International Journal of Systematic and Evolutionary Microbiology (2007), 57, 2259–2261 DOI 10.1099/ijs.0.64915-0

EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences

Jongsik Chun,[1,2] Jae-Hak Lee,[1] Yoonyoung Jung,[1] Myungjin Kim,[2] Seil Kim,[2] Byung Kwon Kim[2] and Young-Woon Lim[2]

Correspondence
Jongsik Chun
jchun@snu.ac.kr

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, 56-1 Shillim-dong, Kwanak-gu, Seoul 151-742, Republic of Korea

[2]School of Biological Sciences and Institute of Microbiology, Seoul National University, 56-1 Shillim-dong, Kwanak-gu, Seoul 151-742, Republic of Korea

16S rRNA gene sequences have been widely used for the identification of prokaryotes. However, the flood of sequences of non-type strains and the lack of a peer-reviewed database for 16S rRNA gene sequences of type strains have made routine identification of isolates difficult and labour-intensive. In the present study, we generated a database containing 16S rRNA gene sequences of all prokaryotic type strains. In addition, a web-based tool, named EzTaxon, for analysis of 16S rRNA gene sequences was constructed to achieve identification of isolates based on pairwise nucleotide similarity values and phylogenetic inference methods. The system developed provides users with a similarity-based search, multiple sequence alignment and various phylogenetic analyses. All of these functions together with the 16S rRNA gene sequence database of type strains can be successfully used for automated and reliable identification of prokaryotic isolates. The EzTaxon server is freely accessible over the Internet at http://www.eztaxon.org/

INTRODUCTION

It is undisputed that information held in 16S rRNA gene sequences has played a vital role in microbiology (Rosselló-Mora & Amann, 2001) and that it can be utilized in many ways in various disciplines, notably taxonomy and ecology. Pairwise nucleotide similarity values of 16S rRNA gene sequences have been used routinely for the delineation of prokaryotic species (Stackebrandt & Goebel, 1994) and have been widely accepted. When a 16S rRNA gene sequence is applied for the identification of prokaryotic isolates, the general process includes: (i) similarity search against public domain nucleotide databases, (ii) retrieval of sequences for type strains with validly published names, (iii) calculation of pairwise nucleotide similarity values between sequences of the isolate and phylogenetically neighbouring type strains and (iv) phylogenetic analysis. This process requires expertise in prokaryotic taxonomy and can be complicated and labour-intensive, as public databases have been flooded with sequences of non-type strains including environmental clones and may contain erroneous and mislabelled sequences. Furthermore, there are no curated databases for 16S rRNA gene sequences of type strains of prokaryotic species. In this study, we present a database containing 16S rRNA gene sequences of all prokaryotic type strains and a web-based tool for analysis of 16S rRNA gene sequences, allowing automation of the process indicated above.

METHODS

Construction of 16S rRNA gene sequence database. The validly published names of prokaryotic species were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ) website (http://www.dsmz.de/). 16S rRNA gene sequences of type strains were extracted from the NCBI GenBank database. The validity and authenticity of sequence entries were checked by manual inspection by reviewing culture collection catalogues and relevant publications. Sequences that are a part of longer sequences (e.g. genome sequence) were trimmed to obtain only the 16S rRNA gene coding region. The names and corresponding 16S rRNA gene sequences will be updated on a monthly basis following the publication of the Validation and Notification Lists of the *International Journal of Systematic and Evolutionary Microbiology*.

Inputs for sequence similarity search. The formats of input data for sequence similarity searches are either text or tracer files generated from automated sequencers. The PHRED program was used for the base-calling and trimming of uploaded tracer files (Ewing & Green, 1998; Ewing *et al.*, 1998).

Finding phylogenetic neighbours using the EzTaxon server. The BLAST program (Altschul *et al.*, 1997) was employed for the initial

similarity search. The parsed BLAST hits are stored and can be sorted by BLAST scores or e values.

Calculation of pairwise sequence similarity values. The phylogenetically closest neighbours identified by BLAST search were then selected for the calculation of pairwise nucleotide sequence similarity using the algorithm of Myers & Miller (1988). The alignment gap was not considered in the similarity calculation.

Multiple sequence alignment and phylogenetic inference. The current version of the EzTaxon server provides multiple sequence alignment by CLUSTAL W (Thompson *et al.*, 1994). The resultant sequence alignment can then be used for the neighbour-joining (Saitou & Nei, 1987), maximum-parsimony (Fitch, 1971), and maximum-likelihood (Felsenstein, 1981) methods using the PHYLIP package (Felsenstein, 2005); all were implemented within the server. The alignment can be exported for use by external programs including jPHYDIT (Jeon *et al.*, 2005), PAUP (Swofford, 2002), PHYLIP, MEGA (Kumar *et al.*, 2004) and MrBayes (Ronquist & Huelsenbeck, 2003).

Operating system and programming languages. All databases and computer programs were generated using MySQL, JSP and JAVA under Linux operating system.

Availability. The EzTaxon server is freely accessible over the Internet at http://www.eztaxon.org/

RESULTS AND DISCUSSION

The database, named TYP16S, consisted of more than 7000 sequence entries together with the validly published names of the taxa concerned, GenBank/EMBL/DDBJ accession numbers and type strain designations. We have found that there are still more than 250 species with validly published names for which 16S rRNA gene sequences are not available.

The overall process for the identification of isolates using the EzTaxon server is given in Fig. 1. The final goal of the system is to find taxonomically meaningful phylogenetic relatives from either sequences or raw tracers (chromatograms). The BLAST search is based on local alignment that cannot be used for the calculation of overall sequence similarity. On the other hand, pairwise global sequence alignment using the algorithm of Myers & Miller (1988) guarantees the identification of the most similar sequence, but requires a higher computing cost. The EzTaxon server utilizes two methods sequentially to obtain accurate search results at reasonable computing cost. Such a combinatory strategy should guarantee the identification of the closest phylogenetic neighbour to the query sequence.

The 16S rRNA gene sequence similarity value has played an important role in delineating novel taxa and in the identification of isolates. Stackebrandt & Goebel (1994) suggested that a 16S rRNA gene sequence similarity of 97 % should become the boundary for delineation of prokaryotic species, which has been well accepted among microbiologists. More recently, Stackebrandt & Ebers (2006) proposed a more relaxed cut-off value of 98.7–99 %, after inspection of a large amount of recently published data. Even though this new proposal requires further validation and discussion, it is evident that high quality of sequencing should be the prerequisite to the use of lower similarity cut-off values for the identification of prokaryotes.

Similarity values tend to vary depending on the alignment algorithms employed and how gaps are considered. In addition, calculations based on multiple sequence alignment might produce different similarity values depending on the sequences included in the analysis. Therefore, to achieve a reliable and conservative measure of pairwise similarity values, we propose that the calculation should be carried out using rigorous pairwise global alignment algorithms such as that developed by Myers & Miller

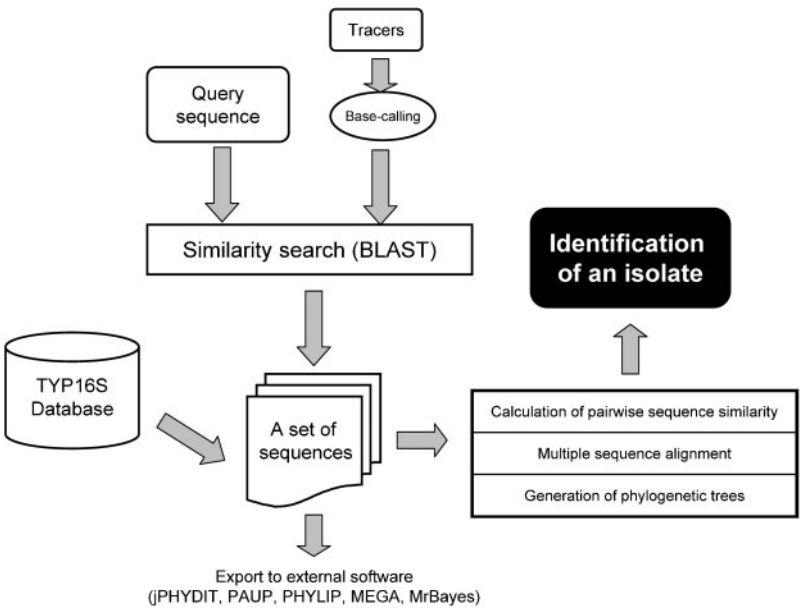

Fig. 1. Overview of the process for analysing 16S rRNA gene sequences using the EzTaxon server (http://www.eztaxon.org/). The websites from which the programs can be obtained are jPHYDIT (http://plaza.snu.ac.kr/~jchun/jphydit/), PAUP (http://paup.csit.fsu.edu/), PHYLIP (http://evolution.genetics.washington.edu/phylip.html), MEGA (http://www.megasoftware.net/) and MrBayes (http://mrbayes.csit.fsu.edu/).

(1988), and that alignment gaps should not be considered. The calculation of pairwise sequence similarity values from multiple alignments should be avoided.

EzTaxon was successfully developed to include the collection of all reference sequences, and to provide various functions, including a similarity search engine, calculation of pairwise similarity, multiple sequence alignment and phylogenetic treeing algorithms at the server side. This would allow users in places where only poor computing facilities are available to carry out most of the necessary bioinformatic analyses of 16S rRNA gene sequences. It could also be useful in various levels of classes for educational purposes.

In future, the EzTaxon server will be upgraded to include more analysis tools for phylogenetics, taxonomy and ecology of prokaryotes. We also plan to implement parallel computing versions of time-consuming methods such as multiple sequence alignment and maximum-likelihood inference of phylogenetic trees.

ACKNOWLEDGEMENTS

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the National Research Lab. Program funded by the Ministry of Science and Technology (no. M10500000110-06J0000-11010). We are grateful to the Korea Bioinformatics Center (http://www.kobic.re.kr/) for supporting hardware.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res* **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res* **8**, 175–185.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368–376.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package), version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, USA.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool* **20**, 406–416.

Jeon, Y. S., Chung, H., Park, S., Hur, I., Lee, J. H. & Chun, J. (2005). jPHYDIT: a JAVA-based integrated environment for molecular phylogeny of ribosomal RNA sequences. *Bioinformatics* **21**, 3171–3173.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Myers, E. W. & Miller, W. (1988). Optimal alignments in linear space. *Comput Appl Biosci* **4**, 11–17.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rosselló-Mora, R. & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39–67.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

Stackebrandt, E. & Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* **33**, 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.

Swofford, D. L. (2002). PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sunderland, MA: Sinauer Associates.

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.