Optimal DNA Sequence Assembly via Sparse String Graphs

Ilan Shomorony[1], Thomas Courtade[1], and David Tse[2]

[1]UC Berkeley, [2]Stanford University

Modern DNA sequencing technologies are based on a two-step process. First, tens or hundreds of millions of fragments from random and unknown locations of the DNA sequence are read via *shotgun sequencing*. These fragments, called reads, are then merged to each other with the goal of recovering the true underlying genome. The task of reconstructing the original sequence from a large number of short reads is known as the Assembly Problem and is one of the fundamental algorithmic problems in bioinformatics.

While the assembly problem has been studied for over twenty years, in practice, high-quality complete genome assemblies are difficult to obtain, and most available assembly tools tend to generate a large number of disconnected fragments. However, the recent development of several *long-read* sequencing technologies is promising to change this picture. By reducing the ambiguity resulting from repeated sequences, which occur frequently in most genomes, long read technologies provide the potential for the generation reference-quality *de novo* assemblies. In this context, the previous generation of assembly algorithms based on the *de Bruijn* graph framework seem unsuited to fully exploit the power of long reads. Instead, read-overlap based approaches -- and in particular string graphs -- are expected to play a central role in the next generation of assemblers, allowing near perfect assembly of whole genomes.

A fundamental challenge in performing assembly using string graphs is that the true sequence corresponds to a (generalized) Hamiltonian path on the graph. Because of that, under most formulations, the problem of achieving perfect assembly; i.e., extracting a single (correct) sequence from the graph connecting all the reads, becomes NP-hard. However, leaving the computational complexity issue aside, a more fundamental question can be asked from an information-theoretic point of view: when does the set of reads contain enough information to allow correct and unambiguous reconstruction of the true sequence?

In this work, utilizing insights from these basic informational considerations, we devise an algorithm to construct a sparse string graph that contains all information required for assembly. We describe sufficient conditions under which the assembly problem becomes an Eulerian path problem on the sparse string graph, and can thus be solved in linear time. By considering a probabilistic model for the sampling of the reads, these conditions can be translated into read length and coverage depth requirements, which are shown to nearly match information-theoretic lower bounds. We conclude that most instances of the assembly problem that are informationally feasible are also efficiently solvable.