Towards Peta-Scale Methods for Taxonomic Assignment of Environmental DNA – Overview

Eske Willerslev
2023 Balzan Prize for Evolution of Humankind:
Ancient DNA and Human Evolution

Balzan GPC Adviser: Peter Suter
Deputy Supervisors: Rasmus Nielsen, Thorfinn Sand Korneliussen
Institution: Section for GeoGenetics, Globe Institute, University of Copenhagen
Period: 2024-2027

Eske Willerslev is Prince Philip Professor in Ecology and Evolution at the University of Cambridge and the Professor in Evolution at Copenhagen University. He is director of the Centre of Excellence for Ancient Environmental Genomics (CAEG) and of the Lundbeck Foundation GeoGenetics Centre at the University of Copenhagen.

Background
Environmental DNA (eDNA) has emerged as a revolutionary tool for biodiversity monitoring, enabling non-invasive detection of species from environmental samples such as water, soil, and air. Recent advancements in molecular techniques and bioinformatics have led to an exponential growth in eDNA data, offering unprecedented insights into ecological dynamics and species distributions in contemporary as well as past ecosystems (Kjær et al., 2022).

Since the dawn of life, nature has repeatedly responded to the Earth's climate changes and external disturbances. Natural ecosystems have adapted to these major changes through species composition alterations and genetic adaptations. Owing to our ability to retrieve and analyze ancient eDNA over the course of time from sediment records going millions of years back in time, this information can help us understand past ecosystems in various climates to learn how contemporary ecosystems will respond to the climate crises we are facing.

An important early step in the analysis of eDNA data is species identification of each sequence read. Since the onset of high-throughput sequencing technologies, the global reservoir of sequence data is ever growing. This creates computational challenges in searching the immense databases and accurately assigning taxonomy to novel sequences. Sequence database search and taxonomic assignment are, however, central to many bioinformatics applications, from metagenomics to evolutionary studies. A plethora of approaches have been tried, such as alignment-based and k-mer-based strategies, but they continue to be a step behind the data growth (Marchet et al., 2021).

Ancient DNA (aDNA), be it from individual specimens, or from environmental samples, presents a unique set of challenges in this context (Orlando et al., 2021; Willerslev & Cooper, 2004). Unlike modern DNA, aDNA is often highly degraded and fragmented, resulting in very short sequences with high sequencing error rates. This altered base composition can distort taxonomic assignment, as it may artificially distance an ancient genome from its modern

counterpart or erroneously associate it with unrelated taxa. Additionally, ancient samples often harbor significant or even dominating microbial contamination, as microbes colonize specimens post-mortem.

Proposal

Striking a balance between taxonomic accuracy and computational efficiency remains an ongoing challenge for tools focused on sequence search and taxonomic assignment. This project proposes to work on two aspects of these challenges: scalability to large datasets, and robustness to sequencing errors. Advancements in tool efficiency will be of general interest for the eDNA community, while the robustness to sequencing errors will be a more specialized aspect for researchers working on aDNA.

Promising avenues to achieve these aims are to extend and refine existing k-mer-based approaches. Numerous advanced methods have been devised in the field, such as spaced k-mers, minimizers, and syncmers to reduce redundancy in the computation (Shaw & Yu, 2022), as well as specialized variants of Bloom filters (Seiler et al., 2021) to reduce the memory requirements. To the best of our knowledge however, no tool yet has been implemented that leverages distributed-memory systems (e.g., computer clusters), for instance via MPI, to enable database sizes that do not fit on a single compute node. Furthermore, approaches from amplicon clustering such as exact matching of sequence microvariants (Mahé et al., 2015, 2021) could be leveraged to achieve error robustness (up to two errors per k-mer) without degrading into an expensive sequence alignment problem.

Application

The application of these tools will then be proven on the datasets produced by Section for GeoGenetics that are currently too vast to be analyzed time- and cost-effectively. In particular, solutions to analyze data in the order of billions of short (ancient) eDNA reads derived from soil – a rich repository of ecological interactions and evolutionary footprints – are currently being sought. With the tools proposed here, it would finally be possible to map, align, and taxonomically assign these reads against our comprehensive database encompassing 30,000 vertebrate whole genomes, or potentially even beyond that.

The aim of the project is not only to identify the species origin of these ancient fragments, but also to position them within the intricate web of vertebrate evolutionary history. The approach promises insights into past biodiversity, ecological dynamics, and evolutionary trajectories, effectively transforming soil-derived ancient eDNA into a window to look deep into the biological past of our planet. This project, if successful, will hence allow us to integrate data from ancient genomics, ecology, and evolutionary biology, offering unprecedented depth and resolution in our understanding of ancient terrestrial ecosystems.

References

Kjær, K. H., Winther Pedersen, M., De Sanctis, B., De Cahsan, B., Korneliussen, T. S., Michelsen, C. S., Sand, K. K., Jelavić, S., Ruter, A. H., Schmidt, A. M. A., Kjeldsen, K. K., Tesakov, A. S., Snowball, I., Gosse, J. C., Alsos, I. G., Wang, Y., Dockter, C., Rasmussen, M., Jørgensen, M. E., … Willerslev, E. (2022). A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. *Nature*, *612*(7939), 283–291.

Mahé, F., Czech, L., Stamatakis, A., Quince, C., de Vargas, C., Dunthorn, M., & Rognes, T. (2021). Swarm v3: towards tera-scale amplicon clustering. *Bioinformatics* . https://doi.org/10.1093/BIOINFORMATICS/BTAB493

Mahé, F., Rognes, T., Quince, C., De Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*. https://peerj.com/articles/1420/

Marchet, C., Boucher, C., Puglisi, S. J., Medvedev, P., Salson, M., & Chikhi, R. (2021). Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Research*, *31*(1), 1–12.

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1), 1–26.

Seiler, E., Mehringer, S., Darvish, M., Turc, E., & Reinert, K. (2021). Raptor: A fast and space-efficient pre-filter for querying very large collections of nucleotide sequences. *iScience*, *24*(7), 102782.

Shaw, J., & Yu, Y. W. (2022). Theory of local k-mer selection with applications to long-read alignment. *Bioinformatics* , *38*(20), 4659–4669.

Willerslev, E., & Cooper, A. (2004). Review Paper. Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1558), 3–16.

Involvement of International Young Researchers

Within the scope of the proposed project, the main focus will be on methods and software development. Subsequent analysis of the dataset described above will involve multiple research groups at the Section for GeoGenetics and collaborating institutions, due to the scope of the data.

Hence, the aim is to employ a high-profile bioinformatics software engineer, in particular, Dr. Lucas Czech, currently a senior postdoc in the Moi Lab at the Carnegie Institution for Science in Stanford, USA. He has shown great interest in the project and agreed to join the Section for GeoGenetics. Dr. Czech will be leading the research and method development; he will also design and implement the software of the proposed tools.

Furthermore, a software developer to support the implementation of the tools will be hired with the aim of finding a strong candidate with a background in either bioinformatics or high-performance computing at the Masters, PhD, or postdoc level.

The project activities will be conducted in close collaboration with the bioinformatics team headed by Thorfinn Sand Korneliussen, a young assistant professor at the Section for GeoGenetics, University of Copenhagen, and students and postdocs in the group of Rasmus Nielsen at the University of California, Berkeley, USA. The project will further involve international collaborations, aiming in particular at working with the group of Richard Durbin at the University of Cambridge, UK.

Publications Derived from the Research

Due to the ambitious aims of the project, and the interest of the scientific community in the proposed tools, results will be published in high-profile journals. Firstly, the tools themselves will be of interest for top-tier bioinformatics journals. Secondly, there will be several downstream projects of the institute that will rely on analyzing their data with the proposed software. Due to their scope, these projects will be published in prestigious high-impact journals such as *Science* and *Nature*. Furthermore, results will be presented at international

conferences, in order to increase visibility and impact.

Finally, it is the intention to publish a book for the international general audience on the personal and scientific journey of Eske Willerslev, from the conception and discovery of environmental DNA, through the establishment and development of a new scientific field, to the nomination as scientific breakthrough of the year in 2020, and beyond.