

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к магистерской диссертации

**«Применение методов машинного обучения в анализе
транскриптомов растений»**

Цалапова Мария Дмитриевна

Научный руководитель – доктор физико-математических наук,
профессор Тузиков А.В.

Минск, 2020

Реферат

Магистерская диссертация, 50 страниц, 25 рисунков, 15 источников.

СЕКВЕНИРОВАНИЕ ДНК, СБОРКА ТРАНСКРИПТОМА DE NOVO, АННОТИРОВАНИЕ ТРАНСКРИПТОМА, МАШИННОЕ ОБУЧЕНИЕ, ЦЕПЬ МАРКОВА ПЯТОГО ПОРЯДКА, ФУНКЦИОНАЛЬНОЕ АННОТИРОВАНИЕ

Объект исследования – сборка и аннотирование транскриптома de novo, модели и алгоритмы машинного обучения для решения данной задачи.

Цель работы – исследование методов и моделей сборки и аннотации транскриптома, разработка пайплайна для получения готового транскриптома по данным секвенирования следующего поколения.

Методы работы – методы сборки транскриптома de novo, методы машинного обучения для аннотирования транскриптома без референсного генома, такие как цепи Маркова высших порядков, функциональное аннотирование.

Результаты – пайплайн для получения транскриптома из данных секвенирования при отсутствии референсного генома. В частности, сборка транскриптома de novo, аннотирование на уже изученные гены, и определение потенциально новых генов, отсутствующих в базах данных. Произведено сравнение результата работы пайплайна для двух образцов *P. Sylvestris*.

Область применения – агрикультурные предприятия, занимающиеся промышленным разведением древесных пород и селекцией для получения более устойчивых к холоду сортов *P. Sylvestris*.

Abstract

Master thesis, 50 pages, 25 figures, 15 sources.

DNA SEQUENCING, TRANSCRIPTOME DE NOVO ASSEMBLY, TRANSCRIPTOME ANNOTATION, MACHINE LEARNING, FIFTH ORDER MARKOV CHAIN, FUNCTIONAL ANNOTATION

The object of research – de novo assembly and annotation of transcriptome, machine learning models and methods to solve current problem.

The aim of research – to study methods and models for transcriptome assembly and annotation and to implement pipeline to get transcriptome from next generation sequencing data.

Methods of the research – methods of transcriptome de novo assembly, machine learning methods for transcriptome annotation without referencing genome, such as higher order Markov chains, functional annotation.

The results – pipeline to get transcriptome from sequencing data without reference genome. In particular, transcriptome de novo assembly, annotation of transcriptome on already studied genes, and detection of potentially new genes, that are not present in databases. Then were compared pipeline outputs for two samples of *P. Sylvestris*.

Application area – agricultural enterprises engaged in industrial cultivation of tree species and breeding of more cold resistant strains of *P. Sylvestris*.