International Standard

ISO 24480

Biotechnology — Validation of database used for nucleotide sequence evaluation

Biotechnologie — Validation de la base de données utilisée pour l'évaluation de la séquence nucléotidique

First edition
2024-11

Reference number
ISO 24480:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 276, *Biotechnology*.

Any feedback or questions on this document should be directed to the user’s national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

A valid database is important for nucleotide sequence evaluation. The development of inclusivity and exclusivity panels for diagnostics and surveillance using community genomic databases, e.g., Genbank, has been evaluated[1],[2],[3]. However, a specific validation procedure for the databases has yet to be provided.

Considering the current database quality, inclusivity and exclusivity are almost impossible to be validated with ideal accuracy. Therefore, in this document a practical procedure for evaluating the quality of nucleotide sequence database to be used for the development of inclusivity and exclusivity panels is comprehensively described. The degree of data accuracy to be used is determined according to the user's intended test purpose. This evaluation can become a part of the validated diagnostic or surveillance method. Ensuring the quality of the database improves its sufficiency for validating the whole measuring system.

In polymerase chain reaction (PCR) and DNA microarray technologies, nucleotide sequence is used as primers or probes to detect the target nucleic acids. Those technologies utilize initially the hybridization of two single strand DNA molecules with complementary sequences. During the design process of the primers or probes, nucleotide sequence database is used for evaluating specificity and exclusivity of probes or primers. In general, target DNA sequences can be confirmed to match the intended sequences but not others by similarity (homology) search on nucleotide databases with computer tools, for example BLAST.

The validated databases can be used for evaluating specificity of probe or primer sequences and ensuring the selectivity of the qualification and quantification measurement system.

Validation of the entire nucleotide sequence database is not appropriate for the database providers because there are wide varieties of purpose of uses by users. It is almost impossible for the users, however, to evaluate the quality of each data entry especially in huge sequence databases. The database can reflect the fitness for the intended test purpose of users.

This document provides the minimum requirements of a practical procedure for the validation of database used for nucleotide sequence evaluation.

Biotechnology — Validation of database used for nucleotide sequence evaluation

1 Scope

This document describes a practical procedure for nucleotide sequence database evaluation and validation. This document describes minimum requirements for the validation of a nucleotide sequence database. This document is applicable only for databases consisting of entries of nucleotide sequences.

This document is not applicable to the general evaluation of the entire database quality including the quality of each data entry.

EXAMPLE The use of the validated database is for confirming a representative sequence specificity including primers or probes for qualification and quantification of target nucleic acids by conventional polymerase chain reaction (PCR), quantitative polymerase chain reaction (qPCR), digital polymerase chain reaction (dPCR) and microarray technologies.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 20691, *Biotechnology — Requirements for data formatting and description in the life sciences*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
nucleotide sequence specificity
capacity to exclusively recognize a specific nucleic acid target sequence, distinguishing it from other nucleic acids and contaminants

Note 1 to entry: It describes the degree of similarity to specifically match to the nucleotide sequence to be searched by distinguishing it from other nucleotide sequences, and the tendency for a primer or probe with the matched nucleotide sequence to hybridize with its intended target and not hybridize with other non-target sequences.

Note 2 to entry: “sequence specificity” can be considered to be the combination of *inclusivity* (3.4) and *exclusivity* (3.5)

3.2
selectivity
extent to which a method can determine particular analyte(s) in a mixture(s) or matrice(s) without interferences from other components of similar behaviour

Note 1 to entry: Selectivity is the recommended term in analytical chemistry to express the extent to which a particular method can determine analyte(s) in the presence other components. Selectivity can be graded. The use of the term “specificity” for the same concept is to be discouraged as this often leads to confusion.

Note 2 to entry: Sequence specificity in molecular biomarker analysis is differentiated from chemical analyte selectivity.

[SOURCE: ISO 16577:2022, 3.3.73]

3.3
sequence similarity
proportion of matched number of units, including nucleotides and amino acids, to the number of units in specified regions between two nucleic acids or proteins

Note 1 to entry: gap or deletion can be considered to compare the unit sequence of the regions between two nucleic acids or proteins.

Note 2 to entry: “sequence similarity” can be evaluated simply based on a proportion of matched number of units, whereas the term “homology” contains biological meaning in the comparison of two nucleic acids or proteins.

3.4
inclusivity
property of a nucleotide sequence to show high *sequence similarity* (3.3) specifically with intended target nucleotide sequence

Note 1 to entry: The term “inclusivity” is used as same meaning of “sensitivity” in some cases [4].

3.5
exclusivity
property of a nucleotide sequence to show low *sequence similarity* (3.3) with those excluding intended target nucleotide sequences

3.6
nucleic acid test
NAT
technique used to detect or quantify a target nucleic acid with specific sequence, by using of oligonucleotide as a primer or probe

3.7
representative sequence
group of nucleotide sequence data containing one or more target sequences in a complete or partial sequence intended for detection or quantification

3.8
undesirable sequence
group of nucleotide sequence data containing one or more nucleotide sequences, which are potentially either influencing or intentionally excluded, or both, for detection or quantification

3.9
intended test purpose
purpose of nucleic acid detection or quantification using oligonucleotide, e.g., primers or probes, whose design is evaluated by using validated nucleotide sequence databases containing representative and *undesirable sequences* (3.8)

3.10
exploring key
data used for examining data entries stored in database

EXAMPLE Key words, sequence data, taxon data, tissue name etc.

3.11
inclusivity database
database used for evaluating *inclusivity* (3.4) of a specified nucleotide sequence

3.12
exclusivity database
database used for evaluating *exclusivity* (3.5) of a specified nucleotide sequence

3.13
provenance information
information that documents the history of a described object and related described activities, and that contains information about the origin or source of the described object, any changes that can have taken place since it was originated, and who has had custody of it since it was originated

[SOURCE: ISO/TS 23494-1:2023, 3.13]

3.14
finalized provenance information
provenance information transformed into a representation specified by the common provenance model, and which is prepared to be conserved or archived and which is considered as being immutable

Note 1 to entry: Finalized provenance information is a subset of provenance information.

[SOURCE: ISO/TS 23494-1:2023, 3.5]

3.15
basic local alignment search tool
BLAST
sequence comparison algorithm optimized for speed that is used to search sequence databases for optimal local alignments to a query

[SOURCE: ISO 20813:2019, 3.1 modified — Notes to entry have been deleted.]

3.16
massively parallel nucleotide sequencing
next generation sequencing
NGS
high throughput nucleotide sequencing method capable of determining multiple DNA sequences simultaneously and in parallel

Note 1 to entry: The data from a single massively parallel sequencing analysis comprises of millions of sequences and the output is a file containing all sequences.

[SOURCE: ISO 16577:2022, 3.7.10, modified —"whole genome sequencing" and "WGS" have been deleted.]

4 General

There are significant differences between the inclusivity and exclusivity confirmation roles. A database used for the inclusivity analysis shall cover all intended sequence entries. In addition, other recognized unintended sequence entries should be contained to show the sequence similarity is specific to the intended sequence entries in a recognized extent, although exclusivity cannot be confirmed only with the recognized undesirable sequences. Entries with high and reliable quality of both intended sequences and recognized undesirable sequences should be included in the database[3]. Quality of the sequence entries for inclusivity analysis is described in the validation of inclusivity database.

On the other hand, a database used for exclusivity analysis should include as many sequence entries as possible including related and not likely related ones. Even though the entries are not likely related sequences, the database should present those sequences because they can contain sequences that can be unintentionally hybridized by the primers or probes in non-specific manner and rise the amplification background. The quality of the sequence entries for the exclusivity analysis is described in the validation of the exclusivity database (for details, see 7.3). One example for uses of a validated database is for an *in silico* analysis evaluating nucleotide sequence specificity in designing primers or probes for nucleic acid measurement including various PCR-based methods and microarray analysis, i.e., inclusivity and exclusivity analyses. Databases for the inclusivity analysis are used for evaluation on how the designed primers and probes can work to distinguish a target in the nucleic acid measurement methods. A database for the exclusivity analysis is used for evaluating the possibility of the designed primers and probes showing non-specific detection or quantification in the nucleic acid measurement methods. Therefore, when users validate a database, i.e., confirm that the requirements of the database for a specific intended use have been fulfilled, they shall validate the database by fulfilling the requirements with those two roles, namely inclusivity and exclusivity confirmations.

Thus, users can specify two independent databases fitting to the two roles in many cases. There are specific requirements for databases used for inclusivity evaluation (inclusivity database) and databases used for exclusivity evaluation (exclusivity database) depending on the roles (see Clause 6 and Clause 7). It does not limit, however, to place both roles in one database. In some cases, the inclusivity database and the exclusivity database can be the same, for example, when the nucleic acid measurement method is used in specified environment where the available nucleotide sequences are well characterized and limited.

5 Common requirements of the database

Databases should implement the FAIR principles[8]. When constructing databases, they shall be constructed in accordance with ISO 20691. Each entry shall be identified with appropriate identifier(s), for example, scientific name, accession number for registered sequence in public database, unique number with authorship for the non-registered sequence. The data format of the database shall be machine-readable and generated in as accessible format[7] for nucleotide sequence analysis, for example, fasta or fastq format (see Annex A). The database shall be accessible by search functions, for example, local BLAST search, taxon search, text search of gene names.

NOTE In some cases, inter-jurisdictional consideration is important, depending on the characteristics of data entries.

6 Inclusivity database: database for inclusivity evaluation

6.1 Quality criteria

Quality criteria for each data entry in inclusivity database shall be determined by the user, considering the intended test purpose of the NAT.

An example of a whole validation process for an inclusivity database is described in Annex B.

6.2 Requirements of an inclusivity database

High quality target sequences, which are intended to be detected or quantified by the NAT measurement system, shall be sufficiently populated in the inclusivity database with numbers of the representative sequences to cover sequence discrimination by the measurement system.

Entries shall include representative sequences of the target and sequences that are undesirable to be detected by the measurement system, for showing that the sequence similarity is specific to the intended sequence entries to a recognized extent (see Annex E as an example for dataset verification commands). Representative sequences of the target should be stored in multiple entries, for example, the sequences of a target analysed by several different laboratories.

NOTE Redundancy of the sequences in the database can be allowed.

The inclusivity database shall contain sequences with high sequence similarity (species specific and non-specific; target to be included or excluded) based on criteria determined by users.

Users should take into account to include sequences with point mutations when applicable. In some cases, sequences from the same taxonomic rank such as genus, species, subspecies, homologous genes, or variants can be selected as sequences with high sequence similarity.

6.3 Individual data quality indicators

6.3.1 Data provenance and updates

The inclusivity database can be updated periodically. The updated database shall be validated by following the procedures described in this document. Date and time of the update shall be documented.

When database entries are documented using finalized provenance information according to the ISO 23494 series, the collected finalized provenance information can be used for quality assessment of the entries. The finalized provenance information can be the most significant indicator for the quality assessment.

6.3.2 Length of the entries

The entries in the database shall have appropriate length, which needs to be longer than the minimum length of the nucleotide sequence specified by the database user within the quality criteria.

NOTE Data entries with short sequences, such as raw data of NGS can be impeditive to the interpretation of results of the evaluation for validation and actual use of the database, for example, primer and probe design.

Long nucleotide sequences, such as long contigs and whole genome sequence data, which are outside of the maximum quality criteria should be excluded. Although longer sequences are useful for the exclusivity evaluation, which is the main purpose of exclusivity database usage (see below), they are less applicable for validation of inclusivity database, primer and probe design and the evaluation of their inclusivity.

6.3.3 Number of unidentified nucleotides (N)

The nucleotide sequence quality should meet certain criteria when incorporating NGS data into the inclusivity database because NGS data, especially raw data, have not only a short length of the nucleotide sequence (see 6.3.2) but also a higher possibility of deletion and ambiguous data (for example, unassigned data marked as “N”).

NOTE Quality Value in FASTQ data format can be used for estimating whether database entries can be used in the inclusivity database.

6.4 Validation of the inclusivity database

The validation plan of the inclusivity database shall be established, implemented and documented.

The validation of the inclusivity database is the result of confirming whether each data entry is appropriate or not to be included in the database. Therefore, each data entry in the inclusivity database should be confirmed by human curation to ensure that it fulfils the determined quality criteria. In cases where human curation is used, it shall be performed at the early stage of the inclusivity database validation to verify data entries, i.e., target genes and species, entries format, length of sequence, and literature references. When the human curation is eliminated in the validation, an alternative procedure to confirm the conformity of each data entry shall be used and documented.

During validation, the quality of the inclusivity database can be confirmed by searching it with representative sequences and evaluating the number of correct and incorrect best matches[5]. Some quality indicators and methods for the evaluation are described in the previous report[6].

The validation search procedure for the inclusivity database shall be able to retrieve correct best matches of the representative sequence and related entries.

The inclusivity database shall be accessible to users and can be retrieved in a popular format with correct header, nucleotide sequence, length of sequence which can be analysed using tools such as local BLAST search. Thereby, the quality of sequence entries, i.e., perfect match and mismatch of the representative sequence such as highly similar sequences (>99 % similarity), can be confirmed.