OSAC 2021-S-0006 Standard for the Use of GenBank for Taxonomic Assignment of Wildlife

Wildlife Forensic Biology Subcommittee
Biology Scientific Area Committee
Organization of Scientific Area Committees (OSAC) for Forensic Science

OSAC Proposed Standard

OSAC 2021-S-0006
Standard for the Use of GenBank for Taxonomic Assignment of Wildlife

Prepared by
Wildlife Forensic Biology Subcommittee
Version: 2.0
November 2021

Disclaimer:

This OSAC Proposed Standard was written by the Wildlife Forensic Biology Subcommittee of the Organization of Scientific Area Committees (OSAC) for Forensic Science following a process that includes an open comment period. This Proposed Standard will be submitted to a standards developing organization and is subject to change.

There may be references in an OSAC Proposed Standard to other publications under development by OSAC. The information in the Proposed Standard, and underlying concepts and methodologies, may be used by the forensic-science community before the completion of such companion publications.

Any identification of commercial equipment, instruments, or materials in the Proposed Standard is not a recommendation or endorsement by the U.S. Government and does not imply that the equipment, instruments, or materials are necessarily the best available for the purpose.

To be placed on the OSAC Registry, certain types of standards first must be reviewed by a Scientific and Technical Review Panel (STRP). The STRP process is vital to OSAC's mission of generating and recognizing scientifically sound standards for producing and interpreting forensic science results. The STRP shall provide critical and knowledgeable reviews of draft standards or of proposed revisions of standards previously published by standards developing organizations (SDOs) to ensure that the published methods that practitioners employ are scientifically valid, and the resulting claims are trustworthy.

The STRP panel will consist of an independent and diverse panel, including subject matter experts, human factors scientists, quality assurance personnel, and legal experts, which will be tasked with evaluating the proposed standard based on a comprehensive list of science-based criteria.

For more information about this important process, please visit our website at: https://www.nist.gov/topics/organization-scientific-area-committees-forensic-science/scientific-technical-review-panels

Standard for the Use of GenBank for Taxonomic Assignment of Wildlife

Foreword

This standard defines the requirements that shall be met when comparing evidentiary sequences to those in GenBank for taxonomic assignment of non-human samples. The aim is to provide a framework that will result in consistency in the wildlife forensic DNA community. Use of these standards is expected for forensic scientists with a working understanding of DNA sequencing.

This standard was developed by the Biology/ Wildlife Forensic Biology Subcommittee of the Organization of Scientific Area Committees. This standard is intended to assist those using GenBank for the taxonomic identification of wildlife in forensic casework.

All hyperlinks and web addresses shown in this document are current as of the publication date of this standard.

Keywords: GenBank, BLAST, DNA, Public sequence databases, Taxonomic identification, Wildlife

Table of Contents

1. Scope **5**

2. Normative References **5**

3. Terms and Definitions **5**

4. Requirements **7**

Annex A (informative) Bibliography **11**

Standard for the Use of GenBank for Taxonomic Assignment of Wildlife

1. Scope

This standard covers the requirements and recommendations for analysis and selection of DNA sequences retrieved from the National Center for Biotechnology Information's GenBank and their subsequent use as reference material for taxonomic identification of wildlife[1]. This standard does not cover the use of DNA sequences from other public sequence databases (*e.g.,* BOLD, UNITE), the protocol for downloading sequences from GenBank for inclusion in in-house databases, or the use of custom BLAST searches against GenBank. However, the criteria can be conceptually applied to other sequence databases.

2. Normative References

NCBI Field Guide Glossary available at
https://www.ncbi.nlm.nih.gov/Class/FieldGuide/glossary.html#

Madden T. (2013). "The BLAST Sequence Analysis Tool." In: *The NCBI Handbook*, *2nd ed*. Bethesda, MD. Available from https://www.ncbi.nlm.nih.gov/books/NBK153387/

ANSI/ASB Standard 019, First Edition. Wildlife Forensics General Standards, 2019.

ANSI/ASB Standard 029, First Edition. Report Writing in Wildlife Forensics: Morphology and Genetics, 2019

3. Terms and Definitions

For purposes of this document, the following definitions and acronyms apply:

3.1
alignment
An arrangement of two or more nucleotide or protein sequences that is used to illustrate similarity among those sequences.

3.2
Basic Local Alignment Search Tool
BLAST
The a) BLAST algorithm, and b) a suite of database search programs that implement variations of this algorithm to generate alignments between a nucleotide or protein sequence in a query, and nucleotide or protein sequences within a database.

3.3
expectation value
e-value
The number of distinct alignments expected by chance; the default sorting metric in BLAST search results.

[1] For the purposes of this document, "wildlife" species are defined as non-human multicellular animals and plants, whether wild, captive-bred, or domesticated.

3.4
GenBank
A public repository of DNA sequences maintained by the National Center for Biotechnology Information, part of the U.S. National Institutes of Health.

3.5
hit(s)
Sequence(s) returned from GenBank when performing a BLAST search. Also known as a "subject sequence."

3.6
interspecific
Between members of different species.

3.7
intraspecific
Between members of the same species.

3.8
National Center for Biotechnology Information
NCBI
The U.S. National Center for Biotechnology Information (NCBI) is located in Bethesda, Maryland and is part of the United States National Library of Medicine (a branch of the National Institutes of Health). NCBI houses a series of databases relevant to biotechnology and biomedicine and provides several bioinformatics tools for searching and analyzing the housed data.

3.9
phylogram
A branching diagram that illustrates relationships amongst organisms. Phylograms are typically generated using genetic sequences and/or morphological characters.

3.10
query
(n) The nucleotide or protein sequence that has an unknown source (*i.e.*, evidence sequence), or (v) the action of searching an unknown sequence against a database.

3.11
query coverage
The percent of the query sequence length that is included in the aligned segment with a hit.

3.12
sequence identity
The percentage or number of nucleotides or amino acids that are identical between two sequences.

3.13
subject sequence(s)
A nucleotide or protein sequence(s) returned from a GenBank BLAST search. Also known as a "hit".

3.14
taxonomic identification
Analyses to establish the classification of biological evidence to family, genus, species, etc. These analyses are based on class characters (*e.g.*, morphological, genetic) that are diagnostic for the taxonomic level in question.

3.15
topology
The branching structure of a phylogram.

3.16
voucher specimen
Biological specimen that is representative of its species in accordance with the relevant taxonomic authority and is therefore valid for comparative purposes. Voucher specimens are of known identity, and are curated with available associated geographic, field collection, and life history data.

4. Requirements

Details about the operation of BLAST can be found in Madden (2013), and detailed information on the terms in the BLAST output can be found in the NCBI Field Guide Glossary.

The following requirements and recommendations address criteria for the preparation and submission of evidentiary query sequences (4.1) and evaluation and interpretation of BLAST results from GenBank (4.2, 4.3), which should take into account whether the returned hit(s) is attributed to the correct species and whether the hit(s) is a close enough match for the taxon in question, appropriate level assignment (4.4) and reporting results from GenBank (4.5).

4.1 Prior to performing a BLAST search, evidentiary query sequences:

4.1.1 Shall be prepared by removing non-template flanking regions (*e.g.,* primer);

4.1.2 Shall meet sequence quality criteria as defined by the laboratory. Thus, laboratories are responsible for having these criteria clearly defined and ensuring their analysts follow these recommendations.

4.1.3 Shall be examined to ensure it does not contain premature stop codons (*e.g.,* by translation).

4.2 To ensure that a hit(s) on which conclusions are based are of high quality, an initial assessment of the BLAST results:

4.2.1 Shall ensure the hit(s) belongs to the expected broader taxonomic group (*e.g.*, macerated plant tissue returns matches to sequences from the plant kingdom, not the bacterial kingdom).

NOTE: In situations involving a complete unknown, it may not be possible to complete this assessment.

4.2.2 Shall ensure that any hit(s) that is an anomaly among the returned results is not used. This would be indicated by being the only representative of its species interleaved among many in a different taxonomic group. This could be an indication of human error in sequence labeling during sequence preparation prior to GenBank upload.

4.2.3 Shall ensure the hit(s) does not originate from an environmental sample (*e.g.*, bulk soil extraction, bacterial swab) or low copy sample.

NOTE: The original publication can often be consulted to determine the source of the sequence. In some instances, this determination may not be possible.

4.2.4 Should include a review for descriptors or characteristics that indicate the sequence was not reviewed prior to uploading in GenBank.

NOTE: Sequences that have not been reviewed for quality may include descriptors such as "NGS", "MPS", "EST", "shotgun", "library", and "WGS"; these may have been batch uploaded directly from the sequencing platform. Unedited sequences may also have a higher number of "Ns" or degenerate bases at the ends, or contain non-template flanking (*e.g.*, primer, adapter) sequences.

4.2.5 Should include a review for ambiguous bases.

NOTE: Ambiguous bases should be treated with caution, as they can indicate poor-quality sequence, but they can also indicate heteroplasmic sites within a high-quality sequence.

4.2.6 Shall ensure the hit(s) from a protein coding region does not contain premature stop codons.

4.3 Any hit(s) on which conclusions are based shall be evaluated to determine if the returned sequence is attributed to the correct species based on the criteria listed below. This section is to determine if returned sequences are appropriate for interpretations as outlined in Section 4.4. These criteria confer either strong or moderate support to the attribution. If the returned sequence(s) does not meet at least the moderate criteria, they shall not be used for taxonomic assignment to the species level.

4.3.1 Strong criteria (not all of these criteria have to be met, see section 4.5 for more information about how to evaluate relevant criteria):

a) Sequence(s) is derived from a voucher specimen that bears a unique identifier.

b) Sequence(s), when downloaded, aligned with sequences from closely-related species and used to construct a phylogram, results in a species-level topology concordant with expectations from the peer-reviewed literature.

c) Sequence(s) is from a study published in a peer-reviewed journal; the study addresses the phylogeny or taxonomy of the taxon of interest and the publication or accompanying metadata makes it clear that the source specimen(s) was morphologically identified by a taxonomic expert.

d) Sequence(s) is part of a population genetic study for the given species published in a peer-reviewed journal.

NOTE: Typically a population genetic study characterizes numerous individuals from the studied species in order to explore intraspecific variation (sample sizes will vary based on genetic variability and rareness of the species in question; published studies will have sample sizes that are appropriate for the species in question). The individuals may either be from the same geographic region, or from distinct populations within the known distributional range.

4.3.2 Moderate criteria (not all of these criteria have to be met, see section 4.5 for more information about how to evaluate relevant criteria):

a) Sequence(s) is from a study published in a peer-reviewed journal; the study includes additional data establishing species identity (*e.g*., morphological evidence, museum specimen), but it is not clear that the source specimen was a voucher (4.3.1a) or was morphologically identified by a taxonomic expert (4.3.1c).

b) Sequence(s) is from a phylogenetic study in a peer-reviewed journal; the study addresses phylogeny or taxonomy of the taxon of interest and:

i. includes most or all members of the genus in question, and

ii. the locus shows resolution at the species level (see 4.4.2).

c) Sequence(s) is one of multiple identical or near-identical sequences for the same locus and species from different submitters or geographic locations.

d) Sequence(s) is not from a peer-reviewed study on the taxon of interest, but is accompanied by additional metadata concerning the source individual (*e.g*., location life history stage, name of collector, name of taxonomic expert who rendered the source individual's identification).

4.4 The following should be evaluated to determine the appropriate level for taxonomic assignment:

4.4.1 Whether all likely candidate species in the taxonomic group in question are represented amongst the returned hit(s).

NOTE: Complete taxon sampling is ideal, but often not feasible. If relevant taxa are missing, other loci or additional reference material should be considered. Species that are distantly related based on published phylogenies or those that do not occur in the geographic area of interest may be exempted from the comparison if sequences are not available. See section 4.5.2 in ASB 019 and section 3.5 in ASB 029.

NOTE: Peer-reviewed literature or internal validation for the species/marker of interest provides the foundation for evaluating whether hits are appropriate and comprehensive enough to provide accurate interpretation for reporting.

4.4.2 Whether the interspecific distance for the taxonomic group of interest at the surveyed locus is greater than intraspecific distance.

NOTE: If inter- and intraspecific distances are similar, one should consider using a different locus or limiting identification to a higher taxonomic level.

4.5 Reporting from BLAST results

4.5.1 It is appropriate to report to the species level when all of these criteria are met:

a) The evidentiary sequence(s) has been prepared as outlined in 4.1,

b) The hit(s) on which conclusions are to be based:

i. meets the quality criteria as defined in 4.2;

ii. meets at least two strong support criteria (as defined in 4.3.1), or at least one strong and one moderate (as defined in 4.3.2) support criteria;

iii. has been evaluated against the criteria defined in 4.4;

iv. and when aligned to the evidentiary query sequence, shows 99–100% identity (inclusive).

NOTE: 99% is a conservative threshold, to be applied in instances where no other information is available for the target taxon. For most species, intraspecific distance will be greater than 1%; in cases where additional information (*e.g*., other loci, taxonomies based on morphological features) indicates species are well-separated, identities lower than 99% may still warrant a species level identification.

NOTE: By default, BLAST results are sorted by E-value, which preferentially weights matches with higher query coverage, and max-score, based on sequence similarities. This can result in shorter sequences with higher percent identity being displayed after longer sequences with lower percent identity. The list may be sorted by the identity value to reveal the highest-similarity matches. It is critical to consider both the percent identity and the length of the match when evaluating BLAST results.

4.5.2 It is appropriate to report to a higher taxonomic level when all of these criteria are met:

a) The evidentiary sequence(s) has been prepared as outlined in 4.1,

b) The hit(s) meets the quality criteria as defined in 4.2,

c) The hit(s) has been evaluated against the criteria defined in 4.4,

d) The hit(s) does not meet the support criteria given in 4.5.1(b)ii, but is from a peer-reviewed publication and:

i. The most similar sequences returned by a query are <99% identical and there is little definitive information on interspecific distance.

OR

ii. All top hits represent a single taxonomic level (*i.e*., genus, family, order), but there is a discrepancy at a lower taxonomic level (*e.g*., hits represent different species, but they all belong to a single genus).

Annex A (informative)

This is not meant to be an all-inclusive list as the group recognizes other publications on this subject may exist. At the time this standard was drafted, these were the publications available for reference. Additionally, any mention of a particular software tool or vendor as part of this bibliography is purely incidental, and any inclusion does not imply endorsement.

Bibliography

1] Altschul SF. (2014). "BLAST Algorithm." In: *eLS, John Wiley & Sons, Ltd (Ed.)*. doi: 10.1002/9780470015902.a0005253.pub2.

2] ANSI/ASB Standard 019, Wildlife Forensics General Standards, First Edition, 2019.

3] ANSI/ASB Standard 029, Report Writing in Wildlife Forensics: Morphology and Genetics, First Edition, 2019.

4] ANSI/ASB Standard 048, Wildlife Forensic DNA Standard Procedures, First Edition, 2019.

5] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. (2013). "GenBank." Nucleic Acids Research 41(D1):D36-42. Available from: https://www.ncbi.nlm.nih.gov/genbank/.

6] BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK279690/.

7] Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 16, Molecular Phylogenetics. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21122/.

8] International Organization for Standardization. (2017). "ISO/IEC 17025:2005 General Requirements for the Competence of Testing and Calibration Laboratories." 28 pp.

9] Lee TRC, Anderson SJ, Tran-Nguyen LTT, Sallam N, Le Ru BP, Conlong D, Powell K, Ward A, Mitchell A. 2019. Towards a global DNA barcode reference library for quarantine identifications of lepidopteran stemborers, with an emphasis on sugarcane pests. *Scientific Reports* 9: 7039. Doi: https://doi.org/10.1038/s41598-019-42995-0.

10] Lorenz JG, Jackson WE, Beck JC, Hanner R. (2005). "The problems and promise of DNA barcodes for species diagnosis of primate biomaterials." *Philosophical Transactions of the Royal Society B* 360, 1869–1877.

11] Madden T. (2013). "The BLAST Sequence Analysis Tool." In: *The NCBI Handbook*, *2nd ed*. Bethesda, MD. Available from https://www.ncbi.nlm.nih.gov/books/NBK153387/.