*Bioinformatics*, 35(15), 2019, 2699–2700
doi: 10.1093/bioinformatics/bty1026
Advance Access Publication Date: 24 December 2018
Letter to the Editor

OXFORD

Sequence analysis

# Reply to the paper: Misunderstood parameters of NCBI BLAST impacts the correctness of bioinformatics workflows

**Thomas L. Madden*, Ben Busby and Jian Ye**

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

**Contact:** madden@ncbi.nlm.nih.gov

Received and revised on November 30, 2018; editorial decision on December 10, 2018; accepted on December 19, 2018

Dear Editor,

A recent letter by Shah *et al.* (2018) addressed the use of a command-line parameter in BLAST (Altschul *et al.*, 1997; Camacho *et al.*, 2009). BLAST is a very popular tool, so it is not surprising that this topic has provoked a great deal of interest. The authors have, however, conflated three different issues. One is a bug that will be fixed in the BLAST+ 2.8.1 release due out in December 2018, another is simply how BLAST works and the third might be viewed as a shortcoming of our implementation of composition-based statistics (CBS). Here, we address these issues and describe some new documentation about the BLAST process.

We briefly discuss how BLAST works in order to enable the rest of the discussion. First, BLAST examines all sequences in a database looking for (short) word matches that it can use as a seed. Second, BLAST performs a gap-free extension with seeds found in the first step. Third, if the score of the gap-free extension exceeds a certain threshold (set so that about 1 in 50 sequences pass), it executes a gapped extension. If the expect value of the gapped extension is low enough, BLAST saves it to a list for further processing. Finally, BLAST performs a more careful gapped alignment, based on the list from the last step and returns results to the user. The first step listed above examines all sequences, but the work expended per sequence is small. Each succeeding step examines fewer sequences, but the work per sequence is greater. This description is intended as a brief overview and omits some details.

Shah et al. (2018) did not provide their own example in the letter, but later provided one at https://github.com/shahnidhi/BLAST_maxtargetseq_analysis. At the NCBI, we examined the new example and it became clear that the demonstrated behavior was a bug, resulting from an overly aggressive optimization, introduced in 2012 for BLASTN and MegaBLAST (DNA–DNA alignments). This bug has been fixed in the BLAST+ 2.8.1 release, due out in December 2018. The aberrant behavior seems to occur only in alignments with an extremely large number of gaps, which is the case in the example provided by Shah and collaborators.

BLAST does process every sequence in its search set. It does not, as Shah *et al.* (2018) state, simply return the 'first N hits that exceed the specified E-value threshold' even if they are not the highest scoring hits. This distinction is important as it means that BLAST returns the most significant matches, based on expect value and score, given the input parameters. If two or more matches are equivalent, meaning that they have the same score and expect value, the order of the sequences in the database is used as the tie-breaker. This is exactly the behavior reported by Shah et al. (2018). We do not consider this a bug as the hits are equivalent. If only one result is requested, there is no alert that there may be additional equivalent matches. This can lead to understandable confusion about the results. The newest BLAST+ release (scheduled for December 2018) will issue a warning if the user requests fewer than five matches. Future releases may also issue a warning if further equivalent matches are not shown. However, we want to emphasize that, as with any database search, there is always the possibility of equivalent results being returned and it rests upon the user to determine which one(s) to choose depending on the user's purpose.

We have created new documentation describing how BLAST works, and it is available at https://go.usa.gov/xPVqM. There are a few key points that a reader should take away from this documentation. One is that BLAST increases, internally, the number of matches processed to guard against changes in the ranking of results at different stages of the search. The use of CBS (Schaffer *et al.*, 2001), available for protein–protein alignments (e.g. BLASTP), takes the composition of the subject sequence into account as the final step of

the alignment. This adjustment can change the expect value for a subject sequence with an anomalous composition, so BLAST increases the internal expect value cutoff to make results robust against such behavior. Unfortunately, the application of the final step of CBS is time-consuming and only applied to a limited number of sequences. This would appear to be the origin of the complaint from Sujai Kumar documented by Peter Cock in his blog post (https://blastedbio.blogspot.com/2015/12/blast-max-target-sequences-bug.html). It should not be confused with how BLAST handles ties or the newly found bug discussed above. We are currently examining ways to complete the final steps in a search with CBS more quickly in order to eliminate the behavior found by Sujai Kumar.

It is also important to be clear that the BLAST website at blast.ncbi.nlm.nih.gov uses the same software libraries as the BLAST+ executables and should produce the same results when the same parameters are used.

We would like to thank Shah and collaborators for identifying this bug and providing a well-documented case that allowed us to quickly fix the problem. We invite users with feedback on the behavior of BLAST or our documentation (https://www.ncbi.nlm.nih.gov/books/NBK279690/) to write to blast-help@ncbi.nlm.nih.gov or use the portal at https://support.nlm.nih.gov/knowledgebase/category/?id=CAT-01239.

## Acknowledgements

The authors would like to thank Peter Cock and Sujai Kumar for useful discussions. They also would like to thank Peter Cooper, Christiam Camacho, Wayne Matten, Dave Arndt, Tao Tao and Scott McGinnis for a careful reading of our manuscript.

## Funding

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

*Conflict of Interest*: none declared.

## References

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Camacho,C. *et al*. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Schaffer,A.A. *et al*. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*., **29**, 2994–3005.

Shah,N. *et al*. (2018) Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/bty833.