Summary

Computational Approaches to Ancient Genome Analysis

Kay Prüfer

Recent advancements in sequencing technologies have made it possible to sequence large quantities of the DNA preserved in ancient remains. This DNA differs from modern DNA in three key aspects. First, the DNA is fragmented into short molecules. Second, the DNA is damaged leading to misreads bases during sequencing. Last, molecules from the ancient species are mixed with molecules from microbial species. These three properties of ancient DNA pose a challenge to the analysis of the sequences gathered from ancient remains.

In this thesis, I worked on two related problems for the study of ancient DNA sequence data. In the first part, I investigated how well endogenous molecules can be identified using a selection of comparison genomes of close to distant divergences. Using simulated sequence data I also tested how sensitive the calculation of divergence estimates are to the correct identification and alignment of ancient DNA sequences, and determined filtering parameters to optimize the accurate calculation of divergence. I found that closely related comparison genomes are a prerequisite to identify a large fraction of endogenous molecules. Stable divergence estimates can be achieved through the use of two comparison genomes and appropriate filtering to ensure comparison between orthologous sequences and by excluding misread bases due to ancient DNA damage. This finding had direct relevance for the analysis of the recently published Neandertal genome sequence.

The second part of my thesis addressed a major challenge in the sequencing of ancient genomes, namely, how to reduce the amount of sequencing of non-target, microbial molecules. In order to address this challenge I investigated the microbial sequences and devised an approach to deplete libraries of these molecules. I found that motifs with a high fraction of Guanine and Cytosine and in particular "CG" dinucleotides are overrepresented in the microbial sequences. These findings lead to an experimental way of depleting microbial molecules based on restriction enzymes. I tested a database of restriction enzyme sites to identify enzymes with the highest discriminatory power between microbial and endogenous sequences. Based on this ranking, experiments were carried out by Adrian Briggs that show that specific sets of enzymes can be used to deplete microbial contamination and enrich for endogenous molecules

after sequencing. This method was applied to generate sequence data for the first draft of the Neandertal genome sequence.

For the mapping of restriction enzyme recognition motifs Udo Stenzel and I developed a specific algorithm that is described in the second part of my thesis. The algorithm is a modification of the Aho-Corasick on-line matching algorithm. Our modification allows for the search of approximate matches and can process queries with ambiguity codes that are often present in restriction enzymes. Apart from the application to restriction enzyme sites, the implementation was successfully applied by our and other groups to search for short RNA sequences and evaluate microarray probes.