МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Кафедра системного анализа и компьютерного моделирования

ФИЛИПЧИК
Алексей Олегович

РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ ДЛЯ ПРЕДСКАЗАНИЯ СОБЫТИЙ АЛЬЕТЕРНАТИВНОГО СПЛАЙСИНГА ОНКОГЕНОВ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Аннотация (реферат) к дипломной работе

Научный руководитель:
доцент кафедры системного анализа и компьютерного моделирования, к. ф.-м. наук,
Яцков Николай Николаевич

Допущен к защите
«___» _____ 2025г.
Зав. кафедрой системного анализа и компьютерного моделирования
кандидат физ.-мат.наук, доцент В.В. Скакун

Минск, 2025

РЕФЕРАТ

Дипломная работа: 55 страницы, 15 рисунков, 4 таблицы, 37 источников, 2 приложения.

Ключевые слова: альтернативный сплайсинг, РНК, экзоны, транскрипты, RNA-seq, нейронные сети, LSTM, CNN, метод главных компонент, иерархическая кластеризация, онкоген RUNX1/RUNX1T1, биоинформатика.

Объект исследования: транскрипты РНК или события альетернативного сплайсинга онкогенов, полученные с использованием высокопроизводительного секвенирования (RNA-seq), включающие экзоны и транскрипты, а также химерный онкоген RUNX1/RUNX1T1, связанный с онкогенезом.

Предмет исследования: Алгоритмы и модели интеллектуального анализа данных для предсказания альтернативных транскриптов РНК онкогенов человека.

Цель работы: Разработка и реализация программного средства на основе алгоритмов интеллектуального анализа данных, включая нейронные сети, для предсказания событий альтернативного сплайсинга онкогенов, таких как RUNX1/RUNX1T1, с целью улучшения диагностики онкологических заболеваний.

Методы исследования:

- Метод главных компонент для снижения размерности признакового пространства экзонов.
- Иерархическая агломеративная кластеризация с использованием метода Уорда и евклидовой метрики.
- Сравнение транскриптов с применением расстояния Левенштейна и коэффициента Жаккара.
- Моделирование последовательностей транскриптов с использованием рекуррентных (LSTM) и сверточных (CNN) нейронных сетей.

Результаты: Разработано программное средство для анализа данных RNA-seq, интегрирующее методы кластеризации и нейронные сети. Проведен вычислительный эксперимент на модельных генах, включая анализ 10 пар генов и химерного онкогена RUNX1/RUNX1T1. Достигнута высокая точность классификации транскриптов (92–98% для классического метода, 90–96% для нейронного подхода; метрики нейронной модели: accuracy=0.9904, F1=0.8800, ROC AUC=0.9751). Определены 10 наиболее вероятных транскриптов RUNX1/RUNX1T1, связанных с онкогенезом, с коэффициентом Жаккара 0.92–1.00. Результаты подтверждают применимость разработанных алгоритмов для предсказания событий альтернативного сплайсинга и их потенциал для персонализированной диагностики онкологических заболеваний.

РЭФЕРАТ

Дыпломная работа: 55 старонкі, 15 ілюстрацый, 4 табліцы, 37 крыніц, 2 пракладання..

Ключавыя словы: альтэрнатыўны сплайсінг, РНК, экзоны, транскрыпты, RNA-seq, нейронныя сеткі, LSTM, CNN, метад галоўных кампанентаў, іерархічная класіфікацыя, онкаген RUNX1/RUNX1T1, біяінфарматыка.

Аб’ект даследавання: Транскрыпты РНК або альтэрнатыўным сплайсінгу онкагенаў, атрыманыя з дапамогай высокапрадукцыйнага секвеніравання (RNA-seq), уключаючы экзоны і транскрыпты, а таксама хімэрны онкаген RUNX1/RUNX1T1, звязаны з онкагенезам.

Прадмет даследавання: Алгарытмы і мадэлі інтэлектуальнага аналізу даных для прагназавання альтэрнатыўных транскрыптаў РНК онкагенаў чалавека.

Мэта працы: Распрацоўка і рэалізацыя праграмнага сродку на аснове алгарытмаў інтэлектуальнага аналізу даных, уключаючы нейронныя сеткі, для прагназавання падзей альтэрнатыўнага сплайсінгу онкагенаў, такіх як RUNX1/RUNX1T1, з мэтай паляпшэння дыягностыкі анкалагічных захворванняў.

Метады даследавання:

- Метад галоўных кампанентаў для зніжэння памернасці прасторы прыкмет экзонаў.
- Іерархічная агламератыўная класіфікацыя з выкарыстаннем метаду Уорда і эўклідавай метрыкі.
- Параўнанне транскрыптаў з выкарыстаннем адлегласці Левенштэйна і каэфіцыента Жакара.
- Мадэляванне паслядоўнасцей транскрыптаў з дапамогай рэкурэнтных (LSTM) і згорткавых (CNN) нейронных сетак.

Вынікі: Распрацаваны праграмны сродак для аналізу даных RNA-seq, які інтэгруе метады класіфікацыі і нейронныя сеткі. Праведзены вылічальны эксперымент на мадэльных генах, уключаючы аналіз 10 пар генаў і хімэрнага онкагена RUNX1/RUNX1T1. Дасягнута высокая дакладнасць класіфікацыі транскрыптаў (92–98% для класічнага метаду, 90–96% для нейроннага падыходу; метрыкі нейроннай мадэлі: дакладнасць=0.9904, F1=0.8800, ROC AUC=0.9751). Вызначаны 10 найбольш верагодных транскрыптаў RUNX1/RUNX1T1, звязаных з онкагенезам, з каэфіцыентам Жакара 0.92–1.00 [7]. Вынікі пацвярджаюць прымянімасць распрацаваных алгарытмаў для прагназавання падзей альтэрнатыўнага сплайсінгу і іх патэнцыял для персаналізаванай дыягностыкі анкалагічных захворванняў.

STRUCTURAL ABSTRACT

Thesis: 55 pages, 15 illustrations, 4 tables, 37 sources, 2 appendices.

Keywords: alternative splicing, RNA, exons, transcripts, RNA-seq, neural networks, LSTM, CNN, principal component analysis, hierarchical clustering, oncogene RUNX1/RUNX1T1, bioinformatics.

Object of Study: Experimental transcriptomic data obtained via high-throughput sequencing (RNA-seq), including exons and transcripts involved in alternative splicing, and the chimeric oncogene RUNX1/RUNX1T1 associated with oncogenesis.

Subject of Study: Algorithms and machine learning models for predicting alternative RNA transcripts of human oncogenes.

Objective: To develop and implement a software tool based on machine learning algorithms, including neural networks, for predicting alternative splicing events of oncogenes, such as RUNX1/RUNX1T1, to enhance cancer diagnostics.

Methods:

- Principal component analysis for dimensionality reduction of exon feature space.
- Hierarchical agglomerative clustering using Ward's method and Euclidean distance.
- Transcript comparison via Levenshtein distance and Jaccard coefficient.
- Sequence modeling with long short-term memory (LSTM) and convolutional neural networks (CNN).

Results: A software tool integrating clustering and neural network methods was developed for RNA-seq data analysis. A computational experiment was conducted on model genes, including 10 gene pairs and the chimeric oncogene RUNX1/RUNX1T1. High classification accuracy was achieved (92–98% for classical methods, 90–96% for neural approaches; neural model metrics: accuracy=0.9904, F1=0.8800, ROC AUC=0.9751). Ten highly probable RUNX1/RUNX1T1 transcripts linked to oncogenesis were identified, with Jaccard coefficients of 0.92–1.00 [7]. The results demonstrate the applicability of the developed algorithms for predicting alternative splicing events and their potential for personalized cancer diagnostics.