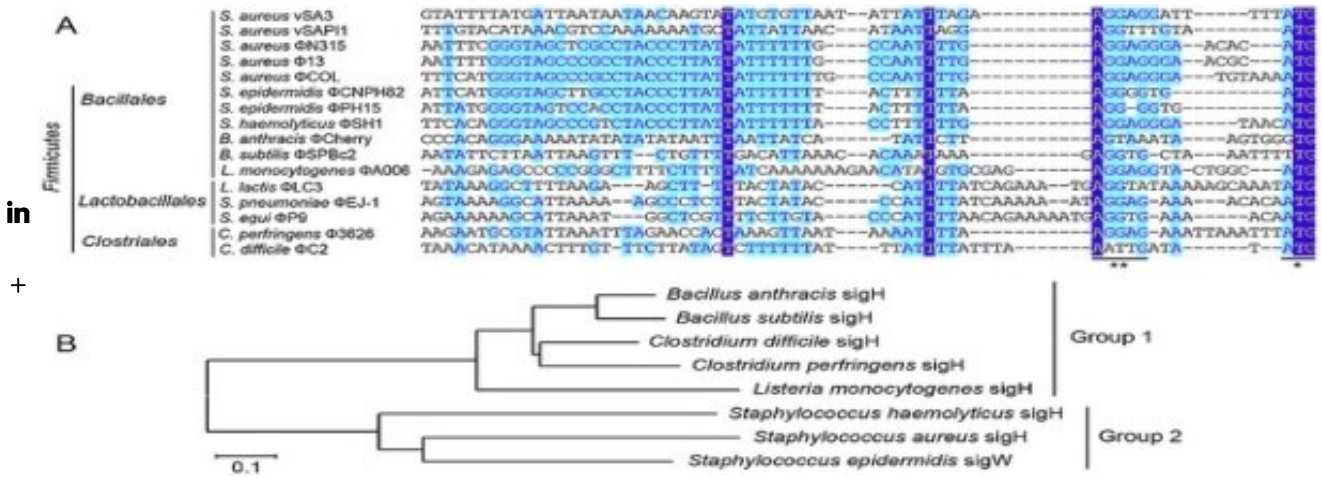
A Crash Course in BLAST Searching



Simple BLAST searching is pretty straightforward to many of us. Just plug in your sequence, select the species genome, and hit search! But have you ever wondered what it takes to run a BLAST query using these mammoth-sized (no pun intended!) sequence databases?

BLAST searching can produce dozens, hundreds, or even thousands of candidate alignments. The results of BLASTing your favorite gene or protein can differ substantially, depending on the exact query sequence used and the parameters of your search (such as database size and e-value cut-off, to name a few). On top of that is the long list of values in the results page. What are these? And what do they mean?

Most importantly, which of the dozens (or hundreds, or thousands) of results are the most accurate? If you're already feeling lost by some of the jargon used above, don't despair! Keep reading this article, and I will help you answer these questions, making you a BLAST-whiz in no time!

Why Do We BLAST?

For bioinformaticians, homology is the main clue to predict gene and protein function. But how does one predict homology? The answer is to examine similarity between two or more sequences. As a general rule of thumb, you should expect at least 25 % sequence similarity for DNA sequence homologues and 70 % sequence similarity for protein homologues if your query included more than 100 nucleotides or amino acids.

The easiest way to assess sequence similarity between two or more sequences is to perform a sequence alignment. BLAST and FASTA are very common tools for doing just this.

FASTA is a sequence search algorithm that flourished in the mid 1980s, but it was and still is time consuming. Since it's debut in 1990, BLAST has fast become the most widely used sequence search program, functioning as a revolutionary tool to search against big sequence databases such as those at NCBI (https://www.ncbi.nlm.nih.gov/) (for example, Nucleotide Collection (nr/nt), Expressed Sequence Tags (EST), Protein Data Bank (PDB)).

The Principle Behind BLAST Searching

BLAST makes a list of 'words' (i.e., a list of short sequences) where the nucleotides/amino acids constitute the letters that make up the query sequence. These words are then screened in the database that has scores over a threshold value (T), as they are easy to track down due to their short length.

The scores for each word out of the list are calculated using a scoring matrix (for example BLOSUM62 (https://en.wikipedia.org/wiki/BLOSUM)). Next, those words above the threshold value act as seeds to widen the alignment between the query and the target sequence. This can be via a gapped or un-gapped alignment extension in either direction to get sequence pairs with high scores or HSPs (High Scoring Segment Pairs). This process is called seeding.

Those HSPs that are above a cutoff score (S) are reported in the BLAST output and the extension process is terminated when the HSPs are below cut-off. This is followed by a trace back procedure to work out the position of insertions, deletions and matches together with some additional computations.

What Is the 'E-Value'?

The percentage similarity between two or more sequences alone is not enough to ensure trustworthy alignments. Bioinformatics buffs tend to regard the e-value as the most informative parameter when looking for true matches.

To keep it simple, the e-value takes into account the number of hits one can expect to obtain by chance when searching a given database, and represents the probability that a given sequence is a significant match. Generally, the lower the e-value obtained, the more significant the alignment. The e-value of an alignment alone may be a useful tool for you to rapidly search different databases against your query. The user must be aware that identical alignments when searching different databases may not receive the same e-value. This is because of the difference in the number of sequences across databases.

Bit score is an important measure that gives an indication about the statistical significance of an alignment. In simple terms, the higher the bit score, the more similar the two sequences are. Bit scores below 50 are generally assumed to be untrustworthy.

Using Word Size and Low Complexity Region Filters

The default BLAST settings for word size are 3 and 11 for protein and nucleotide sequence searches, respectively. The word size can be lessened to 2 for short stretches of amino acids. However, you can either increase to 15 or reduce to 7 to improve your BLAST output in the case of nucleotides. In general, reducing the word size leads to more accurate results but is time consuming.

Low complexity regions are stretches of amino acids or nucleotides that are commonly found with low information content which may have statistical, but not biological, significance. For example, ATATATATATATAT, PPPPPPPPPPPPPPP or Alu repeats may sometimes appear redundant in query sequences, and you can filter these by choosing the 'low complexity region filter' option. This will shorten your query further.

How to BLAST

Once you enter the BLAST page (https://blast.ncbi.nlm.nih.gov/Blast.cgi), select the desired BLAST tool (blastn or blastp). Then, you will need to enter the query sequence, choose the desired algorithm, and set search parameters.

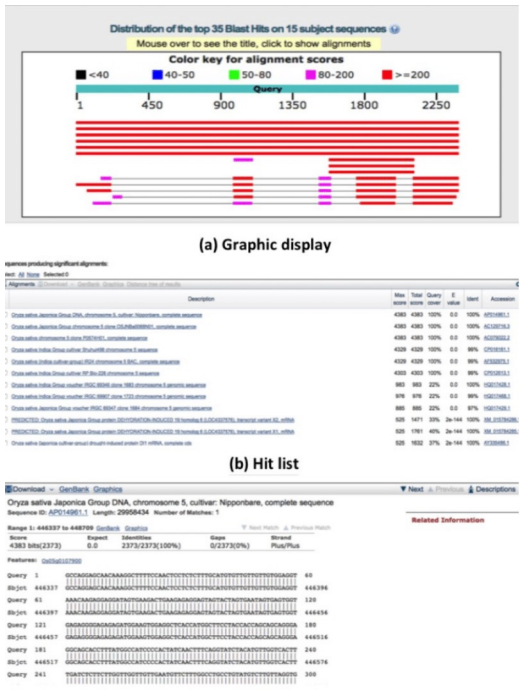
1. **Choose Search Set:** Here, you have the choice of genomic plus transcripts and other databases. You can also create a custom database.
2. **Program Selection:** Here, you have the opportunity to select the intended BLAST algorithm.
3. **Algorithm Parameters:** Lastly, you'll need to set some parameters for your chosen algorithm. Here,

you may consider the e-value, word size and the low-complexity regions filter.

BLAST Results Page

Let me take you on a detour to the BLAST results page (Figure 1). The output in ordered sequence includes:

- **A graphical display:** This provides a brief summary of the alignments ranked according to alignment scores.
- **List of hits:** The hit list tabulates the target sequences that produce significant alignments, along with their corresponding bit score, query coverage (percentage of query sequence aligned), e-value, identity % and assigned accession number.
- **Individual alignments (along with calculated parameters):** Individual alignments provide details on the parameters, e-value, bit score and identity % calculated for each alignment. You will also find the gaps and clusters of identical sequences within individual alignment(s) in this panel.
- **Figure 1. Example of BLAST results output.** (a) graphical display: The color key signifies the alignments to the query (thickest individual horizontal line at top scaling 1-2250) against target database sequences (thin lines). The red-colored lines indicate the best match whereas pink and green lines indicate acceptable matches, (b) hit list, and (c) individual alignments.



(https://bitesizebio.com/wp-content/uploads/2017/11/Figure-2.png)

Figure 1. Example of BLAST results output.

Parting Words of Advice

Bear in mind that you may not get the exact same results when you run the same blast query on two separate occasions. Updates to the contents of your server can lead to change in the results, so it's worthwhile keeping an eye on when updates occur. For example, there could have been 100 sequences in the target database previously and there may now be 250. This change can bring additional hits to your BLAST query results.

More often than not, it is better to stick to the 'default' specifications when running sequence alignments. If this isn't feasible, I recommend paying close attention to the search results at the bottom of the results page, and use what you've learned in this article to decide whether or not a hit is likely to be a good match or not.

Do bear in mind that there is no one magic cut-off to identify true matches for all your query sequence(s). And to be a snappy user, you must be able to confidently tweak the parameters discussed here for optimal results.

I do hope that this helps to demystify the inner workings of BLAST! Have a BLAST!

Further Reading and Resources

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool (https://www.sciencedirect.com/science/article/pii/S0022283605803602). J Mol Biol. Oct 5;215(3):403-10.
2. NCBI BLAST help page (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs).
3. Lesk, A., 2013. *Introduction to bioinformatics*. Oxford University Press. ISBN-13: 978-0199208043.

Has this helped you? Then please share with your network.

Image Credit: Phylogeny Figures (https://www.flickr.com/photos/123636286@N02/14117384356/in/photolist-nvvhwy-nCRbtj-r2WF3-nw37sx-nwA2r5-nxiPCx-8Z4P1B-neVBtd-nouJQJ-nwgJYQ-ne4RDC-nvjKwJ-nefzf2-nGKDwT-nwEohM-niKPXN-nG7NK1-JooTop-nqvsko-nxnWZg-nycTsv-SLAHUf-nf4Rao-nMr2TB-nvAGi9-HTtbUm-nCTL1W-nopQd6-nf7LEp-ne37uT-nNGJ6J-nopniU-7xADCS-nuXpMT-nGWFvD-nuW5Gt-nxyggt-nMprf9-nEXiy5-nYasJa-ne3NgA-nxypwM-nuoPEA-nxh5xe-ne3nKd-nwfssK-nw15AQ-nwfZ7j-ne1Pb7-ne6V1o) Image Credit: ()
Written by Vivek Thiruvettai (https://bitesizebio.com/profile/vthiruvettai/)

Leave a Comment

Comment

Name (required)

Email (will not be published) (required)

Website

Submit Comment

This site uses Akismet to reduce spam. Learn how your comment data is processed (https://akismet.com/privacy/).

Make meaningful connections with researchers Explore Bitesize Bio

with turnkey inbound marketing at Bitesize Bio

Media Kit & Ads (https://bitesizebio.com/life-science-and-biotechnology-content-marketing/)

View Audience (https://bitesizebio.com/life-science-and-biotechnology-content-marketing/)

Meet Our Mentors (https://bitesizebio.com/mentors/)

Webinars (https://bitesizebio.com/webinars/)

eBooks (https://bitesizebio.selz.com/)

Copyright © 2017 Science Squared - all rights reserved

Privacy Policy (https://bitesizebio.com/privacy-policy/)

Cookie Policy (https://bitesizebio.com/cookie-policy/)

Terms of Use (https://bitesizebio.com/terms-of-use/)