Bioinformatics, 2024, **40(4)**, btae175
https://doi.org/10.1093/bioinformatics/btae175
Advance Access Publication Date: 1 April 2024
Applications Note

OXFORD

Genome analysis

POCP-nf: an automatic Nextflow pipeline for calculating the percentage of conserved proteins in bacterial taxonomy

Martin Hölzer [1,*]

[1]Genome Competence Center (MF1), Robert Koch Institute, 13353 Berlin, Germany

*Corresponding author. Genome Competence Center (MF1), Method Development and Research Infrastructure (MFI), Robert Koch Institute, 13353 Berlin, Nordufer 20, Germany. E-mail: hoelzerm@rki.de

Associate Editor: Pier Luigi Martelli

Abstract

Summary: Sequence technology advancements have led to an exponential increase in bacterial genomes, necessitating robust taxonomic classification methods. The Percentage Of Conserved Proteins (POCP), proposed initially by Qin et al. (2014), is a valuable metric for assessing prokaryote genus boundaries. Here, I introduce a computational pipeline for automated POCP calculation, aiming to enhance reproducibility and ease of use in taxonomic studies.

Availability and implementation: The POCP-nf pipeline uses DIAMOND for faster protein alignments, achieving similar sensitivity to BLASTP. The pipeline is implemented in Nextflow with Conda and Docker support and is freely available on GitHub under https://github.com/hoelzer/pocp. The open-source code can be easily adapted for various prokaryotic genome and protein datasets. Detailed documentation and usage instructions are provided in the repository.

1 Introduction

Advances in sequencing technologies have driven the genomics era and led to an unprecedented influx of bacterial genomes. Taxonomic classification of these genomes is crucial for understanding microbial diversity and evolutionary relationships. Various metrics are employed to delineate the taxonomy of bacteria, each providing unique insights into their genomic characteristics and evolutionary history. Typical metrics include Average Nucleotide Identity, digital DNA-DNA hybridization, and 16S rRNA gene sequence similarity (Hayashi Sant'Anna *et al.* 2019). These metrics offer valuable information but may exhibit limitations, especially in prokaryotes with high genomic plasticity. As such, researchers need to employ a combination of metrics to comprehensively assess the evolutionary relationships between bacterial taxa.

One such metric is the Percentage of Conserved Proteins (POCP), a genome-based measure of taxonomic diversity originally proposed by Qin *et al.* (2014). POCP quantifies the degree of protein conservation between two genomes, providing a measure of genomic similarity. Unlike metrics based solely on nucleotide sequences, POCP focuses on functional elements, offering a more biologically relevant perspective on genomic relatedness. Thus, POCP is particularly well-suited for assessing genus boundaries, a challenging task in prokaryotic taxonomy. By considering the conservation of proteins, which are critical players in cellular function, POCP offers a nuanced understanding of the genomic distinctions between bacterial genera. The metric complements other methods and contributes to a more comprehensive characterization of microbial taxonomy. In the past, POCP calculations have been used in combination with other metrics in various studies to assess the genus boundaries of prokaryotes (Pannekoek *et al.* 2016, Harris *et al.* 2017, Leclercq *et al.* 2019, Ormeño-Orrillo and Martínez-Romero 2019, Suresh *et al.* 2019, Esquivel-Elizondo *et al.* 2020, Lalucat *et al.* 2020, Miyake *et al.* 2020, Xu *et al.* 2020, Joshi *et al.* 2021, Meng *et al.* 2021, Pan *et al.* 2021, Vorimore *et al.* 2021, Azpiazu-Muniozguren *et al.* 2022), in metagenomic contexts (Lagkouvardos *et al.* 2016, Zou *et al.* 2019, Wylensek *et al.* 2020, Amulyasai *et al.* 2022), and even fungi (Wibberg *et al.* 2021). In all of these studies, the POCP calculations were implemented and carried out slightly differently, mainly using the criteria defined in the original publication by Qin *et al.* (2014).

To harmonize calculations for the assessment of genus boundaries and to make the results more reproducible and comparable, I present a Nextflow pipeline for the automatic calculation of POCP values called POCP-nf. The pipeline's modular design allows seamless integration into larger analysis workflows, enabling researchers to leverage POCP alongside other metrics for a holistic exploration of bacterial evolutionary relationships. Through this contribution, I aim to enhance the accessibility and utility of POCP as a straightforward yet powerful metric in the rapidly evolving field of microbial genomics.

2 Pipeline description

The POCP-nf pipeline is developed in Nextflow (Di Tommaso *et al.* 2017), a workflow management system that

Received: 15 December 2023; **Revised:** 15 March 2024; **Editorial Decision:** 26 March 2024; **Accepted:** 28 March 2024

ensures portability and scalability across different computing environments. The pipeline accepts bacterial genome or protein datasets in standard FASTA format as input, one per bacterial species. Protein-coding genes are predicted by Prokka (Seemann 2014). If protein sequences are provided, the protein annotation step is skipped.

The pipeline identifies orthologous proteins between species using the blastp subcommand and 'ultra-sensitive' mode of DIAMOND (Buchfink *et al.* 2015). Per default, the proteomes of two strains are compared by bidirectional all-vs-all orthology searches. The user can define an optional target genome or protein FASTA to switch to one-vs-all comparisons when needed and save runtime. Those proteins of the query genome that have a hit with an e-value of $<1e{-}5$, an identity of $>40\%$, and an alignable region of $>50\%$ are called conserved based on the original POCP definition (Qin *et al.* 2014). Although the user can customize these parameters, I recommend sticking to the original parameters as defined by Qin *et al.* (2014) and otherwise clearly indicating any changed parameter options along with the version of POCP-nf used when sharing POCP results. The pipeline displays a warning if nonstandard parameters are used.

Each POCP value corresponds to the sum of the conserved proteins of two genomes divided by the sum of the total number of proteins of both genomes. A POCP of 50% was originally proposed as the genus limit, but it should be noted that the difference in proteome size between two strains influences the POCP value (Hayashi Sant'Anna *et al.* 2019).

The final output is a tab-separated table with all calculated pairwise POCP values and summary statistics to assess the results further. The modular design of the pipeline allows customization for specific datasets and enables integration into larger analysis workflows.

Calculating alignments, BLASTP versus DIAMOND: Please note that in the original POCP publication Qin *et al.* (2014) used BLASTP (Altschul *et al.* 1997) for calculating the alignments. However, DIAMOND is not only faster, which is an advantage when calculating POCP values for larger input datasets, but also achieves the sensitivity of BLASTP (Buchfink *et al.* 2021), especially when using the 'ultra-sensitive' mode, which is activated by default in POCP-nf. Another study comparing different alignment programs found that DIAMOND offered the best compromise between speed, sensitivity, and quality when a sensitivity option other than the default setting was selected (Hernández-Salmerón and Moreno-Hagelsieb 2020). I compared BLASTP and DIAMOND in ultra-sensitive mode within POCP-nf (v2.3.1) on five bacterial datasets with 15 to 167 genomes. I found an average difference in the percentage values of the calculated POCP of ~0.16%. The runtime (protein input) for 44 *Enterococcus* genomes is halved from 10 h 12 m (POCP-nf with BLASTP) to 5 h 30 m (POCP-nf with DIAMOND) on a laptop with eight cores. Further details can be found in the GitHub manual. I have, therefore, decided to use DIAMOND as a more modern solution for calculating the alignments in POCP-nf.

3 Installation and usage

Only Nextflow and either Conda, Mamba, Docker, or Singularity for dependency handling are needed to run the POCP-nf pipeline. The pipeline can be installed and executed with just two commands:

```
# install pipeline
nextflow pull hoelzer/pocp

# run selected release on input genomes
nextflow run hoelzer/pocp -r 2.3.1\
--genomes '<path/to/genomes/*fasta>'\
-profile local,docker
```

The repository's documentation provides detailed instructions, more advanced commands, and dependencies. Customization options and parameters are documented to accommodate different input formats and analysis environments.

4 Example analysis

To showcase the pipeline performance and output, I reanalyzed genomic data of 15 species from a study about the genus delineation of *Chlamydiales* species, where the authors used POCP values to justify the reunifying of the genera *Chlamydia* and *Chlamydophila* into one single genus *Chlamydia* (Pannekoek *et al.* 2016). I obtained the genome FASTAs from NCBI based on Supplementary Table S1 from the previously mentioned study. The pipeline in version 2.3.1 ran 26 min on a Linux laptop with eight cores, using <2 GB RAM. Figure 1 shows the calculated POCP values from the study in 2016 (upper triangle) compared to the re-calculated POCP values using the Nextflow pipeline (bottom triangle). The POCP values differ slightly, most likely due to differences in the protein annotation used in 2016 and with POCP-nf, and underlines the importance of a uniform method for calculating comparable POCP values. Note that the same results can only be achieved if the same protein FASTAs are used as input for the same method (same tools, tool versions, and parameters). However, the resulting POCP values correspond to those calculated and published in 2016.

5 Conclusion

POCP can serve as a robust genomic index for defining genus boundaries for prokaryotic groups. However, it is also important to emphasize that POCP is only one genomic metric among others. Researchers must interpret the results in the context of additional analyses for a holistic understanding of prokaryotic taxonomy. For example, POCP with a standard cutoff of 50% was not suitable for delimiting taxa of the family *Bacillaceae* at the genus level (Aliyu *et al.* 2016) and cannot yield a single criterion for dividing the genus *Borrelia* into two genera (Gupta 2019).

In this context, I also want to mention Protologger (Hitch *et al.* 2021), an all-in-one genome description tool designed to simplify the data collection required to generate protologues. The software, available for local installation and as a Galaxy (Afgan *et al.* 2022) tool, can calculate various metrics, including POCP values. However, while Protologger is a comprehensive software package for various computations comprising taxonomic placement, functional annotations, and ecological analyses, applying it only for POCP calculations on larger datasets, integrating it with other pipelines, or running it on a high-performance cluster or in the cloud can be challenging. In addition, Protologger is associated with a

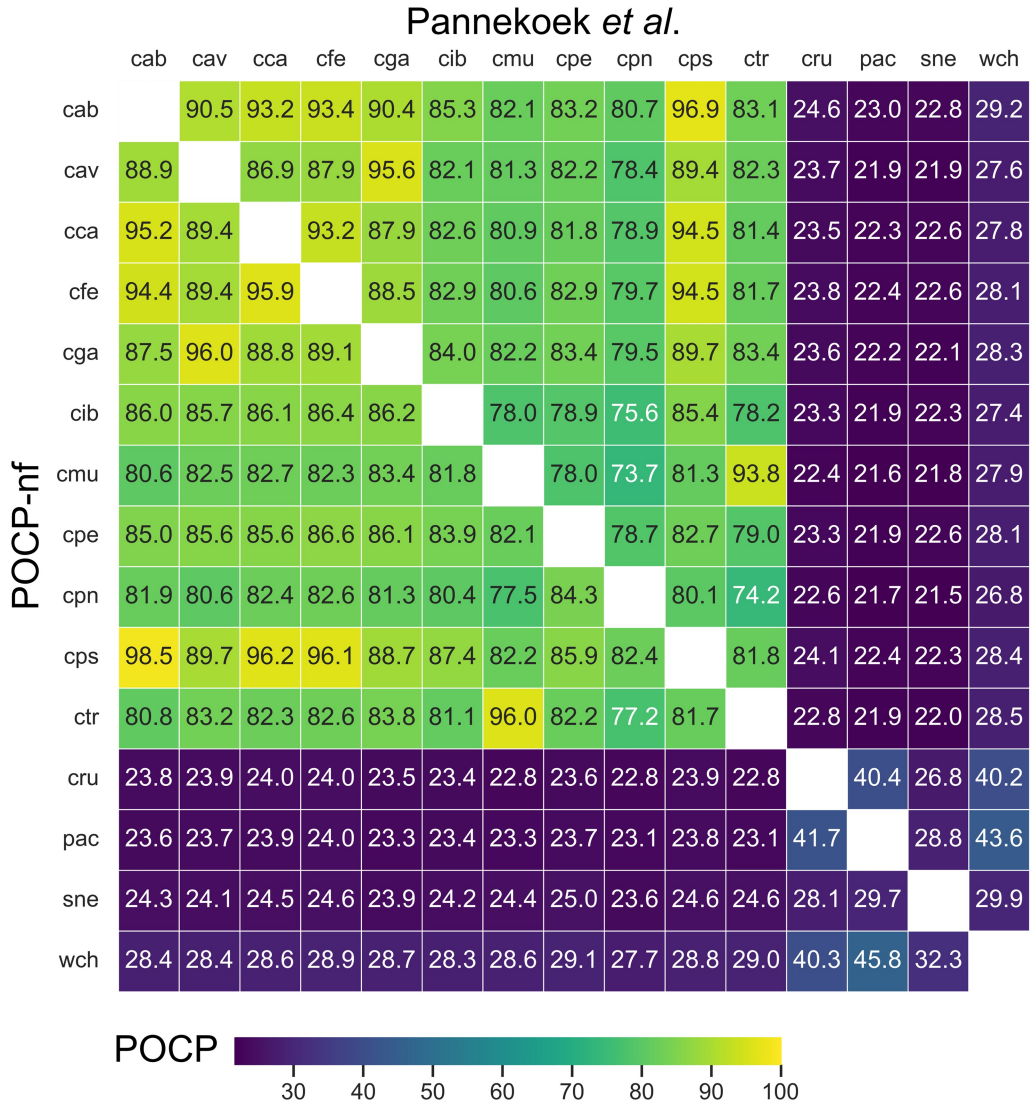

Figure 1. Pairwise POCP values from the original study of Pannekoek *et al.* (2016) (upper triangle) and recalculated with POCP-nf (lower triangle) of *Chlamydia* strains and outgroups. The average difference of percentage is 1.72% between all POCP values. *Chlamydia* (*C.*) *abortus* (cab), *C.avium* (cav), *C.caviae* (cca), *C.felis* (cfe), *C.gallinacea* (cga), *C.ibidis* (cib), *C.muridarum* (cmu), *C.pecorum* (cpe), *C.pneumoniae* (cpn), *C.psittaci* (cps), *C.trachomatis* (ctr), *Parachlamydia acanthamoebae* (pac), *Simkania negevensis* (sne), *Waddlia chondrophila* (wch), *Candidatus Rubidus massiliensis* (cru)

high computing effort, a long installation routine, and high memory requirements if the user is only interested in POCP values. Another alternative for calculating POCP values is provided in the web service EDGAR3.0 (Dieckmann *et al.* 2021), a comparative genomics and phylogenomics platform hosted via Galaxy. EDGAR3.0 is also an easy-to-use web service, especially for nonexperts, but it is subject to restrictions like those mentioned above. For users unfamiliar with the command line interface, I recommend using web services such as Protologger and EDGAR3.0 for the POCP calculation. However, I would encourage them to use POCP-nf on the command line as the necessary installations are already reduced to a minimum by using Nextflow (see example above and GitHub manual).

The POCP-nf pipeline fills a crucial gap by providing a user-friendly, lightweight, locally installable, and automated tool for calculating and harmonizing the percentages of conserved proteins. By facilitating efficient taxonomic classification, researchers can leverage the pipeline to gain insights into genus boundaries based on genomic data.

Acknowledgements

I thank Konrad Sachse for introducing me to bacterial taxonomy and POCP values. I thank Christian Brandt and Adrian Viehweger for introducing me to Nextflow and pushing the boundaries of rapid workflow development and automation via containers. I thank all the GitHub users who were interested in the pipeline and made suggestions for improvement, which ultimately motivated me to provide a Nextflow implementation and write this paper. Special thanks to Grégoire Michoud (@michoug) for suggesting using DIAMOND over Blast for faster protein searches and for providing example code. Thanks also to ChatGPT, which I used to create a basic framework for this paper and a Python script for the heatmap plot. Finally, I thank Matt Huska and Hugues Richard for their constructive comments while revising the pipeline and the manuscript.

Conflict of interest

None declared.

Funding

None declared.

Data availability

The pipeline is freely available at https://github.com/hoelzer/pocp. Input data and additional results for the example analysis are available at https://osf.io/2tzd9. A comparison between BLASTP and DIAMOND for POCP calculation is available in the GitHub manual.

References

Afgan E, Nekrutenko A, Grüning BA *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* 2022;**50**:W345–51.

Aliyu H, Lebre P, Blom J *et al.* Phylogenomic re-assessment of the thermophilic genus *Geobacillus*. *Syst Appl Microbiol* 2016;**39**:527–33.

Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.

Amulyasai B, Anusha R, Sasikala C *et al.* Phylogenomic analysis of a metagenome-assembled genome indicates a new taxon of an anoxygenic phototroph bacterium in the family *Chromatiaceae* and the proposal of "*Candidatus* thioaporhodococcus" gen. nov. *Arch Microbiol* 2022;**204**:688.

Azpiazu-Muniozguren M, García M, Laorden L *et al. Anianabacter salinae* gen. nov., sp. nov. ASV31^T, a facultative alkaliphilic and extremely halotolerant bacterium isolated from brine of a millennial continental saltern. *Diversity* 2022;**14**:1009.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;**18**:366–8.

Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;**35**:316–9.

Dieckmann MA, Beyvers S, Christian Nkouamedjo-Fankep R *et al.* EDGAR3.0: comparative genomics and phylogenomics on a scalable infrastructure. *Nucleic Acids Res* 2021;**49**:W185–92.

Esquivel-Elizondo S, Bağcı C, Temovska M *et al.* The isolate *Caproiciproducens* sp. 7D4C2 produces n-caproate at mildly acidic conditions from hexoses: genome and rBOX comparison with related strains and chain-elongating bacteria. *Front Microbiol* 2020;**11**:594524.

Gupta RS. Distinction between *Borrelia* and *Borreliella* is more robustly supported by molecular and phenotypic characteristics than all other neighbouring prokaryotic genera: response to margos' *et al.* "the genus *Borrelia* reloaded". *PLoS ONE* 2019;**14**:e0221397.

Harris HMB, Bourin MJB, Claesson MJ *et al.* Phylogenomics and comparative genomics of *Lactobacillus salivarius*, a mammalian gut commensal. *Microb Genom* 2017;**3**:e000115.

Hayashi Sant'Anna F, Bach E, Porto RZ *et al.* Genomic metrics made easy: what to do and where to go in the new era of bacterial taxonomy. *Crit Rev Microbiol* 2019;**45**:182–200.

Hernández-Salmerón JE, Moreno-Hagelsieb G. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. *BMC Genomics* 2020;**21**:741.

Hitch TCA, Riedel T, Oren A *et al.* Automated analysis of genomic sequences facilitates high-throughput and comprehensive description of bacteria. *ISME Commun* 2021;**1**:1–16.

Joshi A, Thite S, Karodi P *et al. Alkalihalobacterium elongatum* gen. nov. sp. nov.: an antibiotic-producing bacterium isolated from lonar lake and reclassification of the genus *Alkalihalobacillus* into seven novel genera. *Front Microbiol* 2021;**12**:722369.

Lagkouvardos I, Pukall R, Abt B *et al.* The mouse intestinal bacterial collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol* 2016;**1**:16131–15.

Lalucat J, Mulet M, Gomila M *et al.* Genomics in bacterial taxonomy: impact on the genus *Pseudomonas*. *Genes (Basel)* 2020;**11**:139.

Leclercq A, Moura A, Vales G *et al. Listeria thailandensis* sp. nov. *Int J Syst Evol Microbiol* 2019;**69**:74–81.

Meng D, Liu Y-L, Gu P-F *et al. Chelativorans alearense* sp. nov., a novel bacterial species isolated from soil in Alear, China. *Curr Microbiol* 2021;**78**:1656–61.

Miyake S, Ding Y, Soh M *et al. Muribaculum gordoncarteri* sp. nov., an anaerobic bacterium from the faeces of C57BL/6J mice. *Int J Syst Evol Microbiol* 2020;**70**:4725–9.

Ormeño-Orrillo E, Martínez-Romero E. A genomotaxonomy view of the *Bradyrhizobium* genus. *Front Microbiol* 2019;**10**:1334.

Pan X, Li Z, Li F *et al. Thermohalobaculum xanthum* gen. nov., sp. nov., a moderately thermophilic bacterium isolated from mangrove sediment. *Antonie Van Leeuwenhoek* 2021;**114**:1819–28.

Pannekoek Y, Qi-Long Q, Zhang Y-Z *et al.* Genus delineation of *Chlamydiales* by analysis of the percentage of conserved proteins justifies the reunifying of the genera *Chlamydia* and *Chlamydophila* into one single genus *Chlamydia*. *Pathog Dis* 2016;**74**:ftw071.

Qin Q-L, Xie B-B, Zhang X-Y *et al.* A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* 2014;**196**:2210–5.

Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9.

Suresh G, Lodha TD, Indu B *et al.* Taxogenomics resolves conflict in the genus *Rhodobacter*: a two and half decades pending thought to reclassify the genus *Rhodobacter*. *Front Microbiol* 2019;**10**:2480.

Vorimore F, Hölzer M, Liebler-Tenorio EM *et al.* Evidence for the existence of a new genus *chlamydiifrater* gen. nov. inside the family *Chlamydiaceae* with two new species isolated from flamingo (*Phoenicopterus roseus*): *Chlamydiifrater phoenicopteri* sp. nov. and *Chlamydiifrater volucris* sp. nov. *Syst Appl Microbiol* 2021;**44**:126200.

Wibberg D, Stadler M, Lambert C *et al.* High quality genome sequences of thirteen *Hypoxylaceae* (*Ascomycota*) strengthen the phylogenetic family backbone and enable the discovery of new taxa. *Fungal Divers* 2021;**106**:7–28.

Wylensek D, Hitch TCA, Riedel T *et al.* A collection of bacterial isolates from the pig intestine reveals functional and taxonomic diversity. *Nat Commun* 2020;**11**:6389.

Xu L, Sun C, Fang C *et al.* Genomic-based taxonomic classification of the family *Erythrobacteraceae*. *Int J Syst Evol Microbiol* 2020;**70**:4470–95.

Zou Y, Xue W, Luo G *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* 2019;**37**:179–85.

Bioinformatics, 2024, 40, 1–4
https://doi.org/10.1093/bioinformatics/btae175
Applications Note