Математическая биология и биоинформатика
2020. Т. 15. № 2. С. 455–470. doi: 10.17537/2020.15.455

========== **БИОИНФОРМАТИКА**==========

Технология структурирования и обработки транскриптомных данных на основе гибридного использования RDBMS и NoSQL подходов

Мухин А.М.[1,2], Генаев М.А.[1,2,3], Рассказов Д.А.[1,2], Лашин С.А.[1,2,3], Афонников Д.А.[1,2,3]

[1]*Федеральный исследовательский центр Институт цитологии и генетики СО РАН, г. Новосибирск, Россия*
[2]*Курчатовский геномный центр, Институт цитологии и генетики Сибирского отделения Российской академии наук, г. Новосибирск, Россия*
[3]*Новосибирский государственный университет, г. Новосибирск, Россия*

Аннотация. Эксперимент по секвенированию транскриптома (RNA-seq) стал практически рутинной процедурой для изучения как модельных организмов, так и для сельскохозяйственных культур. В результате биоинформатической обработки таких экспериментов получаются объемные разнородные данные, представленные нуклеотидными последовательностями транскриптов, аминокислотными последовательностями и их структурно-функциональной аннотацией. Полученные данные важно представить широкому кругу исследователей в виде баз данных (БД). В работе предложен гибридный подход к созданию молекулярно-генетических баз данных, которые содержат информацию о последовательностях транскриптов и их структурно-функциональной аннотации. Сущность подхода в одновременном хранении в БД информации как структурированного типа, так и слабо структурированных данных. Технология использована для реализации БД транскриптомов сельскохозяйственных растений. В работе рассматриваются особенности реализации такого подхода и примеры формирования как простых, так и сложных запросов к такой базе данных на языке SQL. База данных OORT реализована для пяти сельскохозяйственных растений, она находится в свободном доступе по адресу: https://oort.cytogen.ru/.

Ключевые слова: *база данных, индексация, запросы, растения, SQL, RDBMS, NoSQL, транскриптомы, сельскохозяйственные культуры*.

ВВЕДЕНИЕ

Изучение транскриптомов растений с помощью высокопроизводительного секвенирования (секвенирование РНК, RNA-seq) широко используется в настоящее время для решения таких задач как оценка экспрессии генов для разных генотипов и в разных условиях среды, идентификация последовательностей РНК (для не модельных организмов), поиск маркеров к функционально важным генам [1, 2]. Эксперимент RNA-seq стал практически рутинной процедурой для изучения как модельных организмов (*Arabidopsis thaliana*) [3], так и для сельскохозяйственных культур (томат, кукуруза, ячмень, пшеница и др.) [4]. Результаты транскриптомного эксперимента представляют собой короткие фрагменты нуклеотидных последовательностей и лишь биоинформатическая обработка, включающая несколько стадий [5–8], позволяет получить на их основе последовательности транскриптов и их функциональную аннотацию. Именно результаты биоинформатической обработки представляют интерес

для биолога и могут быть интерпретированы в терминах функций генов, их продуктов, уровней экспрессии, генетических вариаций и т.п. [9, 10]. Необходимо отметить, что в результате транскриптомного эксперимента получается большое количество данных (десятки и сотни тысяч последовательностей), свободный и удобный доступ к которым важно предоставить широкому кругу биологов, далеких от рутинной биоинформатической обработки результатов секвенирования. Этой цели служат базы данных, имеющие удобный пользовательский интерфейс и организующие связи между биологическими последовательностями и их функциональной аннотацией. Среди таких баз можно указать Expression Atlas Европейского института биоинформатики [11], EGENES, базу данных информации о метаболических путях генов, основанную на транскриптомных данных [12], базы данных по экспрессии генов для определенных видов организмов: TodoFirgene для пихты *Abies sachalinensis* [13]; атлас экспрессии генов для розы [14]; базу аннотированных транскриптомов приморской сосны EuroPineDB [15] и др.

Результаты обработки RNA-seq экспериментов, представленные в таких базах данных, являются комплексными и включают: последовательности транскриптов, их локализацию в геноме, классификацию по типам генов (мРНК, днкРНК, миРНК, тРНК и пр.), функциональную аннотацию транскриптов, оценку уровней экспрессии, оценку вариантов изоформ транскриптов. Эти результаты представлены в виде бинарных и текстовых файлов в различных форматах. Это могут быть файлы последовательностей (форматы FASTA, FASTQ), выравниваний (форматы BAM, SAM, PSL и т.д.) или разметки (BED, GFF, GTF) [16–19]. Результаты анализа дифференциальной экспрессии генов обычно представляют в виде таблиц (форматы TSV, XLS) [20, 21]. Подобные хорошо структурированные данные удобно описывать в виде классической реляционной модели отношений RDBMS (Реляционная система управления базами данных, англ. Relational DataBase Management System): например, у одного гена может быть несколько изоформ, в эксперименте рассматриваются несколько образцов ткани разных различных особей и т.д.

Отметим, однако, что наряду с хорошо структурированными данными, в результате анализа RNA-seq экспериментов генерируются слабо структурированные и неструктурированные данные, которые не могут быть описаны с помощью реляционной модели. Сложности могут возникать в силу разнородности данных, получаемых в процессе выполнения биоинформатических вычислительных конвейеров. Эта разнородность обусловлена разнообразием методов, которые вовлечены в конвейеры биоинформатической обработки транскриптомных данных [7, 22]. Узлами в конвейерах являются различные программы, которые реализуют методы обработки данных. На практике обычно бывает так, что в существующий вычислительный конвейер для решения какой-то задачи в процессе работы вносятся изменения: например, происходит замена некоторых узлов на новые, которые реализуют более точные или более производительные методы обработки данных. Поэтому заранее невозможно полностью декларировать структуру данных, которая будет получена в результате их обработки. Описание таких слабоструктурированных данных удобнее делать с использованием технологий NoSQL (не только SQL, англ. not only SQL) [23, 24].

В настоящей работе мы предлагаем комплексный подход к описанию транскриптомных данных, который заключается в использовании элементов RDBMS при описании хорошо структурированных данных и NoSQL для описания слабо структурированных данных, полученных в результате широкомасштабного биоинформатического анализа транскриптомных экспериментов у пяти сельскохозяйственных растений (кукурузы, риса, ячменя, томата и картофеля). Анализ этих данных был направлен на идентификацию новых транскриптов, которые либо не

выравниваются на референсный геном растения, либо выравниваются на его неаннотированные участки и, таким образом, представляют собой новую, ранее неисследованную часть транскриптома. На примере задачи массового анализа транскриптомов сельскохозяйственных растений мы предлагаем наборы реляционных отношений для описания основных сущностей: исследование, эксперимент, нуклеотидные и белковые последовательности. В то же время, для каждой из этих сущностей мы предлагаем вводить возможность для аннотации наборами слабоструктурированных данных, формат представления которых может быть заранее неизвестен. На основе предложенного гибридного подхода разработана база данных OORT (Out Of Reference Transcripts), которая позволяет пользователям, с помощью поисковых запросов, извлекать информацию о структуре и функциях ранее неаннотированной части транскриптомов сельскохозяйственных растений, в частности: идентифицировать новые гены устойчивости к заболеваниям и абиотическому стрессу, длинные некодирующие РНК, последовательности мРНК, получать оценки уровня экспрессии этих транскриптов. База данных построена на основе анализа 1241 транскриптомных экспериментов и содержит информацию о 20440228 нуклеотидных и 4055996 аминокислотных аннотированных последовательностях. Функциональные возможности базы данных OORT демонстрируются на примере нескольких запросов.

МАТЕРИАЛЫ И МЕТОДЫ

Исходные данные

В качестве исходных данных использованы архивы SRA (Sequence Read Archive), которые хранят «сырые» данные секвенирования транскриптомных библиотек. Архивы загружены с сайта ENA [25, 26]. Каждый архив в базе данных ENA сопровождается метаинформацией, структура которой включает описание: идентификатор биологического проекта, в рамках которого проводилось секвенирование (BioProject) идентификатор исследования (study), к которому относится библиотека, образец, для которого получено секвенирование транскриптома. Метаданные для исследования содержат также его краткое описание, список публикаций в которых результаты транскриптомного эксперимента были опубликованы, данные об исследуемых образцах (BioSample, sample), например, вид, пол, ткань или орган, геометка и т.д. Из каждого образца может быть получено несколько препаратов для секвенирования (experiment), в метаданных эксперимента описан метод выделения РНК, экспериментальная платформа секвенирования. Эксперимент включает один или несколько SRA архивов, каждый из которых соответствует набору прочтений, полученных в результате секвенирования образца на конкретном секвенаторе. Каждый запуск секвенатора соответствует в терминах NCBI SRA одному файлу архива. Детальное описание структуры отношений между указанными уровнями описания данных по секвенированию РНК представлено на сайте NCBI [27].

При формировании базы данных OORT рассматривались эксперименты для пяти видов сельскохозяйственных растений: *Hordeum vulgare* (ячмень, taxonomy ID 4513); *Oryza sativa* (рис, taxonomy ID 4530); *Solanum lycopersicum* (томат, taxonomy ID 4081); *Solanum tuberosum* (картофель, taxonomy ID: 4113); *Zea mays* (кукуруза, taxonomy ID 4577). Для анализа были отобраны SRA архивы с библиотеками RNA-seq, которые соответствовали следующим критериям: платформа секвенирования Illumina HiSeq 2000 или Illumina HiSeq 2500; данные полученные методами PolyA, Inverse rRNA или cDNA; длина прочтений библиотеки не менее 75. В итоге, запрос для получения списка данных из базы SRA формулировался следующим образом:

"instrument_platform == 'ILLUMINA' & library_strategy == 'RNA-Seq' & library_source == 'TRANSCRIPTOMIC' & (instrument_model=='Illumina HiSeq 2000' | instrument_model=='Illumina HiSeq 2500') & sra_has_ftp == True & mean_read_len>=75 &

Математическая биология и биоинформатика. 2020. Т. 15. № 2. doi: 10.17537/2020.15.455

(library_selection == 'cDNA' | library_selection == 'PolyA' | library_selection == 'Inverse rRNA')".

Данному запросу удовлетворяло 3883 SRA архива, 200 исследований, 3419 образцов и 3578 экспериментов. Из полученного списка файлов мы отобрали данные, на которые есть ссылки в публикациях. В результате мы получили список из 69 исследований. Эти данные включают 1395 SRA файла общим размером 695 ГБ.

Конвейер биоинформатической обработки транскриптомных данных БД OORT

Первый этап формирования базы данных OORT заключался в *de novo* сборке последовательностей транскриптов из сырых прочтений. Для этого был разработан конвейер, который состоит из четырех последовательных шагов: (1) извлечение прочтений из SRA файла с помощью пакета SRA Toolkit [28]; (2) подготовка данных (сырые данные были подвергнуты фильтрации при помощи программы fastp [29]); (3) сборка, с использованием программы Trinity-v2.6.6. [5]; (4) оценка уровня экспрессии транскриптов с помощью программы Kallisto [21]. В качестве количественной меры экспрессии транскриптов мы использовали Transcripts Per Million (TPM).

Всего нами было обработано 1298 SRA файлов – библиотек коротких прочтений. Для классификации транскриптов в собранных нами транскриптомах мы использовали программу rnaQUAST v. 1.5 [30]. С использованием выравнивания транскриптов на референсный геном и известную его аннотацию, программа rnaQUAST классифицирует все транскрипты на 5 групп: (1) Unaligned – невыравненные транскрипты на геном; (2) Multiply aligned – транскрипты, которые выравниваются на два и более участка референсного генома; (3) Misassembled – транскрипты, выравнивание, которых имеет разногласия с аннотацией; (4) Uniquely aligned – транскрипты, имеющие ровно одно выравнивание на референсный геном, которые при этом не содержат разногласий с аннотацией; (5) Unannotated – транскрипты, которые выровнены на референсный геном на его неаннотированные участки. Для выравнивания транскриптов на референсный геном в программе rnaQUAST использовался пакет gmap v. 2018-07-04 [31]. Мы использовали последовательности и аннотации референсных геномов пяти растений, загруженные с сайта Ensembl Plants [32,33]. Используемые версии сборок геномов и аннотаций представлены в таблице 1.

Таблица 1. Используемые версии сборок геномов и аннотаций.

Название организма	Версия сборки	Идентификатор сборки и аннотации
Ячмень	v. 42	Hordeum_vulgare.IBSC_v2.42
Рис	v. 40	Oryza_sativa.IRGSP-1.0.40
Томат	v. 40	Solanum_lycopersicum.SL2.50.40
Картофель	v. 40	Solanum_tuberosum.SolTub_3.0.40
Кукуруза	v. 40	Zea_mays.AGPv4.40

Для контигов (набор перекрывающихся сегментов ДНК), которые были классифицированы как Unaligned и Unannotated, с помощью программы TransDecoder v5.5.0. [5] были получены предполагаемые белок-кодирующие последовательности. Для аннотации аминокислотных последовательностей использовалась программа InterproScan v.5.36-75 [34].

Общий объем полученных таким образом данных составил 20440228 нуклеотидных последовательностей транскриптов суммарной длиной 9659793403 нуклеотидов и 4055996 аминокислотных последовательностей суммарной длиной 877229075 аминокислот. В дальнейшем эти данные, а также данные об исследованиях,

экспериментах и результаты аналитических программ были экспортированы в базу данных OORT.

Содержание БД OORT

Содержание БД OORT включает:

- метаинформацию о библиотеках транскриптомных экспериментов;
- нуклеотидные последовательности транскриптов, полученные в результате *de novo* сборки;
- аминокислотные последовательности, полученные в результате трансляции нуклеотидных последовательностей транскриптов, кодирующих белки;
- аннотации нуклеотидных последовательностей (предсказание кодирующего потенциала, выравнивание на референсный геном, оценку уровня экспрессии);
- аннотации аминокислотной последовательности (предсказание функциональных доменов, ассоциированные с последовательностью термины онтологии генов).

При работе с базой данных для пользователя важно решать ряд поисковых задач, связанных с идентификацией последовательностей в базе данных по метаинформации об эксперименте, принадлежности к организму, гомологии, функциональным характеристикам, уровню экспрессии транскрипта. Подобный поиск может осуществляться как для нуклеотидных, так и для аминокислотных последовательностей. При этом следует принимать во внимание отношения между аминокислотной последовательностью и нуклеотидной последовательностью транскрипта, из которого она была получена.

Формирование базы данных, индексация

В качестве СУБД при формировании базы данных OORT мы использовали PostgreSQL версии 12 [35, 36]. С ее помощью была построена реляционная схема данных связывающая исследования, эксперименты, контиги и белки (см. раздел “Структура базы данных”). База данных позволяет определять реляционные отношения с помощью первичных и внешних ключей.

С помощью языка программирования Python данные из файлов с результатами биоинформатической обработки (см. раздел “Содержание БД OORT”) были преобразованы в структуру типа “словарь”, которую затем конвертировали в формат JSON с помощью библиотеки SQLAlchemy [37]. SQLAlchemy позволяет транслировать код на языке Python в команды на языке SQL и передавать их на выполнение СУБД, получая результаты запросов в виде объектов языка Python. Далее эти данные были преобразованы в структуры формата JSONB для дальнейшего хранения и доступа.

Создаваемая нами схема базы данных разрабатывалась для поиска информации, представленных как в типизированных (хорошо структурированных), так и в слабоструктурированных полях. Типизированные поля предназначены для описания сущностей в БД, типы которых известны и определены однозначно. Например, последовательность представлена в виде текстовой строки, длина последовательности описана полем типа int, название организма описано в виде текстового поля. Для поиска информации в типизированных данных (int, real, и text) проводилась индексация с помощью структуры B-tree, которая в дальнейшем позволяла выполнять, в том числе, и операции поиска с учетом отношений “равно”, “больше”, “меньше” [38]. Структура B-tree позволяет оптимизировать выполнение запросов с указанными операциями.

Помимо типизированных, хорошо структурированных данных, содержимое базы включало слабо структурированную информацию. К таким данным относились аннотации аминокислотных последовательностей в терминах онтологии генов (GO) [39] и белковых доменов [34]; результаты выравнивания контигов на референсные

геномы; результаты оценки уровня экспрессии транскриптов и т.д. Данные подобного сорта представлялись в виде текстовых строк в формате JSON: записи вида ключ-значение без определенных заранее ограничений. СУБД PostgreSQL 12 позволяет хранить информацию в формате JSON в структуре реляционной базы как поля специального типа JSONB [40]. Это данные в формате JSON, представленные в бинарном виде для быстрого обращения без повторного синтаксического разбора, при этом ключи и значения в этом поле не описываются в формате языка SQL. Поверх ключей поля JSONB можно создавать индексы для быстрого поиска данных, также некоторые индексы позволяют использовать новые поисковые опции над значениями, например, поиск искомого объекта в массиве. Индексация полей формата JSON проводилась с помощью GIN индексов [40]. Такая индексация позволяет быстро производить поиск текстовых строк в указанном массиве терминов, например, терминов онтологии и белковых доменов.

Для индексации координат контигов в геноме использовалась структура “материализованное представление”, специальная техника, реализованная в СУБД PostgreQSL, которая позволяет сохранить в виде отдельной таблицы базы данных результат определенного запроса [41]. Это позволяет, после выполнения запроса в виде “материализованного представления”, обращаться к нему как к отдельной таблице реляционной БД. Такая технология позволяет существенно ускорить выполнение запросов (особенно сложных) к базе данных, поскольку в случае повторного запроса обращение производится только к этой таблице, а не к реальным полям БД.

Доступ к информации в БД OORT

Доступ к извлечению и модификации данных в БД OORT производился на языке запросов SQL. Типичный запрос на поиск данных включает, как правило, следующие секции этого языка:

- SELECT – секция выбора полей таблиц. Здесь через запятую нужно указать имена колонок, которые нужно отобразить, или * чтобы отобразить все поля.
- FROM – секция таблиц. В этой секции следует указать таблицы, из которых нужно получить результаты. Также в этой секции может использоваться оператор JOIN, с помощью которого выполняется объединение таблиц.
- WHERE – секция условия. В ней пользователь указывает условия поиска и объединяет их с помощью логических операторов AND, OR и NOT (условие не выполняется).

Следует отметить, что СУБД PostgreSQL версии 12 расширяет диалект языка SQL:2016 за счет дополнительных операторов, которые служат для извлечения данных из полей формата JSONB [40]. К этим операторам относятся оператор '−>', который используется для извлечения вложенного JSON-значения по ключу, и оператор '−>>', который используется для получения строковых значений по ключу. Использование этих операторов продемонстрировано ниже в разделе “Примеры поисковых запросов”.

Интерфейс SQL к разработанной БД OORT может быть реализован с помощью таких приложений, как DataGrip [42], pgAdmin [43] или консольного приложения psql СУБД PostgreSQL.

База данных размещена на сервере ЦКП “Биоинформатика” по адресу: https://oort.cytogen.ru/.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Структура базы данных

Реляционная схема базы данных включает описание метаданных транскриптомного эксперимента и результатов сборки транскриптов *de novo* в виде следующих объектов:

исследование (study), эксперимент (exp), нуклеотидная последовательность (contig) и аминокислотная последовательность (pep) (рис. 1). Как было указано в разделе “Формирование базы данных, индексация”, при создании ссылок между таблицами БД мы использовали первичные и вторичные ключи, которые были созданы следующим образом:

- каждая из таблиц БД хранит поле id, которое является первичным ключом;
- таблица exp содержит вторичный ключ study_id, который ссылается на таблицу study;
- таблица contig содержит вторичный ключ exp_id, который ссылается на таблицу exp;
- таблица pep содержит вторичные ключи contig_id и exp_id, которые ссылаются на таблицы contig и exp, соответственно.

В результате, между указанными таблицами были созданы реляционные связи, позволяющие эффективно осуществлять поиск информации по эксперименту, исследованию и полученных в результате эксперимента последовательностях. Структура отношений между таблицами БД OORT представлена на рисунке 1.

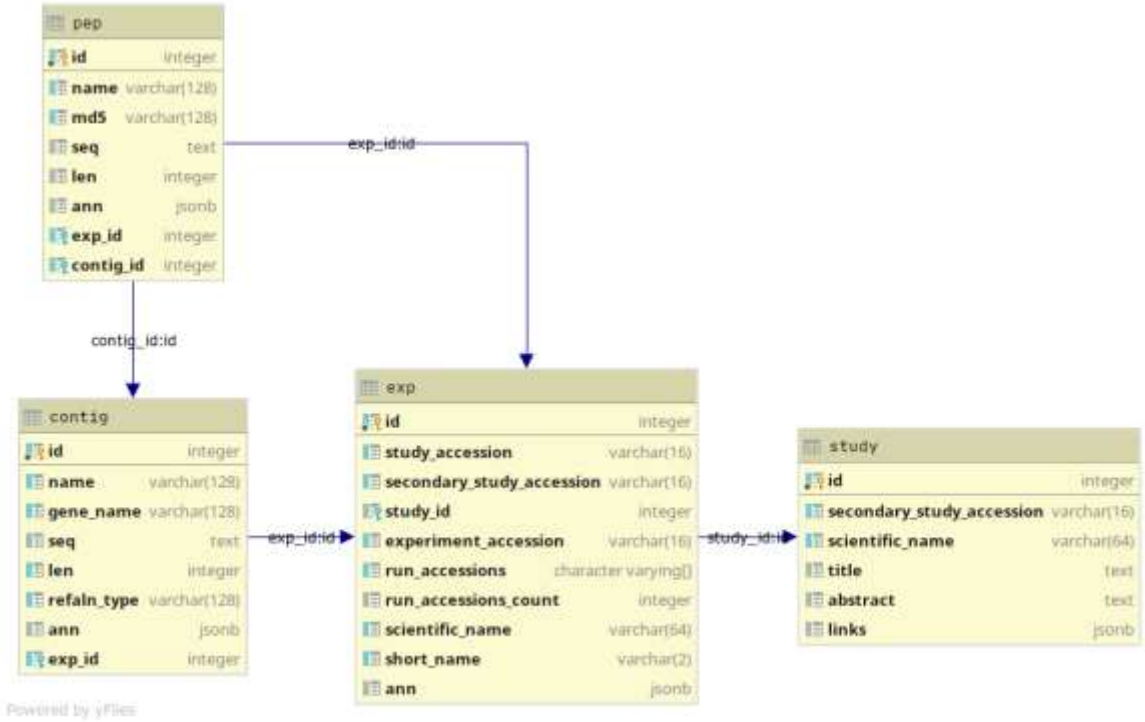

Рис. 1. Структура реляционной базы данных OORT. Показаны таблицы БД и отношения между ними.

Таблица 2. Описание полей в таблице БД OORT ‘study’, которая описывает исследования, в рамках которого было осуществлено секвенирование транскриптома

Название поля	Тип данных	Описание поля
secondary_study_accession	строка	Альтернативный код доступа исследования
scientific_name	строка	Название организма, транскриптом которого был проанализирован
title	текст	Название исследования
abstract	текст	Описание исследования
links	jsonb	Ссылки на внешние ресурсы, связанные с дополнительной информацией об исследовании

Таблица 3. Описание полей в таблице БД OORT ‘exp’, которая описывает эксперимент, в рамках которого была получена библиотека прочтений транскриптома

Название поля	Тип данных	Описание поля
study_accession	строка	Код доступа к записи по исследованию
secondary_study_accession	строка	Альтернативный код доступа записи по исследованию
study_id	число	Идентификатор исследования, вторичный ключ
experiment_accession	строка	Код доступа для эксперимента
run_accessions	массив строк	Список SRA файлов, полученных в результате исследования (каждый файл - результат одного запуска процесса секвенирования (run)
run_accessions_count	число	Количество SRA файлов в эксперименте
scientific_name	строка	Таксономическое название организма, транскриптом которого был секвенирован
short_name	строка	Краткое название организма
ann	jsonb	Аннотация эксперимента

Таблица 4. Описание полей в таблице БД OORT ‘contig’, которая описывает нуклеотидные последовательностей транскриптов, полученных в результате *de novo* сборки из коротких прочтений

Название поля	Тип данных	Описание поля
name	строка	Идентификатор нуклеотидной последовательности (транскрипта)
gene_name	строка	Название гена (для контигов, которые являются его изоформами)
seq	текст	Нуклеотидная последовательность транскрипта
len	число	Длина нуклеотидной последовательности
refain_type	строка	Тип последовательности: невыровненнаяна референсные геном, или неаннотированная
ann	jsonb	Аннотация для нуклеотидной последовательности транскрипта (контига)
exp_id	число	Внешний ключ к записи в таблице ‘exp’ эксперимента, из библиотеки которого получена последовательность

Таблица 5. Описание полей в таблице БД OORT ‘pep’, которая описывает аминокислотную (пептидную) последовательность, полученную в результате трансляции из последовательности транскрипта

Название поля	Тип данных	Описание поля
name	строка	Название аминокислотной последовательности
md5	строка	Хэш-код md5 для аминокислотной последовательности. Необходим для быстрой проверки идентичных последовательностей
seq	текст	Аминокислотная последовательность
len	число	Длина аминокислотной последовательности
ann	jsonb	Аннотация аминокислотной последовательности
contig_id	число	Внешний ключ к записи в таблице ‘contig’ для транскрипта (контига), из которой аминокислотная последовательность была получена
exp_id	число	Внешний ключ к записи в таблице ‘exp’ эксперимента, из библиотеки которого получена последовательность

Детальное описание полей БД показано в таблицах 2–5. Количество данных различного типа в БД OORT представлено в таблице 6. Общий объем данных составил 45.5 гигабайт.

Таблица 6. Количество записей и размеры таблиц базы данных OORT. Размер данных приведен в мегабайтах (МБ)

Название таблицы	Количество записей	Размер данных
Описание исследований	2766	3.0 МБ
Описание экспериментов	1241	7.4 МБ
Нуклеотидные последовательности	20440228	39000.0 МБ
Аминокислотные последовательности	4055996	6495.0 МБ

Пример описания слабоструктурированных данных

Рассмотрим подробнее способ описания слабо структурированных данных в БД OORT. Эти данные в указанных выше таблицах ‘exp’, ‘pep’ и ‘contig’ реализованы в виде текстовых полей типа jsonb, которые содержат результаты аннотации последовательностей. На рисунке 2 представлена часть содержимого поля ‘ann’ для таблицы ‘contig’ транскрипта `TRINITY_DN3631_c0_g1_i1`, которая описывает оценку уровня экспрессии, полученную в результате использования программы kallisto [21]. Эту информацию в поле ‘ann’ можно выделить по названию ключа “kallisto” (первая строка записи). Программа kallisto дает оценку уровня экспрессии как для гена, к которому относится транскрипт (раздел записи, который соответствует ключу “g”, вторая строка на рис. 2), так и для собственно транскрипта, представляющего одну из изоформ данного гена (раздел записи, который соответствует ключу “i”, седьмая строка на рис. 2). Для каждой такой оценки уровня экспрессии программа kallisto выдает три параметра: “tpm”, оценка уровня экспрессии в величинах "транскрипт на миллион”, TPM; “eff_length”, эффективная длина последовательности; “est_counts”, оценка числа прочтений, выравненных на данный транскрипт. Все эти параметры приводятся в соответствующих значениях для ключей, указанных в формате JSON (рис. 2).

```
01    "kallisto": {
02        "g": {
03          "tpm": 8.44,
04          "eff_length": 341.61,
05          "est_counts": 35.0
06        },
07        "i": {
08          "tpm": 8.43596,
09          "eff_length": 341.614,
10          "est_counts": 35.0
11        }
12    }
```

Рис. 2. Пример описания результатов оценки уровня экспрессии транскрипта `TRINITY_DN3631_c0_g1_i1` программой kallisto в формате JSON в поле ‘ann’ таблицы ‘contig’ БД OORT.

Аналогичным образом описывается аннотация аминокислотных последовательностей. Описание результатов аннотации белковых доменов программой Interproscan [34] содержит разделы:

- tsv – двумерный массив вывода программы InterproScan;
- go_ann – массив строк с GO аннотациями белка;

- interpro_acc – массив строк с функциями из InterPro для белка;
- pathways_ann – массив строк со списком сигнальных и метаболических путей, в которых участвует белок.

Описание результатов предсказания открытой рамки считывания и ее параметров программой TransDecoder [5] содержит разделы:

- type – строка с обозначением типа предсказанной рамки трансляции (полная, от старт до стоп-кодона, либо ее фрагмент);
- chain – строка с обозначением направления цепи, с которой происходит трансляция (“+” или “–”);
- score – число, характеризующее надежность идентификации открытой рамки трансляции в нуклеотидной последовательности;
- cds_end – число: координата конца открытой рамки трансляции в нуклеотидной последовательности;
- cds_start – число: координата начала открытой рамки трансляции.

Примеры реализации поисковых запросов к БД OORT

Приведем примеры ряда запросов на языке СУБД PostgreSQL 12 к БД OORT. Код для выполнения этих запросов показан на рисунках 3–7.

```
# Запрос 1: Получить последовательности генов для вида 'Oryza sativa'
01 select contig.seq
02 from contig join exp on exp.id = contig.exp_id
03 where scientific_name = 'Oryza sativa'

# Запрос 2: Найти все гены, с уровнем экспрессии в пределах от 3.0 до
7.0 TPM
01 select *
02 from contig
03 where ((ann -> 'kallisto' -> 'g' ->> 'tpm')::real)
04 between 3.0 and 7.0
```

Рис. 3. Реализация запросов 1 и 2 к базе данных OORT на языке SQL.

Первый запрос (рис. 3) заключается в получении всех последовательностей транскриптов, полученных из библиотек риса (*O. sativa*). Данный запрос для выполнения использует только стандартные операторы SQL обращения к полям БД.

```
# Запрос 3: Вывести список генов, кодирующих идентичные пептиды с разным
уровнем экспрессии в разных экспериментах
01 select subseq.exp_id, subseq.study_id,
02 max(subseq.contig_g_tpm) as contig_g_tpm,
03 study.scientific_name, tissue_type
04 from (
05    select (contig.ann->'kallisto'->'g'->'tpm')::real
06    as contig_g_tpm, contig.id as contig_id, pep.id as
07    pep_id, contig.exp_id, e.study_id, pep.md5,
08    e.ann->'ebi'->'tissue_type' as tissue_type
09    from pep join contig on pep.contig_id = contig.id
10    join exp e on contig.exp_id = e.id
11    where pep.md5 = '857c1634328c5333ab73efcc11a45038'
12 )
13 subseq join study on subseq.study_id = study.id
14 group by exp_id, study_id, study.scientific_name,
15 tissue_type
```

Рис. 4. Пример кода, реализующего запрос 3 к базе данных OORT на языке SQL (см. текст).

Второй запрос заключается в получении списка идентификаторов всех генов, уровень экспрессии которого находится в пределах от 3 до 7 TPM. Этот запрос использует обращение к полю JSON ‘ann’ и извлекает из него значение по ключу “kallisto” с использованием операторов −> и −>> (рис. 3).

Третий запрос заключается в поиске генов, аминокислотные последовательности которых идентичны (т.е. имеют одинаковые значения хэш-кода md5), при этом уровень их экспрессии в разных экспериментах различается.

Запрос 4 заключается в поиске аминокислотных последовательностей, которые в своей аннотации содержат определенные термины онтологии генов, предсказанные программой InterproScan. Так как информация об аннотации не типизирована, то перед выполнением этого запроса необходимо произвести GIN индексацию поля ‘ann’ таблицы ‘pep’ для значений ключа “go_ann” (строки 1 и 2, рисунок 5). Это позволяет искать записи уже в индексном массиве (строки 3–5, рисунок 5).

```
# Запрос 4: Найти все аминокислотные последовательности с определенным набором терминов генной онтологии
# создание индекса для выполнения запроса
01 create index ix_pep_ann_interpro_go_ann on pep using
02 gin (((ann -> 'interpro'::text) -> 'go_ann'::text))
# поиск белков согласно аннотации генной онтологии на основе созданного индекса
03 select * from pep
04 where (ann -> 'interpro' -> 'go_ann')
05 ?& array['GO:0055085', 'GO:0006811']
```

Рис. 5. Пример кода, реализующего запрос 4 к базе данных OORT на языке SQL (см. текст).

Запрос 5 заключается в поиске всех исследований, представленных в БД OORT, которые в описании содержат термин ‘Categorizing’. Для его реализации на первом этапе создается GIN индекс на вектора слов из поля ‘abstract’ (рисунок 6, 2 строка, метод to_tsvector), где записаны краткие описания исследования. Далее выполняется запрос к этим векторам (строки 5–6, рисунок 6).

```
# Запрос 5: Поиск слова в описаниях исследования
# создание индекса для полнотекстового поиска
01 create index study_abstract_idx on study
02 (to_tsvector('english'::regconfig, abstract));
# выполнение полнотекстового поиска
03 select *
04 from study
05 where to_tsvector(study.abstract)
06 @@ plainto_tsquery('Categorizing')
```

Рис. 6. Пример кода, реализующего запрос 5 к базе данных OORT на языке SQL (см. текст).

Запрос 6 заключается в поиске белков, которые синтезируются с транскриптов, выравненных на участок генома с заданными координатами программой gmap. Так как ключ “gmap” имеет структуру двумерного массива, то на первом этапе эти массивы нормализуются в “материальном отображении” таким образом, чтобы в одной строке таблицы была одна строка из исходного массива (рисунок 7, строки 1–8). Далее индексируются три позиции в полученной таблице (строки 9–12) и выполняется поиск искомых последовательностей (строки 13–24).

Приведенные примеры показывают, что к существующей БД можно организовывать достаточно сложные запросы, которые могут включать как информацию о последовательности, ее локализации в геноме, аннотации. Запросы

могут учитывать результаты сравнения различных характеристик последовательностей, описанных в БД.

```
# Запрос 6. Найти все белки, которые транслируются из контигов в геноме Zea mays во 2 хромосоме между 2000 и 5000 нуклеотидами.
# реализация материального представления contig_gmap_view_sqlalchemy
01 create materialized view
02 contig_gmap_view_sqlalchemy as
03 select db.id, jsonb_array_elements(
04 db.ann -> 'gmap'::text) AS gmap
05 from (
06 select contig.id, contig.ann
07 from contig ORDER BY contig.id
08 ) db;
# создание индексов поверх колонок GMAP
09 create index gmap_start_end_chr_idx
10 on contig_gmap_view_sqlalchemy
11 (((gmap ->> 15)::integer), ((gmap ->> 16)::integer),
12 (gmap ->> 13));
# поиск белков, синтезирующихся со 2 хромосомы с 2000 по 5000
# позиции в геноме
13 select pep.id AS pep_id
14 from exp join contig ON exp.id = contig.exp_id
15 join contig_gmap_view_sqlalchemy
16 on contig.id = contig_gmap_view_sqlalchemy.id
17 join pep ON pep.contig_id = contig.id
18 where
19 cast(contig_gmap_view_sqlalchemy.gmap ->> 15
20 as integer) >=2000 AND
21 cast(contig_gmap_view_sqlalchemy.gmap ->> 16
22 as integer) <= 5000
23 and(contig_gmap_view_sqlalchemy.gmap ->> 13) = '2'
24 and exp.scientific_name like 'Zea mays'
```

Рис. 7. Пример кода, реализующего запрос 6 к базе данных OORT на языке SQL (см. текст).

Созданная нами база данных на основе гибридного метода описания комплексных молекулярно-биологических данных в виде как хорошо структурированных (типизированных) данных, так и слабо структурированной информации, обладает рядом преимуществ. Прежде всего, такая организация хранения данных облегчает их сопровождение: при изменении набора программ аннотации, их версий или набора выводимых ими параметров, вместо того, чтобы модифицировать реляционную схему базы и повторно загружать данные в базу, разработчик может составлять слабосвязанные структуры в формате JSON.

Предложенная архитектура БД сохраняет связность таблиц при помощи методов первичных и вторичных ключей. Следует отметить, что структура JSON-описания нетипизированных данных строго не определена и для разных записей даже одной таблицы может отличаться. В конечном итоге, эта структура зависит от того, какие данные были внесены в конкретное поле БД. Поддержание однородности данных в формате JSON при такой организации лежит на разработчике: контроль структуры полей JSON происходит либо в момент экспорта данных, и\или путем реализации функций-обработчиков, которые запускаются при каждом обновлении и добавлении данных.

ВЫВОДЫ

В работе предложен гибридный подход к созданию баз молекулярно-генетических данных, которые содержат информацию о последовательностях транскриптов и их

структурно-функциональной аннотации. Сущность подхода заключается в одновременном хранении в БД информации как структурированного типа, так и слабо структурированных данных. Данная схема обладает рядом преимуществ в сопровождении данных: вместо того, чтобы модифицировать реляционную схему базы и повторно загружать данные в базу, разработчик может составлять слабосвязанные структуры в формате JSON, и также производить индексацию поля внутри этой структуры.

С помощью предложенной технологии была создана БД OORT, которая содержит информацию о транскриптомах сельскохозяйственных растений. База данных состоит как из строго структурированных полей (идентификаторы исследований, эксперимента, нуклеотидные и аминокислотные последовательности, таксономическая информация), так и слабоструктурированных (уровень экспрессии транскриптов, локализация транскрипта в референсном геноме, аннотация на основе терминов онтологии генов и функциональных доменов белков). Для работы с этими данными (добавление, удаление, модифицирование, поиск данных) пользователю требуется реализовать скрипты на языках запросов SQL или использовать сторонние библиотеки, которые автоматизируют работу с базой данных.

Работа была поддержана грантом РНФ №18-14-00293 (разработка структуры и информационное содержание БД OORT) и бюджетным проектом №0324–2019-0040-С-01 (предоставление вычислительных ресурсов ЦКП “Биоинформатика” для реализации БД и их системная поддержка).

СПИСОК ЛИТЕРАТУРЫ

1. Martin L.B.B., Fei Z., Giovannoni J.J., Rose J.K.C. Catalyzing plant science research with RNA-seq. *Frontiers in Plant Science*. 2013. V. 4. P. 66.
2. Usadel B., Fernie A.R. The plant transcriptome-from integrating observations to models. *Frontiers in Plant Science*. 2013. V. 4. P. 48.
3. Klepikova A. V., Kasianov A.S., Gerasimov E.S., Logacheva M.D., Penin A.A. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant Journal*. 2016. V. 88. № 6. P. 1058–1070.
4. Strickler S.R., Bombarely A., Mueller L.A. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American Journal of Botany*. 2012. V. 99. № 2. P. 257–266.
5. Haas B.J., Papanicolaou A., Yassour M., Grabherr M., Blood P.D., Bowden J., Couger M.B., Eccles D., Li B., Lieber M. et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*. 2013. V. 8. № 8. P. 1494–1512.
6. Kim D., Langmead B., Salzberg S.L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*. 2015. V. 12. № 4. P. 357–360.
7. Bryant D.M., Johnson K., DiTommaso T., Tickle T., Couger M.B., Payzin-Dogru D., Lee T.J., Leigh N.D., Kuo T.H., Davis F.G. et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. *Cell Reports*. 2017. V. 18. № 3. P. 762–776.
8. Bolger M.E., Arsova B., Usadel B. Plant genome and transcriptome annotations: From misconceptions to simple solutions. *Briefings in Bioinformatics*. 2018. V. 19. № 3. P. 437–449.
9. Glagoleva A.Y., Shmakov N.A., Shoeva O.Y., Vasiliev G. V., Shatskaya N. V., Börner A., Afonnikov D.A., Khlestkina E.K. Metabolic pathways and genes identified by RNA-seq analysis of barley near-isogenic lines differing by allelic state of the Black lemma and pericarp (Blp) gene. *BMC Plant Biology*. 2017. V. 17. № S1. P. 182.

10. Shmakov N.A., Vasiliev G. V., Shatskaya N. V., Doroshkov A. V., Gordeeva E.I., Afonnikov D.A., Khlestkina E.K. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq. *BMC Plant Biology*. 2016. V. 16. № 3. P. 119–138.
11. Papatheodorou I., Moreno P., Manning J., Fuentes A.M.P., George N., Fexova S., Fonseca N.A., Füllgrabe A., Green M., Huang N. et al. Expression Atlas update: From tissues to single cells. *Nucleic Acids Research*. 2020. V. 48. № D1. P. D77–D83.
12. Masoudi-Nejad A., Goto S., Jauregui R., Ito M., Kawashima S., Moriya Y., Endo T.R., Kanehisa M. EGENES: Transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. *Plant Physiology*. 2007. V. 144. № 2. P. 857–866.
13. Ueno S., Nakamura Y., Kobayashi M., Terashima S., Ishizuka W., Uchiyama K., Tsumura Y., Yano K., Goto S. TodoFirGene: Developing transcriptome resources for genetic analysis of abies sachalinensis. *Plant and Cell Physiology*. 2018. V. 59. № 6. P. 1276–1284.
14. Dubois A., Carrere S., Raymond O., Pouvreau B., Cottret L., Roccia A., Onesto J.P., Sakr S., Atanassova R., Baudino S. et al. Transcriptome database resource and gene expression atlas for the rose. *BMC Genomics*. 2012. V. 13. № 1. P. 638.
15. Fernández-Pozo N., Canales J., Guerrero-Fernández D., Villalobos D.P., Díaz-Moreno S.M., Bautista R., Flores-Monterroso A., Guevara M.Á., Perdiguero P., Collada C. et al. EuroPineDB: A high-coverage web database for maritime pine transcriptome. *BMC Genomics*. 2011. V. 12. № 1. P. 366.
16. Barnett D.W., Garrison E.K., Quinlan A.R., Stŕmberg M.P., Marth G.T. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*. 2011. V. 27. № 12. P. 1691–1692.
17. Quinlan A.R., Hall I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010. V. 26. № 6. P. 841–842.
18. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. V. 25. № 16. P. 2078–2079.
19. Pertea G., Pertea M. GFF Utilities: GffRead and GffCompare. *F1000Research*. 2020. V. 9. P. 304.
20. Anders S., Huber W. Differential expression of RNA-Seq data at the gene level-the DESeq package. *Heidelberg, Germany: European Molecular Biology Laboratory* (*EMBL*). 2012. V. 10. P. f1000research.
21. Bray N.L., Pimentel H., Melsted P., Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*. 2016. V. 34. № 5. P. 525–527.
22. Gunbin K. V., Suslov V. V., Genaev M.A., Afonnikov D.A. Computer System for Analysis of Molecular Evolution Modes (SAMEM): Analysis of molecular evolution modes at deep inner branches of the phylogenetic tree. *In Silico Biology*. 2011. V. 11. № 3. P. 109–123.
23. Han J., Haihong E., Le G., Du J. Survey on NoSQL database. In: *ICPCA 2011*: *6th International Conference on Pervasive Computing and Applications*. 2011. P. 363–366.
24. Gabetta M., Limongelli I., Rizzo E., Riva A., Segagni D., Bellazzi R. BigQ: A NoSQL based framework to handle genomic variants in i2b2. *BMC Bioinformatics*. 2015. V. 16. № 1. P. 415.
25. *ENA Portal*. URL: https://www.ebi.ac.uk/ena/portal/api/ (accessed: 23.10.2020).
26. Harrison P.W., Alako B., Amid C., Cerdeño-Tárraga A., Cleland I., Holt S., Hussein A., Jayathilaka S., Kay S., Keane T. et al. The European Nucleotide Archive in 2018. *Nucleic Acids Research*. 2019. V. 47. № D1. P. D84–D88.

27. *Submit your project and biological samples*. URL: https://www.ncbi.nlm.nih.gov/sra/docs/submitbio/ (accessed: 23.10.2020).
28. Staff S.R.A.S. Using the SRA Toolkit to convert .sra files into other formats. *National Center for Biotechnology Information*. 2011.
29. Chen S., Zhou Y., Chen Y., Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018. V. 34. № 17. P. i884–i890.
30. Bushmanova E., Antipov D., Lapidus A., Suvorov V., Prjibelski A.D. RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*. 2016. V. 32. № 14. P. 2210–2212.
31. Wu T.D., Watanabe C.K. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005. V. 21. № 9. P. 1859–1875.
32. *Ensembl Plants*. URL: https://plants.ensembl.org/index.html (accessed: 23.10.2020).
33. Kersey P.J., Allen J.E., Allot A., Barba M., Boddu S., Bolt B.J., Carvalho-Silva D., Christensen M., Davis P., Grabmueller C. et al. Ensembl Genomes 2018: An integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*. 2018. V. 46. № D1. P. D802–D808.
34. Jones P., Binns D., Chang H.Y., Fraser M., Li W., McAnulla C., McWilliam H., Maslen J., Mitchell A., Nuka G. et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*. 2014. V. 30. № 9. P. 1236–1240.
35. *PostgreSQL: The world's most advanced open source database*. URL: https://www.postgresql.org/ (accessed: 23.10.2020).
36. Schönig H.-J. Schönig H.-J. *Mastering PostgreSQL 11: Expert techniques to build scalable, reliable, and fault-tolerant database applications*. Birmingham: Packt Publishing Ltd., 2018. 448 p.
37. *SQLAlchemy - The Database Toolkit for Python*. URL: https://www.sqlalchemy.org/ (accessed: 23.10.2020).
38. *PostgreSQL: Documentation: 12: 11.2. Index Types*. URL: https://www.postgresql.org/docs/12/indexes-types.html (accessed: 23.10.2020).
39. Carbon S., Douglass E., Dunn N., Good B., Harris N.L., Lewis S.E., Mungall C.J., Basu S., Chisholm R.L., Dodson R.J. et al. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*. 2019. V. 47. № D1. P. D330–D338.
40. Petković D. JSON integration in relational database systems. *Int J Comput Appl*. 2017. V. 168. № 5. P. 14–19.
41. Kaur M., Shaik B. Kaur M., Shaik B. *PostgreSQL Development Essentials* Birmingham: Packt Publishing Ltd., 2016. 210 p.
42. *DataGrip:* кросс-платформенная среда разработки для баз данных и SQL. URL: https://www.jetbrains.com/ru-ru/datagrip/ (accessed: 23.10.2020).
43. *pgAdmin - PostgreSQL Tools*. URL: https://www.pgadmin.org/ (accessed: 23.10.2020).

Рукопись поступила в редакцию 26.10.2020, переработанный вариант поступил 14.12.2020.
Дата опубликования 28.12.2020.

RDBMS and NOSQL Based Hybrid Technology for Transcriptome Data Structuring and Processing

Mukhin A.M.[1,2], Genaev M.A.[1,2,3], Rasskazov D.A.[1,2], Lashin S.A.[1,2,3], Afonnikov D.A.[1,2,3]

[1]*The Federal Research Center Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*
[2]*Kurchatov Genomics Center, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*
[3]*Novosibirsk State University, Novosibirsk, Russia*

Abstract. The transcriptome sequencing experiment (RNA-seq) has become almost a routine procedure for studying both model organisms and crops. As a result of bioinformatics processing of such experimental output, huge heterogeneous data are obtained, representing nucleotide sequences of transcripts, amino acid sequences, and their structural and functional annotation. It is important to present the data obtained to a wide range of researchers in the form of databases. This article proposes a hybrid approach to creating molecular genetic databases that contain information about transcript sequences and their structural and functional annotation. The essence of the approach consists in the simultaneous storing both structured and weakly structured data in the database. The technology was used to implement a database of transcriptomes of agricultural plants. This paper discusses the features of implementing this approach and examples of generating both simple and complex queries to such a database in the SQL language. The OORT database is freely available at https://oort.cytogen.ru/

Key words: *database, indexing, queries, plants, SQL, RDBMS, NoSQL, transcriptomes, crops.*