*Folia Biologica et Oecologica* 18: 66–71 (2024)

Acta Universitatis Lodziensis

# Simple approach to bacterial genomes comparison based on Average Nucleotide Identity (ANI) using fastANI and ANIclustermap

KLAUDIA MUSIAŁ *, LEON PETRUŃKO, DAWID GMITER

Jan Kochanowski University of Kielce, Faculty of Natural Sciences, Institute of Biology, Department of Microbiology, Uniwersytecka 7, 25–406 Kielce, Poland
E-mail: musial.klaudia0805@gmail.com

ABSTRACT

The Average Nucleotide Identity (ANI) was proposed as a standard for taxonomic affiliation of newly sequenced bacterial genomes. However, usage of ANI value as a means of strains phenotypic diversity offers a relatively easy way for studding bacterial phylogeny. Here we present a simple approach to bacterial genomes comparison based on ANI using fastANI and ANIclustermap. Both programs are available as an open-source tools and can be run using simple command lines. We present protocol for programs installation as a conda packages, that facilitate it utilization. Further, we explain how to prepare commands to perform the analysis. We believed our work could be useful for young scientist that begin their experience with bioinformatics.

**KEYWORDS:** bacterial genomes comparison, bacterial phylogeny, Average Nucleotide Identity (ANI), fastANI, ANIclustermap

## Introduction

Growing accessibility of bacterial genomes sequences through recent development on next generation sequencing (NGS) technology result in higher number of studies exploited this data (Buermans and den Dunnen, 2014; Edwards and Holt, 2013; Gmiter *et al*., 2021; Hodkinson and Grice, 2015). Use of bacterial whole genome sequences shed a new light on our understanding of bacterial diversity, evolution and mechanisms of virulence and environmental adaptation (Deurenberg *et al*., 2017; Gmiter *et al*., 2021; Kobras *et al*., 2021). However, use of NGS data might be challenging, as it requires not only knowledge about microbial genetics, but also appropriate hardware and, more importantly, at least basic computational skills (Edwards and Holt, 2013; Gmiter *et al*., 2021). Much software is shared as an open-source tools, that based on usage of relatively simple, but not always intuitive, commands (Edwards and Holt, 2013; Gmiter *et al*., 2021). The programs usage might be problematic, especially from the point of view of young scientist who have

DOI: https://doi.org/10.18778/1730-2366.18.10

just started they journey with bioinformatics.

Previously, we presented a short review of programs used for pan-genome analysis with a beginner's guide of how to work with them. We believe that it might be good introduction into the studies focused on bacterial pan-genomes (Gmiter *et al*., 2021).

Within this paper we would like to present a simple approach to bacterial genomes comparison based on Average Nucleotide Identity (ANI) using fastANI (Jain *et al*., 2018) and ANIclustermap (https://github.com/moshi4/ANIclustermap). The ANI was proposed as a standard for taxonomic affiliation of newly sequenced genomes. It is a similarity index between a given pair of genomes that can be applicable to prokaryotic organisms independently of their G+C content, and a cut-off score of > 95% indicates that they belong to the same species (Figueras *et al*., 2014) Nevertheless, the usage of ANI value as a mean of strains phenotypic diversity offers a relatively easy way for studding bacterial phylogeny. The proposed programs can be used for study the phylogeny of complete as well as DRAFT bacterial genomes. The biggest advantage of the programs is their relative simplicity in use. However, programs allow for basic phylogenetic analysis, and do not consider the differences between coding and non-coding regions or recombination regions. More detailed analysis will require another approach.

**Average Nucleotide Identity (ANI)**

As mentioned, ANI is generally used to confirm the affiliation of new genome to proposed genus. A general rule of 95% cut-off of ANI similarity closely reflects the traditional microbiological concept of DNA–DNA hybridization relatedness for defining species, where recommended cut-off point is 70% (Goris *et al*., 2007; Jain *et al*., 2018). As it is based on comparison of multiple genes it provides higher resolution comparing to other standard methods, such as 16S rRNA sequence comparison (Arahal, 2014). Online tools for calculation ANI value might be used, however, they not always offer flexibility in terms of data curation. Therefore, we propose the usage of fastANI and ANIclustermap, depending on a need of the researcher.

**Installation as a conda packages**

The presented programs are open-source tools, which means they are available for download free of charge. Their use by the PC (Windows) users requires the installation of the latest version of Ubuntu (Linux program based on Debian), which we strongly recommend. It can be installed as a dual boot with Windows (Gmiter *et al*., 2021). On the others hand, both programs can be used also on the computers with the macOS.

The easiest way for installing the programs is through their installation as a conda packages. Conda is an open-source package and environment management system that works on Windows, Linux and macOS. For more details about conda installation please see (Gmiter *et al*., 2021). After conda is properly installed, you can simply download and install fastANI and ANIclustermaps by typing following commands typed into the Terminal (Fig. 1):

```
conda install bioconda::fastani
```

or

```
conda install bioconda/label/cf201901::fastani
```

and

```
conda install -c conda-forge -c bioconda
  aniclustermap
```

**Usage protocol**

For presented work, we use 10 complete genome sequences of

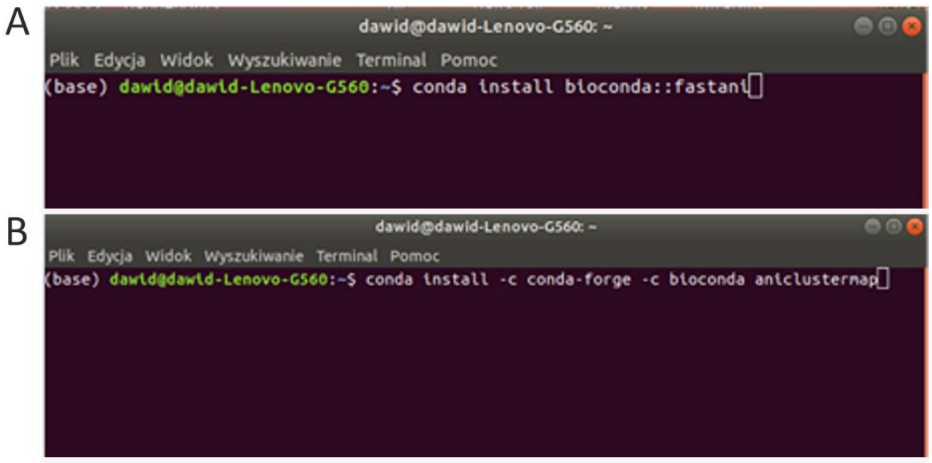


**Figure 1.** The command line required for installation of (A) fastANI and (B) ANIclustermap as conda packages on PC with Linux system.

*Escherichia coli* strains obtained from National Centre of Biotechnological Information (NCBI) (Availability: 04.10.2024) – see Table 1.

**Table 1.** *Escherichia coli* genome sequences used in this work.

| Strain | ID |
|---|---|
| 91H1 | NZ_CP149810.1 |
| 2023CK-01687 | NZ_CP149850.1 |
| 11128 | AP010960.1 |
| EH53 | NZ_CP146512.1 |
| JS01 | NZ_CP148986.1 |
| LF82 | CU651637.1 |
| MDP9-27 | NZ_CP146515.1 |
| SLDY13 | NZ_CP149967.1 |
| Z1322HEC0001 | NZ_CP148583.1 |
| Z1323CEC0007 | NZ_CP148463.1 |

## fastANI

The fastANI allows for the one-to-one, one-to-many and many-to-many genomes comparison. For one-to-one analysis simply type following command:

```
fastANI -q [QUERY_GENOME] -r
    [REFERENCE_GENOME] -o
    [OUTPUT_FILE]
```

The **-q** flag (flags provides the software the options to be used) stands for the query genome, a newly sequenced genome. Meanwhile, **-r** and **-o** flags provide the reference genome and localization of output files (results of analysis), respectively. It is expected that the input files will be genome assemblies in FASTA or multiFASTA format. To inform the program about localization of files, one must simply type the pathway to the file, for example:

```
-q /user/Desktop/genomes/new_genome.fasta
```

For one-to-many and many-to-many genomes comparison users must provide a pathway to text file containing directory paths to reference and/or query genomes, one per line:

```
/user/Desktop/genomes/genome1.fasta
/user/Desktop/genomes/genome2.fasta
/user/Desktop/genomes/genome3.fasta
```

In this situation, the flags **--ql** and **--rl** are required. It means that users must provide a path to the localization of genome list:

```
--ql or--rl /user/Desktop/geneme_list.txt
```

Users might use different or the same genome list as query and reference. To run the analysis the following commands should be used:

```
fastANI -q [QUERY_GENOME] --rl
  [REFERENCE_LIST] -o [OUTPUT_FILE]
fastANI --ql [QUERY_LIST] --rl
  [REFERENCE_LIST] -o [OUTPUT_FILE]
```

In all above cases, OUTPUT_FILE will contain tab delimited row(s) with query genome, reference genome, ANI value, count of bidirectional fragment mappings, and total query fragments. Alignment fraction (wrt. the query genome) is simply the ratio of mappings and total fragments.

Optional, to use the fastANI as a tool of genomes similarity investigation, users might supply **–matrix** parameter, which generate identity values arranged in a PHYLIP-formatted lower triangular matrix. The result of analysis of used *E. coli* genome sequences is presented as a similarity matrix in Table 2. The fastANI generates matrix in .txt format, which might be used in publication after manual correction (e.g. deletion of paths to genome sequences is required). The fastANI offers some option to be modified and the possible parameters are available after usage of **-h** flag. For more details, please see https://github.com/ParBLiSS/FastANI (Availability: 04.10.2024).

**ANIclustermap**

The fastANI provides only a numerical data, which are useful in many cases. However, ANI values might be visualized as a heat map of all-vs-all microbial genomes to better insight using ANIclustermap (https://github.com/moshi4/ANIclustermap). When ANIclustermap is used, ANI values are calculated by fastANI and clustermap is drawn using Seaborn. Additionally, ANIclustermap generate Newick format clustering dendrogram.

The usage of ANIclustermap requires following basic command:

```
ANIclustermap -i [Genome fasta directory] -o
  [output directory]
```

However, in contrary to fastANI, where input data is provided as a direct path to genome file or genomes list, here the path to folder containing all studied genome sequences is expected. ANIclustermap outputs 3 types of files:

- **ANIclustermap.[png|svg]** – ANI clustermap result figure,
- **ANIclustermap_matrix.tsv** – Clustered all-vs-all ANI matrix,
- **ANIclustermap_dendrogram.nwk** – Newick format clustering dendrogram.

Example of basic command required to run ANIclustermap is presented below:

```
ANIclustermap -i /user/Desktop/genomes -o
  /user/Desktop/genomes_results
```

**Table 2.** Matrix presenting ANI similarity between 10 used *E. coli* genome sequences obtained with fastANI.

| Strain | ANI Values | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 91H1 | | | | | | | | | |
| 2023CK-01687 | 98.88 | | | | | | | | |
| 11128 | 98.20 | 98.14 | | | | | | | |
| EH53 | 98.35 | 98.27 | 98.75 | | | | | | |
| JS01 | 98.34 | 98.33 | 98.83 | 99.37 | | | | | |
| LF82 | 96.80 | 96.65 | 96.45 | 96.72 | 96.71 | | | | |
| MDP9-27 | 98.33 | 98.30 | 98.79 | 99.72 | 99.45 | 96.70 | | | |
| SLDY13 | 99.04 | 98.85 | 98.22 | 98.23 | 98.27 | 96.85 | 98.28 | | |
| Z1322HEC0001 | 98.36 | 98.26 | 98.87 | 99.06 | 99.05 | 96.67 | 99.05 | 98.27 | |
| Z1323CEC0007 | 99.04 | 98.77 | 98.00 | 98.22 | 98.30 | 96.69 | 98.16 | 99.34 | 98.20 |

The resulted ANI clustermap of the analysis of 10 *E. coli* sequences is presented on Figure 3. Obtained heat map can be easily modified. Program offers possibility to alter map colours, its size, addition of annotation (drawing ANI values) by simple modification of command line. All parameters are available with **-h** flag. For more details, please see https://github.com/moshi4/ANIclustermap (Availability: 04.10.2024).

## Conclusions

In this work we presented a simple protocol for utilization of fastANI and ANIclustermap as tools allowing direct approach to study biodiversity of bacterial genome sequences. We understand that even though both programs are relatively easy in performance, their usage might be problematic for people without previous bioinformatics experience. We believe that this review and protocol will be a solid introduction in the issue.

## Acknowledgements

The realization of this work was supported by the Polish National Science Centre Grant 2019/32/T/NZ1/00515 (D.G.).

## Reference

Arahal, D.R. 2014. Whole-genome analyses: Average nucleotide identity. [In:] Methods in Microbiology, Vol. 41, pp. 103–122.

Buermans, H.P.J., den Dunnen, J.T. 2014. Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta – Molecular Basis of Disease, 1842(10), 1932–1941.

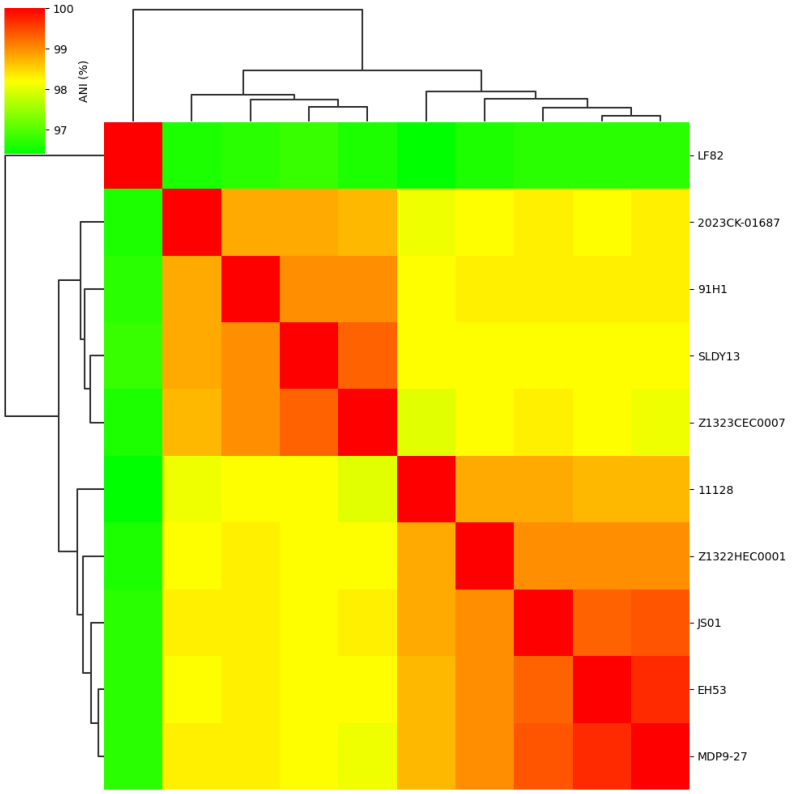


**Figure 3.** Results of ANIclustermap analysis of 10 *E. coli* genome sequences diversity.

Deurenberg, R H., Bathoorn, E., Chlebowicz, M.A., Couto, N., Ferdous, M., García-Cobos, S., Kooistra-Smid, A.M. D., Raangs, E.C., Rosema, S., Veloo, A.C. M., Zhou, K., Friedrich, A.W., Rossen, J.W.A. 2017. Application of next generation sequencing in clinical microbiology and infection prevention. Journal of Biotechnology, 243, 16–24.

Edwards, D.J., Holt, K.E. 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microbial Informatics and Experimentation, 3(1), 2.

Figueras, M.J., Beaz-Hidalgo, R., Hossain, M.J., Liles, M.R. 2014. Taxonomic affiliation of new genomes should be verified using Average Nucleotide Identity and Multilocus Phylogenetic Analysis. Genome Announcements, 2(6), e00927-14-e00927-14.

Gmiter, D., Nawrot, S., Pacak, I., Zegadło, K., Kaca, W. 2021. Towards a better understanding of the bacterial pan-genome. Acta Universitatis Lodziensis. Folia Biologica et Oecologica, 17, 84–96.

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. International Journal of Systematic and Evolutionary Microbiology, 57(1), 81–91.

Hodkinson, B.P., Grice, E.A. 2015. Next-Generation Sequencing: A Review of technologies and tools for wound microbiome research. Advances in Wound Care, 4(1), 50–58.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications, 9(1), 1–8.

Kobras, C.M., Fenton, A.K., Sheppard, S.K. 2021. Next-generation microbiology: from comparative genomics to gene function. In Genome Biology (Vol. 22, Issue 1). BioMed Central Ltd.