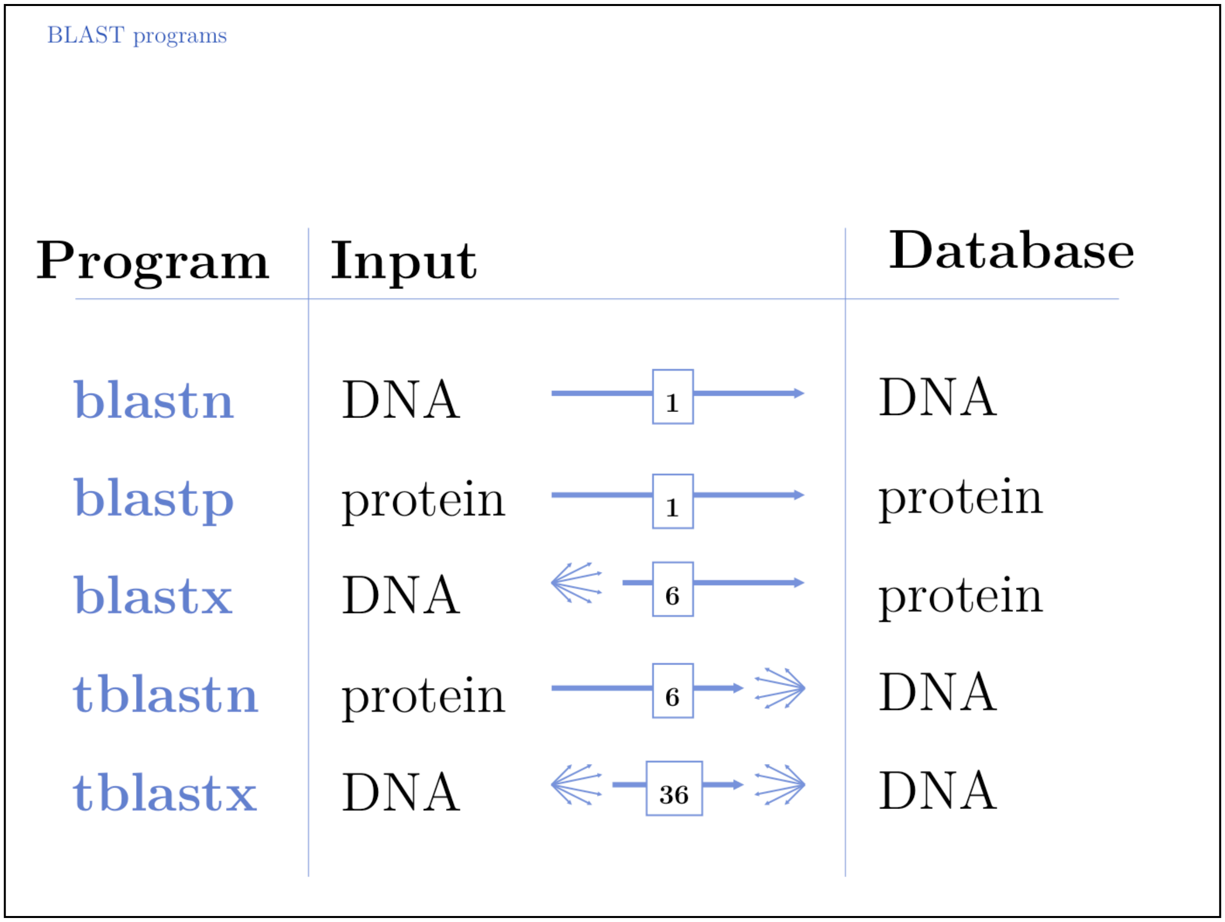
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using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

The **BLAST** home page offers a number of different BLAST *flavours*.



BLAST programs

Program	Input		Database
blastn	DNA	1	DNA
blastp	protein	1	protein
blastx	DNA	6	protein
tblastn	protein	6	DNA
tblastx	DNA	36	DNA

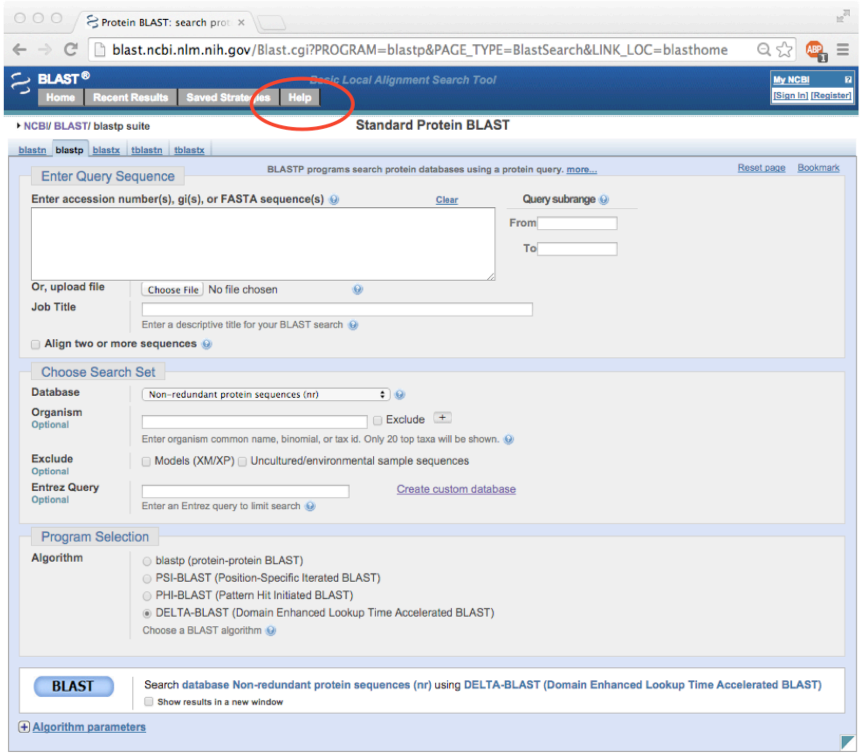
blastn and **blastp** search for data in nucleotide and protein databases, respectively.

blastx starts from a nucleotide sequence, translates it in all six reading frames, and searches all six peptide sequences gainst a protein database. This is useful to discover translated sequence in unannotated DNA sequence (6 translated subject sequences).

tblastn searches for a specific protein sequence in all six reading frames of a DNA sequence, such as an unannotated genome (6 translated queries).

tblastx compares two DNA sequences by aligning all six open reading frames against each other (36 alignments). This is useful to detect homology in unannotated DNA sequences that are so far diverged that DNA comparisons don't yield statistically significant similarities. The strategy exploits the fact that protein sequence is much more highly conserved than DNA sequence.

databases



blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

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

Enter Query Sequence

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

Query subrange

From

To

Or, upload file: No file chosen

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): □ 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): 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

Search database Non-redundant protein sequences (nr) using DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

□ Show results in a new window

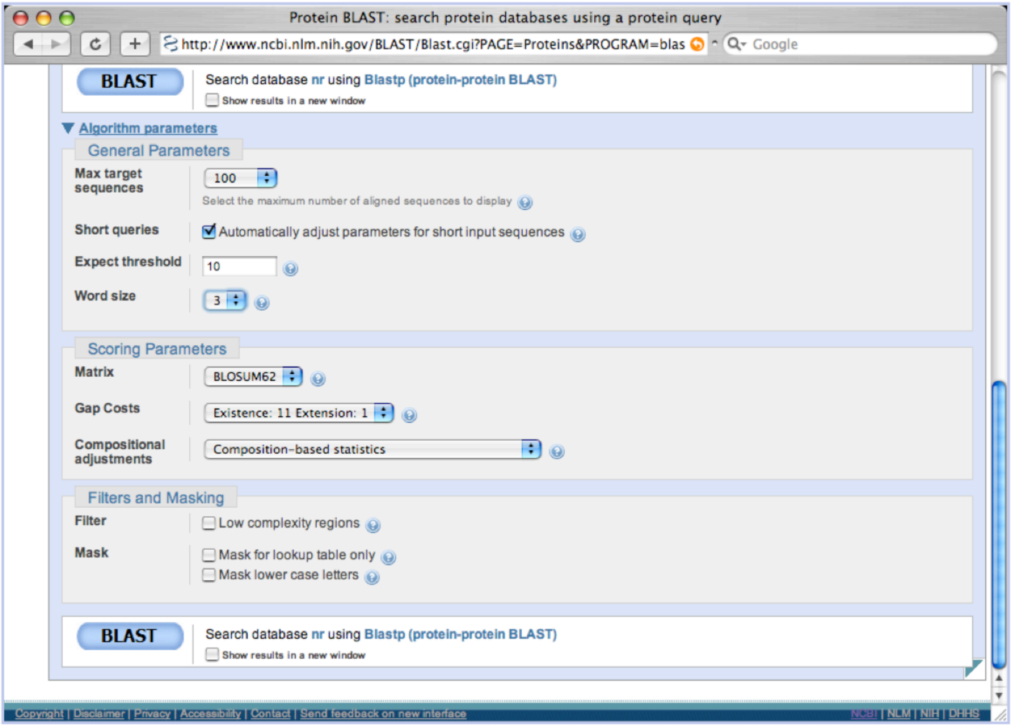
Extensive help is available (and should be read!) for each of the options. Take the time to read the **Web BLAST options document**[1] and be sure to understand how to format input, what databases are available and how the choice of database influences the results. If you are not confident with the document, ask on the course list.

If there is no well-understood reason to do otherwise, use refseq protein as your database of first choice. If you are specifically looking for structural annotations, use the PDB sequences.

The Help page contains detailed guides to the **search interface** and the **report output**!

[1] http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml

algorithm


Protein BLAST: search protein databases using a protein query
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blas
Google
BLAST
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS

Search database nr using Blastp (protein-protein BLAST)
□ Show results in a new window

▼ 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

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Composition-based statistics

Filters and Masking

Filter: □ Low complexity regions

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

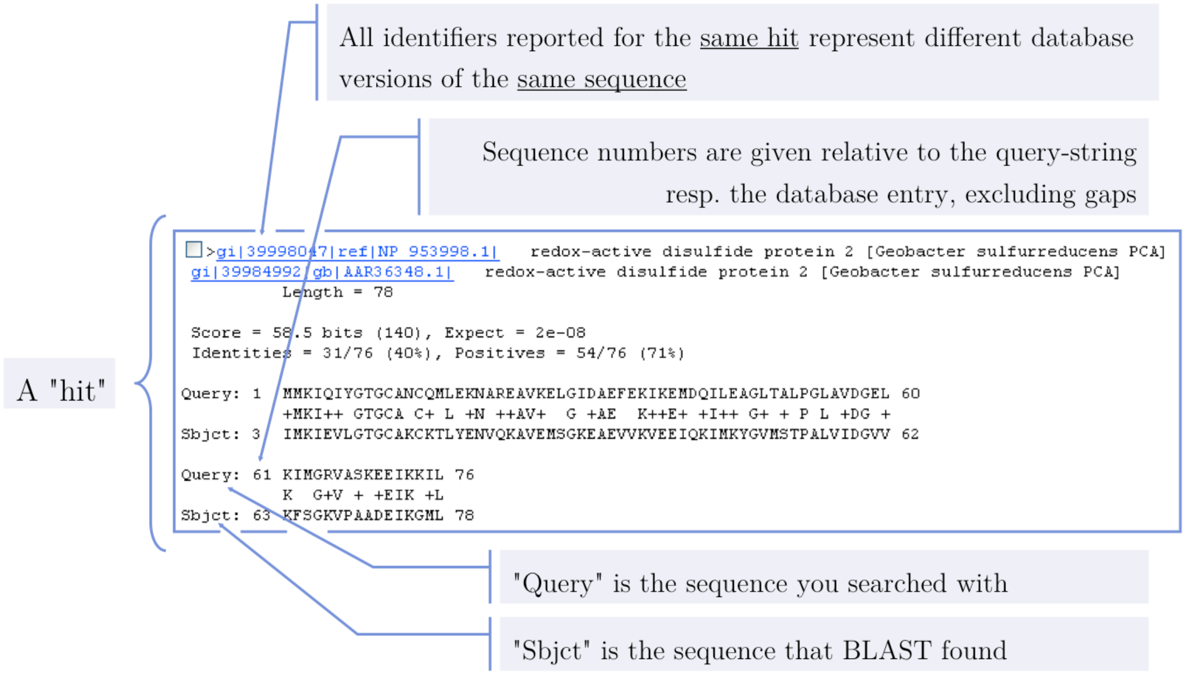
Search database nr using Blastp (protein-protein BLAST)
□ Show results in a new window

BLAST parameters: algorithm

Be sure to understand the choices and their consequences for **Composition-based statistics**[1] and for **Filtering and Masking** segments of low complexity in your query. Filtering is an important option to consider especially for PSI-BLAST searches!

[1] http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#compositional_adjustments
[2] http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#filter

results



All identifiers reported for the same hit represent different database versions of the same sequence

Sequence numbers are given relative to the query-string resp. the database entry, excluding gaps

A "hit"

```
□>gi|39998047|ref|NP 953998.1|   redox-active disulfide protein 2 [Geobacter sulfurreducens PCA]
 gi|39984992|gb|AAR36348.1|   redox-active disulfide protein 2 [Geobacter sulfurreducens PCA]
          Length = 78

 Score = 58.5 bits (140), Expect = 2e-08
 Identities = 31/76 (40%), Positives = 54/76 (71%)

Query: 1  MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKEMDQILEAGLTALPGLAVDGEL 60
          +MKI++ GTGCA C+ L +N ++AV+  G +AE  K++E+ +I++ G+ + P L +DG +
Sbjct: 3  IMKIEVLGTGCAKCKTLYENVQKAVEMSGKEAEVVKVEEIQKIMKYGVMSTPALVIDGVV 62

Query: 61 KIMGRVASKEEIKKIL 76
          K  G+V + +EIK +L
Sbjct: 63 KFSGKVPAADEIKGML 78
```

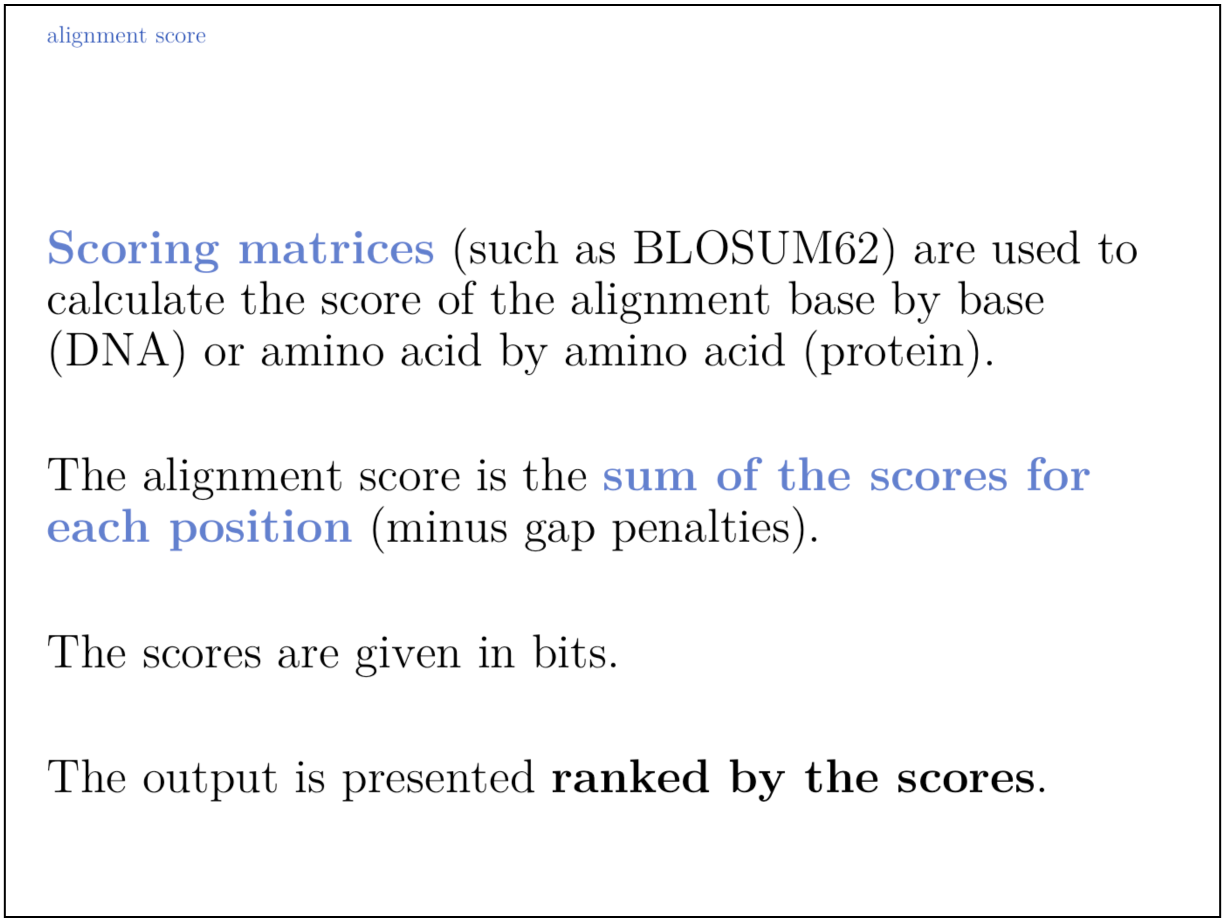
"Query" is the sequence you searched with

"Sbjct" is the sequence that BLAST found

Each Blast **hit** represents an alignment that can contain one or more HSPs (High Scoring Segment Pairs). Note: If a hit is followed by a second hit and no new GI number, it identifies a second region of similarity in the **same sequence**.

Remeber that BLAST is a local alignment algorithm. If two sequences are homologous over their whole length, but there are regions of sequence that separate conserved domains that cannot be well aligned, BLAST will return separate alignments to the domains. These are sorted by the E value of the better match in the list of outputs.

In Mbp1 homologues we see this frequently – the APSES domain and the ankyrin domains often give rise to such separated blocks of alignment.



alignment score

Scoring matrices (such as BLOSUM62) are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).

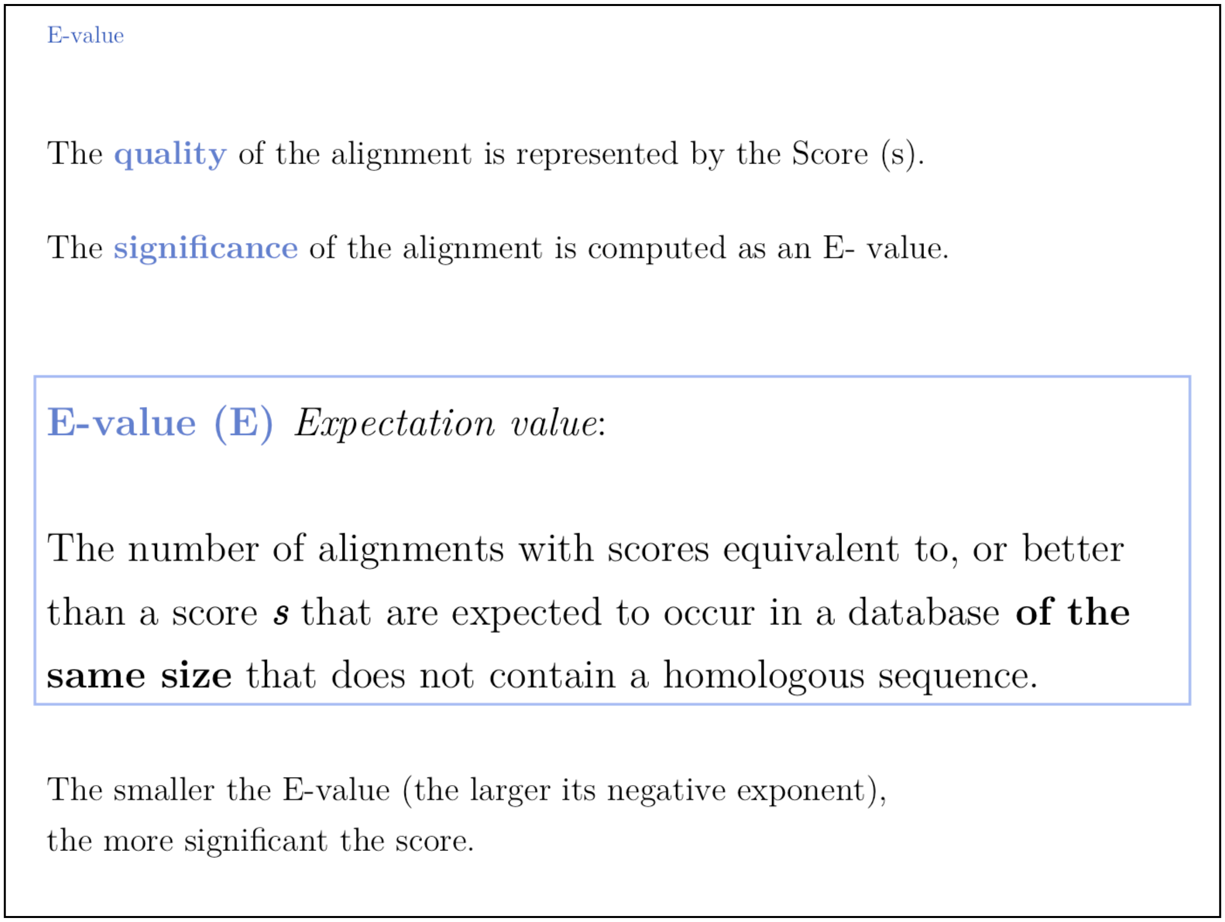
The alignment score is the **sum of the scores for each position** (minus gap penalties).

The scores are given in bits.

The output is presented **ranked by the scores**.

Normally scores depend on the scoring matrix that was used and can't be compared between different matrices and after applying different gap penalties. However the NCBI matrices have been normalized in bits, thus the scores between alignments with different matrices **can** be compared, (this is not generally the case with other matrices). In addition the percentage of *Identical* and *similar ("positives")* residues and the gap fraction are given.

%-Identities and gap fraction are often used to conclude whether two sequences are homologous, the percentage of positives is not usually used since it depends on the matrix.



E-value

The **quality** of the alignment is represented by the Score (s).

The **significance** of the alignment is computed as an E- value.

E-value (E) *Expectation value*:

The number of alignments with scores equivalent to, or better than a score ***s*** that are expected to occur in a database **of the same size** that does not contain a homologous sequence.

The smaller the E-value (the larger its negative exponent), the more significant the score.

The E-value is a statistically well founded metric that allows us to conclude the likelihood of a spurious alignment. Computing E-values is possible for HSPs since the statistics of gap-less alignments are analytically tractable, whereas gapped alignments have no theoretical description of the distribution of expected scores.

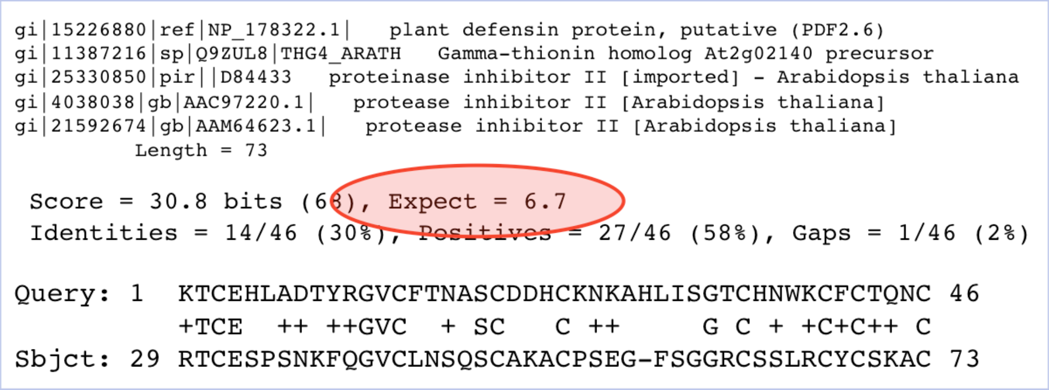
Note that E-values do not represent an assertion about the retrieved sequence, but an assertion about the score and its relation to the expected distribution of scores. Or, to rephrase this, a large E-value does not mean that your hit is not a homologue, but it means that an irrelevant sequence has a a high chance of having just as high a score due to chance similarities. To repeat: a large E-value does not mean your hit is not a homologue. However a small E-value does indeed mean that a chance alignment is unlikely.

It is important to realize that the E-value depends on the database size. Obviously, you would expect randomly high-scoring hits more often in a large database than in a small one. Thus an alignment with the **same score** will have a **smaller E-value** when searched against a specific genome than if you search it against the entire “nr” dataset of GenBank.

More detail in the NCBI tutorial “The Statistics of Sequence Similarity Scores”. (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html) interpretation

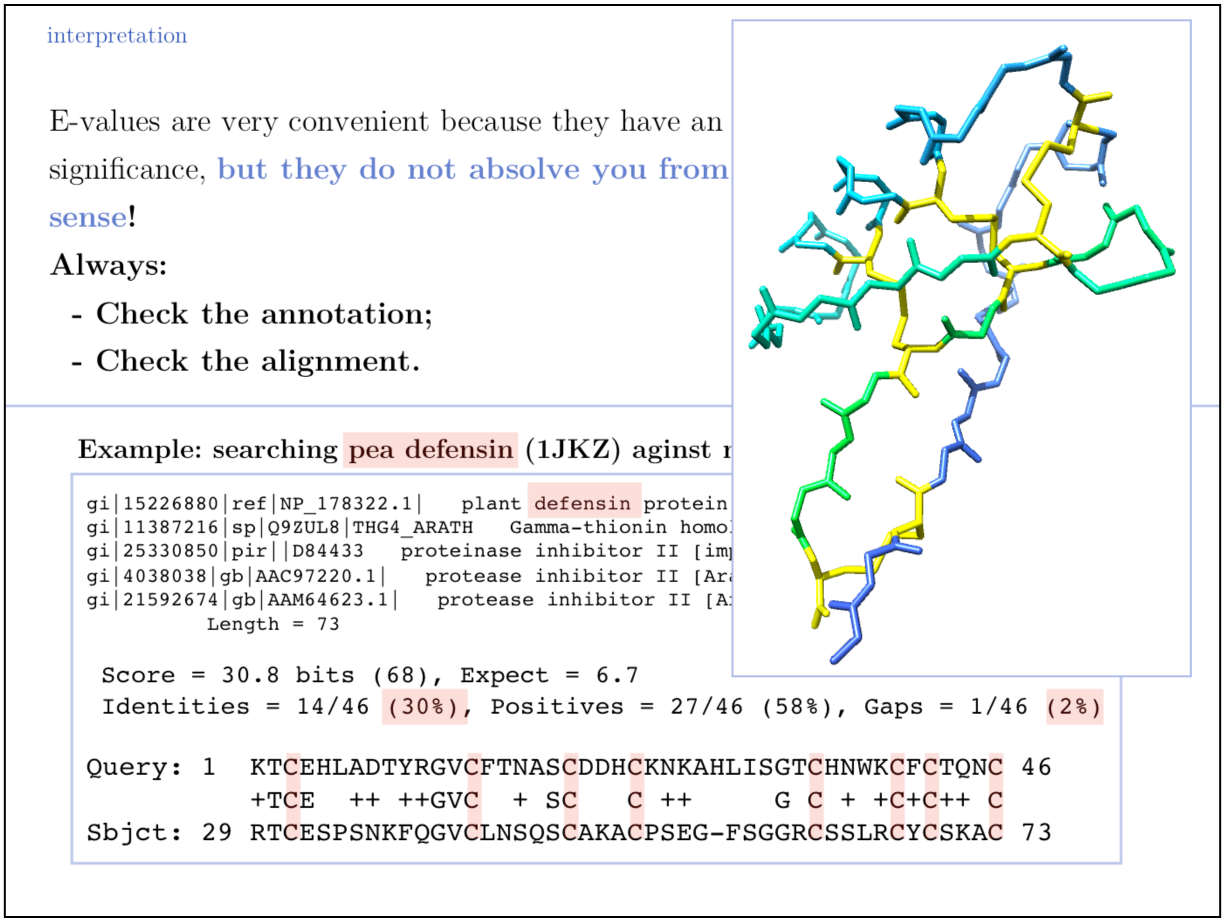
E-values are very convenient because they have an obvious interpretation of significance, **but they do not absolve you from using biological common sense!**

Example: searching pea defensin (1JKZ) aginst nr: are these sequences homologous?



```
gi|15226880|ref|NP_178322.1|   plant defensin protein, putative (PDF2.6)
gi|11387216|sp|Q9ZUL8|THG4_ARATH   Gamma-thionin homolog At2g02140 precursor
gi|25330850|pir||D84433   proteinase inhibitor II [imported] - Arabidopsis thaliana
gi|4038038|gb|AAC97220.1|   protease inhibitor II [Arabidopsis thaliana]
gi|21592674|gb|AAM64623.1|   protease inhibitor II [Arabidopsis thaliana]
          Length = 73

 Score = 30.8 bits (68), Expect = 6.7
 Identities = 14/46 (30%), Positives = 27/46 (58%), Gaps = 1/46 (2%)

Query: 1  KTCEHLADTYRGVCFTNASCDDHCKNKAHLISGTCHNWKCFCTQNC 46
          +TCE  ++ ++GVC  + SC   C ++     G C + +C+C++ C
Sbjct: 29 RTCESPSNKFQGVCLNSQSCAKACPSEG-FSGGRCSSLRCYCSKAC 73
```

In the example above, the BLAST search of a pea defensin - PDB structure 1JKZ - achieved an E-value of only 6.7.



interpretation

E-values are very convenient because they have an
significance, but they do not absolve you from
sense!

Always:

- **Check the annotation;**
- **Check the alignment.**

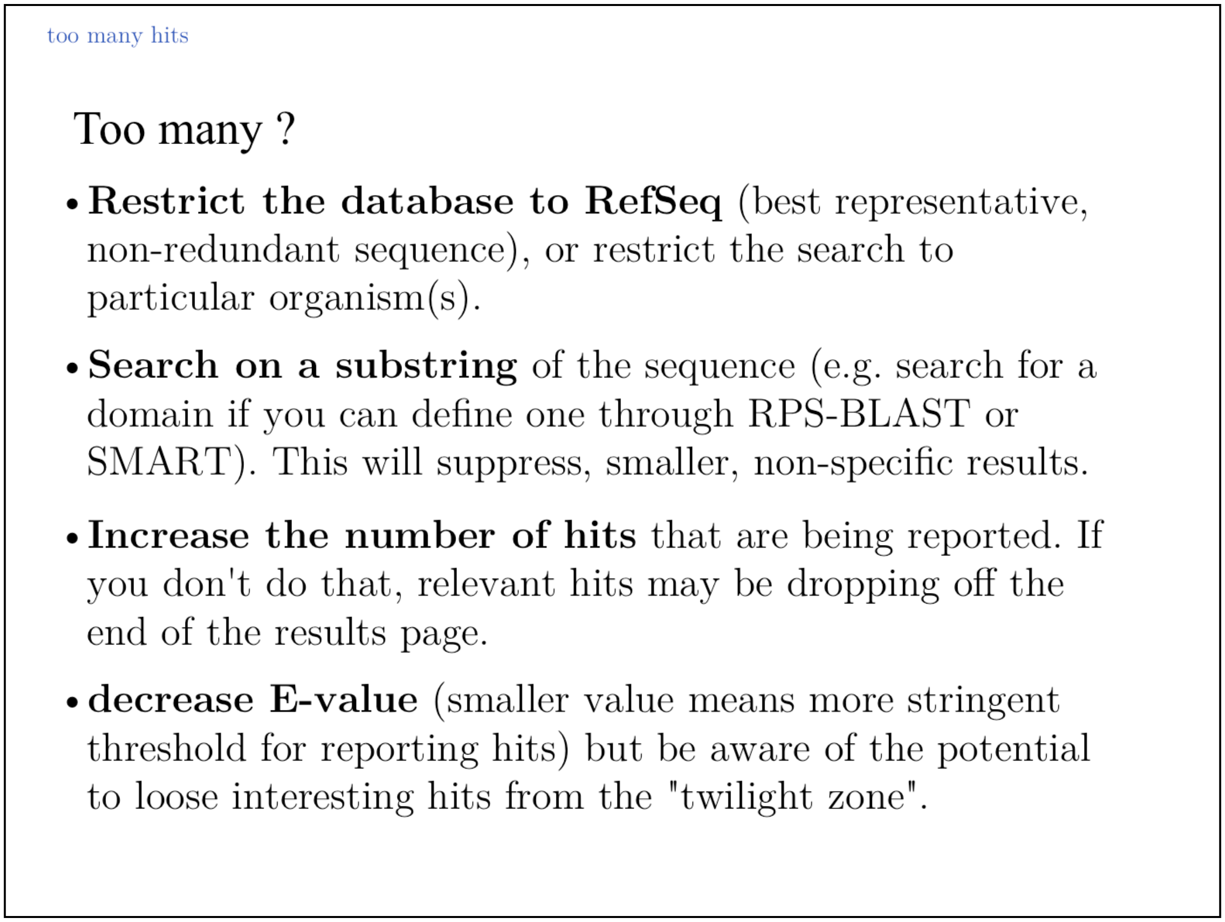
Example: searching pea defensin (1JKZ) aginst r

```
gi|15226880|ref|NP_178322.1|   plant defensin protein
gi|11387216|sp|Q9ZUL8|THG4_ARATH   Gamma-thionin homo]
gi|25330850|pir||D84433   proteinase inhibitor II [imp
gi|4038038|gb|AAC97220.1|   protease inhibitor II [Ara
gi|21592674|gb|AAM64623.1|   protease inhibitor II [A:
          Length = 73

 Score = 30.8 bits (68), Expect = 6.7
 Identities = 14/46 (30%), Positives = 27/46 (58%), Gaps = 1/46 (2%)

Query: 1  KTCEHLADTYRGVCFTNASCDDHCKNKAHLISGTCHNWKCFCTQNC 46
             +TCE   ++ ++GVC   + SC    C ++      G C + +C+C++ C
Sbjct: 29 RTCESPSNKFQGVCLNSQSCAKACPSEG-FSGGRCSSLRCYCSKAC 73
```

However the hit that was retrieved:

- is annotated as an *arabidopsis* **defensin**;
- has **30% identity** over the entire domain, albeit the domain is small;
- requires only **one single gap** for alignment; and
- **has each and every single cysteine conserved, when compared to the query**!

Each of these additional observations alone could have led you to conclude homology. It should be obvious for example that the aligned cysteines are extremely unlikely to be due to a random similarity of unrelated sequences! The large E-value is primarily due to the fact that the protein sequences are quite short.



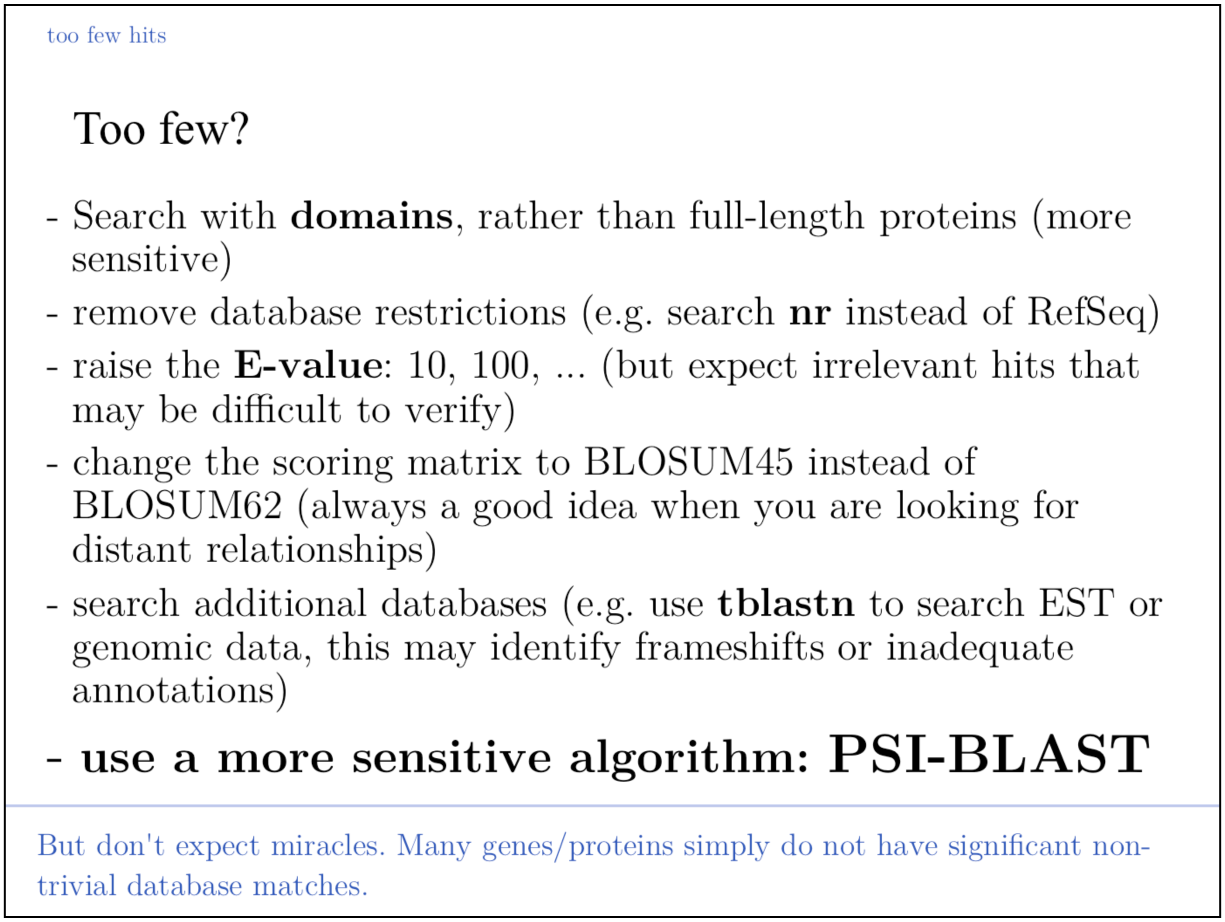
too many hits

Too many ?

- **Restrict the database to RefSeq** (best representative, non-redundant sequence), or restrict the search to particular organism(s).
- **Search on a substring** of the sequence (e.g. search for a domain if you can define one through RPS-BLAST or SMART). This will suppress, smaller, non-specific results.
- **Increase the number of hits** that are being reported. If you don't do that, relevant hits may be dropping off the end of the results page.
- **decrease E-value** (smaller value means more stringent threshold for reporting hits) but be aware of the potential to loose interesting hits from the "twilight zone".

How can there be too many hits, when *lots-of-hits* is what you are looking for? Either you find redundant sequences or trivially similar sequences that are obscuring the rare, interesting similarities you are looking for (GFP or other fusion proteins and ankyrin domains come to mind, for example), or you are searching in a database section that contains redundant sequences.

Note that restricting by organism does not restrict the search, but only the list of results that are being reported. The search takes just as long. Only the specialized genome search pages and some non-NCBI databases of model-organism genome projects offer BLAST searches on reduced datasets. These searches are faster.



too few hits

Too few?

- Search with **domains**, rather than full-length proteins (more sensitive)
- remove database restrictions (e.g. search **nr** instead of RefSeq)
- raise the **E-value**: 10, 100, ... (but expect irrelevant hits that may be difficult to verify)
- change the scoring matrix to BLOSUM45 instead of BLOSUM62 (always a good idea when you are looking for distant relationships)
- search additional databases (e.g. use **tblastn** to search EST or genomic data, this may identify frameshifts or inadequate annotations)
- **use a more sensitive algorithm: PSI-BLAST**

But don't expect miracles. Many genes/proteins simply do not have significant non-trivial database matches.

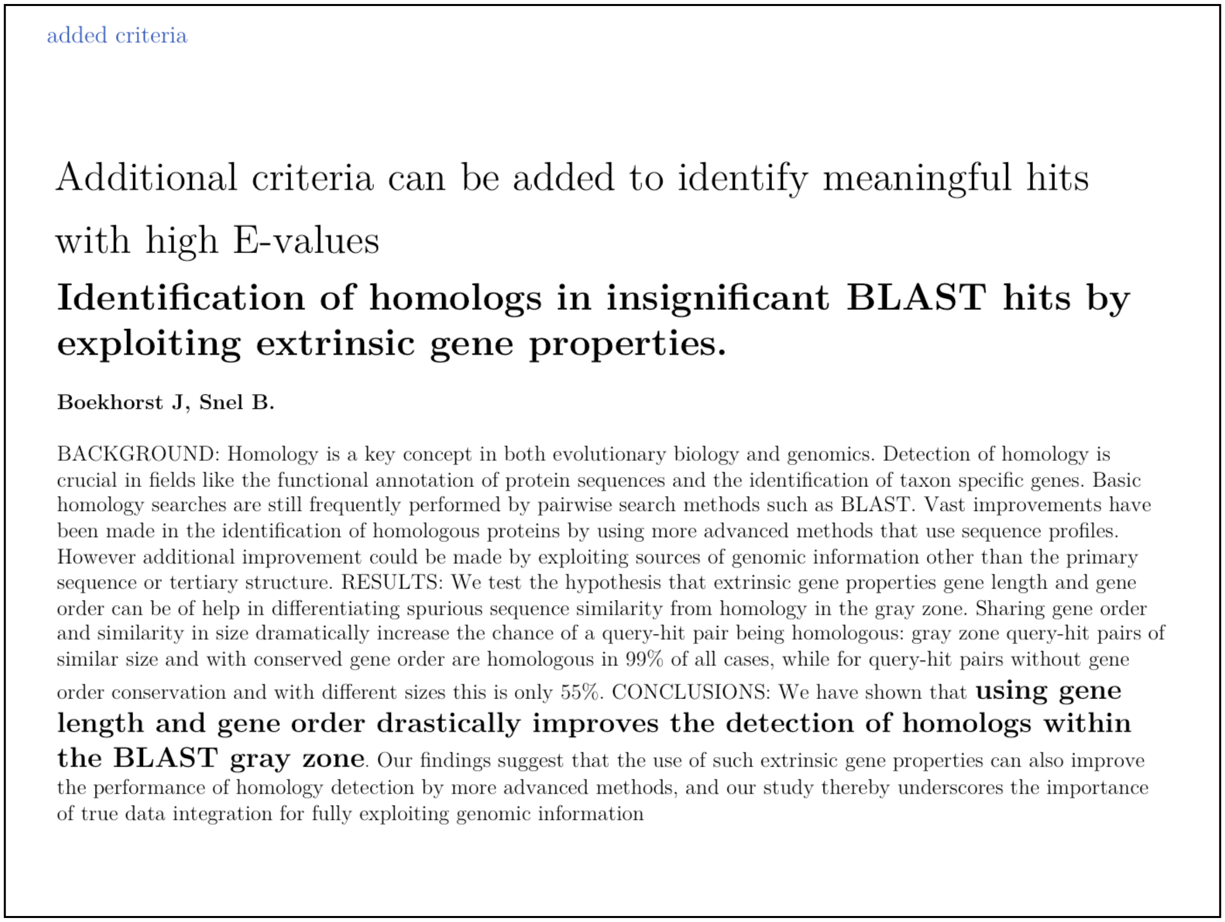
How many genes have no homologues? That depends. Unknown genes (or "ORFans") may comprise a significant (albeit diminishing) fraction of genomes.

In general, between 10 and 30% of sequences may fall into this category and it is likely that even the most closely related species have sequences that are unique.

See Siew&Fischer (2003)[1]and a discussion of the role of viral horizontal gene transfer in ORFans by Yin and Fischer (2006)[2]

[1] http://www.ncbi.nlm.nih.gov/pubmed/12517334

[2] http://www.ncbi.nlm.nih.gov/pubmed/16914045



added criteria

Additional criteria can be added to identify meaningful hits with high E-values

Identification of homologs in insignificant BLAST hits by exploiting extrinsic gene properties.

Boekhorst J, Snel B.

BACKGROUND: Homology is a key concept in both evolutionary biology and genomics. Detection of homology is crucial in fields like the functional annotation of protein sequences and the identification of taxon specific genes. Basic homology searches are still frequently performed by pairwise search methods such as BLAST. Vast improvements have been made in the identification of homologous proteins by using more advanced methods that use sequence profiles. However additional improvement could be made by exploiting sources of genomic information other than the primary sequence or tertiary structure. RESULTS: We test the hypothesis that extrinsic gene properties gene length and gene order can be of help in differentiating spurious sequence similarity from homology in the gray zone. Sharing gene order and similarity in size dramatically increase the chance of a query-hit pair being homologous: gray zone query-hit pairs of similar size and with conserved gene order are homologous in 99% of all cases, while for query-hit pairs without gene order conservation and with different sizes this is only 55%. CONCLUSIONS: We have shown that **using gene length and gene order drastically improves the detection of homologs within the BLAST gray zone**. Our findings suggest that the use of such extrinsic gene properties can also improve the performance of homology detection by more advanced methods, and our study thereby underscores the importance of true data integration for fully exploiting genomic information

Not surprisingly, the more information you can add to your query, the more sensitive your search can become.

http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA