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.

Introduction

This report outlines the progress and ongoing efforts in the Balzan Research Project of Eske Willerslev. The primary goal of this project is to develop scalable and robust methods for the taxonomic assignment of environmental DNA (eDNA), particularly ancient DNA (aDNA), which presents unique challenges due to its degraded and fragmented nature. Accurate and efficient species identification from sequence reads is crucial for understanding past ecosystems and their responses to climate changes. The project focuses on addressing the computational challenges associated with the immense databases of sequence data, aiming to improve both the speed and accuracy of taxonomic assignments.

Current Progress and Methods

As a first step, efforts have been concentrated on scaling the existing approaches to handle larger datasets. Our current aDNA mapping pipeline uses bwa mem or bowtie to map reads from ancient samples against a large reference catalog containing high-quality sequences from NCBI and other sources, covering the full taxonomic range of the NCBI database. This catalog is currently 15TB in size, making a single-step mapping infeasible. In the current setup, the reference catalog is split into around 300 subsets, termed "shards", and each sequence read in a sample is mapped against all shards to find the reference sequences to which the read maps best (and which are hence the putative species of origin of the read). However, for obvious reasons, most of these shards do not contain the actual references of interest, namely those that are good matches to the read. Mapping against the references in those shards is hence computationally wasteful – but necessary in the current pipeline, as it is not known a priori which shards are good candidates that potentially contain fitting references.

To address this, the main part of the current effort in this project is to develop a pre-filter system to assign each read to a relevant subset of shards and thus accelerate the mapping process. This filter is based on a large k-mer index built over all reference sequences, mapping k-mers to shards. Due to the scale of the reference catalog,

traditional index data structures were insufficiently scalable for this task. Instead, a novel approach using canonical k-mers as indices into a large array containing shard assignments has been developed and implemented. This array requires significant memory (1-2TB for k-mer size 20), but its size is fixed and does not grow with the reference catalog, thus ensuring future scalability. The array lookup is extremely fast, requiring only a single lookup per k-mer, and hence scales to our large sample collections.

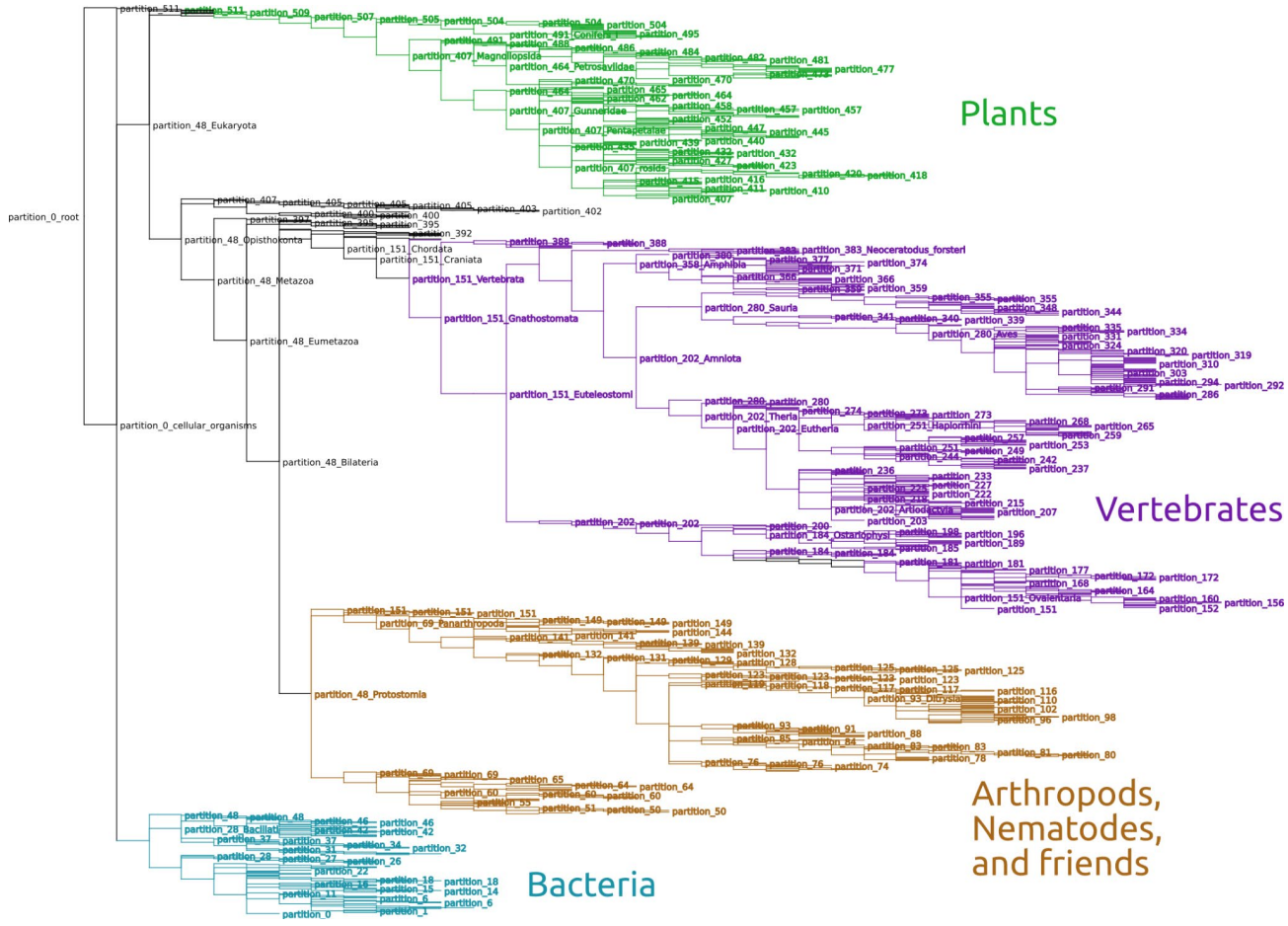

Figure 1: The NCBI taxonomy, broken down into shards. This shows the top-level clades of the taxonomy, where each tip of the taxonomic tree represents a shard (the tips represent the taxonomy groups that we use as shards). The figure is colored by recognizable clade names for visual purposes.

The shard assignment process incorporates further novel aspects, leveraging and adapting existing k-mer-based ideas. A key concept here is the use of colored k-mers, where in our application, colors represent subsets of shards a particular k-mer occurs in. While not all possible subsets can be represented, we developed a heuristic approach to construct a sufficient representative subset of colors. The lookup of a k-mer in the array thus yields its color, which identifies at least the shards the k-mer occurs in and possibly others. That is to say, we might have false positives due to the heuristic of how the set of potential colors is constructed. This leads to mapping more shards than strictly necessary - but this can easily be dealt with, as we are currently also mapping too many shards. Then, by processing all k-mers in a read, a list of matching shards that contain k-mers of the read is obtained for mapping.

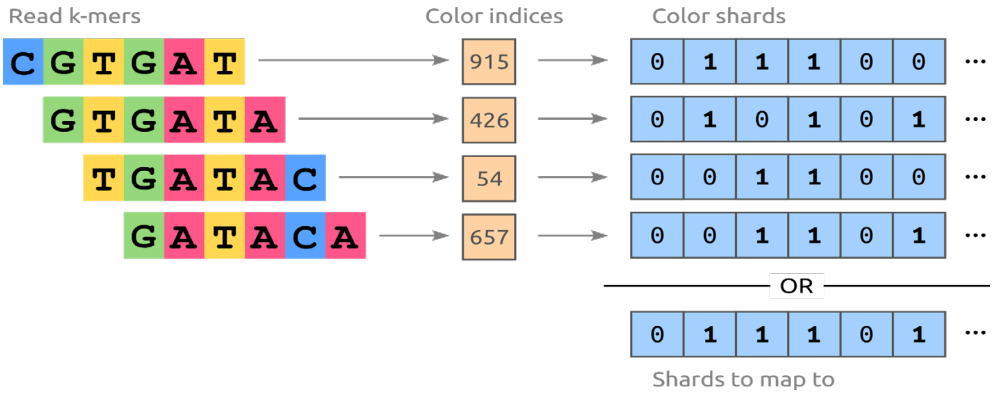

Figure 2: High-level view of the current approach. The sequence read in a sample is broken down into its k-mers. For each k-mer, a lookup is performed into the large array which contains the color indices (orange boxes) of each possible canonical k-mer. The colors correspond to subsets of shards (blue boxes), indicating if a given k-mer occurs in a shard (value 1) or not (value 0). To get all the shards that we need to map the read to, we use a logical OR operation, obtaining the set of all the shards that share k-mers with the read.

This approach has been implemented in a working prototype. Even at the prototype stage, the implementation is fast and scalable - it can process reads at the speed at which the file can be read from the disk in the first place. Hence, the first aim of the proposal (speeding up the processing) is generally well on track as far as the implementation itself is concerned. However, first results indicate that the filter is currently too broad, selecting too many shards, meaning that it is not yet saving much computation downstream. However, it does ensure that no false negatives can occur, meaning that no shard that could potentially contain matching references will be falsely omitted from mapping.

We are currently working on refining the approach by improving the heuristic to select colors and the filter of shards applied to these colors, in order to eliminate more shards for downstream mapping.

The second aim of the proposal is concerned with error and mismatch robustness. Our approach is based on k-mers and hence provides some inherent robustness to errors in the ancient DNA sequences. Any mismatching bases in the reads (which can occur due to sequencing errors, degradation of the DNA, or even genome divergence over the evolutionary timescales often involved in aDNA research) will simply lead to mismatching k-mers and hence ignored, while any matching k-mers will still be usable for matching the sequence to its shards. In the future though, as mentioned in the proposal, we also want to implement an explicit step to allow for mismatches by iterating the microvariant neighborhood of the reads. This will further allow working with severely damaged DNA which might not have sufficient k-mer matches in the current approach.

Outlook and Follow-up Grants

Current efforts focus on fine-tuning the shard assignment algorithm to improve precision and specificity. The developed pre-filter will then be integrated into the existing bioinformatics pipeline to map ancient reads against the large reference collection. It will act as a first step, determining which shards each read needs to be mapped against. While precise speedup is difficult to quantify at this stage, a five- to tenfold or greater improvement is anticipated.

Future development beyond this project will include a full aDNA mapping tool with specific alignment capabilities. We are also planning to move beyond the current Least Common Ancestor (LCA) assignment to a system that assigns confidence values for each taxon, allowing for more detailed filtering and analysis. Additionally, potential contamination of the database and reads with, e.g., bacterial or human genomic data, will be addressed to avoid skewed results. Contamination has turned out to be a large obstacle in practice and is hence a major topic for future research.

To this end, the writing of several follow-up grant proposals is in progress, aiming to build upon this project. These proposals will further develop the approach into a comprehensive aDNA mapping tool and explore more advanced methods for taxonomic assignment with confidence values.

Other Outcomes

At the current stage of active development, there are not any publications describing the results yet. However, plans are underway to have at least one major publication on the results of this research in a high-impact bioinformatics journal, as well as publications on the empirical analyses enabled by the tool. Furthermore, the prototype has been presented in a seminar series for the AEGIS project, a large-scale ancient environmental DNA project of our institute.