Searching nearest neighbours in metric microbiome spaces using phylogenetic distance measures.

Andreas Henschel*

Masdar Institute, Khalifa University of Science and Technology, United Arab Emirates

Accepted on October 09, 2017

Commentary

The deposition of metagenomic data in large scales bears great potential to understand universal mechanisms and environmental factors of microbial community assembly. Notable efforts are the Human Microbiome Project, the Earth Microbiome Project, IMNGS and Qiita/QiimeDB [1-4]. We have thus entered a new era, in which it is in principle possible to discover commonalities between microbial communities from entirely different ecosystems. E.g., samples from below the ocean floor were surprisingly similar to non-marine communities due to methanogens [5]. However, with the current data deluge from Next Generation Sequencing projects it is becoming increasingly difficult to perform manual exhaustive searches to find the most similar microbial communities. Suitable search algorithms for microbiomes are required and solutions for automated microbiome search are beginning to appear, e.g., Meta-Storms [6] and Visibiome [7].

A popular approach to describe compositional features of microbial communities is through marker gene sequencing, in particular using the hypervariable regions of the 16S rRNA gene. These sequences can then be clustered into Operational Taxonomic Units (OTUs). In turn, a sample is represented as a vector of relative abundances of OTUs, spanning the microbiome search space. The total number of considered OTUs (either from a reference library or picked de novo) determines the dimensionality of the microbiome search space. The advent of deep environmental sequencing takes microbiome search literally to another dimension: many low abundance OTUs are now above the detection threshold. Moreover, the previously unappreciated diversity of different, novel bacterial OTUs in the environment [8] exacerbates this curse of dimensionality. From the classical Nearest Neighbour similarity searches, only few algorithms like GNAT and AESA can handle such high dimensions in the order of tens of thousands (or above) [9, 10]. Their complexities are suitable for the size of the abovementioned sample databases. Nearest Neighbour search commonly requires a metric distance measure (e.g., fulfilling the triangle inequality). Weighted UniFrac [11], a popular tool for measuring distances in microbiomes, indeed is a metric while also appreciating phylogenetic relations between OTUs. Visibiome therefore deploys an Earth Mover Distance based implementation of weighted UniFrac [12], optimized for sparse vectors (since not every sample contains every OTU). The entire search algorithm is capable of sublinear searches in high-dimensional metric spaces.

In order to scale to high demands while simultaneously providing user friendly access to microbiome research, Visibiome leverages a scalable, modular and distributed architecture that combines web framework technology, task queuing and scheduling, cloud computing and a dedicated database server.

In analogy to sequence similarity search tools like BLAST [13] that facilitate annotation transfer, Visibiome matches novel microbial communities to other well annotated samples and can thus provide clues about the function of a particular community at hand. Extending the analogy, the equivalent of BLAST's query-subject sequence alignment is in Visibiome a series of comparative stacked bar charts that show corresponding abundances of compositional taxa in query and subject, on various, user selected taxonomic levels.

In conclusion, novel search engines for microbial communities are poised to cope with the demand created by the data deluge in microbiome research. Visibiome in particular is a convenient, scalable and efficient framework to search microbiomes against a comprehensive database of environmental samples. It confirmed the atypical composition of the abovementioned ocean floor sample. The search engine leverages a phylogeny based distance metric, while providing advantages over existing tools.

References

1. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project: Exploring the microbial part of ourselves in a changing world. Nature. 2007 Oct 18: 449 (7164):804**.**
2. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC biology. 2014 Aug 22: 12(1):69.
3. Lagkouvardos I, Joseph D, Kapfhammer M, et al. A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. Scientific reports. 2016: 6:33721.
4. Qiita. http://qiita.microbio.me. Accessed: 2017-03-16.
5. Inagaki F, Hinrichs KU, Kubo Y, et al. Exploring deep microbial life in coal-bearing sediment down to 2.5 km below the ocean floor. Science [Internet]. American Association for the Advancement of Science (AAAS): 2015: 349(6246):420–4.
6. Su X, Xu J, Ning K. Meta-Storms: Efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics. 2012: 28(19):2493-501.

7. Azman SK, Anwar MZ, Henschel A. Visibiome: An efficient microbiome search engine based on a scalable, distributed architecture. BMC bioinformatics. 2017: 18(1):353.
8. Rideout JR, He Y, Navas-Molina, et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ. 2014: 2:e545.
9. Brin S. Near neighbor search in large metric spaces. 21th International Conference on Very Large Data Bases.1995:11-15.
10. Micó ML, Oncina J, Vidal E. A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements. Pattern Recognition Letters. 1994: 15(1):9-17.
11. Lozupone C, Knight R. UniFrac: A new phylogenetic method for comparing microbial communities. Applied and environmental microbiology. 2005: 71(12):8228-35.
12. McClelland J, Koslicki D. EMDUnifrac: Exact linear time computation of the Unifrac metric and identification of differentially abundant organisms. arXiv preprint arXiv: 2016:1611.04634.
13. Stephen F Altschul, Thomas L Madden, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 1997 25(17):3389–3402.

***Correspondence to:**

Andreas Henschel
Masdar Institute
Khalifa University of Science and Technology
United Arab Emirates
Email: ahenschel@masdar.ac.ae
Tel: +971 2 810 9222