NATURAL LANGUAGE PROCESSING TOOL FOR AUTOMATED CURATION AND QUALITY ASSESSMENT OF REFERENCE DATABASES; A CASE STUDY USING 16S rRNA REPOSITORIES

S.J. Kern[1]; J. Portman[1]; C. Webster[1]; I. Cernyte[1]; Y. Kiridooshi[2]; W. Suda[3]; C. Mueller[1]; R. Cardwell[1]

[1]Enthought; [2] JSR-Keio University Medical and Chemical Innovation Center, Tokyo, Japan; [3]Riken Laboratory for Microbiome Sciences, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

PRESENTED AT THE 69TH ANNUAL MEETING OF THE AMERICAN SOCIETY OF HUMAN GENETICS, OCTOBER 16, 2019, HOUSTON, TEXAS

Introduction

As sequencing technology continues to advance, providing context of newly sequenced regions, or higher quality re-sequencing of challenging regions, is fundamental to advancing our scientific knowledge and ultimately yielding better medical outcomes. Reference databases are a foundational element when evaluating genomic sequence origin or anomalies, *i.e.* mutations.

These databases, or repositories, are accumulated collections of sequence data linked with annotations, classifications and/or extended information. Open-access, publicly available databases are commonly available, but proprietary, licensed databases may also be sources of valuable scientific evidence to guide future studies. Each domain-specific database, such as the Ribosomal Database Project (RDP)[1] covering *Archaea* and *Bacteria* for rRNA gene sequences or ClinVar[2] aggregating information about genomic variation and its relationship to human health, may be differentially and inconsistently quality checked, updated and released to reflect current knowledge. To prevent delays or errors, an individual researcher may wish to curate and utilize an available database to suit their need quickly and repeatedly.

An automated curation tool allows application of particular rules to curate the database, fitting the intended use, and adjust to new releases and knowledge quickly. Herein, we present the development of an automated curation tool using multiple, open source repositories of bacterial 16S rRNA gene sequences and taxonomic classifications.

Methods

To use the Python-based automated curation tool, any number of sequence databases may be provided along with an identified "gold standard" reference database to which annotations or classifications will be matched to. Depending on the database type(s), computational rules are designed and prioritized to provide assignment proposals for use in later analysis and applications. New rules can be developed *ad hoc* and added to the set of rules in use. Likewise, rules can be removed from use or re-ordered to reflect changing priorities.

Using the auto-curated approach, a case study mapping four (4) publically-available repositories containing 16S rRNA sequences, CORE[3], RDP[1], GRD[4] and RefSeq (Release 94)[5], to NCBI taxonomies was used to demonstrate database quality variations and entries flagged for further review (Figure 1). Matching rules are defined by providing a Python class with two required methods; an **`is_applicable`** method which returns **`True`** if the rule is applicable to a given taxonomy entry and **`False`** otherwise and a **`get_proposal`** method, which returns a proposal taxonomy for the input entry. Select curation rules make use of the preprocessed NCBI taxonomy database to find matches (exact, case-insensitive, fuzzy nearest neighbors). Fuzzy matching is enabled by use of natural language processing (NLP) techniques. In short, species names are preprocessed with a feature extraction algorithm to convert a collection of text species names into a matrix of *n*-gram occurrences. For poor matches, suggestions are given to the user for manual review. Potentially mis-classified sequences are identified and flagged using two outlier detection approaches; either within a cluster of taxonomy assignments (Taxonomy-based Outliers) or taxonomy similarity within a cluster of like sequences (Sequence-based Outliers). The resulting combined database was then verified using an experimental dataset of bacterial genome sequences encoding the v1-v2 hypervariable region (HVR) of 16S rRNA (n=618) of unknown phylogeny. The measured sequences were queried against the previously curated database (~2011) and the recently auto-curated database (2019) to assign taxonomies.

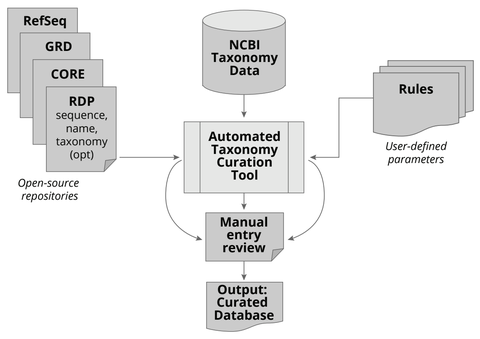


Figure 1: Overview of automated taxonomy curation using multiple sequence repositories

Results

Using automated curation, 97.8% of entries in the 4 databases were matched to the NCBI taxonomy assignment using rules not requiring additional user review [exact match, lowercase match, normalized match (lowercase with punctuation removed)]. The majority of entries in the RDP, RefSeq and GRD databases were matched using the "exact match" rule. The use of specifically crafted rules allowed for the automated curation of large fractions of the less widely used CORE database such as the "pattern match" identifying 52 of the 1,262 entries for further review. A subset of the rules (n = 24) used for matching across databases is shown in Table 1.

***Table 1:** Using a set of rules to identify inconsistencies or entry errors, reference databases are quickly compared for quality. A subset of rules to identify NCBI taxonomy by database is shown. The majority of entries in all databases were identified by "exact match".*

	Database			
Rule Type	CORE	GRD	RDP	RefSeq
Exact match	621	12,883	274,971	20,055
Subspecies infraspecific rank match	0	0	4,438	0
Genus monomial rank match	0	0	549	5
Third party genus rank match	358	81	0	0
Normalized match	0	5	245	19
Pattern match	52	0	0	0
Fuzzy match	0	0	63	0
Total entries in database	**1,262**	**12,990**	**281,261**	**20,100**

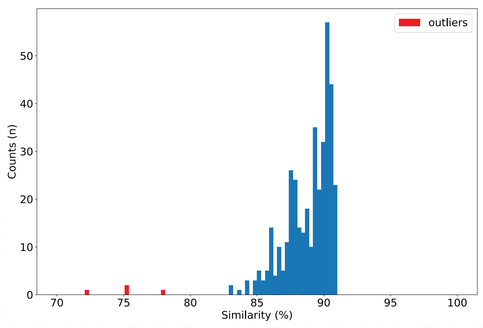
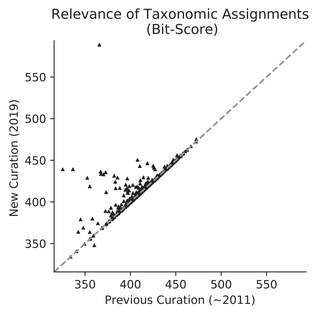


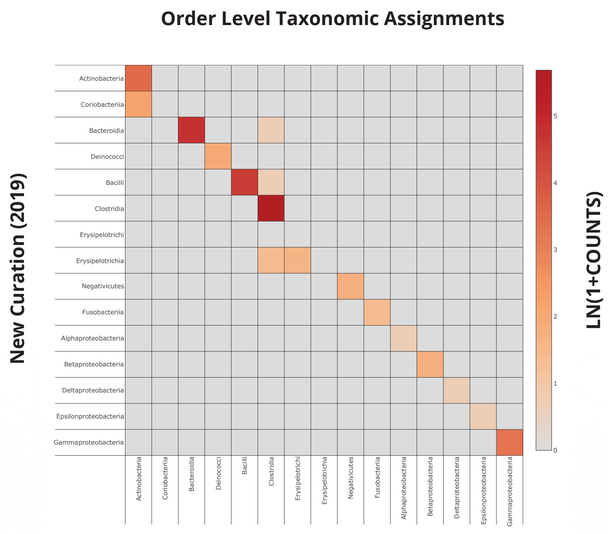
***Figure 2: Identification of sequence outliers based on taxonomy classification at genus level.** Within a Genus-level classification, genomic sequence similarity is used to flag outlier entries. Here, 28 entries within the* Acidithiobacillus *genus are identified as outliers for further review using the default parameters (p25 - 1.5*IQR, p75 + 1.5*IQR). The entries assigned to this genus derive from the GRD, RDP, and RefSeq databases.*

Given taxonomy matches or classifications, the database-provided sequences were then assessed for similarity by taxonomic assignment and outliers flagged. As shown in Figure 2, eight (8) sequence entries within the *Acidithiobacillus* genus are identified as outliers and flagged for further review. While the taxonomic match is equal, the sequence data for these entries may be of low-quality or incorrect (taxonomy-based approach). Similarly, selected sequences may be aligned and the resultant taxonomic classification compared (data not shown), rejecting entries where the assigned taxonomy agrees with less than 50% of clustered sequences.

***Figure 3: Scoring of taxonomic assignments by curation database used.** Values of taxonomic assignments relevance measure (Bit-Score) using previously-curated database as compared to recently auto-curated database. The newly curated database improves assignments (markers above the identity line).*



In Figure 3, the relevance of taxonomic assignment (defined through a scoring system) was improved using the previously curated database (~2011) over the more recently auto- curated database (2019). A higher score is given to the unknown sequence based on similarity to the sequences listed in the database. Figures 4 and 5 present contingency plots assessing the agreement between taxonomies assigned. At the class taxonomic rank, we observe good agreement between assignments with few off-diagonal terms (Figure 4). However, the disagreement increases substantially at the family taxonomic rank (Figure 5-A). A literature search revealed that the phylogenetics of bacteria previously ascribed to the *Clostridiaceae* family are better described by the *Lachnospiraceae*, *Peptostreptococcaceae*, *Ruminococcaceae*, and *Erysipelotrichaceae* families[6]. All of the off-diagonal terms associated with the *Clostridiaceae* family (Figure 5-B) arise from this phylogenetic update proposed in 2013: bacteria previously assigned to solely *Clostridiaceae* are now correctly spread across multiple families. This exercise demonstrates the ability of periodic, automatic curation to integrate new knowledge without the need for manual effort.



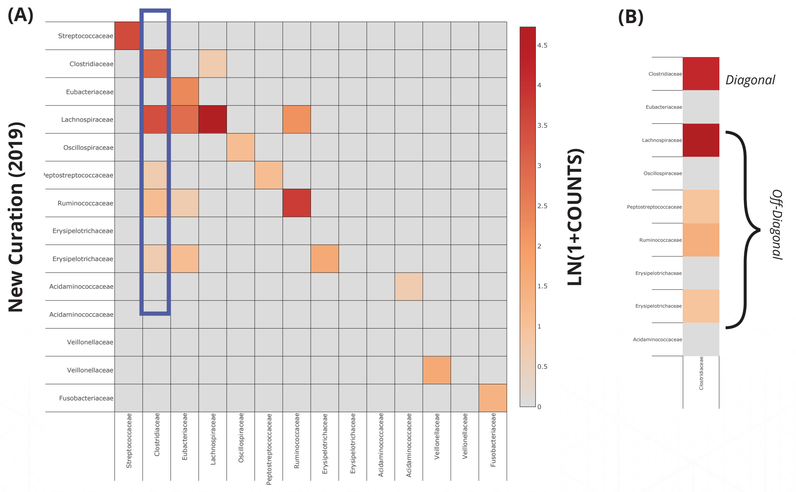


***Figure 4:** Contingency plot of taxonomic assignments between databases curated at different times. At the order taxonomic rank, the assignments agree well with few off-diagonal terms.*

***Figure 5: Identification of taxonomic assignment inconsistencies at family rank.** Contingency plot of taxonomic assignments at the family level between databases curated at different times. The off-diagonal terms indicate inconsistency of the assignments. **(B)** Zoom-in of the slice outlined in **(A)**, highlighting the* Clostridiaceae *family. As discussed in the text, the off-diagonal terms arise from updates within the phylogenetic literature: bacteria previously ascribed to solely* Clostridiaceae *are better described by multiple families[6]. Automatic curation allows such updates to be seamlessly captured.*

Conclusions / Summary

Using computational tools, a user-curated reference database combining inputs from multiple separate source repositories is quickly generated and reflects current knowledge.

- The quality of the sequence data in the repositories may be assessed.
- Curation rules may be added, removed, or rearranged to suit research priorities or intended use.
- Entries flagged for discrepancies or low quality may be manually reviewed.
- Auto-curation tools are general enough to be applicable to a wide range of sequence databases.

Unknown bacterial 16S rRNA sequences have improved taxonomic assignments using the auto-curated database approach.

- Sequence assignment to a taxonomy is improved.
- Outliers, by taxonomic classification or sequence agreement, are identified.
- The provenance of the assigned sequence is tracked to the source database, permitting individual evaluation of the entry.

REFERENCES

1 Cole JR, et al. 2007. *The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data.* Nucleic Acids Res. 2007;35:D169–D172.

2 Landrum MJ, et al. 2014. *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res. 2014 Jan 1; 42(Database issue): D980–D985.

3 Griffen et al. *CORE: A Phylogenetically-Curated 16S rDNA Database of the Core Oral Microbiome.* PLoS One. 2011; 6(4): e19051. Published online 2011 Apr 22.

4 Genomic-based 16S rRNA database (GRD). Laboratory for Integrated Bioinformatics, Center for Integrative Medical Sciences, Riken University. https://metasystems.riken.jp/grd/

5 The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21101 https://www.ncbi.nlm.nih.gov/refseq/

6 Yutin and Galperin. *A genomic update on clostridial phylogeny: Gram-negative spore-formers and other misplaced clostridia.* Environ Microbiol. 2013 Oct; 15(10): 2631–2641. PMCID: PMC4056668

ACKNOWLEDGMENTS

Authors wish to thank Matthew Smarte and Wahiba Taouali for their help in developing this poster.

Poster is available for download at: enthought.com/industries/life-sciences/