

**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ
ТЕХНОЛОГИЙ**

Кафедра системного анализа и компьютерного моделирования

ЦВЕТОВ

Фёдор Сергеевич

**РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ КЛАСТЕРНОГО
АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ**

Аннотация (реферат) дипломной работы

**Научный руководитель:
кандидат физико-математических
наук, доцент Н.Н.Яцков**

Допущена к защите «__» июня 2024 г.

Зав. кафедрой, кандидат физико-математических наук, доцент Н.Н.Яцков

Минск, 2024

РЕФЕРАТ

Дипломная работа: 58 страниц, 26 рисунков, 8 таблиц, 25 источника, 5 приложений.

РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ КЛАСТЕРНОГО АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Ключевые слова: нуклеотидные последовательности; методы кластерного анализа; методы снижения размерности данных; имитационное моделирование; разработка R-пакетов; критерии качества анализа.

Объект исследования: методы кластерного анализа нуклеотидных последовательностей, программный пакет.

Цель работы: разработать методы и программные средства кластерного анализа нуклеотидных последовательностей.

Задачи исследования:

- Обзор литературы по теме исследования.
- Разработка и программная реализация методов имитационного моделирования (ИМ) нуклеотидных последовательностей.
- Разработка и программная реализация методов кластерного анализа нуклеотидных последовательностей.
- Разработка и программная реализация R-пакета. Интегрирование созданных методов.
- Верификация и анализ эффективности разработанных методов на смоделированных и экспериментальных данных.

Методы исследования: марковский метод имитационного моделирования, иерархический метод, метод k-средних, DBSCAN и спектральный метод кластерного анализа.

Полученные результаты: разработана имитационная модель и методы кластерного анализа нуклеотидных последовательностей, программный пакет, содержащий функционал разработанных методов. Проведена верификация и анализ эффективности разработанных методов на смоделированных и экспериментальных данных. В качестве экспериментальных данных использовались кодирующие и не кодирующие молекулы РНК. Протестирована возможность корректной установки разработанного пакета, чтения документации, использования функций.

Область возможного применения: анализ нуклеотидных последовательностей для выявления генетических мутаций и наследственных заболеваний, разработки вакцин и методов лечения различных заболеваний.

РЭФЕРАТ

Дыпломная праца: 58 старонак, 26 малюнкаў, 8 табліц, 25 крыніцы, 5 прыкладанняў.

РАСПРАЦОЎКА ПРАГРАМНЫХ СРОДКАЎ КЛАСТЭРНАГА АНАЛІЗУ НУКЛЕАТЫДНЫХ ПАСЛЯДОЎНАСЦЯЎ

Ключавыя словы: нуклеатыдныя паслядоўнасці; метады кластарнага аналізу; метады зніжэння памернасці дадзеных; імітацыйнае мадэляванне; распрацоўка R-пакетаў; крытэрыі якасці аналізу.

Аб'ект даследавання: метады кластарнага аналізу нуклеатыдных паслядоўнасцяў, праграмны пакет.

Мэта працы: распрацаваць метады і праграмныя сродкі кластарнага аналізу нуклеатыдных паслядоўнасцяў.

Задачы даследавання:

- Агляд літаратуры па тэме даследавання.
- Распрацоўка і праграмная рэалізацыя метадаў імітацыйнага мадэлявання (ІМ) нуклеатыдных паслядоўнасцяў.
- Распрацоўка і праграмная рэалізацыя метадаў кластарнага аналізу нуклеатыдных паслядоўнасцяў.
- Распрацоўка і праграмная рэалізацыя R-пакета. Інтэграванне створаных метадаў.
- Верыфікацыя і аналіз эфектыўнасці распрацаваных метадаў на змадэляваных і эксперыментальных дадзеных.

Метады даследавання: маркаўскі метад імітацыйнага мадэлявання, іерархічны метад, метад k-сярэдніх, DBSCAN і спектральны метад кластарнага аналізу.

Атрыманыя вынікі: распрацавана імітацыйная мадэль і метады кластарнага аналізу нуклеатыдных паслядоўнасцяў, праграмны пакет, які змяшчае функцыянал распрацаваных метадаў. Праведзена верыфікацыя і аналіз эфектыўнасці распрацаваных метадаў на змадэляваных і эксперыментальных дадзеных. У якасці эксперыментальных дадзеных выкарыстоўваліся кадавальныя і некадавальныя малекулы РНК. Пратэсціравана магчымасць карэктнай устаноўкі распрацаванага пакета, чытання дакументацыі, выкарыстання функцый.

Вобласць магчымага прымянення: аналіз нуклеатыдных паслядоўнасцей для выяўлення генетычных мутацый і спадчынных захворванняў, распрацоўкі вакцын і метадаў лячэння розных захворванняў.

ABSTRACT

Diploma thesis: 58 pages, 26 figures, 8 tables, 25 sources, 5 appendices.

DEVELOPMENT OF SOFTWARE TOOLS FOR CLUSTERING ANALYSIS OF NUCLEOTIDE SEQUENCES

Key words: nucleotide sequences; cluster analysis methods; dimensionality reducing methods; imitation modeling; development of R-packages; quality criteria analysis.

Object of work: methods of cluster analysis of nucleotide sequences, software package.

Objective: to develop methods and software for cluster analysis of nucleotide sequences.

Research objectives:

- Literature review on the research topic.
- Development and software implementation of methods for imitation modeling (IM) of nucleotide sequences.
- Development and software implementation of methods for clustering analysis of nucleotide sequences.
- Development and software implementation of an R package. Integration of created methods.
- Verification and analysis of the effectiveness of the developed methods using simulated and experimental data.

Research methods: Markov simulation method, hierarchical, *k*-means, DBSCAN and spectral clustering methods.

Results obtained: imitation model and methods for clustering analysis of nucleotide sequences, a software package containing the functionality of the developed methods, have been developed. The effectiveness of the developed methods was verified and analyzed using simulated and experimental data. Experimental data included encoding and non-coding RNA molecules. The ability of correctly install of the developed package, reading documentation, and using functions was tested.

Area of possible application: analysis of nucleotide sequences to identify genetic mutations and hereditary diseases, development of vaccines and methods of treating various diseases.

}