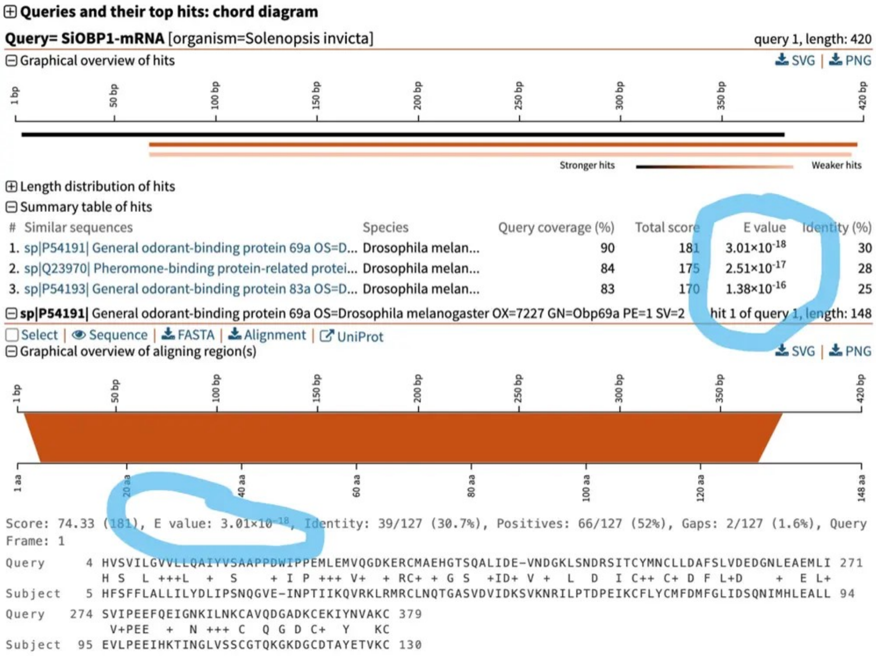
BLAST E-values: how they are calculated and what they mean

A crucial measure that accompanies every hit sequence that BLAST identifies is the E-value (Expectation value). Here, we'll walk through:

- what an E-value is,
- how it is calculated,
- how to interpret it,
- and how to get the most power & sensitivity from your BLAST analysis.



Query= SiOBP1-mRNA [organism=Solenopsis invicta] query 1, length: 420

#	Similar sequences	Species	Query coverage (%)	Total score	E value	Identity (%)
1.	sp|P54191| General odorant-binding protein 69a OS=D...	Drosophila melan...	90	181	3.01×10^{-18}	30
2.	sp|Q23970| Pheromone-binding protein-related protei...	Drosophila melan...	84	175	2.51×10^{-17}	28
3.	sp|P54193| General odorant-binding protein 83a OS=D...	Drosophila melan...	83	170	1.38×10^{-16}	25

sp|P54191| General odorant-binding protein 69a OS=Drosophila melanogaster OX=7227 GN=Obp69a PE=1 SV=2 hit 1 of query 1, length: 148

```
Score: 74.33 (181), E value: 3.01×10^-18, Identity: 39/127 (30.7%), Positives: 66/127 (52%), Gaps: 2/127 (1.6%), Query
Frame: 1
Query     4 HVSVILGVVLLQAIYVSAAPPDWIPPEMLEMVQGDKERCMAEHGTSQALIDE-VNDGKLSNDRSITCYMNCLLDAFSLVDEDGNLEAEMLI 271
            H S  L +++L  +  S    + I P +++ V+  + RC+ + G S  +ID+ V +  L  D  I C++ C+ D F L+D    +  E L+
Subject   5 HFSFFLALLILYDLIPSNQGVE-INPTIIKQVRKLRMRCLNQTGASVDVIDKSVKNRILPTDPEIKCFLYCMFDMFGLIDSQNIMHLEALL 94
Query   274 SVIPEEFQEIGNKILNKCAVQDGADKCEKIYNVAKC 379
             V+PEE  +  N +++ C  Q G D C+  Y   KC
Subject  95 EVLPEEIHKTINGLVSSCGTQKGKDGCDTAYETVKC 130
```

SequenceServer BLAST result highlighting that E-values are shown in two places in the BLAST result report: in the table of of all hits, and as part of the alignment of each hit.

What is an E-value?

The BLAST E-value is:

- not a p-value.
- not the exact number of times a sequence was found due to chance.

Instead, it is an **estimate of the expected number of random alignments** with a particular score or better that could be found by chance in a given database search. In other words, it represents the likelihood that a specific sequence alignment is due to chance rather than a true biological relationship between the sequences.

Interpreting E-values

E-values should be interpreted in the context of the specific research question and alongside other factors like alignment length, sequence identity, and biological context or question. In general:

- Lower (i.e., stronger) E-values indicate more significant alignments, suggesting a higher probability that the sequences share a common evolutionary origin.
- A higher (i.e., weaker) E-value indicates that the alignment might be a random event.

E-values are not fixed thresholds for determining the significance of an alignment. Always consider the biological context.

In many cases, BLAST analysis is just a first step. In particular, a stronger E-value does not necessarily imply a stronger evolutionary relationship.

Interpreting it like that is a common mistake! To understand relationships across sequences, you should typically also perform multiple sequence alignment followed by phylogenetic reconstruction. Additional evidence also helps (e.g., understanding sequence conservation and domain architecture).

How is the BLAST E-value calculated?

The E-value is calculated based on the alignment score (S), the search space size (m × n), and the parameters derived from the scoring system and the database composition, such as the Karlin-Altschul parameters (K and λ). The formula for E-value is:

E-value = $K \times m \times n \times e^{-\lambda S}$

Where:

- m is the length of the query sequence.
- n is the length of the database (i.e., the sum of all the lengths of all the sequences in the database).
- K and λ are the Karlin-Altschul parameters. They can be estimated from large sets of random sequence alignments. The λ parameter normalizes the alignment score, while the K parameter scales the E-value based on the database and sequence lengths.
- S is the alignment score. It is calculated based on the selected scoring matrix and the given sequence alignment. The score reflects the sum of substitution and gap scores for the aligned residues.

The E-value thus depends on the database size. Larger databases have more chances of producing the alignment you see by chance… so E-values for the same amount of similarity end up being weaker (higher).

So how should I tweak my BLAST analysis to get the most power?

1. Use the **appropriate database**. If you're looking for a particular gene in humans… only BLAST against the human genome… not against a database that is orders of magnitude greater. Doing so would make it less likely for you to get strong E-values, even if the gene is present in the human genome. And the BLAST analysis would also take much longer.
2. Use the **appropriate BLAST algorithm** for your biological question and evolutionary distance. Consider that nucleotides diverge faster than protein sequences. So:
    - if you're comparing highly similar sequences (e.g., to help identify intron-exon boundaries, or allelic differences), use BLASTN.
    - if you're identifying orthologs across species, use BLASTP. To be certain that a gene is absent from a species, use TBLASTN.
3. Use an **appropriate scoring matrix**. BLOSUM62 is used by default. But for longer evolutionary timescales, the PAM250 is more appropriate.

Aren't these kinds of adjustments "E-value hacking"? No. If done appropriately it's just using the right tool for the job.

Support
Demo
Blog
Affiliate program
Privacy Policy
Terms of Service
Cookie Policy

LinkedIn Twitter Mastodon

OPEN SOURCE SEQUENCESERVER

Install SequenceServer on your own server

Documentation
Github Repo
Community Support Forum