Article id: jbg.2017.3.5.1
Journal of Bioinformatics and Genomics, 2017, 3-5
doi: 10.18454/jbg.2017.3.5.1
Advance Access Publication Date: 13 November 2017

SEQUENCE ANALYSIS

TAXONDC: CALCULATING THE SIMILARITY VALUE OF THE 16S RRNA GENE SEQUENCES OF PROKARYOTES OR ITS REGIONS OF FUNGI

Acknowledgements

We are grateful to Dr. Taras V. Shevchuk for helpful feedback on the manuscript.

Conflict of Interest

None declared.

Sergey V. Tarlachkov[1,2,*], Irina P. Starodumova[2,3]

[1]All-Russian Collection of Microorganisms (VKM), G.K. Skryabin Institute of Biochemistry and Physiology of Microorganisms, Pushchino, Russia, [2]Branch of Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Pushchino, Russia, [3]Pushchino State Institute of Natural Sciences, Pushchino, Russia

*To whom correspondence should be addressed

Associate editor: Giancarlo Castellano
Received on 07 August 2017, revised on 20 August 2017, accepted on 25 August 2017.

Abstract

The TaxonDC program (Taxon Distance Calculator) performs pairwise sequence alignment followed by determining the similarity value between two or more sequences of interest. Unlike widely used programs, TaxonDC makes only pairwise alignment of input sequences that allows avoiding different similarity values depending on the sequences included in the analysis. The similarity values calculated with TaxonDC are the same compared to those calculated using popular identification oriented web-based tool EzBioCloud that makes calculated values comparable with previous ones. In addition, to help prevent discrepancy among different researchers, the problem concerning the influence of an order of the input of analyzed sequences on similarity values is specially considered. To our knowledge, TaxonDC is the only software which includes these capabilities in combination, simplifies and widens calculation of similarity values in systematics of prokaryotes and eukaryotes. The program has easy-to-use interface and can be run on Windows and Linux.
Availability and Implementation: *The program is available free of charge at https://tarlachkov.ru/en/software/taxondc.*
Supplementary information: *Supplementary data are available at Journal of Bioinformatics and Genomics online.*

Keywords: *TaxonDC, 16S rRNA, ITS, taxonomy-oriented software, similarity value.*

Contact*: sergey@tarlachkov.ru*

1 Introduction

Analysis of 16S rRNA gene sequences has an important role in systematics of prokaryotes (Kämpfer, Glaeser, 2013; Kim, Chun, 2014; Tindall et al., 2010). There are two ways to use 16S rRNA gene sequences: for phylogenetic analyses following multiple sequence alignments, and for calculating pairwise sequence similarities. The first approach is used to find out evolutionary relationships between related taxa and is important for generic or suprageneric classification. While the second approach provides a simple way for identification and delineation of novel isolates, and is a critical checkpoint at the species level (Kim, Chun, 2014; Stackebrandt, Ebers, 2006; Stackebrandt, Goebel, 1994).

The similarity value between sequences of the prokaryotic isolate and the closest species is calculated using a proper and robust global pairwise sequence alignment algorithm, not using local sequence alignment or fast searching algorithm (Tindall et al., 2010; Kim, Chun, 2014). Suchlike capability is available on a popular web-based tool EzBioCloud. However, in contrast to the first version (Chun et al., 2007), in which it was possible to compare any two sequences, the next versions of this tool (Kim et al., 2012; Yoon et al., 2017) are mainly designed to identify isolates using their own database.

There are a lot of programs for the alignment of the sequences and determination of their similarity values. However, these programs have a number of shortcomings. First, the similarity values calculated with these software do not always match with the values calculated using the widely accepted taxonomy-oriented EzBioCloud. Second, the software usually makes multiple alignment of all the entered sequences and is not suitable for fully independent pairwise comparisons of sequences. For this reason, these programs might produce different similarity values depending on the sequences included in the analysis. Thus, there exists a necessity for software, which uses EzBioCloud algorithm, and allows a comparison of two or more desirable sequences between themselves rather than with pre-set database.

At present, the fungal taxonomy is based primarily on analysis of the internal transcribed spacer (ITS) region of the nuclear ribosomal repeat unit (Peay et al., 2008; Schoch et al., 2012). To facilitate ITS-based molecular identification of fungi, specific databases were created like UNITE (Kõljalg et al., 2004) and EzBioCloud. These tools also provide the possibility of determining the similarity value between sequences of the fungal isolate and the closest species using pre-set database, but do not allow the comparison of a desired set of sequences.

Here, we introduce a new program TaxonDC (Taxon Distance Calculator) for calculating the 16S rRNA gene sequence similarities of prokaryotes or ITS regions of fungi.

2 Methods

TaxonDC is a desktop software written in C++ and using Qt libraries. It could be run under Windows and Linux. This program has easy-to-use interface and could operate in two modes.

In the first mode the sequences to compare could be entered in the 'Pair Input' tab (Fig. 1a) for single comparison of two sequences. Resulting alignment, similarity value, and statistics are shown in 'Pair Output' tab (Fig. 1b).

The second mode is the key feature of proposed TaxonDC that allows performing the independent pairwise alignment of more than two sequences with subsequent calculation of similarity values. In this mode initial sequences are entered in the 'Fasta Input' tab (Fig. 1c) in FASTA format. Results of calculation are shown in 'Fasta Output' tab (Fig. 1d). Summary table with sequence similarity values is located at the top of the 'Fasta Output' tab. If the cell with the similarity value is chosen in this table, then alignment and statistics are displayed in the bottom area for the visual inspection. If the sequence name is selected in the table, all alignments with this sequence will be displayed. The results could be exported in TSV format for Microsoft Excel or OpenOffice/LibreOffice Calc.

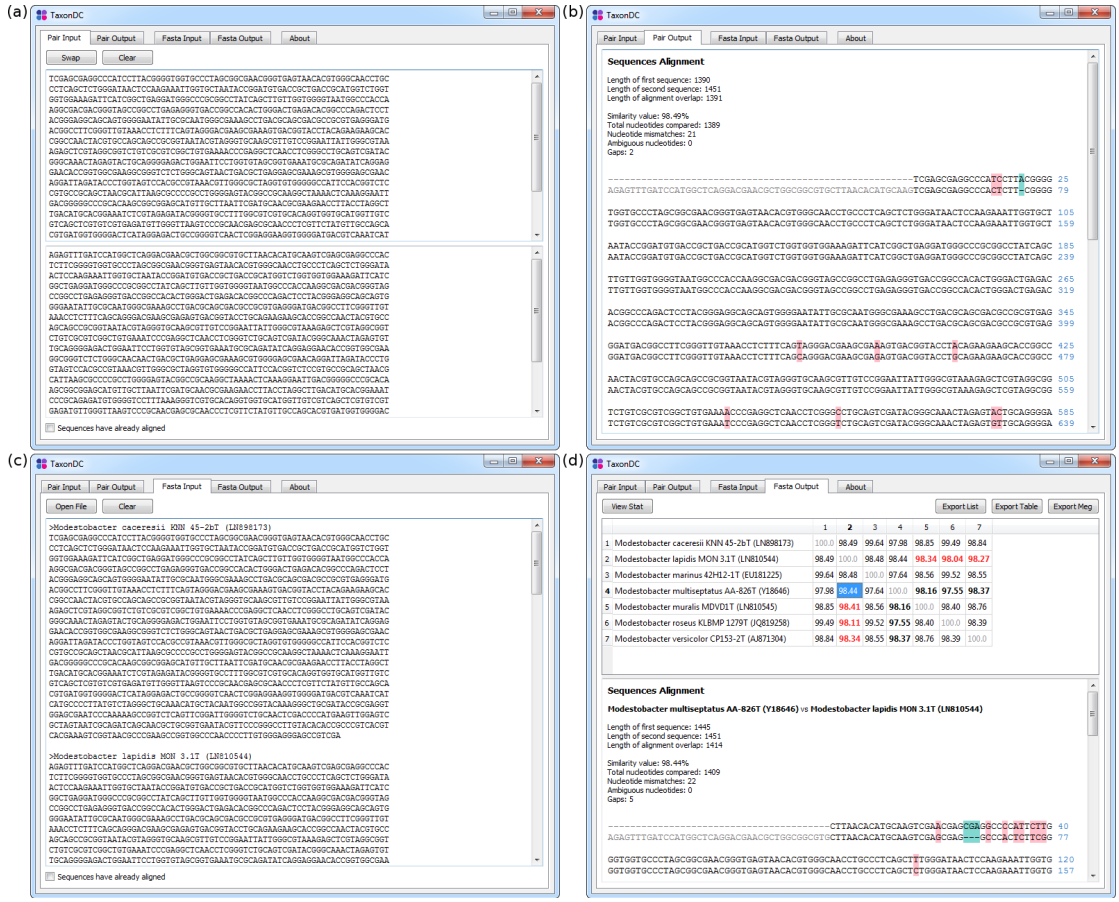

Fig. 1. The interface of TaxonDC. a, b are input and output tabs of single pair comparison mode respectively; c, d are input and output tabs of the mode of the independent pairwise comparison of more than two sequences, respectively.

Selecting the checkbox in the bottom of input tabs allows using pre-aligned sequences in both comparison modes.

To calculate the similarity value, on the first step TaxonDC uses CLUSTALW (Thompson et al., 1994) with default parameters for global sequence alignment. Our program is based on pairwise alignment only, but not on multiple sequence alignment. Therefore, pairs of sequences in the second mode have their own alignment, which do not depend on other sequences. The next step is determining the similarity value (*S.v.*) as:

$$\text{S.v.}\,(\%) = \frac{\text{Matches}}{\text{Mismatches} + \text{Matches}} \times 100$$

The alignment gaps and ambiguous bases (e.g., N) are not considered during calculation of similarity. If ones are considered, similarity values will be lower in most cases, especially in low quality sequences. In the example shown (Fig. 2), the total number of bases compared is 15, and the number of identical bases is 12, so the similarity is calculated as 80%.

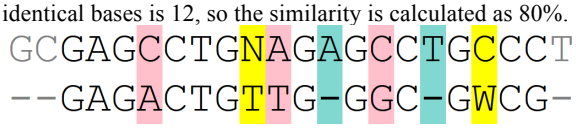

Fig. 2. Two aligned nucleotide sequences. Red color indicates nucleotide substitutions, green color is a gaps, and yellow color is ambiguous positions: N, A/C/G/T; or W, A/T.

3 Results

To verify the TaxonDC capabilities, the sequences of type strains from the previous version of EzBioCloud (Kim et al., 2012) were used. Total of 180 sequence pairs were downloaded from the domains of Bacteria (120), Archaea (30), and the kingdom of Fungi (30) (Supplementary Table S1). Each pair was randomly selected. Further, total of 60 sequence pairs were downloaded from the domains of Bacteria (30) and Archaea (30) from the latest version of tool (Yoon et al., 2016) (Supplementary Table S2). The 16S rRNA gene sequence and ITS region similarities compared were identical in both cases calculated using the TaxonDC and the EzBioCloud (Supplementary Table S1, S2).

The order of entered sequences has an effect on the result of the alignment, thereby affecting the calculated similarity value. This happens due to specific feature of the algorithm of the global sequence alignment. This specific feature is inherent to the EzBioCloud too, and hard to exclude completely keeping the results comparable with the previous ones. For instance, the 16S rRNA gene sequence similarity values of the *Ignicoccus islandicus* (CP006867) and *Ignicoccus pacificus* (AJ271794) pair compared in both directions are 98.35% or 98.27% (Supplementary Fig. S1–S2, Supplementary Table S2), and the ITS region similarity values of the *Geosmithia flava* (HF546291) and *Geosmithia morbida* (FN434081) pair compared in both directions are 98.06% or 98.25% (Supplementary Fig. S3–S4, Supplementary Table S1). Thus, this phenomenon may occur already at the species level. To consider this issue in TaxonDC, each pair of sequences is compared in both directions (A vs B and B vs A) to reveal any possible values difference. In a case, when such difference is found, two alignments and corresponding statistics are displayed. If the order of entered sequences gives different alignment results in the second mode, the corresponding similarity values are marked in bold in the summary table. Furthermore, different values are highlighted by color. For such situation, we recommend to choose a larger value since it corresponds to potentially a smaller number of evolution events. We believe it could help prevent discrepancies among different researchers.

4 Conclusion

TaxonDC was developed to provide an improved tool for calculation of pairwise similarity values between desired sequences. To our knowledge, among publicly available programs, TaxonDC is the only which combines all previously described features: comparability with EzBioCloud, fully independent pairwise alignment and revealing of the effect of the order of sequence input on calculated values. Our program could be useful in phylogenetics, ecology and systematics of prokaryotes and eukaryotes.

References

Chun, J., Lee, J. H., Jung, Y., Kim, M., Kim, S., Kim, B. K., Lim, Y. W. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int J Syst Evol Microbiol*, 57, 2259–2261. doi: 10.1099/ijs.0.64915-0

Kämpfer, P., Glaeser, S. P. (2013). 6 Prokaryote characterization and identification. In The Prokaryotes – Prokaryotic Biology and Symbiotic Associations, pp. 123–147. Edited by E. Rosenberg et al. Berlin Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-30194-0_6

Kim, M., Chun, J. (2014). Chapter 4. 16S rRNA Gene-based identification of Bacteria and Archaea using the EzTaxon server. In Methods in Microbiology. New Approaches to Prokaryotic Systematics, 41, pp. 61–74. Edited by M. Goodfellow, I. Sutcliffe, J. Chun. doi: 10.1016/bs.mim.2014.08.001

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., Park, S. C., Jeon, Y. S., Lee, J. H., Yi, H., Won, S., Chun, J. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol*, 62, 716–721. doi: 10.1099/ijs.0.038075-0

Kõljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F., Tedersoo, L., Vrålstad, T., Ursing, B. M. (2004). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol*, 166, 1063–1068. doi: 10.1111/j.1469-8137.2005.01376.x

Peay, K. G., Kennedy, P. G., Bruns, T. D. (2008). Fungal community ecology: a hybrid beast with a molecular master. *BioScience*, 58, 799–810. doi: 10.1641/b580907

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A*, 109, 6241–6246. doi: 10.1073/pnas.1117018109

Stackebrandt, E., Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today*, 33(4), 152–155.

Stackebrandt, E., Goebel, B.M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol*, 44, 846–849.

Thompson, J. D., Higgins, D. G., Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22, 4673–4680.

Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W., Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol*, 60, 249–266. doi: 10.1099/ijs.0.016949-0

Yoon, S. H., Ha, S. M., Kwon, S., Lim, J., Kim, Y., Seo, H., Chun, J. (2016). Introducing EzBioCloud: A taxonomically united database of 16S rRNA and whole genome assemblies. *Int J Syst Evol Microbiol*, 67(5), 1613–1617. doi: 10.1099/ijsem.0.001755