Third Semester Report

Bernáth Szabolcs (bernathsz@student.elte.hu)
Doctoral School of Physics
Statistical Physics, Biological Physics
and Physics of Quantum Systems Program

Supervisors: Dr. Imre Derényi, Dr. Gergely J. Szöllősi
Thesis title: Understanding somatic evolution and development based on genome-scale sequence data

1. Introduction

During the third semester of my PhD studies, I continued working on my main research topic, which focuses on the analysis of genome-scale sequencing data to explore somatic mutations within a single apricot tree. The work carried out during this period primarily concentrated on technical development, with the goal of improving mutation detection accuracy in complex plant genomes.

2. Description of research carried out in the current semester

Based on my earlier work, I had already developed a computational pipeline capable of identifying raw variants across a large genome. In this context, "raw" refers to the inclusion of all potential variants detected using permissive filtering criteria. While this approach ensures high sensitivity, it also produces a substantial number of false positives, making further refinement necessary.

Accurate identification of true somatic mutations remains a major challenge in genomics. During read alignment, numerous factors complicate variant detection, including misalignments, repetitive DNA regions, microsatellites, and transposable elements. These features significantly increase the difficulty of distinguishing real mutations from technical artifacts. Consequently, the main objective of this semester was to develop new methods to filter the raw variant set more effectively.

As a first step, I tested several existing tools designed to annotate repetitive regions and structural differences in the genome. These included RED, EDTA, ULTRA, and RepeatMasker for repeat detection, as well as nucmer and SyRi for identifying heterozygous regions between the two parental haplotypes. However, neither these tools individually nor their combined application provided a general solution to the problems caused by genomic duplications and heterozygosity in a diploid genome.

Transposable elements introduce a particularly challenging source of ambiguity in variant detection. Because a single point mutation typically has little to no effect on read alignment,

reads carrying the mutation can still align equally well to multiple highly similar genomic copies. As a consequence, if a transposable element is present in two copies in one parental genome but only a single copy in the other, reads originating from the mutated copy may align to both locations when mapped to the genome containing the duplicated element. In this scenario, the same mutation can be identified independently at two genomic positions, each supported by only a fraction of the original reads, despite originating from a single true mutational event.

Such problems are particularly prevalent in plant genomes, where transposable elements can constitute up to 90% of the genomic content.

To address this challenge, I developed a novel method to identify and retain only non-duplicated homozygous regions of the genome. First, I fragmented both parental genomes into 120-base-long artificial reads using a 5-base sliding window. These reads were then aligned back to both parental genomes. By grouping the reads according to the number of successful alignments, I was able to classify genomic regions into three categories:

1. Hemizygous regions, present in only one parental genome;
2. Homozygous regions, present as a single copy in both parental genomes;
3. Multiplicated regions, where at least one parental genome contains multiple copies.

This method has shown promising results so far. The next step will be to identify and remove the remaining heterozygous sites—defined as sequence mismatches between the parental genomes—a task that is considerably simpler once multiplicated regions have already been filtered out.

3. Studies in the current semester

During the semester, I completed the following classes:

- Data Mining and Machine Learning (FIZ/3/084) — graded with a 5
- Evolutionary Game Theory (FIZ/3/059E) — graded with a 5

4. Conferences

During this semester, I attended the XXX. Congress in Biophysics organized by the Hungarian Biophysical Society in Szeged, where I gave a 15-minute oral presentation on the topic of modeling stem cell behavior in trees using mutational patterns.

5. Teaching Activities

I was one of the instructors for the Modern Physics Laboratory course, teaching the ”biolfiz” measurements in 4×45-minute sessions over a period of six weeks.