TECHNISCHE UNIVERSITÄT DRESDEN

CENTER FOR INFORMATION SERVICES
& HIGH PERFORMANCE COMPUTING
PROF. DR. WOLFGANG E. NAGEL

Performance Tuning & Parallelisation of Inchworm

Ankur Sharma

Dresden, March 24, 2014

Contents

1 Introduction

The high demand for low-cost sequencing has driven the development of high-throughput sequencing, which is also termed as Next generation sequencing (NGS). Thousands or millions of sequences concurrently produced in next-generation sequencing process. RNA-sequencing is a technology that uses the capabilities of next-generation sequencing to reveal the snapshot of RNA presence and quality from a genome at given moment in time. A number of assembly programs are available. Although these programs have been generally successful in assembling genomes, transcriptome assembly presents some unique challenges. Whereas high sequence coverage for a genome may indicate the presence of repetitive sequences, for a transcriptome, they may indicate abundance. Trinity is one of those assemblers that has superior quality and represents a novel method for efficient and robust de novo reconstruction reconstruction of transcriptomes from RNA-seq data. Careful performance analysis of *Trinity* software demonstrated that few of the components of the application specially *Inchworm* can be manipulated in order to achieve high performance gain. This document thus discusses the performance optimisations and parallel master slave approach of computing sequence assemblies deployed in *Inchworm* that boosted the performance to a great extent.

1.1 Biological Background

The detailed analysis of *Inchworm* involves familiarity with a lot of biological terms and techniques. This section discusses some of those key biological elements that play crucial role in better understanding of implementation and a deeper insight of how the changes in the assembly algorithm is effecting the final results. The core of *Inchworm* is in the assembly that is deployed presently using a simple but very efficient greedy algorithm. In order to extend a k-mer, the algorithm simply chooses a kmer with k-1 overlap having highest count from the catalog and continues it until the extension is possible. In further sections, some of the important terms and techniques involved in the application are discussed:

1.1.1 Key Terms

Ribo Nuceic Acid (RNA) : Ribonucleic acid (RNA) is a ubiquitous family of large biological molecules that perform multiple vital roles in the coding, decoding, regulation, and expression of genes. Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function whereby Messenger RNA molecules direct the assembly of proteins on ribosomes.

k-mer : It basically refers to n-tuple of nucleic acid and amino acid sequences that can be used to identify certain regions within biomolecules like RNA. Either k-mer strings as such can be used for finding regions of interest, or *k-mer* statistics giving discrete probability distribution of a number of possible k-mer combinations are used. These sequences (string) consist of 4 different type of elements which are A (Adenine), T (Thymine), C (Cytosine) and G (Guanine). One example of *k-mer* is a *25-mer*: ATTAGCATACCCAAGCTAGAACTTA.

Contig : A contig is a set of overlapping DNA segments that together represent a consensus region of DNA. In bottom-up sequencing projects, a contig refers to overlapping sequence data (reads); in top-down sequencing projects, contig refers to the overlapping clones that form a physical map of the genome that is used to guide sequencing and assembly. Contigs can thus refer both to overlapping DNA sequence and to overlapping physical segments (fragments) contained in clones depending on the context.

Sequence Assembly : Sequence assembly refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original sequence. This is needed as DNA sequencing

technology cannot read whole genomes in one go, but rather reads small pieces of between 20 and 30000 bases, depending on the technology used. Typically the short fragments, called reads, result from shotgun sequencing genomic DNA, or gene transcript.

1.2 Technical Background

Along with the basics of biological knowledge, we need to discuss some of the technologies involved it the analysis and implementation of *Inchworm*.