

## ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ПРЕДСКАЗАНИЯ И АННОТАЦИИ ОТКРЫТЫХ РАМОК СЧИТЫВАНИЯ В РНК-ТРАНСКРИПТАХ ЧЕЛОВЕКА

В. В. Скакун<sup>1</sup>, Н. Н. Яцков<sup>1</sup>, П. В. Назаров<sup>2</sup>,  
М. К. Чепелева<sup>1</sup>, В. В. Гринев<sup>1</sup>

<sup>1</sup> *Белорусский государственный университет, Минск, Беларусь,*

<sup>2</sup> *Luxembourg Institute of Health, Strassen, Luxembourg*

*e-mail: skakun@bsu.by*

Разработано веб-приложение ORFhunteR для автоматического определения и биологической аннотации открытых рамок считывания в большом наборе молекул РНК человека. Интерактивный веб-интерфейс позволяет вызывать функции R пакета ORFhunteR и предназначен для удаленной работы в режиме онлайн через глобальную сеть Internet. Разработка может быть использована в фундаментальных и прикладных биомедицинских исследованиях, связанных с изучением транскриптома нормальных и измененных клеток человека.

Ключевые слова: *РНК, открытая рамка считывания, аннотация открытых рамок считывания, программный пакет, веб-интерфейс.*

Методы геномного секвенирования позволяют устанавливать нуклеотидные последовательности молекул РНК и определять их количественное содержание в клетке [1]. Кодировующий потенциал молекул РНК может быть оценен с помощью алгоритмов и программных средств для определения открытых рамок считывания (ОРС) [2], среди которых следует отметить вычислительный подход на основе алгоритмов векторизации и случайного леса [3, 4]. Однако интерактивные, доступные для широкого круга пользователей программные приложения, реализующие разработанные алгоритмы определения ОРС больших наборов молекул РНК, практически отсутствуют или имеют ряд ограничений, главным образом связанных с невысокой точностью определения ОРС и не предоставлением средств автоматизации.

Для решения задачи автоматизации определения ОРС на языке R разработан программный пакет ORFhunteR, интегрирующий алгоритмы поиска и биологической аннотации найденных ОРС [4-7]. Пакет ORFhunteR размещен в открытом доступе в системе контроля версий GitHub, доступен по адресу <https://github.com/rfctbio-bsu/ORFhunteR>, и опубликован в международной рецензируемой базе научных разработок по биоинформатике – платформе Bioconductor: <http://bioconductor.org>. На рисунке 1 представлена функциональная схема пакета. Функция predictORFs() предназначена для предсказания ОРС, функции getSeqORFs(), findPTCs(), translateORFs() и annotateORFs() – для аннота-

ции OPC. Для корректной работы пакета требуется наличие файла `codon.table.txt`, содержащий расшифровку преобразований кодонов и файла `BSgenome`, содержащего данные полногеномных последовательностей. По умолчанию для аннотации OPC в молекулах РНК человека используется пакет «`BSgenome.Hsapiens.UCSC.hg38`».

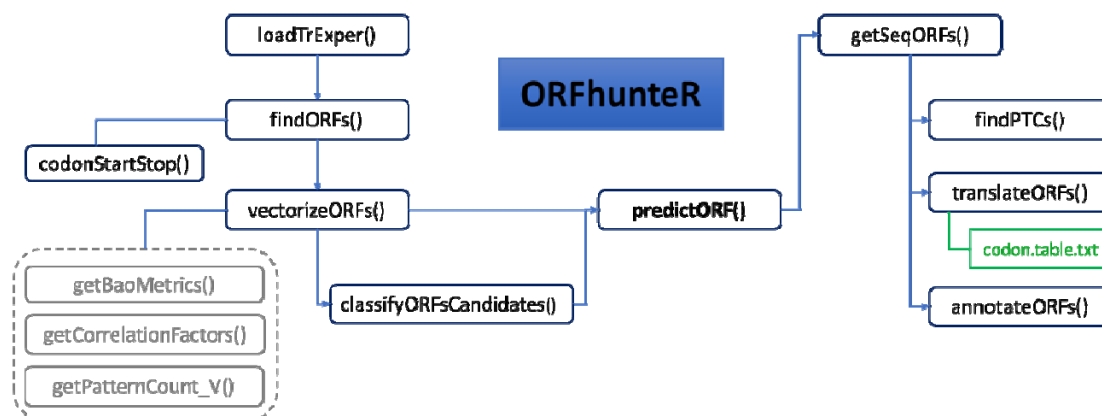


Рис. 1. Функциональная схема пакета ORFhunterR

Использование программного пакета ORFhunterR предполагает знание языка R и основ программирования, что существенно сужает круг пользователей пакета. Для предоставления широкого удаленного доступа к вычислительным возможностям пакета создано веб-приложение ORFhunter ([orfhunter.bsu.by](http://orfhunter.bsu.by)) [5]. В первой версии веб-приложения был реализован только функционал для предсказания OPC (с помощью запуска функции `findORFs()`) [5]. Другие возможности пакета, включая аннотацию полученных рамок считывания, остались нереализованными, что существенно уменьшает привлекательность веб-приложения.

Целью настоящей работы является интеграция возможностей R пакета ORFhunter по автоматической аннотации предсказанных OPC в большом наборе молекул РНК в разработанное веб-приложение для обеспечения полноценной удаленной работы научного сообщества с разработанным пакетом в режиме онлайн через глобальную сеть Internet.

Для создания веб-интерфейса выбрана технология Microsoft ASP.NET Core 2.1. Программирование веб-интерфейса выполнено с использованием языков C#, Razor и JavaScript в среде Visual Studio 2019 по шаблону модель-представление-контроллер (MVC). Для организации асинхронного обмена данными в режиме реального времени между клиентом (веб-браузер) и веб-сервером (без необходимости постоянного обновления веб-страниц) выбрана технология ASP.NET SignalR. Библиотека для разработчиков ASP.NET реализует процесс обмена данными в реальном времени в веб-приложениях. Серверная часть кода SignalR запрограммирована в виде хабов (Hub) – классов, наследуемых от класса Hub. Клиентская часть представляет собой подключаемую библиотеку и код

JavaScript, в котором реализованы процессы инициализации соединения с сервером и запуска обмена данными. Процесс интерактивного взаимодействия со средой R реализован с помощью библиотеки взаимодействия R.Net. Библиотека предназначена для быстрого обмена данными в одном и том же процессе между экземпляром R, выступающим в роли сервера, и клиентом. R.Net предоставляет возможность вызова как одной операции R, так и скрипта, состоящего из многих сотен строк в режиме интерактивного взаимодействия со средой исполнения R. Инициализация соединения с сервером и запуск процесса вычислений выполняется в коде JavaScript на стороне клиента. Хаб реализует серверную логику SignalR, инициализирует процесс запуска вычислений в среде R и возвращает клиенту результаты вычислений. Каждое из сообщений, отправленное сервером клиенту, конвертируется в HTML теги и добавляется в режиме реального времени в структуру DOM веб-страницы. Веб-браузер автоматически перерисовывает изменяющуюся часть веб-страницы.

Созданная в рамках данной работы страница аннотации предсказанных ОРС представлена на рисунке 2. Слева расположена форма загрузки GTF файла и выбора параметров аннотирования. Справа расположен набор вкладок, отображающих результаты аннотирования (для первых 100 транскриптов). Во вкладке «ORF sequences» размещена таблица с последовательностями ОРС, являющаяся результатом работы функции пакета `getSeqORFs()`. На вкладке «PTCs» отображается таблица с определёнными преждевременными стоп-кодонами (`findPTCs()`). Во вкладке «In silico translation» выводятся последовательности белков, транслируемых найденными ОРС (`translateORFs()`). В крайней правой вкладке расположены результаты базовой аннотации предсказанных ОРС (`annotateORFs()`): 5'UTR length, 3'UTR length, start codon, stop codon, ORF start, ORF stop, ORF length, MW, pI, index PPI. Результаты аннотации автоматически сохраняются в файлы для скачивания.

Таким образом, в результате проделанной работы произведена интеграция возможностей пакета ORFhunteR по определению и биологической аннотации ОРС в больших наборах молекул РНК человека в одноименное веб-приложение. Приложение реализует полноценную методику анализа и интерпретации кодирующей информации нуклеотидных последовательностей молекул РНК. Веб-приложение реализует асинхронный быстрый обмен данными между сервером и веб-браузером с помощью технологии SignalR, что позволяет взаимодействовать со средой исполнения R и наблюдать за получением результатов без необходимости постоянной перерисовки всей страницы. Работоспособность разработанного веб-интерфейса проверена на примере анализа файлов молекул РНК из базы данных NCBI RefSeq.

ORFhunterR About us Contact us

## Annotation

To annotate identified Open Reading Frames please upload GTF file then click the button Annotate

**Settings:**

Load GTF (.fasta file) ...  
 Файл не выбран.

File Set trans\_sequences.gtf has been successfully uploaded

**Progress of annotation:**

Initialization of engine...  
 Waiting for resources to be released...  
 Getting ORF sequences is in progress...  
 Getting ORF sequences is complete  
 Getting premature termination codons is in progress...  
 Getting premature termination codons is complete.  
 Getting protein sequences is in progress...  
 Getting protein sequences is complete  
 Basic annotation is in progress...  
 Basic annotation is complete.

Annotation is finished  
 Total execution time: 25.37 s

ORF sequences   PTCs   In silico translation   Basic annotation

**ORF sequences:**

RNA transcript ID	ORF sequence	Width
NM_001005221.2	ATGGATGGAGAGAAT ... AAGAGGATCTCATAA	939
NM_001005277.1	ATGGATGGAGAGAAT ... AAGAGGATCTCATAA	939
NM_001005484.1	ATGGTGAAGTAAATTC ... AGTAAAGTTTAA	918
NM_001160184.1	ATGGGGAACAGCCAC ... GTGCAGTGGATCTGA	1731
NM_001291366.1	ATGGAAAATTTCCAG ... AGCCCGAGCTCCTAG	2373
NM_001291367.1	ATGCCGACCCAGGAC ... AGCCCGAGCTCCTAG	2091
NM_015658.3	ATGGAGCACTGGGGA ... CTTCGGCTCTCTGA	888
NM_021170.3	ATGGGACAGTTGGCA ... CAGCCTCCCAAGTAG	699
NM_032429.2	ATGGGAAACAGCCAC ... GTGCAGTGGATCTGA	1836
NM_152486.2	ATGCCAAAGGGGATC ... CAGCCTCTGTGTGA	2046
NM_198317.2	ATGCAGCCCGGACG ... TCCACCAAGCCTCTGA	1929
XM_005244771.4	ATGGCCGAGACAGC ... CCGTGGCTGCGCTGA	570
XM_006710600.3	ATGCAGCCCGGACG ... TCCACCAAGCCTCTGA	1988
XM_006710601.3	ATGCAGCCCGGACG ... CAGGCGAGTCTTGA	1392
XM_006710944.3	ATGTGTGGGCACGTT ... GTGCAGTGGATCTGA	2196
XM_011542248.2	ATGTGTGGGCACGTT ... GTGCAGTGGATCTGA	2316
XM_011542538.1	ATGCGTAGACACACA ... CGTCACAGGCGCTGA	2253
XM_017002408.1	ATGGATGGAGAGAAT ... AAGAGGATCTCATAA	939
XM_017002409.2	ATGGATGGAGAGAAT ... AAGAGGATCTCATAA	939
XM_017002410.1	ATGGATGGAGAGAAT ... AAGAGGATCTCATAA	939

**About ORFhunterR**

- [ORFhunterR user manual](#)
- [License](#)
- [About us](#)
- [Contact us](#)

**Useful links:**

- [NCBI](#)
- [NCBI Reference Sequence Database](#)
- [Ensembl](#)
- [NCBI ORF Finder](#)

© ORFhunterR - 2020-2021, Belarusian State University, Luxemburg Institute of Health

Рис. 2. Окно аннотации ОРС веб-приложения ORFHunterR

### БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Reuter J. A., Spacek D. V., Snyder M. P. High-throughput sequencing technologies // Mol. Cell., 2015, 58, P. 586-597.
2. Woodcroft B. J., Boyd J. A., Tyson G. W. OrfM: a fast open reading frame predictor for metagenomic data // Bioinformatics. 2016. V. 32 N. 17. P. 2702-2703.
3. Закирова В. Р., Сырокваш Д. А., Гилевский С. В., Назаров П. В., Яцков Н. Н. Разработка алгоритмов и программных средств классификации кодирующих и некодирующих нуклеотидных последовательностей // Информатика. 2019. Т. 16, № 2. С. 111–120.
4. Яцков Н. Н., Скакун В. В., Гринев В. В. Разработка вычислительного подхода для автоматического определения открытых рамок считывания в молекулах РНК человека // Квантовая электроника: Матер. XII Междунар. конф. Мн.: БГУ. 2019, С. 279–281.
5. Скакун В. В., Яцков Н. Н., Назаров П. В., Гринев В. В.. Программный пакет ORFhunterR для автоматического определения открытых рамок считывания в молекулах РНК человека // Компьютерные технологии и анализ данных (СТДА'2020) : матер. II Междунар. науч.-практ. конф. – Минск : БГУ, 2020. С. 20-24.
6. Grinev V. V., Yatskou M. M., Skakun V. V., Chepeleva M. K., Nazarov P. V. ORFhunterR: an accurate approach for the automatic identification and annotation of open reading frames in human mRNA molecules // bioRxiv. 2021. P. 1–13.
7. Grinev V. V., Yatskou M. M., Skakun V. V., Chepeleva M.K, Nazarov P. V. ORFhunterR: An accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules // Software Impacts. 2022. N. 100268.