P91

HyDA-Vista: Towards Optimal Guided Selection of *k*-mer Size for Sequence Assembly

Basir Shariat 1, Narjes Sadat Movahedi Tabrizi 2, Hamidreza Chitsaz 1 and Christina Boucher 1

1 Department of Computer Science, Colorado State University, USA

2 Department of Computer Science, Wayne State University, USA

Abstract

Motivation

Intimately tied to assembly quality is the complexity of the de Bruijn graph built by the assembler. Thus, there have been many paradigms developed to decrease the complexity of the de Bruijn graph. One obvious combinatorial paradigm for this is to allow the value of *k* to vary; having a larger value of *k* where the graph is more complex and a smaller value of *k* where the graph would likely contain fewer spurious edges and vertices. One open problem that affects the practicality of this method is how to predict the value of *k* prior to building the de Bruijn graph. We show that optimal values of *k* can be predicted prior to assembly by using the information contained in a phylogenetically-close genome and therefore, help make the use of multiple values of *k* practical for genome assembly.

Results

We present HyDA-Vista, which is a genome assembler that uses homology information to choose a value of *k* for each read prior to the de Bruijn graph construction. The chosen *k* is optimal if there are no sequencing errors and the coverage is sufficient. Fundamental to our method is the construction of the maximal sequence landscape, which is a data structure that stores for each position in the input string, the largest repeated substring containing that position. In particular, we show the maximal sequence landscape can be constructed in $O(n + n \log n)$-time and $O(n)$-space. HyDA-Vista _rst constructs the maximal sequence landscape for a homologous genome. The reads are then aligned to this reference genome, and values of *k* are assigned to each read using the maximal sequence landscape and the alignments. Eventually, all the reads are assembled by an iterative de Bruijn graph construction method. Our results and comparison to other assemblers demonstrate that HyDA-Vista achieves the best assembly of *E. coli* before repeat resolution or sca_olding.

Availability

HyDA-Vista is freely available at https://sites.google.com/site/hydavista. The code for constructing the maximal sequence landscape and choosing the optimal value of *k* for each read is also separately available on the website and could be incorporated into any genome assembler.