Computational and Structural Biotechnology Journal 19 (2021) 847–851

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

ResiDB: An automated database manager for sequence data

Michaela Hendling *, Rick Conzemius, Ivan Barišić *

Austrian Institute of Technology, Center for Health & Bioresources, Molecular Diagnostics, Giefinggasse 4, 1210 Vienna, Austria

ARTICLE INFO

Article history:
Received 15 September 2020
Received in revised form 14 January 2021
Accepted 15 January 2021
Available online 19 January 2021

Keywords:
DNA database
Antibiotic resistance
Diagnostics
Phylogeny
Assay design

ABSTRACT

The amount of publicly available DNA sequence data is drastically increasing, making it a tedious task to create sequence databases necessary for the design of diagnostic assays. The selection of appropriate sequences is especially challenging in genes affected by frequent point mutations such as antibiotic resistance genes. To overcome this issue, we have designed the webtool resiDB, a rapid and user-friendly sequence database manager for bacteria, fungi, viruses, protozoa, invertebrates, plants, archaea, environmental and whole genome shotgun sequence data. It automatically identifies and curates sequence clusters to create custom sequence databases based on user-defined input sequences. A collection of helpful visualization tools gives the user the opportunity to easily access, evaluate, edit, and download the newly created database. Consequently, researchers do no longer have to manually manage sequence data retrieval, deal with hardware limitations, and run multiple independent software tools, each having its own requirements, input and output formats. Our tool was developed within the H2020 project FAPIC aiming to develop a single diagnostic assay targeting all sepsis-relevant pathogens and antibiotic resistance mechanisms. ResiDB is freely accessible to all users through https://residb.ait.ac.at/.

1. Introduction

Infectious diseases are a major burden for the human healthcare systems. In the last century, this situation was dramatically improved with the development of antibiotics and vaccinations. With the emergence and global dissemination of antibiotic resistances, however, we are starting to face a healthcare crisis again. In addition to the development of new antibiotics, infection control measures are the most effective instruments to fight the spread of antibiotic resistances. Essential tools for the implementation of such measures are fast and cost-effective diagnostic tests. This can be currently well observed during the COVID-19 virus pandemic where the lack of appropriate diagnostic testing capacities limits the implementation of effective infection control measures.

Consequently, molecular diagnostic tests are essential to conduct the required fast identification and characterization of pathogens. However, the high diversity of pathogens, antibiotic resistance mechanisms and virulence factors represents a massive challenge for the development of such tests. For example, more than 1000 different antibiotic resistance genes have been described so far. Thus, the diagnostic tests that are on the market focus only on a small fraction of the clinically relevant targets. Another problem is that the commercial molecular tests are too expensive for routine testing of e.g. all hospitalized patients. This is especially relevant for the developing countries were the antibiotic resistance crisis is most problematic. As a consequence, many hospitals have developed their own cost-effective in-house tests. These have the additional advantage that they can be tailored in respect to the local epidemiological situation. The most popular are DNA-based tests using PCR or LAMP to detect genes of interest.

The biggest challenge for researchers and clinicians to develop such in-house tests concerns the design of appropriate primers and probes. While several streamlined tools for the oligonucleotide design itself have been developed, little attention was brought to the management of the target and non-target DNA sequence data that is required for the assay design. Furthermore, the exponential expansion of sequence data necessitates revolutionary measures for data accessibility, management and analysis [1]. The National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) provide, besides comprehensive collections of biological data, access to helpful data analysis tools [2,3]. These tools can be either accessed online or standalone, including similarity search applications, such as BLAST, or multiple sequence alignment applications, such as Clustal Omega [4,5]. Although many sequence analysis services provide a wide range of different algorithms to

* Corresponding author.
E-mail addresses: Michaela.Hendling@ait.ac.at (M. Hendling), Ivan.Barisic@ait.ac.at (I. Barišić).

https://doi.org/10.1016/j.csbj.2021.01.024

manage data, most of these tools are not interconnected with each other. All-in-one packages, providing a workflow where each tool can communicate with the others, are necessary for efficient work on big data sets. The standalone software environments ARB and Mothur provide such all-in-one analysis workflows [6,7]. ARB users are able to create their own sequence databases and perform analysis using integrated software tools that are directly interacting with one another as well as the database. Mothur allows users to use one tool for the analysis of community sequence data. However, users need to have appropriate computational infrastructure to run these tools. In general, these big data sets require appropriate computational power and sophisticated database models. Additionally, in contrast to online data retrieval systems, downloaded databases need to be updated manually. Downloading a comprehensive set of biological data can be very time-consuming and requires a lot of disk space (e.g. NCBI WGS databases comprise 1.6TB). Therefore, it is suggested to replace local hardware with the cloud for data storing and computing [8].

Here, we introduce resiDB, a web-based sequence database manager for bacteria, fungi, viruses, protozoa, invertebrate, plants, archaea, environmental and whole genome shotgun sequences. It is a tool that enables the streamlined creation of user-defined DNA sequence databases that can be used e.g. for the design of diagnostic assays. It automatically identifies and curates similar sequences to create comprehensive and well-organized custom sequence collections using an up-to-date set of background databases. Annotation data, sequences, cluster information, alignments and consensus sequences can easily be accessed, evaluated, edited and downloaded. A collection of powerful visualization tools facilitates the data analysis without the need of programming knowledge. ResiDB is freely accessible to all users through https://residb.ait.ac.at/.

We demonstrated the tool by creating two bacterial, pathogen-relevant sequence databases used for the design of antibiotic resistance and virulence factor detection assays.

2. Materials and methods

ResiDB is a web-based rapid sequence similarity-dependent database manager, aiming to provide an all-in-one software workflow to generate, access, evaluate, edit and download custom DNA sequence databases. The workflow of resiDB is illustrated in Fig. 1. The following subsections describe the main features of each step in detail.

2.1. Input

The creation of a new database starts with the import of nucleotide sequences or protein sequences and parameters for the sequence similarity search and filtering. The sequences have to be provided in FASTA format, either by file upload or by using a designated input box. ResiDB can handle hundreds of input sequences, but it needs to be considered that the runtime and database size are linked to the used background databases and user-defined input parameters. Each step of the database creation is based on the user-defined input parameters described in the following subsections.

2.2. Similarity search

The first step of our workflow generates a sequence-similarity based matrix using the input sequences and a collection of nucleotide sequence databases, referred to as local GenBank mirror. It is generated by the standalone version of NCBI BLAST version 2.7.0 + . The local GenBank mirror is a collection of publicly available sequence files covering *greater than 200* million sequences, annotation and feature data from bacteria, viruses, fungi, archaea, invertebrates, environmental samples, protozoa, plants and WGS projects. It was downloaded and is regularly updated from the NCBI FTP server (see Table 1). The similarity is based on a user-defined similarity measure, which is either defined by the sequence similarity (e.g. 90%) or the E-value (e.g. 10). This threshold defines which sequences in the local GenBank mirror should be selected and implemented in the subsequent workflow steps.

2.3. Filtering

The sequence entries, which were automatically selected during the similarity search, can be filtered by species and other taxonomic ranks (domain, phylum, genus, etc.). In addition, the sequence data can be filtered by criteria such as the presence of a Pubmed ID, a gene name, a sample collection country, a start and stop codon, a "coding sequence" tag, and a "reference sequence" tag annotated in the original GenBank files. These filters are especially helpful for the automatic curation of the sequence database. Further filtering options can be included by user inquiry.

2.4. Sequence clustering

After filtering, the input sequences are analyzed using NCBI BLAST version 2.7.0+. Here, all input sequences are aligned against each other applying user-defined sequence similarity clusters. The BLAST results are compared and sequences that match the similarity criteria and share at least 50% of their hits are grouped (see Fig. 1 at the Sequence clustering). Sequences from the local GenBank mirror are added to these clusters if they meet the user-defined criteria.

2.5. Alignment and consensus sequences

In the final step, the sequences in each cluster are aligned using MAFFT 7.427 [9]. The resulting alignments are the basis for the calculation of the consensus sequences, which are defined as the sum of the most common nucleotides at each position of their alignment. Users can control the calculation of the consensus sequences by setting base frequency thresholds. For example, if a user sets the frequency threshold to 90%, the base with an abundance of more than 90% at a position of the alignment is selected for the consensus sequence. If the most common base shows an abundance of less than 90% at a position, then the IUPAC nucleotide code annotation is used for the base at this position [10]. Furthermore, 100% similar sequences are grouped in "Alignment groups" for a better visualization of the alignment results.

2.6. Implementation

The database creation workflow runs on a Linux server (64 CPUs, 256GB RAM). Most of the implemented workflow scripts run in parallel to obtain maximum performance and utilization of server capacity. Therefore, only one database creation can run at a time. The main software tools and libraries used for the database creation are BLAST 2.7.0+, MAFFT 7.427 and Python 2.7 together with the Biopython library [11]. The local GenBank mirror and the new sequence database created by resiDB are implemented using MongoDB 3.2.22 (MongoDB, https://www.mongodb.com/de). Here, MongoDB provides a powerful database model for the creation of biological sequence databases due to its flexibility, scalability and performance. Therefore, it is possible to easily generate and access such sequence databases containing annotation, feature and sequence data. A simplified scheme of the resiDB database structure can be seen in Supplementary

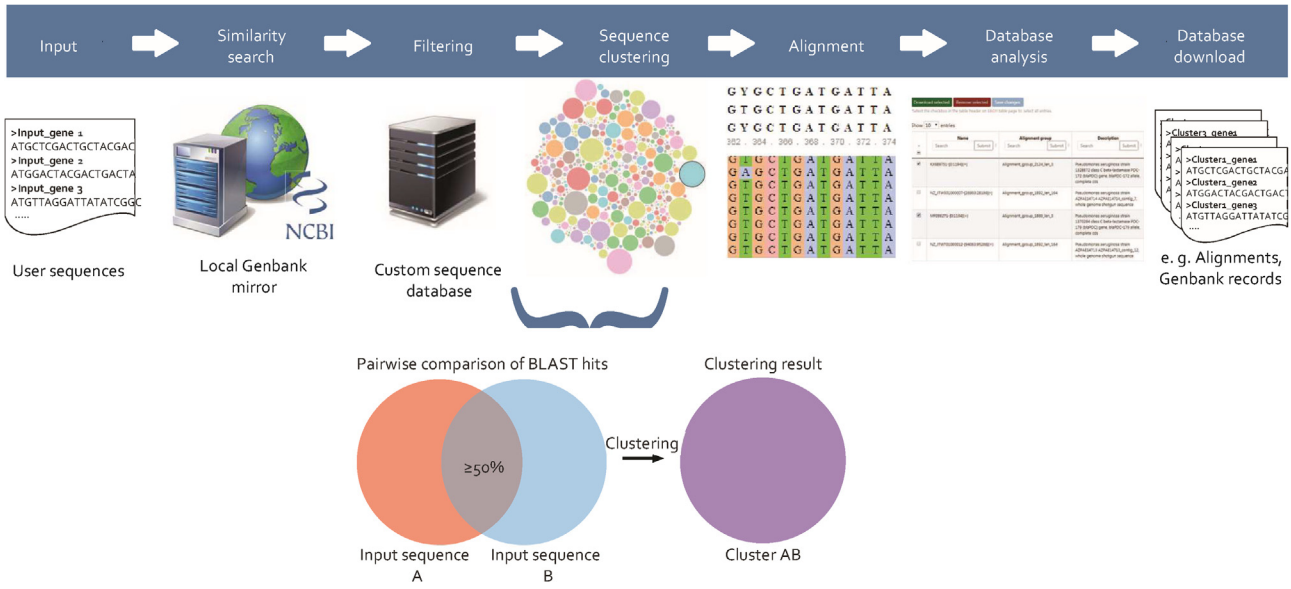

Fig. 1. Illustration of the database creation workflow of resiDB. The workflow starts with the input of *n* DNA sequences, followed by a similarity search, where each sequence is blasted against a selected database of the local GenBank mirror. Subsequently, the resulting hits are filtered by user-defined parameters (see section Filtering for details). Afterwards, the input sequences and database hits are clustered by similarity applying the same similarity threshold as for the similarity search at the beginning. Finally, the sequences of the clusters are each aligned and consensus sequences are calculated. User-friendly analysis tools help the user to visualize and edit the resulting sequence databases before download.

Table 1
NCBI database sources used for the current local GenBank mirror. The number of sequences and their share of the entire data pool are listed.

Source	Number of sequences	Database fraction
Bacteria	20,605,774	9.48%
Environmental samples	8,995,722	4.14%
Invertebrates	93,956,595	43.21%
Patented sequences	41,341,130	19.01%
Plants	4,910,669	2.26%
Viruses	2,551,233	1.17%
Archaea	219,233	0.10%
Fungi	10,109,117	4.65%
Protozoa	5,333,500	2.45%
WGS project sequences	29,442,947	13.54%
Total amount of sequences	217,465,920	100.00%

Fig. S1. The highly responsive graphical user interface is implemented using Bootstrap 4.1.1. The most important plugins and libraries for the visualization and manipulation of the data, are D3.js (D3.js, https://d3js.org/), jsTree (jsTree, https://www.jstree.com/), OpenStreetMap (OpenStreetMap, https://www.openstreetmap.org), Leaflet (Leaflet, https://leafletjs.com/) and the MSAViewer [12]. The high standards of the graphical elements implemented in this tool require web access via browsers with ES6 support. ResiDB is freely accessible to all users through https://residb.ait.ac.at/.

3. Results

With resiDB, annotation data, sequences, cluster information, alignments, and consensus sequences can easily be accessed, evaluated, edited, and downloaded. A collection of powerful visualization tools facilitates the data analysis without the need of programming knowledge. The generated database is accessible on a webpage including six different views: home, cluster view, feature view, alignment view, map view, and download center.

3.1. Home screen

The home screen gives an overview of the calculated similarity-based clusters, the number of similar sequences within each cluster and user parameters. It also provides a search field to get cluster and alignment group information of submitted input sequences. Manual changes in the cluster composition are also visually indicated.

3.2. Cluster view

The cluster view serves as a user interface for the visualization and manipulation of the sequence clusters. This interactive tree view provides insight into clusters and gives the possibility to edit them. Child sequences of the clusters can be moved and removed, and new clusters can be established and renamed. In this view, users can always easily restore the original database.

3.3. Feature view

A table within the feature view shows feature information derived from the underlying GenBank records of selected clusters. Entries can be filtered, removed, downloaded, and directly shown in the GenBank file format.

3.4. Alignment view

The alignment view of resiDB enables the user to view and edit sequence cluster alignments and their consensus sequences easily. Whereas editing involves the changing, insertion, and deletion of bases. This view has a direct link to the feature view to get detailed information on specific alignment sequences.

3.5. Map view

GenBank sequence entries can have additional isolation source information, which lists the physical, environmental, and/or local geographical source of the biological sample from which a

sequence was derived (e.g. human blood, soil, sediment, ocean water, lake water, forest debris, soil from outside a specific chemical factory, gasoline polluted soil, etc.). The map view gives the user the opportunity to select any cluster and view where sequences have been collected over time on an interactive map (Fig. 2).

3.6. Download center

The created sequence database can be downloaded in the FASTA and GenBank file format. Furthermore, alignments and consensus sequences are also available for download.

4. Discussion

ResiDB provides a rapid and user-friendly online database manager. Consequently, researchers do no longer have to manually manage data retrieval, deal with hardware limitations, and run multiple tools, each having its own requirements and file formats for input and output.

4.1. Experimental evaluation

ResiDB was experimentally evaluated during the creation of two bacterial, pathogen-relevant sequence databases used for the design of a five-plex and a comprehensive pathogen characterization assay (810 target genes) respectively. The five-plex assay targets clinically relevant antibiotic genes from five gene clusters: *aad*, *qnrB*, *bla*TEM, *bla*CMY and *bla*SHV. For this purpose, a total of 971 genes from the Comprehensive Antibiotic Resistance Database (CARD) and the Antibiotic Resistance Gene-ANNOTation tool (ARGANNOT) were merged and used as input for resiDB [13,14]. A similarity search was performed using a sequence identity threshold of 90% under the selection of the bacterial and environmental background database. The presence of a start and stop codon as well as a "coding sequence" tag were used as filters to increase the possibility of retrieving complete variants of the antibiotic resistance genes. The database creation was finished within three hours. The resulting consensus sequences with similarity thresholds of 90%, 85%, and 80% were used as target sequences for multiplex oligonucleotide design using oli2go [15].

The alignment view and cluster view were especially useful to curate clusters where no primers and probes could be designed automatically, due to a high number of ambiguous nucleotides. In these cases, the alignments gave an indication of clusters that need to be divided into further clusters and sequences that had to be removed due to poor quality. Furthermore, the alignments were also used to redesign primers manually. The cluster view was used to merge the *aadA24* and *aadA7* clusters to increase the coverage of the assay and rename the chosen clusters. The final data set contained 23 antibiotic resistance gene clusters where the five biggest *aad*, *qnrB*, *bla*TEM, *bla*CMY and *bla*SHV clusters were chosen for the assay. As a result, specific oligonucleotides could be designed successfully using the consensus sequences and alignments of the database generated by resiDB (see Supplementary Figs. S2–S4). The data can be visualized, but not changed in the sample results section at https://residb.ait.ac.at.

The map view can be used to e.g. illustrate the spreading of antibiotic resistances across the world and the clinical relevance for distinct geographic regions. Fig. 2 illustrates the global dissemination of the antibiotic resistance gene *qnrB* between 2002 and 2014. From its first appearance in South America in 2002, it started to spread slowly across the USA and the United Kingdom. However, an increased mobility of this genetic element was observed from 2010 to 2014.

While the five-plex assay was developed as a concise showcase, a main purpose of resiDB is to support the design of comprehensive DNA-based characterization assays. Within the H2020 project FAPIC, we aim to develop an assay targeting all clinically relevant antibiotic resistance genes and virulence factors with a single assay. For this purpose, a total of 3854 genes derived from the Comprehensive Antibiotic Resistance Database and the Antibiotic Resistance Gene-ANNOTation tool (ARGANNOT) were used as input for resiDB and processed as described for the five-plex assay. The resulting consensus sequences (Supplementary material) were used as input for the multiplex assay design tool Oli2go and experimentally evaluated (data not shown). Another use-case of resiDB concerned the creation of sequence databases comprising genes associated with colistin resistance (*mgrB*, *lpxC*, *pmrB*, *phoP*, *phoQ*,

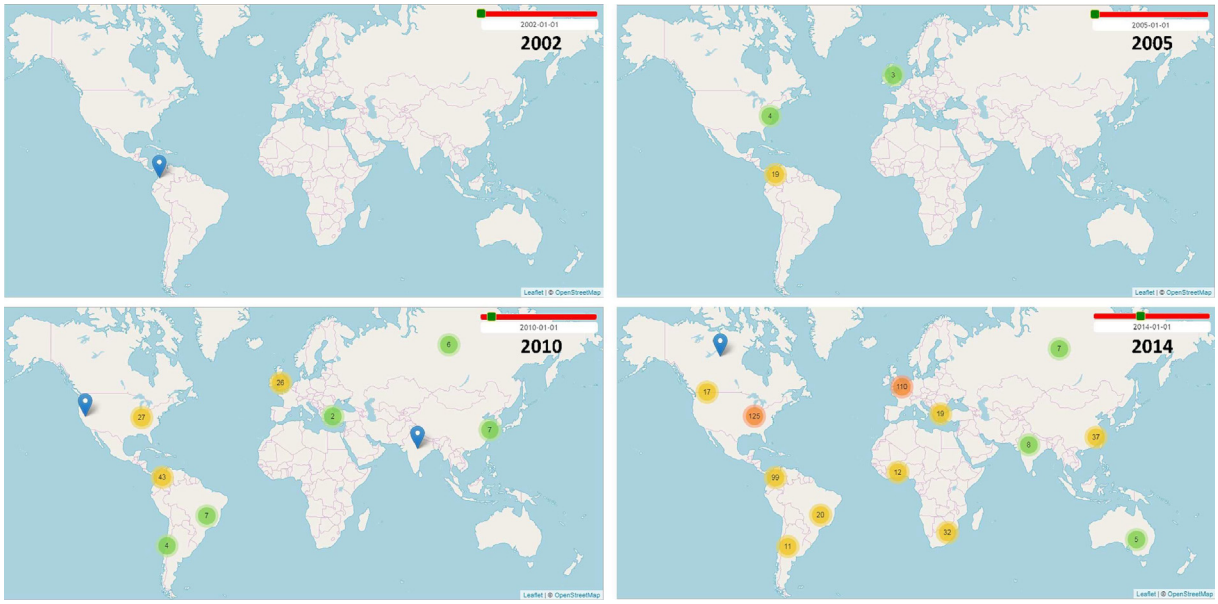

Fig. 2. Sample locations of the antibiotic resistance gene *gnrB* for the years 2002, 2005, 2010 and 2014. In 2002, *qnrB* was only found in South America. In the next three years, it started spreading over the United States and the United Kingdom. In 2010, *qnrB* spread in Europe and Asia. Only four years later, it was also found in Australia and Africa.

crrB). Colistin is a last-resort antibiotic that becomes inactive due to mutations in these genes in more than 90% of the clinical cases. A global phylogenetic analysis of these genes including their geographic distribution could be calculated within a few weeks due to the streamlined sequence database generation facilitated by resiDB [16].

4.2. Outlook

ResiDB has the potential to improve the management of novel and existing DNA sequence databases. This web solution provides a unique combination of an automated all-in-one sequence analysis pipeline, comprehensive background databases and powerful visualization tools accessible via one website. Consequently, researchers do not longer have to manually manage data retrieval, deal with hardware limitations and run multiple independent software tools, each having their own requirements, input and output file formats.

CRediT authorship contribution statement

Michaela Hendling: Conceptualization, Methodology, Software, Writing - original draft. **Rick Conzemius:** Investigation, Writing - review & editing. **Ivan Barišić:** Funding acquisition, Conceptualization, Supervision, Writing - review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Branka Bedenić (University Hospital Center Zagreb, Zagreb, Croatia) and Inge Gyssens (Radboud University Medical Center, Nijmegen, Netherlands) for providing the clinical strains. Furthermore, we would like to thank Timo Schwebs (Austrian Institute of Technology, Center for Health & Bioresources, Vienna, Austria) for the DNA isolation for the experimental evaluation of resiDB, Konrad Peters (University of Vienna, Faculty of Computer Science, Vienna, Austria) for his helpful recommendations during the software development and Ariadne Haunold and Doris Rakoczy for the revision of the manuscript. In the end, we want to thank Marie-Luise Jaufer for the logo design.

Funding

This work was supported by the European Union's Horizon 2020 research and innovation program [634137]. Funding for open access charge: H2020 [634137].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.01.024.

References

[1] Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. Big data: the future of biocuration. Nature 2008;455:47–50.
[2] Tatusova TA, Karsch-Mizrachi I, Ostell JA. Complete genomes in WWW Entrez: data representation and analysis. Bioinformatics 1999;15:536–43.
[3] Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 2015;43:W580–4.
[4] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[5] Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci 2018;27:135–45.
[6] Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H. ARB: a software environment for sequence data. Nucleic Acids Res 2004;32:1363–71.
[7] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 2009;75:7537–41.
[8] Marx V. Biology: The big challenges of big data. Nature 2013;498:255–60.
[9] Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 2018;34:2490–2.
[10] Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations. Nucleic Acids Res 1984 (1985),;13:3021–30.
[11] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3.
[12] Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics 2016;32:3501–3.
[13] Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. expansion and model-centric curation of the comprehensive antibiotic resistance database (2017). Nucleic Acids Res 2017;45:D566–73.
[14] Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014;58:212–20.
[15] Hendling M, Pabinger S, Peters K, Wolff N, Conzemius R, Barišic I. Oli2go: an automated multiplex oligonucleotide design tool. Nucleic Acids Res 2018;46:W252–6.
[16] D'Onofrio V, Conzemius R, Varda-Brkić D, Bogdan M, Grisold A, Gyssens IC, Bedenić B, Barišić I. Epidemiology of colistin-resistant, carbapenemase-producing Enterobacteriaceae and Acinetobacter baumannii in Croatia. Infect Genet Evol 2020;81.