Swarm: robust and fast clustering method for amplicon-based studies

Frédéric Mahé[1,2,3], Torbjørn Rognes[4,5], Christopher Quince[6], Colomban de Vargas[1,2], and Micah Dunthorn[3]

[1]**CNRS, UMR 7144, EPEP – Évolution des Protistes et des Écosystèmes Pélagiques, Station Biologique de Roscoff, 29680 Roscoff, France.**
[2]**Sorbonne Universités, UPMC Univ Paris 06, UMR7144 Station Biologique de Roscoff, Roscoff, France.**
[3]**Department of Ecology, University of Kaiserslautern, Kaiserslautern, Germany.**
[4]**Department of Microbiology, Oslo University Hospital, Rikshospitalet, Oslo, Norway.**
[5]**Department of Informatics, University of Oslo, Oslo, Norway.**
[6]**School of Engineering, University of Glasgow, Glasgow, UK.**

Keywords: environmental diversity, barcoding, molecular operational taxonomic units

High-throughput sequencing technologies are today our best approach to deeply assess the diversity of complex assemblages of microorganisms. Because of the increasing sizes of amplicon (or barcoding) datasets, fast and greedy *de novo* clustering heuristics are the preferred and the only practical approach to produce molecular operational taxonomic units (OTUs) (Edgar, 2010; Ghodsi et al., 2011; Fu et al., 2012). These greedy clustering methods suffer from two fundamental problems. First, they use an arbitrary fixed global clustering threshold. As lineages evolve at variable rates, no single cut-off value can accommodate the entire tree of life. A single global clustering threshold will inevitably be too relaxed for slow-evolving lineages and too stringent for rapidly evolving ones (Stackebrandt and Goebel, 1994; Sogin et al., 2006; Koeppel and Wu, 2013). Second, the input order of amplicons strongly influences the clustering results. Previous centroid selections are not re-evaluated as clustering progresses, which can generate inaccurately formed OTUs, where closely related amplicons can be separated and unrelated amplicons can be grouped (Koeppel and Wu, 2013).

Swarm's rationale

While working on two large scale environmental diversity studies using different markers and different sequencing platforms—the BioMarKs project (e.g. Dunthorn et al., 2014; Logares et al., 2014) and the TARA OCEANS project (e.g. Karsenti et al., 2011)—the limitations of greedy *de novo* clustering methods became salient, leading to erroneous ecological interpretations. To solve these issues, we developed Swarm—a novel method that avoids both fixed global clustering thresholds, and input-order dependency due to centroid selection. Our objective was to implement an exact, yet fast, *de novo* clustering method that produces meaningful OTUs and reduces the influence of clustering parameters.

Swarm can be defined as an agglomerative, unsupervised (*de novo*) single-linkage-clustering algorithm that first computes sequence differences between aligned pairs of amplicons to delineate OTUs, using *k*-mer comparisons (Ukkonen, 1992) and a new and extremely fast global pairwise alignment algorithm (similar to Rognes, 2011); and in a second step, uses amplicon abundance information and OTUs' internal structures to refine the clustering results.

The assumption behind Swarm is that amplicons do not form a continuum. If this condition holds true, then OTUs can be allowed to grow iteratively until they reach their natural limits. Operating in this way, Swarm removes the two main sources of variability inherent in greedy *de novo* clustering methods: the need to designate an OTU center (centroid selection), and the need for an arbitrary global clustering threshold. Swarm outlines OTUs without imposing one particular shape or size, and produces the same OTUs regardless of the initially selected amplicon.

Under certain conditions, when using short or slowly evolving markers for instance, the assumption that amplicons do not form a continuum can be violated. Single-linkage clustering is known to produce chains of amplicons that can potentially link closely related OTUs and decrease clustering resolution (Huse et al., 2010). To solve this issue, Swarm takes a post-clustering step: by representing the internal

structure of the cluster and amplicon abundance values as a graph, Swarm can identify natural breaking points, and delineate higher-resolution OTUs.

Swarm's performances and perspectives

Tests on mock-communities and on real datasets such as BioMarKs or TARA OCEANS show that Swarm is as fast as greedy methods, and produces equally good or better clustering results. Swarm results are stable for a wide range of clustering parameter values, limiting the impact of the choice of clustering threshold on downstream analyzes.

Swarm is efficient enough to deal with today's largest datasets, and several new optimizations are now in development to handle even larger future datasets. For example, using the new 256-bit SIMD instructions of the Intel CPUs can double Swarm's pairwise alignment throughput. We are also testing more efficient parallelization strategies, as well as new filters to avoid un-needed pairwise alignments. These hardware and software evolutions will improve Swarm's scalability even further. In parallel, improvements to our abundance-based chain breaking model will increase Swarm's capacity to produce high-resolution OTUs and meaningful biological results, even for the most intricate species complexes.

Swarm is freely available at https://github.com/torognes/swarm under the GNU Affero General Public License version 3.

REFERENCES

Dunthorn, M., Otto, J., Berger, S. A., Stamatakis, A., Mahé, F., Romac, S., de Vargas, C., Audic, S., The BioMarKs Consortium, Stock, A., Kauff, F., and Stoeck, T. (2014). Placing Environmental Next-Generation Sequencing Amplicons from Microbial Eukaryotes into a Phylogenetic Context. *Molecular Biology and Evolution*, 31(4):993–1009.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, 12(1):271.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, 12(7):1889–1898.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E. G., Sardet, C., Sieracki, M. E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P., and the Tara Oceans Consortium (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biology*, 9(10):e1001177.

Koeppel, A. F. and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Research*, 41(10):5175–5188.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.-M., Decelle, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Gobet, A., Kooistra, W. H. C. F., Mahé, F., Not, F., Ogata, H., Pawlowski, J., Pernice, M. C., Romac, S., Shalchian-Tabrizi, K., Simon, N., Stoeck, T., Santini, S., Siano, R., Wincker, P., Zingone, A., Richards, T. A., de Vargas, C., and Massana, R. (2014). Patterns of Rare and Abundant Marine Microbial Eukaryotes. *Current Biology*, 24(8):813–821.

Rognes, T. (2011). Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinformatics*, 12(1):221.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12115–12120.

Stackebrandt, E. and Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, 44(4):846–849.

Ukkonen, E. (1992). Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92(1):191–211.