International Journal of Systematic and Evolutionary Microbiology (2016), **66**, 1100–1103 **DOI** 10.1099/ijsem.0.000760

OrthoANI: An improved algorithm and software for calculating average nucleotide identity

Imchang Lee,[1,2] Yeong Ouk Kim,[2,3] Sang-Cheol Park[2,3] and Jongsik Chun[1,2,3]

Correspondence
Jongsik Chun
jchun@snu.ac.kr

[1]School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea

[2]Institute of Molecular Biology & Genetics, Seoul National University, Seoul 151-742, Republic of Korea

[3]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Republic of Korea

Species demarcation in *Bacteria* and *Archaea* is mainly based on overall genome relatedness, which serves a framework for modern microbiology. Current practice for obtaining these measures between two strains is shifting from experimentally determined similarity obtained by DNA–DNA hybridization (DDH) to genome-sequence-based similarity. Average nucleotide identity (ANI) is a simple algorithm that mimics DDH. Like DDH, ANI values between two genome sequences may be different from each other when reciprocal calculations are compared. We compared 63 690 pairs of genome sequences and found that the differences in reciprocal ANI values are significantly high, exceeding 1 % in some cases. To resolve this problem of not being symmetrical, a new algorithm, named OrthoANI, was developed to accommodate the concept of orthology for which both genome sequences were fragmented and only orthologous fragment pairs taken into consideration for calculating nucleotide identities. OrthoANI is highly correlated with ANI (using BLASTn) and the former showed approximately 0.1 % higher values than the latter. In conclusion, OrthoANI provides a more robust and faster means of calculating average nucleotide identity for taxonomic purposes. The standalone software tools are freely available at http://www.ezbiocloud.net/sw/oat.

INTRODUCTION

The genome is the ultimate source of information for taxonomic purposes and its use has been accelerated significantly thanks to advances in high-throughput DNA sequencing technologies (Chun & Rainey, 2014). Currently, the major application of genome sequence data in bacterial taxonomy is to measure overall genomic relatedness between two strains, which also serves as the framework for the species concept (Rosselló-Móra & Amann, 2015). The DNA–DNA hybridization (DDH) method has been regarded as the gold standard for the last few decades (Wayne *et al.*, 1987), despite the fact that it is only an indirect measure of genome sequence similarity, error-prone and labour-intensive (Johnson & Whitman, 2007). Since whole-genome sequencing is readily available for general microbiology laboratories, several overall genome relatedness indices (OGRI) have been developed to replace the problematic DDH methods. In general, OGRI algorithms are used to calculate similarity between two genome sequences without gene-finding and functional annotation steps, therefore they tend to be more objective, reproducible, fast and easy-to-implement.

Among various ORGI, average nucleotide identity (ANI) has been the most widely used (Beaz-Hidalgo *et al.*, 2015; Li *et al.*, 2015; Rosselló-Móra & Amann, 2015; Stropko *et al.*, 2014; Yi & Chun, 2015). ANI was first introduced to mimic the process of experimental DDH and thereby also called as digital version of DDH (Goris *et al.*, 2007; Konstantinidis & Tiedje, 2005). ANI values can be obtained using either BLASTn or MUMMER software (Richter & Rosselló-Móra, 2009) and the former is much widely used for taxonomic purposes (Kim *et al.*, 2014; Rosselló-Móra & Amann, 2015; Stropko *et al.*, 2014; Yi & Chun, 2015). Recently, Li *et al.* (2015) suggested that MUMMER is not suitable for ANI calculation. Therefore, we use the term ANI for the technique based on BLASTn in this study.

ANI is calculated from two genome sequences (of the query and subject strains) as follows: First, the genome sequence

Abbreviations: ANI, average nucleotide identity; DDH, DNA–DNA hybridization; OGRI, overall genome relatedness indices.

Four supplementary tables are available with the online Supplementary Material.

of the query strain is divided into 1020 bp-long sequences (fragments). Second, each fragment is searched against the whole genome sequence of the subject strain using NCBI's BLASTn program (Altschul *et al.*, 1997). In this process, the BLASTn program calculates nucleotide identity values between fragments of the query strain and the genome of the subject strain. Average nucleotide identity is the mean of these nucleotide identity values.

It has been known that reciprocal DDH values between two strains are often not the same, therefore not symmetrical, when DDH methods use labelled DNA (Johnson & Whitman, 2007, Tindall *et al.*, 2010). Since the theoretical concept of ANI derives from DDH, this may be also true for ANI. In other words, ANI of strain A (as query) to strain B (as subject) may be different from that of strain B (as query) to strain A (as subject). A reasonable practice would be to use the mean of two reciprocal ANI values, even though there is no theoretical basis for this, or for choosing either value. In this study, we investigate this problem and propose a new algorithm, called OrthoANI (Average Nucleotide Identity by Orthology), which can replace the original ANI.

METHODS

Dataset. A total of 14 745 genome sequences representing members of 10 genera (*Acinetobacter*, *Bacillus*, *Enterococcus*, *Escherichia*, *Mycobacterium*, *Pseudomonas*, *Salmonella*, *Staphylococcus*, *Streptococcus* and *Vibrio*) were selected from the EzBioCloud Genome database (http://www.ezbiocloud.net/) in which low quality and potentially contaminated genomes were checked and excluded. These genera were chosen as they contain the largest numbers of genomes.

Calculation of the original ANI values. Since calculating all possible pairs in our dataset was not computationally possible, we randomly selected genome pairs belonging to the same genus. The final dataset contained 63 690 genome pairs. For the ANI calculation, we used the previously described algorithm (Richter & Rosselló-Móra, 2009) except that NCBI BLASTn+ was used instead of the legacy BLASTn package. The reciprocal ANI values were obtained for each of the genome pairs.

OrthoANI algorithm. The algorithmic schema to calculate OrthoANI between two genomes is given in Fig. 1, which consists of three steps. First, both genome sequences were cut into consecutive 1020 bp-long fragments. Any fragments less than 1020 bp in size were omitted and ignored. Second, all fragments were searched and nucleotide identities were calculated using the BLASTn program. In this study, we used NCBI-BLASTn+ (version 2.2.30) with the following parameters: -task=BLASTn, -dust=no, -xdrop_gap=150, -penalty=−1, -reward=1 and -evalue=$1.0e^{-15}$; the rest of the parameters that could affect the result were set to default. Third, orthologous fragments between two genomes were identified when they showed reciprocal best hit in BLASTn searches. Because BLASTn is based on local alignment, we chose local alignments (also called HSP) with at least 35 % of the total length of the fragment (i.e. 357 bp out of 1020 bp); this cut-off value is set to match the value of 70 % suggested by Goris *et al.* (2007) in which only one genome sequence is fragmented. In contrast, both genome sequences are fragmented for OrthoANI. Since nucleotide identities can be obtained reciprocally, these were averaged to give average nucleotide identity of an orthologous fragment pair. The genome-wide nucleotide identity value was finally calculated as the average of identity values among all orthologous fragment pairs between two genomes.

Statistical analysis. Statistical analysis was performed to investigate the correlation between the original ANI and OrthoANI values using the R package (https://www.r-project.org).

Implementation and availability. The OrthoANI algorithm is implemented in JAVA programming language and is provided as two different software types: OAT (Orthologous ANI Tool) is a graphical user interface program that can be used interactively on personal computer environments and provides the functionality of performing UPGMA clustering. OAT_cmd is a command-line program that can be integrated into the user's own bioinformatics pipeline. In addition to OrthoANI, it also offers the calculation of GGDC genome-relatedness values (Meier-Kolthoff *et al.*, 2013). Both software tools are freely available at http://www.ezbiocloud.net/sw/oat.

RESULTS AND DISCUSSION

Like DDH methods based on labelled DNA, ANI is not symmetrical. Indeed, 55 % of 63 690 genome pairs examined in this study exhibited over 0.1 % discrepancy between reciprocal ANI values. Moreover, 1101 pairs showed more than 1 % discrepancy with the highest being 4.15 % (Tables S1 and S2, available in the online Supplementary Material and Fig. 2). Given that approximately 95–96 % ANI values are considered as the species boundary (Chun & Rainey, 2014; Goris *et al.* 2007; Richter & Rosselló-Móra, 2009), this level of discrepancies is significant enough to affect subsequent taxonomic interpretation. We also obtained reciprocal nucleotide identities values using ANI calculator (http://enve-omics.ce.gatech.edu/ani/) and Jspecies (http://imedea.uib-csic.es/jspecies/) for 100 genome pairs (Table S3). In general, all software tools do not provide exactly identical values, albeit they provide very similar values.

To resolve this problem, we developed a new ANI algorithm, named 'OrthoANI', to include the concept of orthology (Fig. 1). Unlike the original ANI, reciprocal OrthoANI values are always identical because of its algorithmic nature. The correlation between the original ANI and OrthoANI is very high (R^2=0.9998 for whole range and R^2=0.9995 for >90 % OrthoANI range; Fig. 3). OrthoANI values are slightly higher (approximately0.1 %) than the original ANI values in the range of approximately 95–96 %.

The computing time required for calculating OrthoANI between two genomes is 1.3–4-fold less than reciprocal original ANI, when tested on a desktop personal computer (Table S4). The degree of speed-up depends on the number of threads, length of the contigs and the overall genome sizes. In general, more threads and longer contigs result in a higher speed-up while the overall size of the genome is inversely proportional to the speed-up. Therefore, OrthoANI should be better suited to large scale comparison studies.

Several early studies recommended ANI value of approximately 95–96 % as cut-off for species demarcation

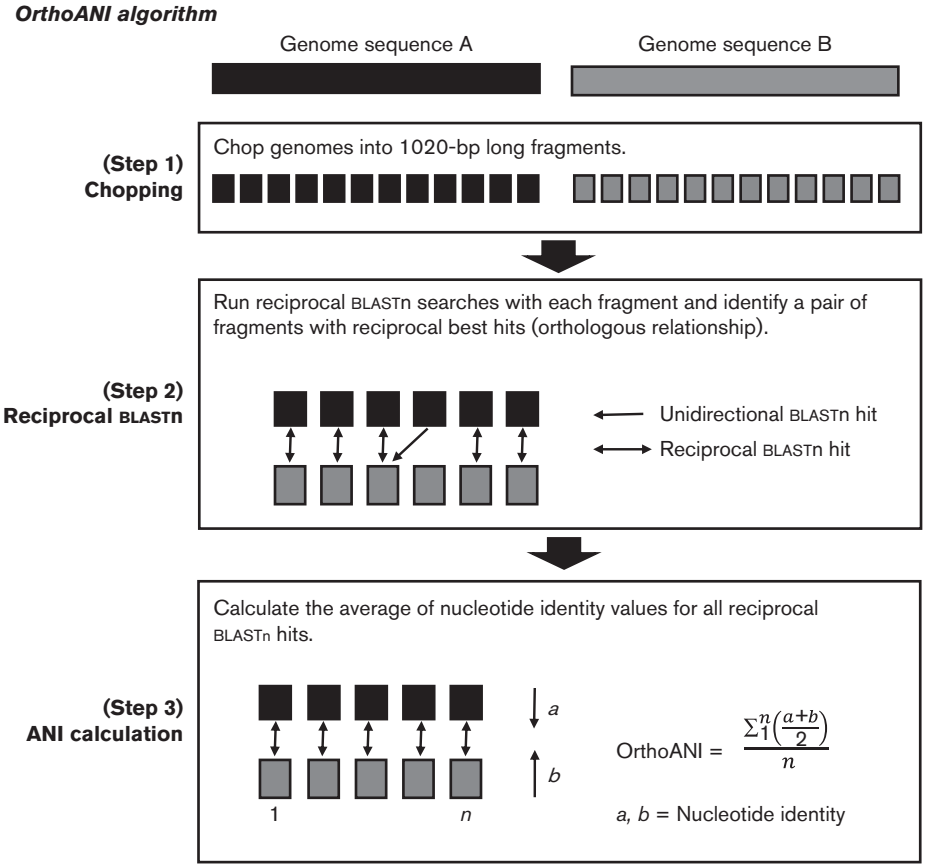

Fig. 1. Schematic diagram for the OrthoANI algorithm. The major differences between ANI and OrthoANI are: (1) in OrthoANI, both genomes are fragmented *in silico*, (2) OrthoANI does not use fragments of less than 1020 bp, and (3) in OrthoANI, only when two fragments are reciprocally searched as best hits using BLASTn program are their nucleotide identity values included in the subsequent computation.

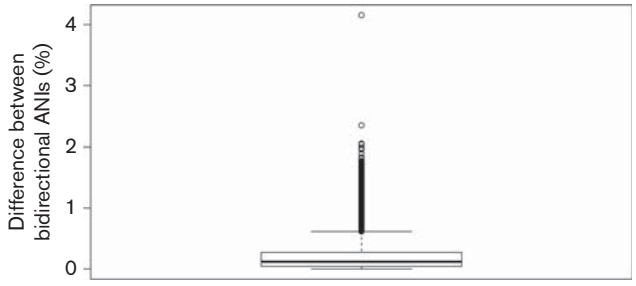

Fig. 2. Box plot showing differences between reciprocal ANI values on the basis of 63 690 pairs of genome sequences.

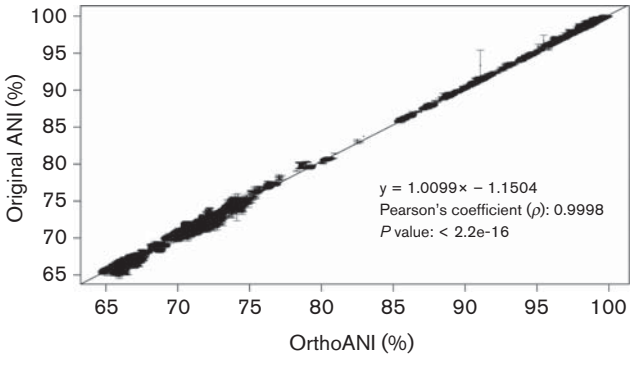

Fig. 3. Correlation between original ANI (average value of two reciprocal values) and OrthoANI identities. For the original ANI both reciprocal values were plotted.

(Goris *et al.*, 2007; Richter & Rosselló-Móra, 2009). Since OrthoANI in this range is only slightly higher than original ANI, we also recommend a similar range of cut-offs. It is also worth noting that ANI and OrthANI do not provide good measures for distantly related genomes (Kim *et al.*, 2014; Rosselló-Móra & Amann, 2015). For example, they should not be used to compare genomes belonging to different genera.

In conclusion, we proposed a modified version of ANI, named OrthoANI, to solve the problem of reciprocal inconsistency of the original ANI algorithm. Moreover, this new measure of genomic relatedness correlates well with the original ANI and can be readily used for taxonomic purposes. Like original ANI, it does not require gene-finding and functional annotation processes, allowing simple, reproducible and standardized procedures for taxonomic uses. With the easy-to-use GUI version and command-line version for large-scale computation, the algorithm should be accessible to all levels of microbiologists and students.

Acknowledgements

This work was supported by the National Science Foundation (grants NRF-2014M3C9A3063541 and NRF-2015R1A2A2A01008404).

References

Altschul, S. F., Madden, T. L., Schäffer, 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.

Beaz-Hidalgo, R., Hossain, M. J., Liles, M. R. & Figueras, M. J. (2015). Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for *Aeromonas* genomes in the GenBank database. *PLoS One* **10**, e0115813.

Chun, J. & Rainey, F. A. (2014). Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. *Int J Syst Evol Microbiol* **64**, 316–324.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007). DNA−DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**, 81–91.

Johnson, J. L. & Whitman, W. B. (2007). Similarity Analysis of DNAs. In *Methods for General and Molecular Microbiology*, pp. 624–652. Edited by C. A. Reddy. Washington, DC: American Society for Microbiology.

Kim, M., Oh, H. S., Park, S. C. & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* **64**, 346–351.

Konstantinidis, K. T. & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**, 2567–2572.

Li, X., Huang, Y. & Whitman, W. B. (2015). The relationship of the whole genome sequence identity to DNA hybridization varies between genera of prokaryotes. *Antonie van Leeuwenhoek* **107**, 241–249.

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. P. & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.

Richter, M. & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126–19131.

Rosselló-Móra, R. & Amann, R. (2015). Past and future species definitions for *Bacteria* and *Archaea*. *Syst Appl Microbiol* **38**, 209–216.

Stropko, S. J., Pipes, S. E. & Newman, J. D. (2014). Genome-based reclassification of *Bacillus cibi* as a later heterotypic synonym of *Bacillus indicus* and emended description of *Bacillus indicus*. *Int J Syst Evol Microbiol* **64**, 3804–3809.

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.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E. & other authors (1987). International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**, 463–464.

Yi, H. & Chun, J. (2015). *Neisseria weaveri* Andersen *et al.* 1993 is a later heterotypic synonym of *Neisseria weaveri* Holmes *et al.* 1993. *Int J Syst Evol Microbiol* **65**, 463–464.