ВЕСТНИК
НОВОСИБИРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Научный журнал
Основан в ноябре 1999 года

Серия: Информационные технологии 2018. Том 16, № 3

СОДЕРЖАНИЕ

VESTNIK
NOVOSIBIRSK STATE UNIVERSITY

Scientific Journal
Since 1999, November
In Russian

Series: Information Technologies 2018. Volume 16, № 3

CONTENTS

Editor in Chief Anatolij M. Fedotov

Vice-Editor A. V. Avdeev

Executive Secretary N. N. Pestereva

Editorial Board of the Series

I. V. Bychkov, professor, academician (Irkutsk), *B. M. Glinsky*, professor (Novosibirsk)
A. N. Gorban', professor (Lester, GB), *E. P. Gordov*, professor (Tomsk)
B. S. Dobronets, professor (Krasnoyarsk), *A. M. Elizarov*, professor (Kazan)
G. N. Erokhin, professor (Kaliningrad), *A. I. Kamyshnikov*, professor (Khanty-Mansijsk)
G. P. Karev, professor (Maryland, USA), *N. A. Kolchanov*, professor, academician (Novosibirsk)
M. M. Lavrentjev, professor (Novosibirsk), *V. E. Malyshkin*, professor (Novosibirsk)
N. N. Mirenkov, professor (Aizu, Japan), *N. M. Oskorbin*, professor (Barnaul)
D. E. Palchunov, professor (Novosibirsk), *T. Pizansky*, professor (Ljubljana, Slovenia)
V. P. Potapov, professor (Kemerovo), *O. I. Potaturkin*, professor (Novosibirsk)
V. A. Serebryakov, professor (Moscow), *A. V. Starchenko*, professor (Tomsk)
S. I. Smagin, professor, corresponding member of RAS (Khabarovsk)
D. A. Tusupov, professor (Astana, Kazakhstan)
V. V. Shajdurov, professor, corresponding member of RAS (Krasnoyarsk)
Yu. I. Shokin, professor, academician (Novosibirsk)

The journal is published quarterly in Russian since 1999
by Novosibirsk State University Press

The address for correspondence

Centre of Information Technologies, Novosibirsk State University

Pirogov Street 2, Novosibirsk, 630090, Russia

Tel. +7 (383) 363 40 28

E-mail address: inftech@vestnik.nsu.ru

On-line version: http://elibrary.ru

РАЗВИТИЕ ОБРАЗОВАНИЯ В БИОИНФОРМАТИКЕ ПО МАТЕРИАЛАМ, ПРЕДСТАВЛЕННЫМ НА СТУДЕНЧЕСКИХ КОНФЕРЕНЦИЯХ МНСК-2018, ШКОЛЕ МОЛЕКУЛЯРНОГО МОДЕЛИРОВАНИЯ И ХАКАТОНЕ В НОВОСИБИРСКЕ

Рубрика «Биоинформатика» содержит материалы, представленные на студенческих конференциях и учебных мероприятиях по биоинформатике в Новосибирске в 2018 г. Прежде всего отметим серию международных молодежных научных Школ «Системная биология и биоинформатика» / «Systems Biology and Bioinformatics (SBB)», организованных Институтом цитологии и генетики СО РАН (академик Н. А. Колчанов) в Новосибирске уже в 10-й раз в 2018 г. [1] Школа SBB-2018 завершила свою работу в августе 2018 г. после международной конференции по биоинформатике и системной биологии BGRS/SB-2018, проходящей в Новосибирске каждые два года с 1998, уже в 11-й раз, в этом году впервые совместно на базе ИЦиГ СО РАН и НГУ. Кроме междисциплинарных научных исследований отметим интерес к вопросам образования, преподавания, студенческих научных обменов, чему была посвящена отдельная секция конференции по образованию в области биоинформатики 23 августа 2018 г. (Н. А. Колчанов, А. Г. Окунев, Д. А. Афонников). Студенческие работы по биоинформатике, особенно выполненные в больших коллективах, становятся все более профессиональными, из года в год повышая требования к студенческим дипломам и аспирантским диссертациям.

Ежегодная Международная научная студенческая конференция (МНСК), проводимая Новосибирским государственным университетом, имеет давнюю историю – в этом году она была 56-й [2]. На секции биоинформатики (в форме отдельной секции и конкурса) были представлены научные доклады студентов и аспирантов на стыке информатики, математики и биологии. В последние годы отмечается рост числа докладов по биоинформатике, обсуждаемых на МНСК, растет и качество работ – научный уровень, наличие собственных публикаций, процитированных в работах студентов, повышается степень программной реализации представленных алгоритмов биоинформатики. Ряд студенческих работ, представленных в данной рубрике, посвящен разработке программного обеспечения для анализа геномных и транскриптомных данных, компьютерному моделированию структуры биологических макромолекул, разработке баз данных и веб-интерфейсов.

Отметим, что параллельно с классическими формами студенческих конференций – обучающих обзорных лекций и докладов самих студентов, появляются новые форматы, связанные с практическими занятиями, программированием, решением задач за короткое время в командах студентов, такие как хакатоны. Хакатоны по биоинформатике проводятся в НГУ уже второй раз, инициированные Лабораторией молекулярной динамики биополимеров (канд. биол. наук А. В. Бакулина) при поддержке САЕ «Синтетическая биология» (Д. О. Жарков, Ю. Л. Орлов). Хакатоны ассоциированы со Школами по молекулярной динамике.

Рассмотрим основные направления работ, представленные на Школе молодых ученых «Компьютерное моделирование структуры и динамики биомолекул» и Хакатоне-2018, прошедшем в НГУ и Технопарке Академгородка 28 апреля – 1 мая 2018 г. Рассматривались задачи геномики, моделирования структуры белка для задач фармацевтики и биотехнологии, задачи анализа биоизображений.

С пленарным докладом «Молекулярное моделирование в медицинской химии» на Школе выступил Д. С. Баев. Затем был проведен Практикум по молекулярному докингу с использованием программы AutoDock Vina. В рамках Школы были организованы секции «Разработка

[1] http://conf.bionet.nsc.ru/sbb2018/.
[2] https://issc.nsu.ru/.

новых биологически активных соединений», «Молекулярная динамика», «Анализ структуры белков» и «Визуализация результатов научных исследований». Каждая секция открывалась лекцией приглашенного специалиста по соответствующей тематике: «Расчет свободной энергии с помощью молекулярной динамики» (А. В. Ким), «Умные и красивые модели в Blender 3D» (З. В. Радькова).

В рамках Школы прошли практикумы: мастер-класс по расчету свойств белка с использованием данных ЯМР (С. С. Марьясина, И. М. Пищенко, Г. Ю. Лаптев), практикум по программе Blender.

Студенты представили свои работы на секциях Школы короткими докладами. Практические направления разработки приложений студентами и молодыми учеными продолжают развиваться, в том числе в сотрудничестве с коммерческими компаниями, такими как UGene.

В целом представляемая рубрика журнала представляет срез работ по приложениям информационных технологий в технических и биоинформационных задачах.

Благодарности. Авторы благодарны Н. А. Колчанову, Д. А. Афонникову, Д. О. Жаркову за организационную поддержку учебных мероприятий. Организация и проведение Школы и Хакатона по биоинформатике были поддержаны НГУ (программа «Топ 5-100») и грантом Министерства образования и науки РФ (28.12487.2018/12.1).

Ю. Л. Орлов, *А. Ю. Бакулина*

Новосибирский государственный университет
orlov@bionet.nsc.ru, bakulina@nsu.ru

УДК 57.084.1:577.171
DOI 10.25205/1818-7900-2018-16-3-7-21

А. О. Брагин [1], **К. А. Табанюхов** [1, 2], **И. В. Чадаева** [1, 3], **А. В. Цуканов** [1, 3]
Р. О. Бабенко [3], **И. В. Медведева** [1], **А. Г. Богомолов** [1, 3]
В. Н. Бабенко [1, 3], **Ю. Л. Орлов** [1, 3]

[1] *Институт цитологии и генетики СО РАН*
пр. Академика Лаврентьева, 10, Новосибирск, 630090, Россия

[2] *Новосибирский государственный аграрный университет*
ул. Добролюбова, 160, Новосибирск, 630039, Россия

[3] *Новосибирский государственный университет*
ул. Пирогова, 2, Новосибирск, 630090, Россия

ibragim@bionet.nsc.ru, ya.cukanton@yandex.ru, ichadaeva@bionet.nsc.ru

КОМПЬЮТЕРНАЯ БАЗА ДАННЫХ ДЛЯ АНАЛИЗА ДИФФЕРЕНЦИАЛЬНО ЭКСПРЕССИРУЮЩИХСЯ ГЕНОВ, СВЯЗАННЫХ С АГРЕССИВНЫМ ПОВЕДЕНИЕМ, НА МОДЕЛЯХ ЛАБОРАТОРНЫХ ЖИВОТНЫХ [*]

Разработка и применение компьютерных средств анализа транскриптомных данных в модельных организмах животных представляет актуальную задачу биоинформатики. Задача исследования экспрессии генов современными методами высокопроизводительного секвенирования в отделах мозга лабораторных животных является исключительно важной для изучения генетических основ поведения в целом. Изучение генетических детерминант агрессивного поведения не только сохраняет актуальность для исследования молекулярных механизмов регуляции поведения, но и имеет широкую практическую составляющую при работе с животными, для решения задач агробиологии. Наследственная предрасположенность животных к агрессивному поведению приводит к появлению различий в строении головного мозга, и сравнение таких различий позволит найти как общие, так и специфичные механизмы регуляции поведения, способствующие проявлению агрессии в провоцирующих условиях среды. Представлены компьютерные программы анализа сплайсинга и прототип базы данных экспрессии генов в отделах мозга лабораторных животных – серых крыс, селектированных по проявлению агрессивного поведения. Выполнена функциональная аннотация генов с повышенной и пониженной экспрессией в экспериментах на крысах, рассмотрены варианты изоформ и альтернативного сплайсинга этих генов.

Ключевые слова: биоинформатика, транскриптомика, сплайсинг, дифференциальная экспрессия, лабораторные животные, генные онтологии, поведение, база данных.

[*] Работа А. О. Брагина и И. В. Чадаевой поддержана РФФИ (проект № 18-34-00496 мол_а). Экспериментальные работы и выполнение транскриптомного секвенирования были поддержаны РНФ в 2014–2016 гг. Работа с лабораторными животными поддержана грантом Минобрнауки РФ (№ RFMEFI62117X0015). Работа Ю. Л. Орлова была поддержана проектом Министерства образования РФ № 28.12487.2018/12.1.

Авторы благодарны А. Л. Маркелю, Р. В. Кожемякиной и С. С. Ковалеву за помощь в работе, предоставление экспериментальных данных и научные консультации по базе данных.

Брагин А. О., *Табанюхов К. А.*, *Чадаева И. В.*, *Цуканов А. В.*, *Бабенко Р. О.*, *Медведева И. В.*, *Богомолов А. Г.*, *Бабенко В. Н.*, *Орлов Ю. Л.* Компьютерная база данных для анализа дифференциально экспрессирующихся генов, связанных с агрессивным поведением, на моделях лабораторных животных // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 7–21.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2018. Том 16, № 3
© А. О. Брагин, К. А. Табанюхов, И. В. Чадаева, А. В. Цуканов, Р. О. Бабенко,
И. В. Медведева, А. Г. Богомолов, В. Н. Бабенко, Ю. Л. Орлов, 2018

Введение

Задача исследования экспрессии генов на основе современных методов высокопроизводительного секвенирования в отделах мозга лабораторных животных является исключительно важной для исследования генетических основ поведения в целом, в том числе агрессивного поведения. Несмотря на огромный научный интерес, направленный на изучение различных психофизиологических механизмов, регулирующих поведение, многие механизмы, отвечающие за проявление агрессии, до сих пор мало изучены [1]. Агрессия – это активная негативная реакция организма на различные проявления внешней среды, один из механизмов борьбы за выживание, в то время как ручное, толерантное, в том числе к человеку, поведение являлось одним из главных направлений искусственного отбора животных в ходе их доместикации, или одомашнивания. Эти вопросы поднимались еще академиком Д. К. Беляевым, основавшим целое направление генетики; по результатам работ проведена «Беляевская конференция – 2017», где широко обсуждались вопросы селекции животных [2]. Исследование основ агрессивного и толерантного поведения при совместном содержании животных может быть применено для практических работ по звероводству, а также служит основой биомедицинских исследований.

Изучение генетических детерминант агрессивного поведения в целом не только сохраняет актуальность в современном мире, но и имеет широкую практическую составляющую при работе с животными, в том числе в пушном звероводстве. На различных моделях животных (см., например, [3]) было показано, что одним из возможных последствий отбора по поведению является изменение ряда фенотипических признаков, в том числе окраса шерсти, что важно для исследования сельскохозяйственных животных, в частности норок и лис в пушном звероводстве. Выявление детерминант поведения важно и для животных мясных пород. Наследственная предрасположенность животных к агрессивному поведению приводит к появлению различий в строении головного мозга, и сравнение таких различий позволит найти как общие, так и специфичные механизмы регуляции поведения, способствующие проявлению агрессии в провоцирующих условиях среды.

Модельные линии крыс являются удобными объектами для анализа факторов, которые влияют на поведение животных. Применение современных технологий высокопроизводительного секвенирования ДНК при работе с моделями лабораторных животных позволяет провести комплексное молекулярно-генетическое и физиологическое исследование [1; 4; 5]. Селекция серых крыс, которая более 30 лет проводится в Институте цитологии и генетики СО РАН (ИЦиГ СО РАН), создает уникальную основу для анализа генетических детерминант поведения. Для анализа генетических основ агрессивного поведения в ИЦиГ СО РАН были выведены две линии серых крыс, *Rattus norvegicus*. Обе линии крыс в течение около 70 поколений селектировались по поведению по отношению к человеку (тест на перчатку – когда рука исследователя в защитной перчатке касается животного в клетке, с последующей видеофиксацией и численной оценкой поведения в баллах). В одну из этих линий в ходе селекции отбирали крыс по дружелюбному поведению по отношению к человеку и другим животным, в другую линию – крыс по агрессивному поведению, оцениваемому по количественной шкале.

Использование в исследованиях нескольких отделов мозга крыс вместе с применением секвенирования транскриптома позволило выявить ряд дифференциально экспрессирующихся генов у агрессивных и ручных (неагрессивных) крыс [4]. В данной работе уточнены механизмы действия генов, представлена кластеризация и компьютерная база данных генов крысы с дифференциальной экспрессией.

Экспериментальная часть исследования включала измерение уровней экспрессии с использованием ПЦР в реальном времени и на основе транскриптомного секвенирования. Проведена оценка экспрессии генов белков, связанных с работой дофамина и серотонина, в вентральной тегментальной области (ВТО) мозга таких крыс на основе РНК-профилирования. Были обнаружены значимые различия в экспрессии генов Th, Ddc, Drd2, Tph2, Htr1a, Htr1b, Htr5b, а также дофаминового (Slc6a3) и серотонинового (Slc6a4) транспортеров между агрессивными и неагрессивными крысами.

Согласно статье [4], был проведен сравнительный транскриптомный анализ РНК из трех отделов мозга агрессивных и неагрессивных (ручных) крыс. В результате определены различия альтернативного сплайсинга некоторых генов синаптической регуляции, в том числе гена глутаматного ионотропного синапса Grin1. Было получено достоверное различие между пропорциями определенных изоформ транскриптов в нескольких генах, связанных с функционированием синапса, что исследователи связывают с особенностями поведения подопытных животных. Показано значение ферментов, усиливающих действие сплайсинга и приводящих к изменениям структуры РНК (энхансеры NOVA1/2, FOX1/2, nSR100/SRM4), а также веществ, обладающих противоположным эффектом (сайленсеры PTB1, PTB2) [4].

Материалы и методы

Эксперименты выполнены на самцах серых крыс (*Rattus norvegicus*), прошедших селекцию на агрессивность («агрессивные» $n = 10$) и на приученность («ручные» $n = 12$). Все крысы содержались в стандартных условиях вивария ИЦиГ СО РАН. Все манипуляции с крысами проводились в соответствии с международными правилами работы с экспериментальными животными.

РНК-секвенирование образцов осуществлялось на платформе Illumina с использованием протокола NEBNext mRNA Library PrepReagent Set for Illumina (NEB, USA). Библиотеки данных РНК-секвенирования крыс находятся в открытом доступе на сайте The European Nucleotide Archive [1] под индексом ERP011250.

В работе использовали 12 образцов нервной ткани: образцы (2 повтора) из 3-х отделов головного мозга агрессивных и ручных крыс – гипоталамус, покрышка среднего мозга (ПСМ) и периакведуктум серого вещества (ПСВ). Для каждой линии крыс анализировали по два повтора. Были рассмотрены гены, где отличия значения экспрессии (fragments per kilobase of transcript per million mapped reads – FPKM) реплик определенного отдела мозга не превышали стандартного отклонения экспрессии, вычисленного по всем 12 образцам.

Метод RNA-Seq является мощным инструментом, позволяющим определить активность генов в изучаемом образце. В данном исследовании образцами являлись ткани разных отделов головного мозга крыс (гипоталамус, покрышка среднего мозга и периакведуктальное серое вещество). Образцы нервной ткани головного мозга крыс обеих линий были изъяты в соответствии со специализированной картой, представленной в Allen Mouse Brain Atlas [2]. После препарирования образцы тканей мозга крыс обрабатывались в компании «Геноаналитика» (Москва), где мРНК были экстрагированы с использованием Dynabeads mRNA Purification Kit (Ambion). Библиотеки кДНК были получены с помощью инструментального набора Illumina (США) в соответствии с протоколом производителя и направлены на секвенирование.

Параметры компьютерной фильтрации прочтений ДНК и картирования библиотек РНК-секвенирования крыс представлены в табл. 1.

Таблица 1

Статистика ридов библиотек РНК-секвенирования агрессивных и неагрессивных крыс

Номер животного в эксперименте	Линия крыс	Число ридов, млн		
		изначально	после фильтрации	картированных
2	неагрессивные	22,97	17,51	16,6
5	неагрессивные	22,01	17,13	16,32
8	агрессивные	22,5	17,33	16,56
11	агрессивные	20,15	15,41	14,68

[1] http://www.ebi.ac.uk/ena/.
[2] http://mouse.brain-map.org/static/atlas/.

Как можно видеть из табл. 1, изначально в библиотеках было по 20–22 млн прочтений ДНК (ридов). После фильтрации программой Trimmomatic было удалено около 23 % ридов. Из оставшихся после фильтрации ридов около 95 % было картировано на референсный геном крысы.

Компьютерные программы
Оценка экспрессии генов

Проведен анализ экспрессии генов крыс. Для компьютерного анализа данных РНК-секвенирования мозга крыс был задействован суперкомпьютер ССКЦ СО РАН. Фильтрованные программой Trimmomatic [6] чтения картировались программой TopHat2 [6] на референсный геном RGSC Rnor_5.0\rn5 с последующей оценкой экспрессии программой Cufflinks [8]. Уровень экспрессии генов оценивался в FPKM, что позволяет учитывать длину генов и общее количество картированных чтений ДНК. Индексы генов анализируемых белков, взаимодействующих с серотонином и дофаминов, брались из баз данных KEGG [9] и RefSeq [10].

Для обработки конечных данных секвенирования были применены программы TopHat2 (картирование на геном крысы) и Cufflinks (расчет уровней экспрессии генов) [8]. По данным экспрессии были определены дифференциально экспрессирующиеся гены.

Разработан и применен конвейер по обработке данных РНК-секвенирования образцов клеток мозга (рис. 1). Проводится оценка экспрессии генов и выявление случаев дифференциальной экспрессии между анализируемыми образцами при помощи программ Cufflinks, согласно методике, рекомендованной авторами.

Рис. 1. Компьютерный конвейер анализа данных

Составление списков геномных мутаций проводилось Samtools [11]. Случаи альтернативного сплайсинга предсказывались программой rMATS, согласно приведенной на сайте программы инструкции [12]. Проводился поиск сверхпредставленных терминов генных онтологий (GO) интернет-ресурсом DAVID [3] и GOseq по спискам дифференциально экспрессирующихся генов. Полученные списки генов, сравнивались со списками генов, связанных с агрессивностью, из литературы.

Конвейер реализован на языке программирования Bash под ОС Linux Redhat, использует многопоточность. На вход конвейера подаются данные секвенирования в формате fastq. Результат представляет собой набор файлов со следующей информацией: экспрессия генов

[3] https://david.ncifcrf.gov/.

в каждом из анализируемых образцов, дифференциально экспрессирующиеся гены, случаи альтернативного сплайсинга, сверхпредставленные GO термины, полиморфизмы в каждом из анализируемых образцов, подтвержденные случаи ассоциации генов с заданным функциональным термином и полиморфизмов.

Результаты

При помощи секвенирования транскриптов удалось выявить различие уровней экспрессии генов при сравнении выборок образцов мозга ручных и агрессивных крыс, а также дифференциальный альтернативный сплайсинг для гена Grin1 (табл. 2).

Таблица 2

Уровни экспрессии гена Grin1
в трех отделах мозга агрессивных и ручных крыс

Отдел мозга	Крысы			
	агрессивные		ручные	
	1	2	1	2
Гипоталамус	23,9613	14,7004	17,5862	12,1724
Вентральная тегментальная область	9,3537	21,3030	5,6566	7,5530
Периакведуктальное серое вещество	1,6186	4,6650	7,9608	15,5895

Согласно полученным данным, во всех рассмотренных отделах мозга прослеживается тенденция к более широкому размаху экспрессии у агрессивных крыс, чем у ручных. Наряду с этим показатели экспрессии в периакведуктальном сером веществе у агрессивных крыс значительно ниже, чем у дружелюбных, а в вентральной тегментальной области наблюдается обратная ситуация, с более высокими уровнями экспрессии у агрессивных крыс. В гипоталамусе такие явные различия не представлены. Разнородность показателей экспрессии в ВТО и ПСВ, предположительно, связана с особенностями функций данных отделов мозга.

Гипоталамус регулирует нейроэндокринную деятельность мозга и ответственен за поддержание гомеостаза всего организма за исключением функции дыхания, сердечного ритма и кровяного давления. Кроме того, гипоталамус связан с большей частью отделов центральной нервной системы. Поэтому сходные уровни экспрессии гена Grin1 у агрессивных и ручных крыс в данном отделе мозга можно объяснить отсутствием патологий работы центральной нервной системы у этих животных. Таким образом, можно предположить, что проявление агрессивного или мирного поведения затрагивает гипоталамус в меньшей степени, нежели формирования мозга, вовлеченные в систему «награждения», такие, как вентральная тегментальная область и периакведуктальное серое вещество.

Второй представленный в табл. 2 отдел мозга – вентральная тегментальная область. Она вовлечена в мезокортикальный и мезолимбический дофаминовые пути. Через ВТО проходят множественные нервные пути, ответственные за систему вознаграждения в поведении. Этот же отдел мозга участвует в формировании зависимости от никотина, героина, кокаина и метамфетамина. Отказ от перечисленных веществ вызывает тревогу, беспокойство и повышение агрессивности. По данным таблицы, экспрессия изучаемого гена в данном отделе мозга у агрессивных крыс в 1,7–2,8 раза выше, чем у ручных. На основе этого можно сделать предположение, что экспрессия гена Grin1, а следовательно, активность рецепторов NMDA и чувствительность к дофамину в вентральной тегментальной области будет выше у агрессивных крыс.

Третьим отделом мозга, показавшим в ходе анализа тенденцию к различиям в уровнях экспрессии у ручных и агрессивных крыс, было периакведуктальное серое вещество. Данный участок головного мозга расположен в непосредственной близости от вентральной тегментальной области и, помимо других функций, участвует в передаче тепловых и болевых им-

пульсов, связывая определенные участки таламуса и сенсорные окончания спинного мозга. Низкий уровень экспрессии гена, кодирующего синаптический рецептор, в данном отделе мозга может означать пониженную восприимчивость к боли и температуре окружающей среды. У агрессивных крыс, по данным табл. 2, показатель экспрессии гена Grin1 в 1,7–9,6 раза ниже, чем у дружелюбных по отношению к человеку.

Путем анализа и обработки данных об альтернативном сплайсинге гена Grin1, представленных на сайте NCBI (Reference Sequence Database, RefSeq), была составлена табл. 3.

Таблица 3

Изоформы гена Grin1 (RefSeq)

RefSeq_id	Изоформа	3'UTR	Экзон 21	Экзон 5
NM_001270606.1	5	короткая	+	+
NM_001270608.1	7	короткая	–	+
NM_001270602.1	1	длинная	+	+
NM_001270603.1	3	длинная	–	+
NM_001287423	9	короткая	+	–
NM_017010	2	длинная	+	–
NM_001270605.1	4	длинная	–	–
NM_001270610.1	8	короткая	–	–

В табл. 3 все восемь изоформ гена Grin1 отсортированы по наличию или отсутствию экзона 5. 3’UTR – это 3’ – нетранслируемая область (3’ – untranslated region), которая является участком мРНК, выполняющим регуляторные функции. Мутации, затрагивающие данный участок, способны отразиться на экспрессии многих генов за счет особенностей связывания транспортных белков с 3’UTR. Кроме того, данный отдел мРНК является заключительным (после него идет полиаденилированный конец цепочки) и терминирует синтез белка.

Согласно типам альтернативного сплайсинга (1. Альтернативная селекция промотора; 2. Альтернативный выбор сайтов расщепления / полиаденилирования; 3. Альтернативный сплайсинг с сохранением интрона; 4. Кассетный экзон (пропуск экзона)), для гена Grin1 представлены такие типы сплайсинга, как пропуск экзона и альтернативный выбор сайта полиаденилирования. Таким образом, самый короткий транскрипционный вариант данного гена составит 3 813 оснований, а самый длинный – 4 343 (по данным http://ncbi.nih.gov).

При анализе изоформ гена Grin1 [4] было установлено, что в гипоталамусе агрессивных крыс наиболее экспрессируемой является изоформа гена, которая представляет собой наиболее короткий транскрипт, а именно отсутствие экзонов 5 и 21, а также короткий вариант экзона 22. У неагрессивных (ручных) крыс эта изоформа экспрессировалась в среднем для всей выборки режиме, в то время как экспрессия изоформы с включенным экзоном 21 была повышена.

Значения экспрессии генов белков, участвующих в работе дофамина, представлены в табл. 4. Видно что, гены ферментов, участвующих в синтезе дофамина, значимо выше экспрессируются у агрессивных крыс (например, тирозин гидролаза *Th*). Экспрессия генов ферментов катаболизма дофамина незначительно выше у неагрессивных крыс. Экспрессия генов дофаминового транспортера *Slc6a3* статистически значимо выше у агрессивных крыс. Экспрессия гена дофаминового рецептора второго типа почти в два раза выше у агрессивных крыс, в то время как гены рецепторов дофамина первого, третьего, четвертого и пятого типов незначительно активнее экспрессируются у неагрессивных крыс.

Было показано, что, гены ферментов синтеза серотонина активнее экспрессируются в ВТО мозга агрессивных крыс. Значимых различий экспрессии генов ферментов катаболизма серотонина между двумя линиями крыс не обнаружено. Ген серотонинового транспортера вдвое активнее экспрессируется у агрессивных крыс, экспрессия генов рецепторов *Htr1a*, *Htr1f*, *Htr5b*, *Htr3a*, *Htr6* у них ниже. Экспрессия генов рецепторов *Htr1d*, *Htr2c*, *Htr1b*, *Htr2a*, *Htr5a*, *Htr4*, *Htr7* выше у агрессивных.

Таблица 4

Экспрессия генов ферментов метаболического пути синтеза
и катаболизма дофамина в ВТО мозга крыс

Функция	Название гена	Уровень экспрессии (FPKM)	
		Неагрессивные	Агрессивные
Синтез дофамина	Тирозин гидролаза (*Th*) NM_012740 *	3,26	26,68
	DOPA-декарбоксилаза (*Ddc*) NM_012545 *	3,61	7,95
Катаболизм	*Моноаминоксидаза* А (*MAOA*) NM_033653	40	39
	Катехол-О-метилтрансфераза (*COMT*) NM_012531	23,7	22,2
Дофаминовые рецепторы	*Drd1* NM_012546.3	1,75	1,21
	Drd2 NM_012547.1 **	4,42	8,59
	Drd3 NM_017140.2	0,17	0,17
	Drd4 NM_012944.2	0,2	0,07
	Drd5 NM_012768.1	1,82	1,44
Транспортер	Дофаминовый транспортер (*Slc6a3*) NM_012694 *	0,92	14,46

* Значимые различия с учетом поправки на множественное сравнение.
** p-value < 0,05.

Описание функций генов

Найденные гены с дифференциальной экспрессией и дифференциальным альтернативным сплайсингом представлены в разработанной компьютерной базе данных. Рассмотрим некоторые их функции. Глутамат – это аминокислота, важная для центральной нервной системы человека и других высших животных, поскольку позволяет проводить нервный импульс через специальные нейронные образования, которые называются синапсами. Рецепторы находятся на постсинаптической мембране и отвечают за соединение мембран синапса после улавливания веществ-активаторов, которые после секреции должны пройти синаптическую щель.

NMDA – рецепторы являются лиганд-зависимыми. Принцип их работы является нестандартным, поскольку, помимо необходимости улавливать один из активаторов-агонистов (глицин или глутамин) для проведения импульса необходимо присутствие ионов магния Mg^{2+}. NMDA (N-метил D-аспартат) и $GABA_A$ (гамма-аминомасляный тип А) рецепторы являются главными ингибиторными ионотропными нейротрансмиттерными рецепторами головного мозга млекопитающих. Их уникальность состоит в возможности одновременно связывать два агонистичных нейротрансмиттера – L-глутамат и глицин вместе с осуществлением заряд-зависимой блокады ионами магния путем активации не-NMDA глутаматных рецепторов в синапсе [13; 14]. Эти рецепторы главным образом представлены в синапсах центральной нервной системы [15].

Наличие возбуждающего постсинаптического потенциала (ВПСП), появляющегося при активации рецептора NMDA, также способствует повышению концентрации ионов кальция в клетке. В свою очередь, ионы кальция могут действовать в качестве вторичного мессенджера в различных сигнальных путях. Этот процесс модулируется рядом эндогенных и экзогенных соединений и играет ключевую роль в широком диапазоне как физиологических (память), так и патологических процессов.

Анализ коэкспрессии генов по данным RNA-seq (WGCNA)

Для проведения анализа WGCNA на данных RNA-seq в образцах агрессивных и ручных крыс количество ридов, выровненных на ген, было подсчитано с помощью функции featureCounts в пакете Subread. Из выборки были исключены гены, на которые не выровнялись прочтения RNA-seq. В дальнейший анализ вошло 18 000 генов. Анализ дифференциальной экспрессии осуществлялся с помощью пакета DESeq2 (пороговое значение 0,01). Дифференциальная экспрессия между образцами считалась для трех моделей: 1) генотип; 2) генотип + отдел мозга; 3) отдел мозга.

В первом случае было идентифицировано 37 генов, экспрессия которых выше у ручных крыс, и 40 генов, экспрессия которых выше у агрессивных крыс. Анализ представленности терминов Gene Ontology и путей KEGG [4] с помощью пакета limma показал, что у ручных крыс преобладают гены, участвующие в метаболизме сахаров, в то время как у агрессивных крыс – гены, участвующие в метаболизме пуринов и при окислительном стрессе. Для каждой модели получено топ 20 дифференциально экспрессирующихся генов (рис. 2, табл. 5, 6).

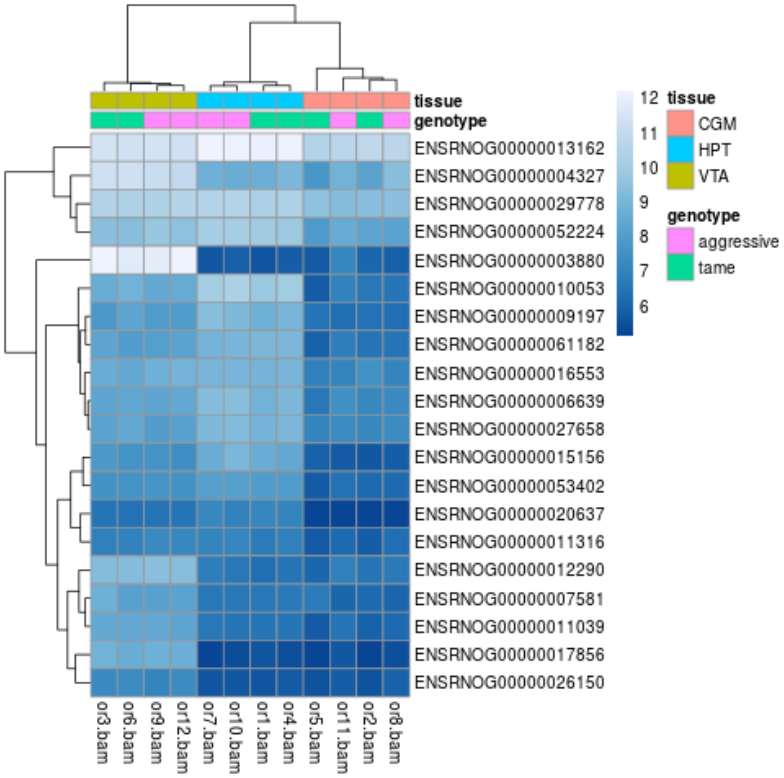


Рис. 2. Наиболее дифференциально экспрессирующиеся гены между агрессивными и ручными крысами

[4] https://www.genome.jp/kegg/.

Таблица 5

Значимые генные онтологии для генов,
экспрессия которых значимо выше либо у агрессивных, либо у ручных крыс

	Категория генной онтологии	*p*-value
Агрессивные	Catalytic activity	6,10E-04
	Transition metal ion binding	6,19E-04
	ATP-dependent DNA helicase activity	7,13E-04
	Bicarbonate binding	1,47E-03
	Carnitine O-acetyltransferase activity	1,47E-03
	Deoxyguanosine kinase activity	1,47E-03
	D-xylose 1-dehydrogenase (NADP+) activity	1,47E-03
Ручные	Monosaccharide binding	8,73E-05
	Glucose binding	0,00016
	Carbohydrate binding	0,000321
	Fructose transmembrane transporter activity	0,001467

Таблица 6

Значимые генные онтологии для генов, дифференциально экспрессирующихся
между линиями крыс с учетом отделов мозга

	Категория генной онтологии	*p*-value
Агрессивные	Regulation of ion transport	1,19E-05
	Plasma membrane part	1,40E-05
	Synaptic transmission, GABAergic	0,000178
	Main axon	0,000683
Ручные	Regulation of parathyroid hormone secretion	6,24E-06
	Oxidoreductase activity	1,25E-05
	Regulation of endocrine process	7,49E-05
	Regulation of developmental growth	0,000117
	Positive regulation of cAMP biosynthetic process	0,000122

Во втором случае было найдено всего 439 дифференциально экспрессирующихся генов, а в третьем – 383, при этом только 83 гена из второго списка не совпало с генами из третьего списка. Поэтому дифэкспрессию именно этих генов мы посчитали ассоциированной с агрессивным / толерантным поведением у крыс. Пересечения списков генов, дифференциально экспрессирующихся только по генотипу и дифференциально экспрессирующихся по генотипу и по отделам мозга, не наблюдалось. GSEA (Gene Set Enrichment Analysis) показал, что эти гены активны при синаптической и гормональной регуляции [5]. Таким образом, было выделено еще две группы генов, влияющих на агрессивное / толерантное поведение.

Анализ взвешенных сетей коэкспрессии генов осуществлялся с помощью пакета R WGCNA. После определения оптимального параметра ($\beta = 6$) алгоритм построения WGCNA был использован, чтобы преобразовать коэффициент корреляции экспрессии между генами в коэффициент смежности. Затем на основе этого коэффициента считалось расхождение в топологической матрице перекрытия (TOM).

Используя рассчитанное расхождение, мы провели иерархическую кластеризацию. В результате было выделено 63 кластера коэкспрессии с количеством генов более 20 в каждом. Средняя связность коэкспрессии генов внутри модулей была больше 0,75. Высокая связность генов внутри модулей показывает значительную биологическую релевантность полученных модулей. Поэтому мы оценили, в каких кластерах коэкспрессируются дифференциально экспрессирующиеся гены. Из первой группы 47 генов попало в модуль «darkseagreen3» и 12 –

[5] http://software.broadinstitute.org/gsea/index.jsp/.

в «brown2». Согласно GSEA, эти модули отвечают за метаболические процессы и за процессинг и связывание РНК соответственно. Большая часть группы генов, экспрессия которых изменяется под влиянием генотипа и особенностей работы отделов мозга, представлена в модулях «mediumorchid», «brown», «brown2». Модуль «mediumorchid» содержит гены, наиболее вероятно участвующие в проведении синаптических импульсов, а модуль «brown» – гены, связанные с поведением и защитными механизмами в нейронах.

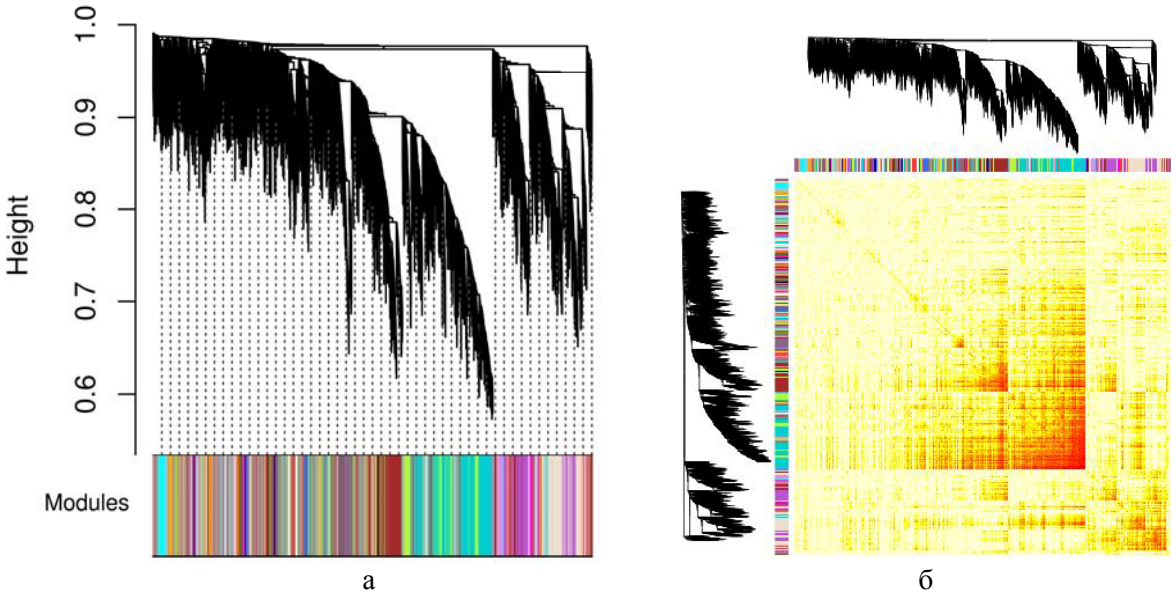


Рис. 3. Распределение генов по модулям коэкспрессии (*а*). Тепловая карта корреляций экспрессии генов (*б*). Значения корреляции варьируются от белого к красному. На карте можно различить модули, гены которых высоко коррелируют между собой

Также для двух групп генов были построены ассоциативные сети с помощью программы ANDVisio с помощью раскладки в силовом поле [16].

Субъединицы NR-2 и NR-3 имеют по несколько гомологов в результате альтернативного сплайсинга, приводящего к разнообразию продуктов транскрипции за счет выборочного удаления одного или нескольких экзонов из иРНК. Альтернативный сплайсинг, обеспечивающий разнообразие вариантов NR-2 и NR-3, обеспечивает и вариабельность работы синапса, а также может иметь определяющее значение для чувствительности рецептора, что, в свою очередь, влияет на скорость и качество проводимого через синапсы нервного импульса.

Заключение и обсуждение

Разработана база данных «ГК-АСАП (Дифференциальный альтернативный сплайсинг генов крыс, селектированных по агрессивному поведению)», которая является компьютерной технологией в области системной биологии животных [17]. База данных представляет варианты дифференциального альтернативного сплайсинга генов крыс, селекционированных на агрессивное поведение, полученных по данным экспериментов транскриптомного секвенирования. База данных разработана для экспериментаторов-селекционеров и исследователей, которые заинтересованы в получении информации о генах лабораторных животных – серых крыс, и их взаимосвязи с поведением. Представлены варианты изоформ генов с различной экспрессией у крыс с агрессивным и толерантным поведением по отношению к человеку. База данных имеет графический веб-интерфейс, который предоставляет пользователям возможность проводить поиск интересующих образцов, планировать эксперименты и обеспечивает хранение данных о различных экспериментах и образцах генов, оптимизируя научно-исследовательский процесс в области молекулярно-биологических и генетических иссле-

дований лабораторных животных. База данных содержит информацию о транскрипте, об уровнях экспрессии данного транскрипта и значимости различий у агрессивных и неагрессивных крыс. Структура базы позволяет пополнять ее данными других транскриптомных экспериментов по изучению агрессии у крыс.

База данных содержит уникальные данные об экспрессии, изоформах и дифференциальном альтернативном сплайсинге генов крыс, селектированных по агрессивному поведению, предоставляя комплементарную информацию к базе Маджин [18].

Проведено исследование кластеризации генов. Исследования [13] показали, в частности, что NMDA-рецептор не является статичной составляющей синапса, появляющейся вместе с нервной клеткой, а производится в дендритах (коротких отростках нейронов) в эндоплазматической сети, после чего через аппарат Гольджи происходит его транспорт через синаптическую мембрану. Это свидетельствует, в первую очередь, о высокой пластичности и многофункциональности такой системы.

При помощи секвенирования транскриптов удалось выявить различие уровней экспрессии генов при сравнении выборок образцов мозга ручных и агрессивных крыс, а также дифференциальный альтернативный сплайсинг для гена Grin1. При анализе изоформ гена Grin1 [4] было установлено, что в гипоталамусе агрессивных крыс наиболее экспрессируемой является изоформа гена, которая представляет собой наиболее короткий транскрипт. У неагрессивных (ручных) крыс эта изоформа экспрессировалась в среднем для всей выборки режиме, в то время как экспрессия изоформы с включенным экзоном 21 была повышена. Работа представляет продолжение исследований поведения на моделях лабораторных мышей [19; 20] и использует современные подходы биоинформатики для анализа транскриптомных данных [21; 22]. База данных экспрессии генов крыс по транскриптомным экспериментам развивает разработку базы данных RatDNA, построенную ранее по данным микрочипов [23].

Список литературы

1. *Кудрявцева Н. Н., Маркель А. Л., Орлов Ю. Л.* Агрессивное поведение: генетико-физиологические механизмы // Вавиловский журнал генетики и селекции. 2014. Т. 18 (4/3). С. 1133–1155.

2. *Orlov Yu. L., Baranova A. V., Markel A. L.* Computational models in genetics at BGRS\SB-2016: introductory note // BMC Genet. 2016. Vol. 17 (Suppl 3). P. 155. DOI: 10.1186/s12863-016-0465-3.

3. *Kulikov A. V., Bazhenova E. Y., Kulikova E. A., Fursenko D. V., Trapezova L. I., Terenina E. E., Mormede P., Popova N. K., Trapezov O. V.* Interplay between aggression, brain monoamines and fur color mutation in the American mink // Genes Brain Behav. 2016. Vol. 15 (8). P. 733–740. DOI: 10.1111/gbb.12313.

4. *Бабенко В. Н., Брагин А. О., Чадаева И. В., Маркель А. Л., Орлов Ю. Л.* Альтернативный сплайсинг в отделах головного мозга селектированных по агрессивности крыс // Молекулярная биология. 2017. Т. 51, № 5. С. 870–880. DOI: 10.7868/S0026898417050159.

5. *Брагин А. О., Сайк О. В., Чадаева И. В., Деменков П. С., Маркель А. Л., Орлов Ю. Л., Рогаев Е. И., Лаврик И. Н., Иванисенко В. А.* Роль генов апоптоза в контроле агрессивного поведения, выявленная с помощью комбинированного анализа ассоциативных генных сетей, экспрессионных и геномных данных по серым крысам с агрессивным поведением // Вавиловский журнал генетики и селекции. 2017. Т. 21 (8). С. 911–919.

6. *Bolger A. M., Lohse M., Usadel B.* Trimmomatic: a flexible trimmer for Illumina sequence data // Bioinformatics. 2014. Vol. 30 (15). P. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

7. *Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg S. L.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions // Genome biology. 2013. Vol. 14 (4). P. R36. DOI: 10.1186/gb-2013-14-4-r36.

8. *Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D. R., Pimentel H., Salzberg S. L., Rinn J. L., Pachter L.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks // Nat. Protoc. 2012. Vol. 7. No. 3. P. 562–578.

9. *Kanehisa M., Goto S.* KEGG: Kyoto encyclopedia of genes and genomes // Nucleic acids research. 2000. Vol. 28 (1). P. 27–30.

10. *Pruitt K. D., Tatusova T., Maglott D. R.* NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins // Nucleic acids research. 2006. Vol. 35 (suppl. 1). P. D61-65.

11. *Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R.* 1000 Genome Project Data Processing Subgroup. The Sequence Alignment / Map format and SAMtools // Bioinformatics. 2009. Vol. 25. No. 16. P. 2078–2079.

12. *Shen S., Park J. W., Lu Z. X., Lin L., Henry M. D., Wu Y. N., Zhou Q., Xing Y.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data // Proc. Natl. Acad. Sci. 2014. Vol. 111. No. 51. P. E5593–E5601.

13. *Wenthold R., Prybylowski K., Standley S., Sans N., Petralia R.* Trafficking of NMDA receptors // Annual Review for Pharmacological Toxicology. 2003. Vol. 43. P 335–358.

14. *Liu X.-B., Murray K., Jones E.* Switching of NMDA Receptor 2A and 2B Subunits at Thalamic and Cortical Synapses during Early Postnatal Development // The Journal of Neuroscience. 2004. Vol. 24 (40). P. 8885–8895.

15. *Furukawa H., Singh S. K., Mancusso R., Gouaux E.* Subunit arrangement and function in NMDA receptors // Nature. 2005. Vol. 438 (7065). P. 185–192.

16. *Ivanisenko V. A., Saik O. V., Ivanisenko N. V., Tiys E. S., Ivanisenko T. V., Demenkov P. S., Kolchanov N. A.* ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology // BMC Syst. Biol. 2015. Suppl. 2. P. S2. DOI 10.1186/1752-0509-9-S2-S2.

17. *Брагин А. О., Чадаева И. В., Бабенко В. Н., Богомолов А. Г., Кожемякина Р. В., Маркель А. Л., Орлов Ю. Л.* Дифференциальный альтернативный сплайсинг генов крыс, селектированных по агрессивному поведению (ГК-АСАП) / Differential alternative splicing of rat genes selected for aggressive behavior (RG-ASAB). Заявка на гос. регистрацию базы данных, июнь 2018 г., ИЦиГ СО РАН.

18. *Медведева И. В., Брагин А. О., Чадаева И. В., Маркель А. Л., Орлов Ю. Л.* База данных генов, ассоциированных с агрессивным поведением (Маджин) / Database of genes associated with aggressive behavior (Maggene). Свидетельство о государственной регистрации базы данных № 2016621672 от 16.12.2016.

19. *Коваленко И. Л., Смагин Д. А., Галямина А. Г., Орлов Ю. Л., Кудрявцева Н. Н.* Изменение экспрессии дофаминергических генов в структурах мозга самцов мышей под влиянием хронического социального стресса: данные RNA-seq // Молекулярная биология. 2016. Т. 50, № 1. С. 184–187.

20. *Smagin D. A., Kovalenko I. L., Galyamina A. G., Bragin A. O., Orlov Y. L., Kudryavtseva N. N.* Dysfunction in ribosomal gene expression in the hypothalamus and hippocampus following chronic social defeat stress in male mice as revealed by RNA-Seq // Neural Plasticity. 2016. Vol. 2016. Article ID 3289187, 6 p. DOI:10.1155/2016/3289187.

21. *Спицина А. М., Орлов Ю. Л., Подколодная Н. Н., Свичкарев А. В., Дергилев А. И., Чен М., Кучин Н. В., Черных И. Г., Глинский Б. М.* Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК // Программные системы: теория и приложения. 2016. Т. 4 (31). С. 3–20.

22. *Fedoseeva L. A., Klimov L. O., Ershov N. I., Alexandrovich Y. V., Efimov V. M., Markel A. L., Redina O. E.* Molecular determinants of the adrenal gland functioning related to stress-sensitive hypertension in ISIAH rats // BMC Genomics. 2016. Vol. 17 (Suppl 14). P. 989.

23. *Орлов Ю. Л., Брагин А. О., Медведева И. В., Гунбин И. В., Деменков П. С., Вишневский О. В., Левицкий В. Г., Ощепков В. Г., Подколодный Н. Л., Афонников Д. А., Гроссе И., Колчанов Н. А.* ICGenomics: программный комплекс анализа символьных последовательностей геномики // Вавиловский журнал генетики и селекции. 2012. Т. 16 (4/1). С. 732–741.

24. *Кожевникова О. С., Мартыщенко М. К., Генаев М. К., Корболина М. К., Муралева Н. А., Колосова Н. А., Орлов Ю. Л.* RatDNA: база данных микрочиповых исследований на крысах для генов, ассоциированных с заболеваниями старения // Вавиловский журнал генетики и селекции. 2012. Т. 16 (4/1). С. 756–765.

Материал поступил в редколлегию 01.08.2018

A. O. Bragin [1], **K. A. Tabanyukhov** [1, 2], **I. V. Chadaeva** [1, 3], **A. V. Tsukanov** [1, 3]
R. O. Babenko [3], **I. V. Medvedeva** [1], **A. G. Bogomolov** [1, 3]
V. N. Babenko [1, 3], **Yu. L. Orlov** [1, 3]

[1] *Institute of Cytology and Genetics SB RAS*
10 Academician Lavrentiev Ave., *Novosibirsk*, *630090*, *Russian Federation*

[2] *Novosibirsk State Agrarian University*
160 Dobrolyubov Str., *Novosibirsk*, *630039*, *Russian Federation*

[3] *Novosibirsk State University*
1 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

ibragim@bionet.nsc.ru, *ya.cukanton@yandex.ru*, *ichadaeva@bionet.nsc.ru*

THE COMPUTER DATABASE FOR THE ANALYSIS OF DIFFERENTIALLY EXPRESSING GENES, RELATED TO AGGRESSIVE BEHAVIOR ON MODELS OF LABORATORY ANIMALS

Development and application of computer means of the analysis of transcriptome sequencing data in model laboratory animals represents an important problem of bioinformatics. The research problem of an expression of genes on the basis of the modern methods of high-throughput sequencing in brain areas of laboratory animals is common basis for a research of genetic background of behavior. Studying of genetic determinants of aggressive behavior in general not only actual for research of molecular mechanisms of behavior regulation, but also has wide practical component for the work with animals and application in agrobiology. The genetic predisposition of animals to aggressive behavior leads to emergence of differences in a brain structure, and comparison of such distinctions allow to find both the common, and specific mechanisms of behavior regulation promoting aggression manifestation in provocative conditions of the environment. Computer programs of the gene splicing analysis, and prototype of the database of an expression of genes in brain areas of laboratory animals – gray rats, selected on manifestation of aggressive behavior are developed. We fulfilled the functional summary of over- and under-expressed genes in experiments on rats, described isoforms and the alternate splicing patterns of these genes.

Keywords: bioinformatics, transcriptomics, splicing, differential expression, laboratory animals, gene ontologies, behavior, database.

References

1. Kudryavtseva N. N., Markel A. L., Orlov Yu. L. Aggressive behavior: Genetic and physiological mechanisms. *Russian Journal of Genetics: Applied Research*, 2015, vol. 5, no. 4, p. 413–429. (in Russ.)

2. Orlov Yu. L., Baranova A. V., Markel A. L. Computational models in genetics at BGRS\SB-2016: introductory note. *BMC Genet*., 2016, vol. 17 (suppl. 3), p. 155. DOI: 10.1186/s12863-016-0465-3.

3. Kulikov A. V., Bazhenova E. Y., Kulikova E. A., Fursenko D. V., Trapezova L. I., Terenina E. E., Mormede P., Popova N. K., Trapezov O. V. Interplay between aggression, brain monoamines and fur color mutation in the American mink. *Genes Brain Behav*., 2016, vol. 15 (8), p. 733–740. DOI: 10.1111/gbb.12313.

4. Babenko V. N., Bragin A. O., Chadaeva I. V., Markel A. L., Orlov Yu. L. Differential Alternative Splicing in Brain Regions of Rats Selected for Aggressive Behavior. *Mol. Biology*, 2017, vol. 51 (5), p. 870–880. DOI: 10.7868/S0026898417050159. (in Russ.)

5. Bragin A. O., Saik O. V., Chadaeva I. V., Demenkov P. S., Markel A. L., Orlov Yu. L., Rogaev E. I., Lavrik I. N., Ivanisenko V. A. Role of apoptosis genes in aggression revealed using combined analysis of ANDSystem gene networks, expression and genomic data in grey rats with aggressive behavior. *Vavilovskii Zhurnal Genetiki i Selektsii* [*Vavilov Journal of Genetics and Breeding*], 2017, vol. 21 (8), p. 911–919. DOI: 10.18699/VJ17.312. (in Russ.)

6. Bolger A. M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, vol. 30 (15), p. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

7. Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 2013, vol. 14 (4), p. R36. DOI: 10.1186/gb-2013-14-4-r36.

8. Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D. R., Pimentel H., Salzberg S. L., Rinn J. L., Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., 2012, vol. 7, no. 3, p. 562–578.

9. Kanehisa M., Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 2000, vol. 28 (1), p. 27–30.

10. Pruitt K. D., Tatusova T., Maglott D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 2006, vol. 35 (suppl. 1), p. D61–65.

11. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009, vol. 25, no. 16, p. 2078–2079.

12. Shen S., Park J. W., Lu Z. X., Lin L., Henry M. D., Wu Y. N., Zhou Q., Xing Y. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci*., 2014, vol. 111, no. 51, p. E5593–E5601.

13. Wenthold R., Prybylowski K., Standley S., Sans N., Petralia R. Trafficking of NMDA receptors. *Annual Review for Pharmacological Toxicology*, 2003, vol. 43, p. 335–358.

14. Liu X.-B., Murray K., Jones E. Switching of NMDA Receptor 2A and 2B Subunits at Thalamic and Cortical Synapses during Early Postnatal Development. The Journal of Neuroscience, 2004, vol. 24 (40), p. 8885– 8895.

15. Furukawa H., Singh S. K., Mancusso R., Gouaux E. Subunit arrangement and function in NMDA receptors. *Nature*, 2005, vol. 438 (7065), p. 185–192.

16. Ivanisenko V. A., Saik O. V., Ivanisenko N. V., Tiys E. S., Ivanisenko T. V., Demenkov P. S., Kolchanov N. A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst. Biol*., 2015, suppl. 2. p. S2. DOI 10.1186/1752-0509-9-S2-S2.

17. Bragin A. O., Chadaeva I. V., Babenko V. N., Bogomolov A. G., Kozhemyakina R. V., Markel A. L., Orlov Yu. L. Differential alternative splicing of rat genes selected for aggressive behavior (RG-ASAB). *Application on State registration of computer database*, June 2018, Institute of Cytology and Genetics SB RAS. (in Russ.)

18. Medvedeva I. V., Bragin A. O., Chadaeva I. V., Markel A. L., Orlov Yu. L. Database of genes associated with aggressive behavior (Maggene). *Certificate of State registration of computer database* No. 2016621672 dated 16.12.2016. (in Russ.)

19. Kovalenko I. L., Smagin D. A., Galyamina A. G., Orlov Yu. L., Kudryavtseva N. N. Changes in the Expression of Dopaminergic Genes in Brain Structures of Male Mice Exposed to Chronic Social Defeat Stress: An RNA-seq Study. *Molecular Biology*, 2016, vol. 50 (1), p. 184–187. (in Russ.)

20. Smagin D. A., Kovalenko I. L., Galyamina A. G., Bragin A. O., Orlov Yu. L., Kudryavtseva N. N. Dysfunction in ribosomal gene expression in the hypothalamus and hippocampus following chronic social defeat stress in male mice as revealed by RNA-Seq. *Neural Plasticity*, 2016, vol. 2016, Article ID 3289187, 6 p. DOI:10.1155/2016/3289187.

21. Spitsina A. M., Orlov Yu. L., Podkolodnaya N. N., Svicharev A. V., Dergilev A. I., Chen M., Kuchin N. V., Chernykh I. G., Glinsky B. M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Program systems: theory and applications*, 2015, vol. 6, no.1 (24), p. 157–174. (in Russ.)

22. Fedoseeva L. A., Klimov L. O., Ershov N. I., Alexandrovich Y. V., Efimov V. M., Markel A. L., Redina O. E. Molecular determinants of the adrenal gland functioning related to stress-sensitive hypertension in ISIAH rats. *BMC Genomics*, 2016, vol. 17 (suppl. 14), p. 989.

23. Orlov Yu. L., Bragin A. O., Medvedeva I. V., Gunbin K. V., Demenkov P. S., Vishnevsky O. V., Levitsky V. G., Oshchepkov D. Y., Podkolodnyy N. L., Afonnikov D. A., Grosse I., Kolchanov N. A. ICGenomics: a program complex for analysis of symbol sequences in genomics.

Vavilovskii Zhurnal Genetiki i Selektsii [*Vavilov Journal of Genetics and Breeding*], 2012, vol. 16 (4/1), p. 732–741. (in Russ.)

24. Kozhevnikova O. S., Martyschenko M. K., Genaev M. A., Korbolina E. E., Muraleva N. A., Kolosova N. G., Orlov Yu. L. RatDNA: database on microarray studies of rats bearing genes associated with age-related diseases. *Vavilovskii Zhurnal Genetiki i Selektsii* [*Vavilov Journal of Genetics and Breeding*], 2012, vol. 16 (4/1), p. 756–765. (in Russ.)

For citation:

Bragin A. O., Tabanyukhov K. A., Chadaeva I. V., Tsukanov A. V., Babenko R. O., Medvedeva I. V., Bogomolov A. G., Babenko V. N., Orlov Yu. L. The Computer Database for the Analysis of Differentially Expressing Genes, Related to Aggressive Behavior on Models of Laboratory Animals. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 7–21. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-7-21

УДК 004.91:577.2.01
DOI 10.25205/1818-7900-2018-16-3-22-36

С. С. Ковалев [1, 2], **Е. Ю. Леберфарб** [1, 2], **Н. В. Губанова** [1], **А. О. Брагин** [1], **А. Г. Галиева** [3]
А. В. Цуканов [1, 3], **В. Н. Бабенко** [1, 3], **Ю. Л. Орлов** [1, 3]

[1] *Институт цитологии и генетики СО РАН*
пр. Академика Лаврентьева, 10, Новосибирск, 630090, Россия

[2] *Новосибирский государственный медицинский университет*
Красный проспект, 52, Новосибирск, 630091, Россия

[3] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

sergey.kovalev.1994@list.ru, orlov@bionet.nsc.ru

КОМПЬЮТЕРНЫЙ АНАЛИЗ АЛЬТЕРНАТИВНОГО СПЛАЙСИНГА ГЕНОВ В КУЛЬТУРАХ КЛЕТОК ГЛИОМ ПО ДАННЫМ RNA-SEQ [*]

Современные постгеномные методы изучения экспрессии генов с помощью транскриптомного профилирования имеют большое значение для фундаментальных биомедицинских исследований в онкологии, поиска новых маркеров развития опухолей на культурах клеток глиом. Такие эксперименты требуют разработки новых компьютерных инструментов анализа объемных данных секвенирования. Цель представленного исследования – компьютерный поиск генов и их изоформ, нарушение экспрессии которых связано с развитием глиобластом, с помощью современных высокопроизводительных технологий секвенирования транскриптом и международных биомедицинских банков данных. Поиск генов – кандидатов для терапевтического воздействия в опухолях, в том числе отдельных изоформ генов, актуален в здравоохранении и современной высокотехнологичной медицине. В данной работе представлены задачи биоинформатики, связанные с разработкой компьютерных конвейеров обработки транскриптомных данных, определения дифференциально экспрессирующихся генов, анализа альтернативного сплайсинга, описания категорий генных онтологий для найденных групп генов. Рассмотрены задачи автоматического поиска и описания функций генов в связи с раковыми заболеваниями, визуализации результатов и разработки биомедицинских баз данных. Представлен прототип базы данных дифференциального альтернативного сплайсинга генов – «Дифференциальный альтернативный сплайсинг генов человека при вторичной глиобластоме (ДАСГГ)», с возможностью работы через веб-сайт, поиска уровней экспрессии отдельных изоформ в глиальной опухоли.

Ключевые слова: биоинформатика, транскриптомика, биомедицинская информатика, глиобластома, альтернативный сплайсинг, базы данных.

Введение

Исследование экспрессии генов в клетках глиобластомы, поиск генов – кандидатов для терапевтического воздействия имеют несомненную актуальность в здравоохранении, совре-

[*] Работа была частично поддержана Министерством образования и науки РФ (28.12487.2018/12.1), разработка базы данных поддержана бюджетным проектом ИЦиГ СО РАН (№ 0324-2018-0019).

Первичный послеоперационный материал был любезно предоставлен А. Л. Кривошапкиным и А. С. Гайтаном (Национальный медицинский исследовательский центр им. академика Е. Н. Мешалкина, Новосибирск).

Авторы благодарны Геннадию Владимировичу Васильеву за организацию транскриптомного секвенирования, Ирине Вадимовне Медведевой, Роману Олеговичу Бабенко и Антону Геннадьевичу Богомолову за помощь в разработке компьютерной базы данных.

Ковалев С. С., Леберфарб Е. Ю., Губанова Н. В., Брагин А. О., Галиева А. Г., Цуканов А. В., Бабенко В. Н., Орлов Ю. Л. Компьютерный анализ альтернативного сплайсинга генов в культурах клеток глиом по данным RNA-seq // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 22–36.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2018. Том 16, № 3
© С. С. Ковалев, Е. Ю. Леберфарб, Н. В. Губанова, А. О. Брагин,
А. Г. Галиева, А. В. Цуканов, В. Н. Бабенко, Ю. Л. Орлов, 2018

менной высокотехнологичной медицине [1]. Для такого сложного объекта, как опухоли мозга, требуется проведение новых исследований, опирающихся на современные клеточные технологии, технологии высокопроизводительного секвенирования, методы современной биоинформатики, использующие интеграцию имеющейся информации из международных баз и банков данных.

Глиальные опухоли составляют большинство первичных опухолей центральной нервной системы у взрослых и включают целый спектр опухолей, различающихся по уровню клеточной дифференциации и злокачественности. Глиобластома является наиболее распространенной (60 % от всех первичных опухолей) и злокачественной (выживаемость около 1 года после постановки диагноза) первичной опухолью центральной нервной системы у взрослых [1; 2]. Глиобластома может развиваться de novo (первичная) или как конечный этап трансформации фибриллярных астроцитом (II степень злокачественности согласно классификации ВОЗ) либо анапластических астроцитом (III степень злокачественности согласно классификации ВОЗ) (вторичная глиобластома) [2]. Первичная глиобластома в большинстве случаев встречается у лиц в возрасте старше 50 лет, и для нее характерен, как правило, короткий анамнез заболевания. Вторичная глиобластома чаще развивается в возрасте до 45 лет, трансформация в глиобластому может длиться от 1 до 10 лет. Злокачественные глиальные опухоли характеризуются ярким инвазивным фенотипом, отсутствием четких границ распространения опухоли и способностью к продолжению роста после хирургического удаления.

Несмотря на анализ отдельных мутаций и генов, далеко не все транскрипты, часто встречающиеся в глиальных опухолях, функционально аннотированы. Таким образом, большую актуальность имеет поиск новых маркеров развития глиом, разработка биоинформационных методов анализа глиом, в том числе разработка компьютерных баз данных по экспериментам транскриптомного профилирования [3]. Основные пути активации генов в опухолевых клетках были изучены ранее, в настоящее время необходимо более детальное исследование отдельных изоформ транскриптов, некодирующей РНК, которое может быть сделано с помощью современных компьютерных программ на основе технологий глубокого секвенирования [3; 4].

Секвенирование полного транскриптома (RNA-seq) с целью определения генов, ответственных за рост опухоли, на культурах клеток является современным молекулярно-биологическим методом исследования, послужившим основой данной работы. Биоинформационная часть исследования связана с использованием наиболее полных международных транскриптомных баз данных и аннотаций транскриптов, анализом отдельных изоформ генов для исследования глиобластомы. В данной работе выполнен перерасчет дифференциального альтернативного сплайсинга, с помощью конвейеров анализа данных [1]. Рассмотрены отдельные изоформы генов с дифференциальной экспрессией в культурах клеток глиом и культурах здоровой ткани.

Создан прототип компьютерной базы данных с возможностью поиска уровней экспрессии отдельных изоформ в глиальной опухоли. База содержит ссылки на международные ресурсы и базы данных генетической информации (в частности OMIM). Подготовлена заявка на госрегистрацию базы данных «Дифференциальный альтернативный сплайсинг генов человека при вторичной глиобластоме (ДАСГГ)» / Database «Differential Alternative Splicing of human Genes in secondary Glioblastome (DASGG)», ИЦиГ СО РАН, Новосибирск, 2018. База данных доступна по запросу к авторам. Представленные в базе данные по дифференциальному альтернативному сплайсингу могут быть использованы в фундаментальных исследованиях по стволовым клеткам рака, а также в разработке диагностики глиом.

Материалы и методы

Используемые базы данных

Использовалась аннотация генома и транскриптома человека Ensembl [1]. Отметим, что, помимо общей аннотации генома, существуют специализированные международные базы данных по экспрессии генов в различных тканях и органах (включая микрочипы и транс-

[1] http://ensembl.org/

криптомные данные GEO NCBI [2], в том числе содержащие данные по экспрессии в клетках опухолей различных типов (The Cancer Gene Atlas [3]), экспрессии генов в компартментах мозга (Allen Brain Atlas), а также по транскриптомному профилированию опухолей, в том числе глиом и глиобластом [4]. В работе использовались общие базы данных белковых взаимодействий, такие как HPRD [5], биохимических реакций KEGG [6], Interactome [7].

Институтом Аллена разработана база Ivy Glioblastoma Atlas Project [8] по данным пациентов, страдающих глиомой. Авторами этой базы данных ранее выделено 343 гена, специфичных для глиом в различных структурах мозга. По запросу ‘Glioma’ в базе данных генов и фенотипов заболеваний OMIM (Online Mendelian Inheritance in Man) получен список генов, ассоциированных с глиомой по литературным данным [9]. Таким образом, были получены базовые списки генов человека для анализа.

Культуры клеток

В работе использовались данные транскриптомного секвенирования на клетках глиом и нормального мозга, полученные в ИЦиГ СО РАН в рамках завершенного проекта РФФИ в сотрудничестве с Национальным медицинским исследовательским центром имени академика Е. Н. Мешалкина, г. Новосибирск (см. прилож.).

Компьютерный конвейер анализа сплайсинга

Определение альтернативных вариантов изоформ генов по данным секвенирования представляет важную задачу биоинформатики. Альтернативный сплайсинг – процесс, в ходе которого экзоны генов, вырезаемые из пре-мРНК, объединяются в различных комбинациях (альтернативно), что порождает различные формы зрелой мРНК. Один ген может порождать не одну, а множество форм белка. Экзон одного варианта сплайсинга может оказаться интроном в альтернативном варианте. Поэтому молекулы мРНК, образованные в результате альтернативного сплайсинга, различаются набором экзонов, что приводит к образованию разных мРНК и, соответственно, разных белков одного первичного транскрипта. Сплайсинг гена может приводить к образованию разных изоформ белка, кодируемого этим геном [5]. На рис. 1 схематически представлены варианты сплайсинга гена и образующиеся изоформы.

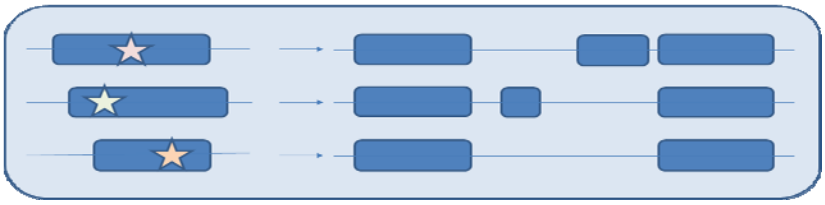


Рис. 1. Варианты сплайсинга и возможные мутации. Адаптировано из [5]

Мутации генов в опухолевых клетках могут приводить к распространению нефункциональных изоформ белка, нарушению клеточной функции и пролиферации опухоли. Ранее считалось, что только мутации в кодирующей ДНК вызывают рак. В настоящее время показано, что изменения в альтернативном сплайсинге также провоцируют заболевание [4]; развиваются соответствующие научные направления поиска специфических паттернов сплайсинга [3], исследования процессинга РНК [6].

[2] http://www.ncbi.nlm.nih.gov/geoprofiles/.
[3] cancergenome.nih.gov/.
[4] https://cghub.ucsc.edu/.
[5] http://hprd.org/.
[6] http://www.genome.jp/kegg/.
[7] http://interactome.org/.
[8] http://glioblastoma.alleninstitute.org/.
[9] http://omim.org/.

Существует несколько программ и конвейеров анализа данных сплайсинга. Нами разработан и применен конвейер по обработке данных РНК-секвенирования образцов глиом и здоровой ткани человека. Конвейер основан на использовании таких программ, как TopHat2 [7], Cuffflinks [8], rMATS [9], Samtools [10], VCFtools [11], GOseq [12], и скриптов на языках программирования Perl, R и Bash. На первом этапе проводится картирование программой TopHat2 данных РНК-секвенирования на референсный геном человека GRCh38 с учетом экзон-интронной структуры (рис. 2).

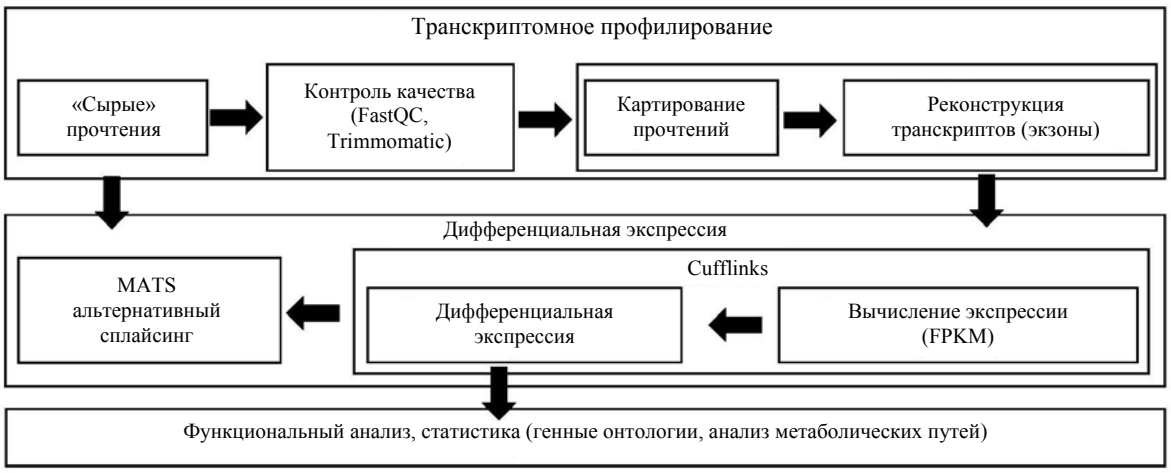


Рис. 2. Компьютерный конвейер обработки транскриптомных данных с учетом контроля качества прочтений [1]

На рис. 2 показаны этапы обработки данных – оценка качества прочтений, картирование с помощью TopHat. Далее проводится оценка экспрессии генов и выявление случаев дифференциальной экспрессии между анализируемыми образцами при помощи программ Cufflinks [8].

Составление списков геномных мутаций проводилось Samtools с последующим нахождением мутаций, встречающихся только в образцах глиомы, программами VCFTools. Найденные мутации сравнивались с мутациями глиомы из базы данных COSMIC [13]. Случаи альтернативного сплайсинга предсказывались программой rMATS согласно инструкции, приведенной на сайте программы.

Проводился поиск сверхпредставленных терминов генных онтологий (GO) интернет-ресурсом DAVID [10] и GOseq по спискам дифференциально экспрессирующихся генов. Полученные списки генов, ассоциированных с глиомой, сравнивались со списками генов, опосредующих развитие глиомы, из литературы.

Конвейер реализован на языке программирования Bash под ОС Linux Redhat, использует многопоточность. На вход конвейера подаются данные секвенирования в формате fastq. Результат представляет собой набор файлов со следующей информацией: экспрессия генов в каждом из анализируемых образцов, дифференциально экспрессирующиеся гены, случаи альтернативного сплайсинга, сверхпредставленные GO термины, полиморфизмы в каждом из анализируемых образцов, подтвержденные случаи ассоциации с глиомой дифференциально экспрессирующихся генов и полиморфизмов.

Результаты

Список дифференциально экспрессирующихся генов в глиоме

В результате работы при сравнении тканей глиомы со здоровой тканью были найдены 8 284 случая дифференциальной экспрессии (ДЭ). ДЭ гены оказались сверхпредставлены

[10] https://david.ncifcrf.gov/.

в процессах, связанных с межклеточным взаимодействием, клеточной гибелью, метаболизмом в клетке, генной экспрессией, что подтверждает связь выявленных случаев ДЭ с раковыми заболеваниями. На основе анализа базы данных OMIM и литературы выявлено 73 ДЭ гена, которые, как ранее показано, опосредуют развитие глиомы. Найдено 38 500 мутаций, специфичных глиоме, из которых более 3 000 представлены в базе данных COSMIC. Выявлено более 18 000 случаев альтернативного сплайсинга.

При сравнении тканей глиомы со здоровой тканью были найдены более восьми тысяч случаев дифференциальной экспрессии. Выявлено 73 дифференциально экспрессирующихся гена, которые опосредуют развитие глиомы.

Построена таблица категорий генных онтологий для генов со значимым дифференциальным сплайсингом в выборках клеток глиобластом и нормального мозга.

Категории генных онтологий генов с дифференциальным альтернативным сплайсингом в исследованных выборках глиом, рассчитанные с помощью инструмента DAVID

Категория	Термин	Число генов	*p*-value	Корректированное значение
GOTERM_CC_DIRECT	Extracellular exosome	26	8,3E-7	1,0E-4
GOTERM_CC_DIRECT	Focal adhesion	10	4,8E-6	2,9E-4
UP_KEYWORDS	Coiled coil	23	3,8E-5	2,4E-3
GOTERM_BP_DIRECT	Intracellular protein transport	6	7,4E-4	1,8E-1
GOTERM_CC_DIRECT	Brush border	4	9,3E-4	3,7E-2
INTERPRO	Proteinase inhibitor I2, Kunitz, conserved site	3	2,2E-3	3,4E-1
INTERPRO	Proteinase inhibitor I2, Kunitz metazoa	3	2,5E-3	2,1E-1
GOTERM_CC_DIRECT	Membrane	11	3,2E-3	9,2E-2
SMART	KU	3	3,3E-3	1,5E-1
KEGG_PATHWAY	Focal adhesion	5	4,0E-3	2,8E-1
INTERPRO	E F-Hand 1, calcium-binding site	5	4,5E-3	2,5E-1

Примечание: корректированное значение статистической значимости дано по критерию Бенджамини – Хохберга.

Таблица показывает связь категорий генных онтологий (рассчитано по DAVID для 73 генов из списка) с клеточной адгезией, мембраной, внеклеточными белками, что указывает на свойства пролиферации опухолей.

Анализ изоформ генов в глиобластоме

Для анализа участия изоформ этих генов в развитии глиом был выполнен поиск опубликованной информации в GenBank, OMIM и PubMed. Подтверждение в литературе об их роли в развитии глиомы имеют 123 гена.

В процессе анализа дифференциального сплайсинга были выявлены достоверные различия профилей сплайсинга в трех генах, связанных с возможной пролиферацией, между клетками нормального мозга и глиобластомы: белок-прекурсор амилоида бета APP (amyloid beta precursor protein), ген предрасположенности к раку CASC4 (cancer susceptibility candidate 4) и известный онкоген – транскрипционный фактор TP53. В частности, в гене TP53 наблюдалась некодирующая изоформа NR_015381 с достоверно большей частотой в клетках глиобластомы.

Короткая изоформа гена APP (NM_201413) высоко экспрессировалась в клетках глиобластомы, тогда как наиболее его длинная изоформа (NM_000484) специфически экспрессировалась в клетках нормального мозга. Для гена CASC4 его короткая изоформа (NM_177974) экспрессировалась в 6 раз выше в клетках глиобластомы, в то же время наиболее длинная изоформа (NM_138423) также имела высокий уровень экспрессии. Сверхэкспрессия CASC4 связана с повышенной экспрессией протоонкогена Her2 при раке яичника и молочной железы [11; 12].

В последние годы определены важные генетические мутации в глиомах. Ведущими мутациями в патогенезе злокачественных глиальных опухолей являются: потеря гетерозиготности (loss of heterozygosity – LOH) в длинном плече хромосомы 10 (LOH 10q), мутация гена PTEN (10q23.3), мутации в различных экзонах гена опухолевого супрессора p53, амплификация гена EGFR, делеция или инактивирующие мутации гена p16, а также гиперметилирование промотора гена MGMT. Эти мутации могут служить новым прогностическим фактором наряду с клиническими факторами прогноза и открывают новые перспективы и подходы в лечении GO [16]. Последовательное изменение генов EGFR/PTEN/Akt/mTOR является основным патогенетическим путем развития первичной глиобластомы [17]. Амплификация гена EGFR встречается в 40 % всех случаев первичных глиобластом и тесно связана с возрастом пациентов [18]. Мутация гена TP53 (p53) является основным событием, играющим роль в развитии вторичной глиобластомы [19].

Анализ роли изоформ в развитии опухолей

В ходе анализа дифференциальной экспрессии изоформ выявлены изоформы со статистически значимой разницей в экспрессии между культурой клеток нормального мозга и глиобластом. Ранее с помощью статистических подходов были выделены главные компоненты экспрессии изоформ гена APP в клетках нормального мозга и глиобластомы [1]. N-APP связывает TNFRSF21, запускающую активацию каспазы и дегенерацию тел нейрональных клеток (через каспазу-3) и аксонов (через каспазу-6). На рис. 3 представлены варианты сплайсинга гена APP.

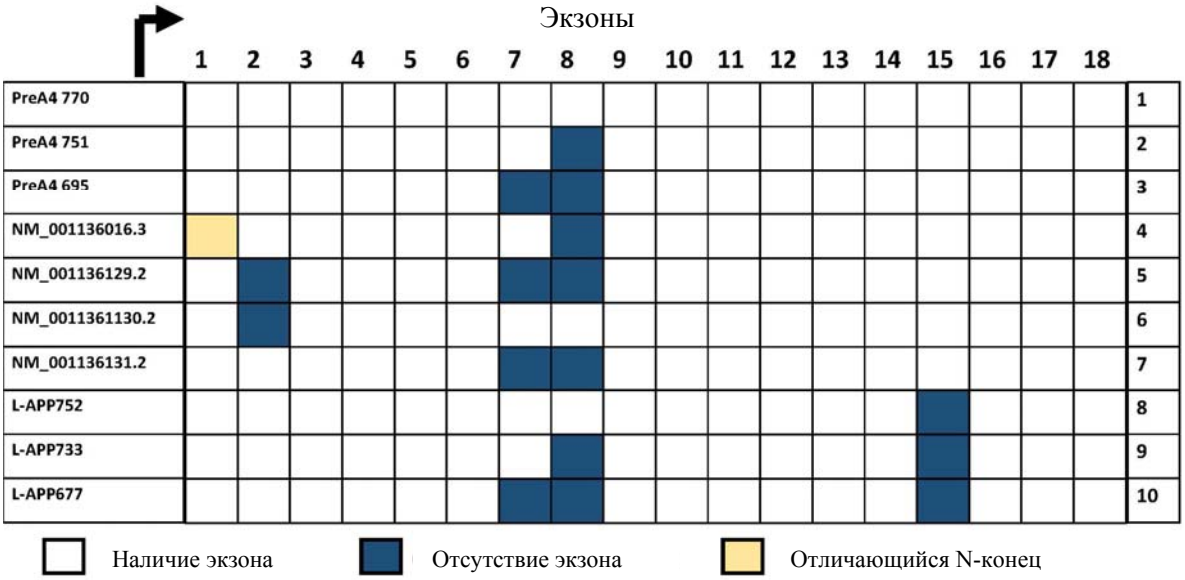


Рис. 3. Альтернативный сплайсинг гена APP

Известно десять вариантов транскрипции (см. рис. 3). Вариант транскрипции (1) представляет собой самый длинный транскрипт и кодирует самую длинную изоформу *a*, также известную как PreA4 770 (NM_000484).

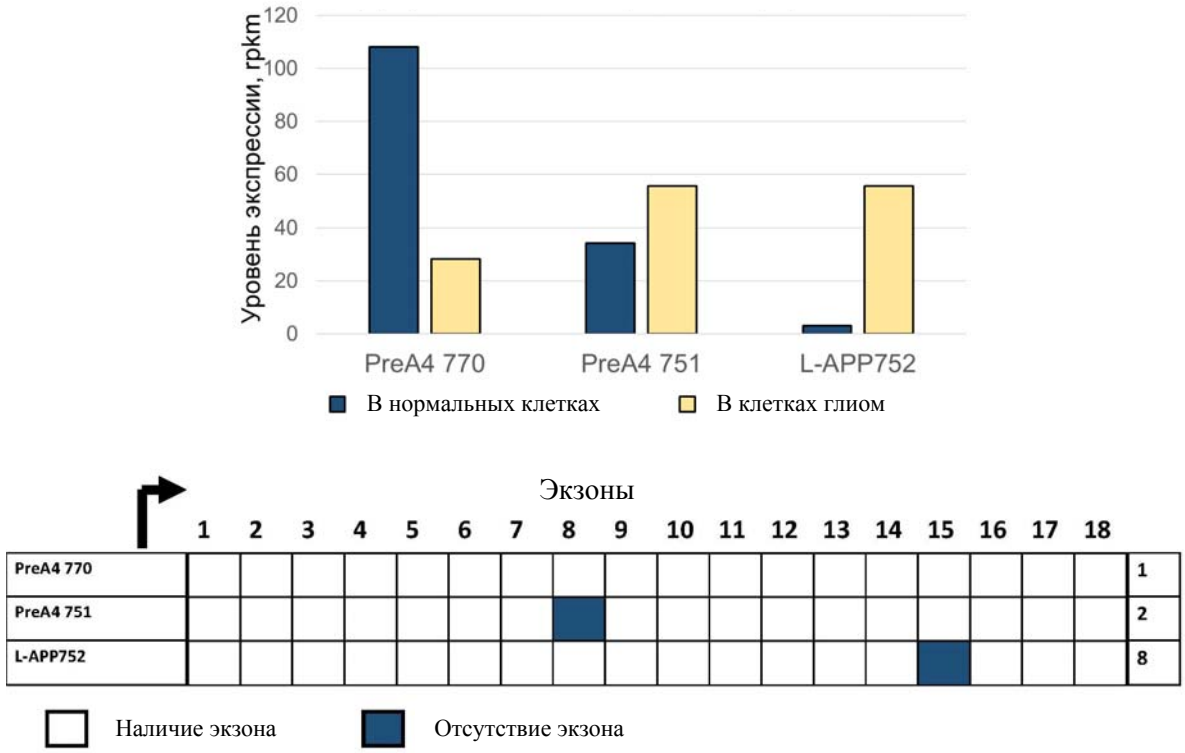


Рис. 4. Экспрессия изоформ гена APP в клетках глиом и нормальных клетках мозга (верхняя панель). Структура изоформ гена APP (нижняя панель). PreA4 751 и L-APP 752 имеют более короткие N- и C-концы и не имеют альтернативного экзона по сравнению с самой длинной изоформой PreA4 770

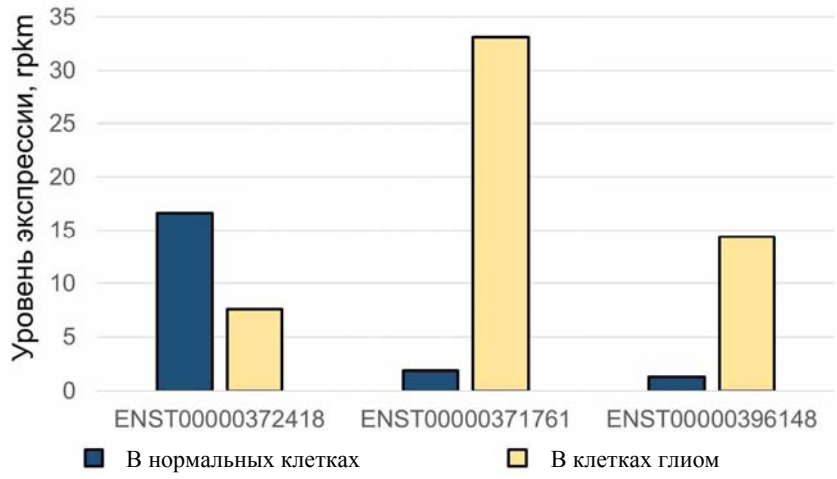


Рис. 5. Экспрессия изоформ гена CDKN2A в нормальных клетках и клетках глиом

Вариант транскрипции (2) не имеет альтернативного экзона в рамке считывания по сравнению с вариантом 1. Данная изоформа *b*, также известная как PreA4 751, имеет те же N- и C-концы, но короче по сравнению с изоформой *a*.

Вариант транскрипции (3) не имеет альтернативного внутрикадрового сегмента по сравнению с вариантом 1. Полученная изоформа *c*, также известная как PreA4 695, имеет те же N- и C-концы, но короче по сравнению с изоформой *a*.

Вариант транскрипции (4) отличается в 5 'UTR и кодирующей последовательности и не имеет альтернативного экзона в кадре по сравнению с вариантом 1. Данная изоформа *d* имеет

более короткий и отличный N-конец и не имеет внутреннего сегмента по сравнению с изоформой *a*.

Вариант транскрипции (5) не имеет трех альтернативных экзонов в той же рамке считывания по сравнению с вариантом 1. Полученная изоформа *e* имеет те же N- и C-концы, но короче по сравнению с изоформой *a*.

У варианта транскрипции (6) отсутствует альтернативный экзонов в той же рамке считывания по сравнению с вариантом 1. Полученная изоформа *f* имеет те же N- и C-концы, но короче по сравнению с изоформой *a*.

Вариант транскрипции (7) отличается в 5 'UTR и кодирующей последовательности и не имеет двух альтернативных экзонов в той же рамке считывания по сравнению с вариантом 1. Полученная изоформа *g* короче на N-конце и не имеет внутреннего сегмента по сравнению с изоформой *a*.

Вариант транскрипции (8) не имеет альтернативного экзона в кадре по сравнению с вариантом 1. Полученная изоформа *h*, также известная как L-APP752, имеет те же N- и C-концы, но короче по сравнению с изоформой *a*. Для этой транскрипции нет полной расшифровки стенограммы.

Вариант транскрипции (9) не имеет двух альтернативных экзонов в той же рамке считывания по сравнению с вариантом 1. Полученная изоформа *i*, также известная как L-APP733, имеет те же N- и C-концы, но короче по сравнению с изоформой *a*. Для этой транскрипции нет полной расшифровки стенограммы.

Вариант транскрипции (10) не имеет трех альтернативных экзонов в кадре по сравнению с вариантом 1. Полученная изоформа *j*, также известная как L-APP677, имеет те же N- и C-концы, но короче по сравнению с изоформой *a*. Для этой транскрипции нет полной расшифровки стенограммы.

В качестве примера представлены диаграммы дифференциальной экспрессии изоформ APP и CDKN2A (рис. 4 и 5).

В ходе анализа экспрессии изоформ APP выявлено, что более короткие изоформы PreA4 751 и L-APP 752 больше экспрессируются в клетках глиом, нежели каноничная изоформа PreA4 770, которая больше экспрессируется в нормальных клетках мозга.

Для гена CDKN2A характерна такая же картина, что для гена APP. Более короткие изоформы ENST00000371761 и ENST00000396148 экспрессируются выше в клетках глиом, в то время как самая длинная изоформа ENST00000372418 экспрессируется выше в нормальных клетках мозга.

Отметим, что для TP53 весь спектр (37 видимых на наших данных изоформ из 43 известных) изоформ, наблюдаемых в клетках глиобластом, отличался по уровню экспрессии от изоформ в нормальных клетках мозга. В частности, некодирующая изоформа NR_015381 относится только к клеткам глиобластом.

База данных экспрессии генов и альтернативного сплайсинга на культурах клеток глиом

В настоящее время в открытом доступе находятся данные экспериментов секвенирования в опухолевых клетках, в том числе по различным типам глиом. Опубликованы статьи и материалы, содержащие информацию о генах, специфично экспрессирующихся в раковых клетках, о профилях метилирования и т. д. Эта информация нуждается в аккумулировании для дальнейшего использования в медицине. Ранее авторским коллективом была разработана компьютерная база данных BROG [11] генов-мишеней онкогенов, определенных по данным экспериментов ChIP-seq (включая транскрипционные факторы ER, MYC, TP53). В базе была представлена разметка сайтов связывания транскрипционных факторов; данные о сайтах связывания были получены по открытым публикациям и ресурсам GEO NCBI. На рис. 6 представлен фрагмент интерфейса этой базы данных.

[11] База представлена на сайте http://lcg.nsu.ru/neuro/rffi/brog_database.html.

BROG: brain oncogenes

Gene name	Gene description	Ensembl ID	NCBI ID	Chromosome	Start	End	Strand
OR4F5	olfactory receptor, family 4, subfamily F, member 5	ENSG00000186092	NM_001005484	1	58953	59871	1
OR4F3	olfactory receptor, family 4, subfamily F, member 16	ENSG00000273547	NM_001005224	1	357521	358460	1
OR4F16	olfactory receptor, family 4, subfamily F, member 16	ENSG00000273547	NM_001005277	1	357521	358460	1
OR4F29	olfactory receptor, family 4, subfamily F, member 16	ENSG00000273547	NM_001005221	1	357521	358460	1
OR4F3	olfactory receptor, family 4, subfamily F, member 16	ENSG00000273547	NM_001005224	1	610958	611897	-1
OR4F16	olfactory receptor, family 4, subfamily F, member 16	ENSG00000273547	NM_001005277	1	610958	611897	-1
OR4F29	olfactory receptor, family 4, subfamily F, member 16	ENSG00000273547	NM_001005221	1	610958	611897	-1
SAMD11	sterile alpha motif domain containing 11	ENSG00000187634	NM_152486	1	850983	869824	1
ER site ID	**Chromosome**	**Position**	**Motif start**	**Direction**	**Affinity**	**Promoterity**	**Amplified in MCF7**
1	1	851602	851589	none	10	5Kb_ups	OK
NOC2L	NOC2-like nucleolar associated transcriptional repressor	ENSG00000188976	NM_015658	1	869445	884542	-1
KLHL17	kelch-like family member 17	ENSG00000187961	NM_198317	1	885829	890958	1
PLEKHN1	pleckstrin homology domain containing, family N member 1	ENSG00000187583	NM_032129	1	891739	900347	1
HES4	hes family bHLH transcription factor 4	ENSG00000188290	NM_021170	1	924206	925333	-1

Рис. 6. Фрагмент разработанной ранее базы данных
генов-мишеней транскрипционных факторов – BROG (Brain Oncogenes)

- Научная задача (Главная страница)
  - -Цели Проекта
  - -Методы и походы, использованные в ходе выполнения Проекта
  - -Актуальность проблемы
  - -Результаты
- Демонстрация работы
- Базы данных
- Публикации

Рис. 7. Структура сайта по проекту анализа глиом,
содержащего обновленную базу данных

Для написания собственного сайта была применена среда разработки WordPress. Сайт располагается по адресу: http://gliomaicigsbras.ru. База пополнена данными по альтернативному сплайсингу, ссылками на международные ресурсы экспрессии генов в глиомах, системой навигации. В настоящее время идет патентование (Заявка на получение свидетельства регистрации Базы данных «Дифференциальный альтернативный сплайсинг генов человека при вто-

ричной глиобластоме (ДАСГГ)» / Database «Differential Alternative Splicing of human Genes in secondary Glioblastome (DASGG)»). На рис. 7 показана структура сайта, представлены блоки содержащие информацию о научной работе, базы данных, публикации и контакты участников проекта.

Структура сайта состоит из трех блоков. Первый блок посвящен информации о научной работе. Здесь можно ознакомиться с целями и задачами научного проекта. В «Методах и подходах, использованных в ходе выполнения проекта» подробно описана методика получения первичных культур, использованных в дальнейшем для RNA-Seq. Там же есть информация о компьютерном конвейере анализа данных секвенирования транскриптом, комплексе программ на языке Java для статистического анализа расположения генов на пространственных топологических доменах и известных частях хромосом, данных по экспрессии генов. На странице «Демонстрация работы» размещены изображения, сделанные в ИЦиГ СО РАН, часть которых была опубликована. В разделе «Публикации» представлен перечень статей, опубликованных в журналах и сборниках.

В следующем блоке даны контакты участников проекта и ссылки на международные базы данных.

Далее располагается база данных, находящаяся в свободном доступе. Здесь предоставлена информация о дифференциальной экспрессии генов.

На рис. 8 представлен скриншот разработанной базы, показывающий уровни экспрессии генов в глиоме и здоровых клетках, и статистические параметры дифференциальной экспрессии. Также на фрагменте таблицы базы данных указаны идентификаторы Ensembl транскрипта, название гена, уровни экспрессии в выборках, статистическая достоверность различий.

test_id	gene_id	gene	locus	NB	NGB	log2(fold_change)	test_stat	p_value	q_value
ENST00000318602	ENSG00000175899	A2M	12:906517	11,1173	30,0908	1,43652	7,19784	5,00E-05	0,000802281
ENST00000316519	ENSG00000081760	AACS	12:125065	3,66967	1,46472	-1,32503	-3,49974	5,00E-05	0,000802281
ENST00000261772	ENSG00000090861	AARS	16:702522	20,116	13,6244	-0,562153	-2,89239	5,00E-05	0,000802281
ENST00000473553	ENSG00000008311	AASS	7:122075€	1,69523	4,99477	1,55894	3,39916	5,00E-05	0,000802281
ENST00000619387	ENSG00000275700	AATF	17:369488	18,0664	6,95821	-1,37652	-6,72055	5,00E-05	0,000802281
ENST00000629058	ENSG00000281376	ABALON	20:316644	18,0576	9,74291	-0,890185	-3,08979	5,00E-05	0,000802281
ENST00000374736	ENSG00000165029	ABCA1	9:104781(	5,54539	1,81719	-1,60958	-8,18991	5,00E-05	0,000802281
ENST00000435803	ENSG00000179869	ABCA13	7:481714!	0,69576	0,03087	-4,49433	-6,08813	5,00E-05	0,000802281
ENST00000295750	ENSG00000115657	ABCB6	2:219209?	4,61558	0,981341	-2,23369	-3,51466	5,00E-05	0,000802281
ENST00000503337	ENSG00000108846	ABCC3	17:506347	0,201968	3,07076	3,9264	3,93899	5,00E-05	0,000802281
ENST00000376887	ENSG00000125257	ABCC4	13:950198	1,26256	4,29679	1,76691	5,92342	5,00E-05	0,000802281
ENST00000427120	ENSG00000114770	ABCC5	3:1839199	0,232827	1,24171	2,41499	3,51256	5,00E-05	0,000802281
ENST00000296577	ENSG00000164163	ABCE1	4:1449671	17,9483	31,5973	0,815955	2,86213	5,00E-05	0,000802281
ENST00000222388	ENSG00000033050	ABCF2	7:151207€	4,58501	9,74052	1,08707	2,55459	5,00E-05	0,000802281

Рис. 8. Фрагмент данных (http://gliomaicgsbras.ru/базы-данных/)

Выводы и обсуждение

При помощи анализа данных современных высокопроизводительных технологий секвенирования транскриптом был выполнен компьютерный поиск генов, нарушение экспрессии которых связано с развитием глиобластом. Показана роль изоформ генов при развитии глиобластомы, дана функциональная аннотация [17–23]. Разработанный компьютерный конвейер может быть использован для решения аналогичных биомедицинских задач, основанных на обработке данных RNA-Seq [3; 4].

Исследование роли альтернативного сплайсинга при глиоме мозга на культурах клеток проводилось по данным RNA-seq, полученным в ИЦиГ СО РАН [1]. Высокопроизводитель-

ное секвенирование полного транскриптома культуры клеток (RNA-seq) с целью определения генов, ответственных за рост опухоли, является современным методом исследования, который должен шире использоваться в медицинской практике [22; 23].

Работа по поиску маркеров развития глиом имеет большую практическую значимость для медицины. База данных ДАСГГ предназначена для медиков и исследователей, которые заинтересованы в получении информации об альтернативном сплайсинге при опухолях на первичных культурах клеток. Представленные данные могут быть использованы в фундаментальных исследованиях по стволовым клеткам глиом и в разработке диагностик.

Список литературы

1. *Babenko V. N., Gubanova N. V., Bragin A. O., Chadaeva I. V., Vasiliev G. V., Medvedeva I. V., Gaytan A. S., Krivoshapkin A. L., Orlov Yu. L.* Computer Analysis of Glioma Transcriptome Profiling: Alternative Splicing Events // J. Integr. Bioinform. 2017. Vol. 14. No. 3. PII: /j/jib.2017.14.issue-3/jib-2017-0022/jib-2017-0022.xml. DOI: 10.1515/jib-2017-0022.

2. *Ohgaki H., Kleihues P.* Genetic Pathways to Primary and Secondary Glioblastoma // Am. J. Pathol. 2007. Vol. 170. No. 5. P. 1445–1453.

3. *Ferrarese R., Harsh G. R. 4th, Yadav A. K., Bug E., Maticzka D., Reichardt W., Dombrowski S. M., Miller T. E., Masilamani A. P., Dai F., Kim H., Hadler M., Scholtens D. M., Yu I. L., Beck J., Srinivasasainagendra V., Costa F., Baxan N., Pfeifer D., von Elverfeldt D., Backofen R., Weyerbrock A., Duarte C. W., He X., Prinz M., Chandler J. P., Vogel H., Chakravarti A., Rich J. N., Carro M. S., Bredel M.* Lineage-specific splicing of a brain-enriched alternative exon promotes glioblastoma progression // J. Clin. Invest. 2014. Vol. 124. No. 7. P. 2861–2876.

4. *Correa B. R., de Araujo P. R., Qiao M., Burns S. C., Chen C., Schlegel R., Agarwal S., Galante P. A., Penalva L. O.* Functional genomics analyses of RNA-binding proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma // Genome Biol. 2016. Vol. 17. No. 1. P. 125.

5. *Thorsen F., Tysnes B. B.* Brain tumor cell invasion, anatomical and biological considerations // Anticancer Res. 1997. Vol. 17. No. 6B. P. 4121–4126.

6. *Marcelino Meliso F., Hubert C. G., Favoretto Galante P. A., Penalva L. O.* RNA processing as an alternative route to attack glioblastoma // Hum. Genet. 2017. Vol. 136. No. 9. P. 1129–1141.

7. *Kim N. K., Jayatillake R. V., Spouge J. L.* NEXT-peak: A normal-exponential two-peak model for peak-calling in ChIP-seq data // BMC Genomics. 2013. Vol. 14. No. 1. P. 349.

8. *Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D. R., Pimentel H., Salzberg S. L., Rinn J. L., Pachter L.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks // Nat. Protoc. 2012. Vol. 7. No. 3. P. 562–578.

9. *Shen S., Park J. W., Lu Z. X., Lin L., Henry M. D., Wu Y. N., Zhou Q., Xing Y.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data // Proc. Natl. Acad. Sci. 2014. Vol. 111. No. 51. P. E5593–E5601.

10. *Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R.* 1000 Genome Project Data Processing Subgroup. The Sequence Alignment / Map format and SAMtools // Bioinformatics. 2009. Vol. 25. No. 16. P. 2078–2079.

11. *Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M. A., Handsaker R. E., Lunter G., Marth G. T., Sherry S. T., McVean G., Durbin R.* 1000 Genomes Project Analysis Group. The variant call format and VCFtools // Bioinformatics. 2011. Vol. 27. No. 15. P. 2156–2158.

12. *Chen C. M., Lu Y. L., Sio C. P., Wu G. C., Tzou W. S., Pai T. W.* Gene Ontology based housekeeping gene selection for RNA-seq normalization // Methods. 2014. Vol. 67. No. 3. P. 354–363.

13. *Bamford S., Dawson E., Forbes S., Clements J., Pettett R., Dogan A., Flanagan A., Teague J., Futreal P. A., Stratton M. R., Wooster R.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website // Br. J. Cancer. 2004. Vol. 91. No. 2. P. 355–358.

14. *Inoue K., Fry E. A.* Aberrant splicing of estrogen receptor, HER2, and CD44 genes in breast cancer // Genet. Epigenetics. 2015. Vol. 1. No. 7. P. 19–32.

15. *Oh J. J., Grosshans D. R., Wong S. G., Slamon D. J.* Identification of differentially expressed genes associated with HER-2/neu overexpression in human breast cancer cells // Nucleic Acids Res. 1999. Vol. 27. No. 20. P. 4008–4017.

16. *Van Meir E. G., Hadjipanayis C. G., Norden A. D., Shu H. K., Wen P. Y., Olson J. J.* Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma // CA. Cancer J. Clin. 2010. Vol. 60. No. 3. P. 166–193.

17. *Kita D., Yonekawa Y., Weller M., Ohgaki H.* PIK3CA alterations in primary (de novo) and secondary glioblastomas // Acta Neuropathol. 2007. Vol. 113. No. 3. P. 295–302.

18. *Ekstrand A. J., Sugawa N., James C. D., Collins V. P.* Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails // Proc. Natl. Acad. Sci. 1992. Vol. 89. No. 10. P. 4309–4313.

19. *Hulleman E., Helin K.* Molecular mechanisms in gliomagenesis // Adv. Cancer Res. 2005. Vol. 94. No. 1 (Suppl). P. 1–27.

20. *Gubanova N. V., Orlov Yu. L., Bragin A. O., Vasiliev G. V., Babenko V. N., Kovalev S. S., Gaytan A. S.* Transcriptome profiling of primary glioma cell cultures // Proc. of MCCMB-2017, Moscow. ISBN 978-5-901158-30-2.

21. *Брагин А. О., Губанова Н. В., Ковалев С. С., Бабенко В. Н., Орлов Ю. Л.* Анализ альтернативного сплайсинга в глиомах с помощью секвенирования транскриптом // II Междунар. науч.-практ. конф. «NGS в медицинской генетике 2017»: Тез. докл. Суздаль, 2017. С. 25.

22. *Gubanova N. V., Bragin A. O., Kovalev S. S., Medvedeva I. V., Babenko V. N., Gaytan A. S., Krivoshapkin A. L., Orlov Yu. L.* Analysis of nuclear pore complex genes in glioblastoma by transcriptome profiling // Abstracts of Symposium «Cognitive Sciences, Genomics and Bioinformatics (CSGB-2016)» / Scientific Research Institute of Physiology and Basic Medicine. Novosibirsk, 2016. P. 21. ISBN 978-5-91291-029-6.

23. *Ковалев С. С., Орлов Ю. Л., Леберфарб Е. Ю.* Дифференциальная экспрессия и альтернативный сплайсинг в культурах клеток глиом по данным RNA-seq // Авиценна-2018: Сб. тез. докл. Новосибирск: Изд-во НГМУ, 2018.

Приложение

РАБОТА С КУЛЬТУРАМИ КЛЕТОК

В работе использовались данные по первичным культурам клеток глиом и здоровых клеток мозга, полученные в ИЦиГ СО РАН (послеоперационный клеточный материал был получен в сотрудничестве с Национальным медицинским исследовательским центром имени академика Е. Н. Мешалкина, Новосибирск). Впервые сравнивались культуры злокачественной и здоровой ткани мозга, полученные в одинаковых условиях [1]. Первичный послеоперационный материал для культур клеток был любезно предоставлен А. Л. Кривошапкиным и А. С. Гайтаном (получено разрешение этического комитета Национального медицинского исследовательского центра им. академика Е. Н. Мешалкина).

Клетки культур были проанализированы методом иммуноокрашивания на маркеры: астроцитов – GFAP (Glial fibrillary acidic protein) и нейронов – beta III tubulin. Вторичная глиобластома была получена от пациента, перенесшего операцию по удалению первичной глиобластомы, а замет прошедшего курс химио- и радиотерапии. Иммуноцитохимическое окрашивание клеток культуры выявило, что все клетки культуры позитивны по нейральному и астроцитарному маркерам. Отбор материала для секвенирования производился, когда количество клеток достигало 8–10 млн, что соответствовало 2–3 пассажу в зависимости количества исходного материала.

Суммарная РНК из клеточных культур выделялась с использованием реагента Trizol (Ambion), для выделения РНК с культурального планшета удалялась среда, немедленно без промывки PBS добавлялось 2 мл Trizol, и клетки лизировались, дальнейшие процедуры проводились согласно протоколу изготовителя, для каждой культуры (здоровых клеток мозга и глиомы) использовались по 3 отдельных культуральных планшета. После выделения проводилась оценка качества РНК на биоанализаторе BA2100 набором RNA Nano.

Для создания бар-кодированных RNA-Seq библиотек было взято по 30 нг РНК, использован набор ScriptSeq™ v2 RNASeq Library Preparation Kit (Epicentre) согласно протоколу изготовителя, при амплификации библиотек использовано 12 циклов ПЦР, финальная очистка проводилась на магнитных шариках AMPure XP. Качество полученных библиотек и их молярность проверены на биоанализаторе BA2100 набором DNA High Sensitivity, перед нанесением библиотеки разбавлялись 1 : 10. Молярность 6-ти полученных библиотек находилась в пределах 78 000–157 000 pMol/l. Аликвоты полученных библиотек переданы в ЗАО «Геноаналитика», где было выполнено секвенирование на приборе Illumina HiSeq (односторонние прочтения размером 50 нт). Создано по 3 библиотеки транскриптомов глиом и здорового мозга. Полученные библиотеки секвенированы с глубиной прочтения по 20 млн прочтений ДНК каждая. Анализ внутреннего контроля ERCC Spike-In Mix показал отсутствие искажений представленности при приготовлении библиотек и секвенировании.

Материал поступил в редколлегию 18.07.2018

S. S. Kovalev [1, 2], **E. Yu. Lieberfarb** [1, 2], **N. V. Gubanova** [1], **A. O. Bragin** [1]
A. G. Galieva [3], **A. V. Tsukanov** [1, 3], **V. N. Babenko** [1, 3], **Yu. L. Orlov** [1, 3]

[1] *Institute of Cytology and Genetics SB RAS*
10 Academician Lavrentiev Ave., *Novosibirsk*, *630090*, *Russian Federation*

[2] *Novosibirsk State Medical University*
52 Krasnyi prospekt, *Novosibirsk*, *630091*, *Russian Federation*

[3] *Novosibirsk State University*
1 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

sergey.kovalev.1994@list.ru, *orlov@bionet.nsc.ru*

COMPUTER ANALYSIS OF GENE ALTERNATIVE SPLICING IN GLIOMA CELL CULTURES BY RNA-seq DATA

Fundamental biomedical research in oncology, the search for new markers of tumor development, modern post-genomic studies of gene expression on cell cultures need glioma transcriptome profiling and analysis of individual gene isoforms. Such experiments, in turn, require development of new computer tools and database for analysis of bulk sequencing data. The aim of our study is a computer search for genes and gene isoforms, the difference of their expression is associated with the development of glioblastoma. The work is based on modern high-throughput sequencing technologies and international biomedical data banks analysis. The search for candidate genes in tumors for therapeutic treatment, including individual gene isoforms, is very relevant in healthcare and modern high-tech medicine. This work presents the bioinformatics problems related to the development of computer pipelines for the processing of transcriptomic data, the revealing of the differentially expressed genes, the analysis of alternative splicing, and the description of the gene ontologies categories for the genes sets found. The tasks of automatic search and description of gene functions in connection with cancer diseases, visualization of results and development of biomedical databases are considered. A prototype database of differential alternative splicing of genes is presented, «Differential Alternative Splicing of Human Genes in Secondary Glioblastoma (DASGG)», with the ability to work through a website, to search for expression levels of individual isoforms in tumor cells.

Keywords: bioinformatics, transcriptomics, biomedical informatics, glioblastoma, alternative splicing, databases.

References

1. Babenko V. N., Gubanova N. V., Bragin A. O., Chadaeva I. V., Vasiliev G. V., Medvedeva I. V., Gaytan A. S., Krivoshapkin A. L., Orlov Yu. L. Computer Analysis of Glioma Trans-

criptome Profiling: Alternative Splicing Events. *J. Integr. Bioinform*., 2017, vol. 14, no. 3. PII: /j/jib.2017.14.issue-3/jib-2017-0022/jib-2017-0022.xml. DOI: 10.1515/jib-2017-0022.

2. Ohgaki H., Kleihues P. Genetic Pathways to Primary and Secondary Glioblastoma. *Am. J. Pathol*., 2007, vol. 170, no. 5, p. 1445–1453.

3. Ferrarese R., Harsh G. R. 4th, Yadav A. K., Bug E., Maticzka D., Reichardt W., Dombrowski S. M., Miller T. E., Masilamani A. P., Dai F., Kim H., Hadler M., Scholtens D. M., Yu I. L., Beck J., Srinivasasainagendra V., Costa F., Baxan N., Pfeifer D., von Elverfeldt D., Backofen R., Weyerbrock A., Duarte C. W., He X., Prinz M., Chandler J. P., Vogel H., Chakravarti A., Rich J. N., Carro M. S., Bredel M. Lineage-specific splicing of a brain-enriched alternative exon promotes glioblastoma progression. *J. Clin. Invest*., 2014, vol. 124, no. 7, p. 2861–2876.

4. Correa B. R., de Araujo P. R., Qiao M., Burns S. C., Chen C., Schlegel R., Agarwal S., Galante P. A., Penalva L. O. Functional genomics analyses of RNA-binding proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma. *Genome Biol*., 2016, vol. 17, no. 1, p. 125.

5. Thorsen F., Tysnes B. B. Brain tumor cell invasion, anatomical and biological considerations. *Anticancer Res*., 1997, vol. 17, no. 6B, p. 4121–4126.

6. Marcelino Meliso F., Hubert C. G., Favoretto Galante P. A., Penalva L. O. RNA processing as an alternative route to attack glioblastoma. *Hum. Genet*., 2017, vol. 136, no. 9, p. 1129–1141.

7. Kim N. K., Jayatillake R. V., Spouge J. L. NEXT-peak: A normal-exponential two-peak model for peak-calling in ChIP-seq data. *BMC Genomics*, 2013, vol. 14, no. 1, p. 349.

8. Trapnell C., Roberts A., Goff L., Pertea G., Kim D., Kelley D. R., Pimentel H., Salzberg S. L., Rinn J. L., Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., 2012, vol. 7, no. 3, p. 562–578.

9. Shen S., Park J. W., Lu Z. X., Lin L., Henry M. D., Wu Y. N., Zhou Q., Xing Y. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci*., 2014, vol. 111, no. 51, p. E5593–E5601.

10. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment / Map format and SAMtools. *Bioinformatics*, 2009, vol. 25, no. 16, p. 2078–2079.

11. Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M. A., Handsaker R. E., Lunter G., Marth G. T., Sherry S. T., McVean G., Durbin R. 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 2011, vol. 27, no. 15, p. 2156–2158.

12. Chen C. M., Lu Y. L., Sio C. P., Wu G. C., Tzou W. S., Pai T. W. Gene Ontology based housekeeping gene selection for RNA-seq normalization. *Methods*, 2014, vol. 67, no. 3, p. 354–363.

13. Bamford S., Dawson E., Forbes S., Clements J., Pettett R., Dogan A., Flanagan A., Teague J., Futreal P. A., Stratton M. R., Wooster R. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, 2004, vol. 91, no. 2, p. 355–358.

14. Inoue K., Fry E. A. Aberrant splicing of estrogen receptor, HER2, and CD44 genes in breast cancer. *Genet. Epigenetics*, 2015, vol. 1, no. 7, p. 19–32.

15. Oh J. J., Grosshans D. R., Wong S. G., Slamon D. J. Identification of differentially expressed genes associated with HER-2/neu overexpression in human breast cancer cells. *Nucleic Acids Res*., 1999, vol. 27, no. 20, p. 4008–4017.

16. Van Meir E. G., Hadjipanayis C. G., Norden A. D., Shu H. K., Wen P. Y., Olson J. J. Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma. *CA. Cancer J. Clin*., 2010, vol. 60, no. 3, p. 166–193.

17. Kita D., Yonekawa Y., Weller M., Ohgaki H. PIK3CA alterations in primary (de novo) and secondary glioblastomas. *Acta Neuropathol*., 2007, vol. 113, no. 3, p. 295–302.

18. Ekstrand A. J., Sugawa N., James C. D., Collins V. P. Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails. *Proc. Natl. Acad. Sci*., 1992, vol. 89, no. 10, p. 4309–4313.

19. Hulleman E., Helin K. Molecular mechanisms in gliomagenesis. *Adv. Cancer Res*., 2005, vol. 94, no. 1 (suppl.), p. 1–27.

20. Gubanova N. V., Orlov Yu. L., Bragin A. O., Vasiliev G. V., Babenko V. N., Kovalev S. S., Gaytan A. S. Transcriptome profiling of primary glioma cell cultures. *Proc. of MCCMB-2017*. Moscow, 2017. ISBN 978-5-901158-30-2.

21. Bragin A. O., Gubanova N. V., Kovalev S. S., Babenko V. N., Orlov Yu. L. Analysis of alternative splicing in glioma using transcriptome sequencing. II International Science-Practical Conference «NGS in medical genetics 2017». Proc. Suzdal, 2017, p. 25. (in Russ.)

22. Gubanova N. V., Bragin A. O., Kovalev S. S., Medvedeva I. V., Babenko V. N., Gaytan A. S., Krivoshapkin A. L., Orlov Yu. L. Analysis of nuclear pore complex genes in glioblastoma by transcriptome profiling. *Abstracts of Symposium «Cognitive Sciences, Genomics and Bioinformatics (CSGB-2016)»*. Scientific Research Institute of Physiology and Basic Medicine. Novosibirsk, 2016, p. 21. ISBN 978-5-91291-029-6.

23. Kovalev S. S., Orlov Yu. L., Lieberfarb E. Yu. Differential expression and alternative splicing in glioma cell culture by RNA-seq data. *Abstract book of students conference "Avicenna-2018"*. Novosibirsk, NSMU Press, 2018. (in Russ.)

For citation:

Kovalev S. S., Lieberfarb E. Yu., Gubanova N. V., Bragin A. O., Galieva A. G., Tsukanov A. V., Babenko V. N., Orlov Yu. L. Computer Analysis of Gene Alternative Splicing in Glioma Cell Cultures by RNA-seq Data. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 22–36. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-22-36

УДК 004.942
DOI 10.25205/1818-7900-2018-16-3-37-50

Н. Л. Подколодный [1, 2], **Д. А. Гаврилов** [3], **Н. Н. Твердохлеб** [1, 3], **О. А. Подколодная** [1]

[1] *Институт цитологии и генетики СО РАН*
пр. Академика Лаврентьева, 10, Новосибирск, 630090, Россия

[2] *Институт вычислительной математики и математической геофизики СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

[3] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

pnl@bionet.nsc.ru

CYTOSCAPE – ПЛАГИН ДЛЯ ПОСТРОЕНИЯ СТРУКТУРНЫХ МОДЕЛЕЙ БИОЛОГИЧЕСКИХ СЕТЕЙ В ВИДЕ СЛУЧАЙНЫХ ГРАФОВ [*]

Современные методы экспериментальных исследований позволяют реконструировать различного типа биологические сети, включая генные и метаболические сети, сети интерактомики, сети коэкспрессии генов, сети заболеваний и т. д. В данной статье представлена разработанная нами система построения структурных моделей биологических сетей в виде набора случайных графов, структурные закономерности которых совпадают со структурными закономерностями исходной биологической сети. Такие структурные модели могут быть использованы для проверки различных статистических гипотез на сетях, в исследовании влияния структурных закономерностей в биологических сетях на их функцию и других задачах.

При генерации структурных моделей в случайных графах могут быть зафиксированы следующие характеристики: распределение степеней вершин, попарное распределение степеней вершин, средняя степень соседних вершин, коэффициент кластеризации, спектр кластеризации, частота структурных мотивов различных размеров и др. Разработанная система построена по архитектуре клиент-сервер и состоит из плагина Cytoscape и удаленного вычислительного сервиса. Взаимодействие между клиентом и сервером реализовано посредством фреймворка gRPC с применением протокола сериализации структурированных данных Protocol Buffers. Система позволяет асинхронно конструировать структурные модели заданных биологических сетей в виде случайных графов посредством программ Random Network Generator и GTrie Scanner. Результат построения может быть загружен для визуализации и анализа средствами пакета Cytoscape. С использованием разработанной системы проведен вычислительный эксперимент по реконструкции структурных моделей ряда биологических сетей, для которых удалось построить алгоритм предсказания времени расчетов структурных моделей.

Ключевые слова: биологические сети, случайные графы, структурные модели, Cytoscape.

Введение

Современные методы экспериментальных исследований в молекулярной биологии позволяют реконструировать различного типа биологические сети, включая генные сети, сети метаболических и сигнальных путей, сети интерактомики (сети взаимодействий белков, РНК,

[*] Разработка программного обеспечения выполнена в рамках государственного задания ИВМиМГ СО РАН по проекту № 0315-2016-0005. Подготовка данных для вычислительного эксперимента и расчеты с использованием вычислительных ресурсов ЦКП «Биоинформатика» проведены при поддержке бюджетного проекта № 0324-2018-0017.

Подколодный Н. Л., *Гаврилов Д. А.*, *Твердохлеб Н. Н.*, *Подколодная О. А.* Cytoscape – плагин для построения структурных моделей биологических сетей в виде случайных графов // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 37–50.

© Н. Л. Подколодный, Д. А. Гаврилов, Н. Н. Твердохлеб, О. А. Подколодная, 2018

ДНК), сети коэкспрессии генов, сети ассоциации заболеваний и генетических мутаций, сети кровеносных и лимфатических сосудов, сети контактов в макромолекулах, ассоциативные и семантические сети и др. [1–11]. Сетевое представление сложных биологических систем и процессов – достаточно продуктивный подход, который в настоящее время получает широкое распространение. Активно развиваются новые методы анализа сетей для решения различных задач, включая выявление ключевых элементов, структурных мотивов и кластеров в биологических сетях, приоритизацию на этой основе генов, белков, мутаций, поиск биомаркеров заболеваний и т. д. [2; 4; 6; 11–13].

Для проверки различных статистических гипотез на сетях в качестве «нулевых гипотез», как правило, используются различные модели случайных графов, которые могут быть описаны распределением вероятностей на графах или случайным процессом, который генерирует случайные графы.

Модели случайных графов, заданные структурные характеристики которых совпадают со структурными характеристиками анализируемой реальной биологической сети, можно рассматривать как структурные модели этой биологической сети. Исследование структурных моделей реальных биологических сетей позволяет выявлять внутренние зависимости между характеристиками сети. Для выявления таких зависимостей осуществляется генерация большого числа случайных графов, имеющих некоторую характеристику Y, и проверка гипотезы, что с большой вероятностью эти графы также имеют другую характеристику X.

Структурные модели позволяют также исследовать, от каких характеристик зависит целевой или функциональный признак биологической сети. Все структурные модели являются эквивалентными с точностью до структурных характеристик, по которым построены эти модели. Если имеется возможность оценки целевого или функционального признака биологической сети по ее структуре, то анализ структурных моделей позволяет выявить структурные закономерности в биологической сети, ответственные за проявление ее функций или целевых свойств. Например, сеть капиллярных сосудов характеризуется большой сложностью и может быть описана сетью, от структуры которой зависит кровоток [9; 10]. Другим примером такого рода целевых свойств могут быть робастность функционирования биологической сети (сохранение функций при удалении вершины сети или ребра) [14; 15].

Структурно-функциональные закономерности организации генной сети и сети белок-белковых взаимодействий могут быть важны для выделения главных компонент молекулярной системы и использоваться для построения математических моделей генных сетей или молекулярно-генетических систем.

Целью настоящей работы является разработка плагина-приложения системы Cytoscape для реконструкции различного уровня приближения структурных моделей исследуемых биологических сетей в виде случайных графов. Мы сосредоточились на одном из типов биологических сетей, описываемых неориентированным графом. К такого рода биологическим сетям относятся сети интерактомики, в частности сети белок-белковых взаимодействий (protein interaction network – PIN). Исследования в этой области привели к осознанию того, что топологические особенности сетей молекулярных взаимодействий могут дать новые знания об их природе и функциональных характеристиках, участвующих в этих взаимодействиях биологических макромолекул.

В статье представлены методы генерации структурных моделей сетей в виде случайных графов и описание распределенной системы генерации структурных моделей, реализованной с использованием клиент-серверной архитектуры (клиент-плагин Cytoscape и вычислительный сервис для реконструкции структурных моделей в виде случайных графов). В заключение представлены результаты вычислительного эксперимента по реконструкции структурных моделей реальных биологических сетей.

Случайные графы и методы генерации моделей

Случайные графы

Случайный граф – это граф, являющийся результатом случайного выбора из некоторого множества графов (генеральной совокупности графов) в соответствии с заданным на этом

множестве вероятностным распределением. Случайные графы можно описать как распределением вероятности на множестве графов, так и случайным процессом, создающим эти графы. Наиболее простыми и распространенными моделями случайных графов являются модели Эрдёша – Реньи [16] и Гилберта [17].

Для определения модели случайных графов Эрдёша и Реньи рассмотрим вероятностное пространство $G_{n,m}$ всех $\binom{N}{M}$ графов, имеющих n вершин и M ребер, где $N = \binom{n}{2}$ – максимальное число всех возможных ребер между n вершинами. Пусть $G(n,M)$ – случайный элемент пространства $G_{n,m}$ (т. е. $G(n,M) \in G_{n,m}$). Будем считать, что все случайные элементы пространства $G_{n,m}$ равновероятны с $p = 1 \Big/ \binom{N}{M}$. Таким образом, заданное вероятностное пространство случайных графов имеет четко выраженную интерпретацию. Другая классическая модель случайных графов введена Гилбертом одновременно с моделью Эрдёша и Реньи в 1959 г. [17]. В модели Гилберта для определения вероятностного пространства графов $G_{n,p}$ задается матрица инцидентности графа в виде массива случайных переменные $\{X_{ij} : 1 \le i < j \le 1\}$, для которых $\Pr(X_{ij} = 1) = p$ и $\Pr(X_{ij} = 0) = 1 - p$. Случайный граф $G(n,p)$, в котором вершины i и j связаны, если случайная величина $X_{ij} = 1$, является случайным элементом вероятностного пространства $G_{n,p}$ (т. е. $G(n,p) \in G_{n,m}$). Вероятность получения случайного графа, имеющего n вершин и M ребер, равна $p^M (1-p)^{\binom{n}{2} - M}$.

Для $M \simeq pN$, где N – максимальное число возможных ребер, эти две наиболее широко используемые модели $G(n,M)$ и $G(n,p)$ почти взаимозаменяемы [18]. Поэтому многие теоретические результаты получены с использованием модели Гилберта. Однако такие простые модели случайных графов пригодны в качестве «нулевых моделей» далеко не для всех статистических гипотез. Возможным обобщением модели случайных графов Эрдёша и Реньи могут быть модели, в которых вероятности случайных элементов вероятностного пространства $G_{n,m}$ различны. Например, часто возникает необходимость использовать модели случайных графов, сохраняющих распределение степеней вершин [19; 20] или другие структурные характеристики, наблюдаемые в реальной биологической сети. Такие модели случайных графов можно рассматривать как структурные модели биологической сети.

Методы рандомизации биологической сети с использованием Марковских цепей

Пусть задана биологическая сеть, которую можно представить в виде простого неориентированного графа $G = (V,E)$, где V – это конечное множество вершин, а E – совокупность пар (v_i, v_j), где $v_i, v_j \in V$. При этом в графе G нет петель, кратных и ориентированных ребер.

В данной работе используются следующие численные характеристики графов.

1 $d(v)$ – степень вершины v или число ребер вершины v.

2 Распределение вероятности степеней вершин графа задается вектором f, где $f(d)$ – вероятность того, что случайно выбранная вершина в графе будет иметь степень d.

3 Совместное распределение вероятности степеней вершин определяется матрицей J_{nn}, где $J(i,j)$ – вероятность того, что случайно выбранное ребро объединяет вершины со степенью i и j.

4 Коэффициент кластеризации $C_i = \frac{2n_i}{d(v_i)(d(v_i)+1)}$, где n_i – число связей, соединяющих всех соседей вершины v_i, иными словами, C_i есть вероятность того, что два ближайших соседа этой вершины сами есть ближайшие соседи.

5 Коэффициент ассортативности $r = \frac{M\sum_{i\in E} j_i k_i - \left(\sum_{i\in E}(j_i + k_i)/2\right)^2}{M\sum_{i\in E}\left(j_i^2 + k_i^2\right)/2 - \left(\sum_{i\in E}(j_i + k_i)/2\right)^2}$ – это коэффициент корреляции Пирсона $(0 \le r \le 1)$ между степенями связанных вершин, где j_i и k_i – степень вершин на концах *i*-го ребра $(i \in E)$, *M* – число ребер в графе. Для ассортативных сетей $(r > 0)$ вершины с большим числом ребер чаще связаны с такого же типа вершинами, а вершины с небольшим числом связей, в свою очередь, чаще связаны с вершинами с небольшой степенью, для дисассортативных сетей $(r < 0)$ вершины с большим количеством связей чаще связаны с вершинами с небольшим количеством связей [21].

6 Спектр кластеризации $c(k)$ – вероятность того, что две вершины, соседних с вершиной степени *k*, будут соседями [22].

7 Структурный мотив – неслучайный подграф, частота появления которого в сети значимо выше, чем по случайным причинам [23; 24].

Одним из подходов для получения случайного графа является рандомизация графа путем перестановки вершин в случайно выбранных парах ребер. В реализованной нами программе RNGmotifs для генерации случайных графов, сохраняющих заданные структурные характеристики, наблюдаемые в исходной биологической сети, используется метод рандомизации сети путем парной перестановки вершин в 2-х случайно выбранных ребрах графа. Процесс рандомизации соответствует Марковской цепи, которая стартует с $G_0 = G,\ E_0 = E$.

Для рандомизации графа с сохранением распределения степеней вершин на каждом *k* шаге выполняется выбор равновероятно двух непересекающихся ребер графа (v_1, v_2) и (v_3, v_4), где $v_1 \# v_3,\ v_1 \# v_4,\ v_2 \# v_3$ и $v_2 \# v_4$. Далее делается равновероятно одна замена из двух вариантов замен ребер в графе:

$$E_k \leftarrow E_{k-1} \cup \{(v_1, v_4), (v_3, v_2)\} / \{(v_1, v_2), (v_3, v_4)\}$$

или

$$E_k \leftarrow E_{k-1} \cup \{(v_1, v_3), (v_2, v_4)\} / \{(v_1, v_2), (v_3, v_4)\}.$$

Если (V, E_k) простой граф (отсутствие петель и параллельных рёбер), то с заданной вероятностью *p* принимается $G_k = (V, E_k)$, иначе $G_k = G_{k-1},\ E_k = E_{k-1}$.

В работе [25] показано, что если размер (n, M) двух графов и распределение степеней вершин совпадает, то один граф может быть трансформирован в другой с помощью конечного числа шагов данной Марковской цепи.

В процессе рандомизации кроме явно задаваемых критериев отбора графов (ограничений) используются разные алгоритмы реализации Марковских цепочек, которые обеспечивают получение равновероятно всех возможных графов $G(n, M) \in G_{n,m}$, которые сохраняют либо распределение степеней вершин, либо совместное распределение степеней вершин.

Для генерации случайных графов с заданными ограничениями можно задавать вероятность замены ребер, зависящую от функционала отклонения модели на очередном шаге Марковского процесса от исходной биологической сети по заданным структурным характеристикам, которые необходимо сохранить в модели, например: коэффициент кластеризации, спектр кластеризации, частота структурных мотивов и т. д.

Одним из наиболее распространенных является алгоритм, подобный оптимизации методом имитации отжига, когда вероятность отбора соответствует распределению $p = \exp\left(-\frac{W}{T}\right)$, где W – функционал отклонения модели от исходной биологической сети, который можно интерпретировать как энергию системы при некоторой температуре T.

Например, в реализованной нами программе RNGmotifs для генерации случайных графов с частотой структурных мотивов, совпадающей с характеристиками исходной биологической сети, используется $W = \sum_{i=1}^{s} \frac{\left(w_{i1} - w_{i0}\right)^2}{s\left(w_{i1} + w_{i0}\right)^2}$, где w_{i1}, w_{i0} – частоты наблюдения мотивов в биологической сети и модели соответственно, s – число мотивов в ограничении. Предполагается, что процесс протекает при постепенно понижающейся температуре $T(k)$ с увеличением шага k Марковской цепи. В этом случае в начале Марковского процесса используется более мягкий критерий отбора (принимаются варианты обмена рёбер в графе, которые могут привести к отклонению значения функционала качества), но далее критерий становится более жестким, что обеспечивает, в конечном счете, сходимость структурных характеристик модели к характеристикам исходного графа. Такой вариант управления сходимостью функционала качества обусловлен необходимостью выйти из возможного локального минимума функционала и обеспечить возможность достижимости всех графов, имеющих заданные характеристики.

Для отбора статистически значимых мотивов в качестве «нулевой» гипотезы могут быть использованы случайные графы, сохраняющие распределение степеней вершин. При другом подходе, предлагаемом нами, используется пошаговый алгоритм, когда для отбора мотивов степени $k+1$ в качестве нулевой гипотезы используются случайные графы, сохраняющие не только распределение степеней вершин, но и частоты неслучайных структурных мотивов размером k. В этом случае структурная модель строится последовательно, начиная от мотивов размером 3 и далее, пошагово включая в структурную модель мотивы большего размера.

Для генерации структурных моделей нами также используется подход, основанный на так называемых *dk*-сериях, реализованный в программной библиотеке Random Network Generator [26]. *dk*-серия является вложенной структурой: каждый следующий уровень (d + 1)k-распределения содержит тот же объем информации об исходном графе, что и *dk*-распределение, но при этом также предоставляет о нем дополнительные сведения посредством включения в список ограничений более строгих правил генерации. Так, нулевой элемент последовательности, 0k-распределение, фиксирует самую грубую из возможных характеристик графа – среднюю степень вершин, что дает наиболее слабое из правил генерации. Следующий элемент, 1k-распределение, сохраняет распределение степеней вершин – более строгое условие, чем сохранение средней степени вершин. 2k-распределение сохраняет уже совместное распределение степеней вершин, т. е. число подграфов размером 2 – другими словами, рёбер – между вершинами со степенями k_1 и k_2. Таким образом, 2k-распределение обозначает попарную корреляцию степеней вершин, а также коэффициент ассортативности графа. 3k-распределение сохраняет 3-совместное распределение степеней вершин, т. е. подграфов в виде замкнутого треугольника и клик, состоящих из трех вершин со степенями k_1, k_2 и k_3, что определяет коэффициент кластеризации, и т. д. [20; 26].

Здесь используются два варианта рандомизации графа: (1) описанный выше алгоритм рандомизации графа с сохранением распределения степеней вершин и (2) алгоритм рандомизации графа с сохранением совместного распределения степеней вершин.

В последнем варианте на каждом k шаге Марковского процесса выполняются следующие операции.

1. Выбираем равновероятно одно ребро графа (v_1, v_2). Далее выбираем равновероятно одну вершину в этом ребре. Пусть это будет v_2.

2. Выбираем равновероятно другую вершину графа v_3 такая, что $d(v_2)=d(v_3)$. Далее выбирается равновероятно соседняя вершина v_4, инцидентная v_3. Это дает второе ребро (v_3,v_4).

3. Замена рёбер в графе $E_k \leftarrow E_{k-1} \cup \{(v_1,v_3),(v_2,v_4)\} / \{(v_1,v_2),(v_3,v_4)\}$.

4. Если (V,E_k) – простой граф, то с заданной вероятностью p принимается $G_k=(V,E_k)$, иначе $G_k=G_{k-1}$, $E_k=E_{k-1}$.

Такая процедура позволяет выбирать равномерно случайно все возможные элементы $G(n,M)\in G_{n,m}$, имеющие заданное распределение степеней рёбер и совместное распределение степеней вершин [27].

Для определения критерия остановки используют различные подходы. Однако теоретические оценки дают слишком большие значения числа шагов, что практически трудно реализуемо для больших графов. В работе [28] предложен более практичный критерий остановки, который находится на основе сходимости распределения вероятностей ряда структурных характеристик графа (глобальный коэффициент кластеризации, диаметр графа, максимальное собственное значение Лапласиана графа) к стационарному распределению, т. е. если распределение структурных характеристик графа практически не меняется при достижении некотором числа шагов, то процесс можно останавливать. В результате рекомендовано использовать число шагов Марковского процесса $N=\alpha * M$, где M – число рёбер в графе, а $5<\alpha<30$.

Говоря о возможностях генерации случайных графов в целом, следует отметить, что Random Network Generator поддерживает работу с *dk*-серией вплоть до уровня 2*k*, включая его расширения в виде 2.1*k* и 2.5*k* [26]. Эти уровни являются частными расширениями 2*k*, полученными в результате добавления правил генерации в виде сохранения коэффициента кластеризации и спектра кластеризации исходного графа соответственно. Таким образом, Random Network Generator может применяться для пошаговой реконструкции структурных моделей согласно *dk*-серии.

Программная реализация системы

Система генерации структурных моделей построена по архитектуре «клиент-сервер», где в качестве клиента выступает плагин-приложение системы визуализации графов Cytoscape [29–31], а в качестве сервера – удаленный вычислительный сервис с модульным подключением программных средств построения структурных моделей.

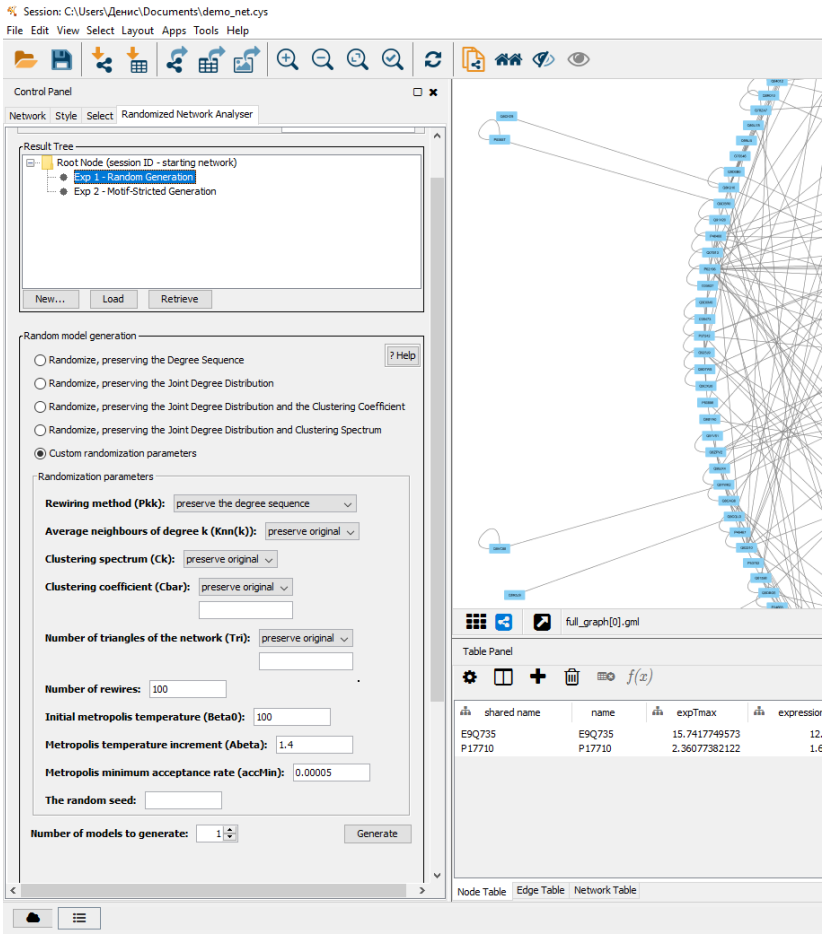
Система Cytoscape [1] – это программная платформа с открытым исходным кодом для визуализации и анализа сложных биологических сетей с возможностью интеграции разного типа биоинформационных данных, таких как функциональная аннотация генов, уровень экспрессии генов и пр. Важной особенностью программной архитектуры Cytoscape является технология расширения функциональности системы путем подключения дополнительных модулей (плагинов), созданных на языке Java сторонними разработчиками. Большая часть плагинов (более 300) доступна на Cytoscape App Store [2], и их можно устанавливать с помощью менеджера приложений App Manager [32].

Большая часть этих плагинов Cytoscape разработаны для решения различных задач биоинформатики: в частности, загрузка биологических сетей из доступных баз данных, реконструкция биологических сетей на основе интеграции гетерогенной информации, визуализация, сравнение и анализ сетей, выявление функциональных модулей в сети, генерация атрибутов узлов сети с использованием доступных баз биологических данных, например базы данных по экспрессии генов, GO (Gene Ontology) аннотации и т. д. Система Cytoscape при реализации проекта дает возможность использовать этот функционал как для получения биологических сетей, так и для анализа структурных моделей.

[1] http://www.cytoscape.org/.
[2] http://apps.cytoscape.org/.

В рамках разрабатываемой системы, плагин-приложение является «тонким» клиентом, т. е. весь функционал по обработке загружаемых сетей ложится на удаленный вычислительный сервис. Плагин-приложение Cytoscape реализовано на языке Java, вычислительный сервис реализован на языке GoLang [33]. Взаимодействие между клиентом и сервером реализовано посредством фреймворка gRPC [34; 35] с применением протокола сериализации структурированных данных Protocol Buffers [36].

При разработке интерфейса использовались основные компоненты библиотеки Java Swing: JPanel, JScrollPane и JTree. В качестве основного элемента интерфейса используется древовидное представление текущей сессии работы с удаленным сервисом. В зависимости от выбранного пользователем элемента дерева, ему предлагается различный набор доступных опций. Например, при выборе узла верхнего уровня, соответствующего одному из доступных методов обработки исходной сети, пользователю предоставляется панель настройки параметров запускаемого алгоритма (см. рисунок).



Скриншот интерфейса системы реконструкции структурных моделей биологических сетей

Разработанная нами система позволяет асинхронно конструировать структурные модели заданных биологических сетей в виде случайных графов посредством программных библиотек Random Network Generator [26] и RNGmotifs, разработанной в ИЦиГ СО РАН на основе модификации пакета GTrie Scanner [37].

Пользовательский интерфейс плагина Cytoscape обеспечивает загрузку биологической сети для анализа, формирование запроса на удаленный вычислительный сервер для реконструкции различных структурных моделей в соответствии с их спецификацией, визуализацию реконструированных структурных моделей и их сравнительный анализ в пакете Cytoscape.

Вычислительный эксперимент: реконструкция и анализ структурных моделей биологических сетей

Целью вычислительного эксперимента является исследование того, как изменяется время вычислений при генерации различных структурных моделей в зависимости от структурных характеристик исходных биологических сетей.

В качестве исходных данных были выбраны следующие биологические сети:

1) сеть взаимодействия заболеваний человека ($N = 1\,419$, $M = 2\,738$), построенная на основе данных об известных ассоциациях «заболевание – ген» и указывает на общее генетическое происхождение многих заболеваний [38];

2) сеть белок-белковых взаимодействий у дрожжей ($N = 2\,361$, $M = 5\,375$) [39];

3) сеть белок-белковых взаимодействий в печени мыши для белков с циркадным изменением скорости трансляции ($N = 5\,753$, $M = 98\,813$) [40];

4) сеть белок-белковых взаимодействий в печени мыши для белков с циркадным изменением скорости трансляции и повышенной скоростью трансляции в начале суток ($T = 0$) ($N = 2\,702$, $M = 25\,893$) [40].

Сеть белок-белковых взаимодействий в печени мыши (PPI) была построена на основе базы данных IID (Integrated Interactions Database, версия от 2017-04) [41]. Далее было выбрано подмножество белков **Pc**, скорость трансляции которых в печени мыши имеет выраженное изменение в течение суток. С использованием этого подмножества взаимодействующих белков с циркадным изменением скорости трансляции была сформирована сеть белок-белковых взаимодействий № 3 (PIN_c) [40]. Если из белков **Pc**, гены которых имеют выраженную суточную динамику трансляции, выбрать для каждого момента времени суток подмножества белков со скоростью трансляции в это время больше, чем среднесуточное значение, то мы получаем множество сетей белок-белковых взаимодействий или динамическую сеть белок-белковых взаимодействий $PIN_c(T)$, зависящую от времени суток T [40]. Таким образом сформирована сеть № 4 $PIN_c(T0)$, которая является подмножеством сети № 3. Исследование циркадных изменений динамической сети белок белковых взаимодействий $PIN_c(T)$ имеет большое значение для выявления главных компонент структуры математической модели циркадного осциллятора.

Используя разработанный нами Cytoscape плагин и реконструированную сеть была проведена генерация структурных случайных моделей различного уровня точности: dk1.0 (сохранение распределения степеней вершин), dk2.0 (сохранение совместного распределения степеней вершин, т. е. распределения и корреляций степеней вершин), dk2.1 (сохранение совместного распределения степеней вершин и коэффициента кластеризации) и dk2.5 (сохранение совместного распределения степеней вершин и спектра кластеризации, т. е. вероятности того, что два узла, соседних с узлом степени k, будут соседями).

Расчеты выполнялись на рабочей станции HP Z800 Xeon 2x X5570QC 2.93ГГц. Сравнительные результаты времени расчетов указанных биологических сетей представлены в таблице.

Время вычислений реконструкции моделей биологических сетей (T/T^*, с) и предсказание времени вычислений по структурным характеристикам сети

Модель	Качество предсказания времени расчета R^2/p-value	Сеть взаимодействия заболеваний человека	Сеть белок-белковых взаимодействий у дрожжей	Сеть белок-белковых взаимодействий в печени мыши	Сеть белок-белковых взаимодействий в печени мыши в начале суток
dk1.0	0,981/0,01	0,05/0,06	0,1/0,075	1/1,38	6/4,83
dk2.0	0.981/0.004	0,3/0,4	1/0,6	150/213	3240/2580
dk2.1	1/0.008	11,6/11,4	75/75.4	10510/11030	227000/219240
dk2.5	0.996/0.002	10,9/13,1	27/20,2	5879/8067	130000/105160

Как показывают результаты вычислительного эксперимента, скорость сходимости может сильно различаться даже на сетях с сопоставимым количеством вершин и рёбер.

Для исследования зависимости времени расчетов от структуры сети нами были оценены различные характеристики сети: размеры сети, распределение степеней вершин, коэффициенты кластеризации и ассортативности. Анализ показал, что распределение степеней вершин $f(d)$ во всех использованных в вычислительном эксперименте сетях имеет зависимость $f(d) = b * e^{-a*\sqrt{d}}$, где a и b – параметры распределения, которые были оценены для каждой сети с коэффициентом детерминации от 0,87 до 0,92 и уровнем значимости p-value $< 10^{-22}$.

Далее для предсказания времени вычисления каждой модели мы использовали метод пошаговой регрессии. Наиболее информативным показателем, учитывающим влияние структуры сети на логарифм времени вычислений, оказались характеристики распределения степеней вершин, в частности параметр a. В таблице представлены результаты предсказания времени расчета с указанием уровня достоверности и коэффициента детерминации (квадрат коэффициента корреляции (R^2) между предсказанным и истинным значением времени расчета).

Предсказание времени выполнения запросов для вычисления структурных моделей биологических сетей важно для планирования вычислительных экспериментов. Безусловно, полученные результаты можно использовать только для сетей, имеющих указанную зависимость распределения степеней вершин. Тем не менее предварительные результаты показали перспективность такого рода исследования для этого класса биологических сетей. При этом накопление вариантов проведенных расчетов моделей для различных биологических сетей такого рода могут автоматически использоваться для уточнения предсказания.

Заключение

Разработана система построения структурных моделей биологических сетей в виде набора случайных графов, структурные закономерности которых совпадают со структурными закономерностями исходной биологической сети. При генерации структурных моделей в случайных графах могут быть зафиксированы следующие характеристики: распределение степеней вершин, попарное распределение степеней вершин, средняя степень соседних вершин, коэффициент кластеризации, спектр кластеризации, частота заданных структурных мотивов различных размеров и т. д.

Система построена по архитектуре «клиент-сервер» и состоит из плагина-приложения Cytoscape и удаленного вычислительного сервиса. Взаимодействие между клиентом и сервером реализовано посредством фреймворка gRPC с применением протокола сериализации структурированных данных Protocol Buffers.

Система позволяет асинхронно конструировать структурные модели заданных биологических сетей в виде случайных графов посредством программных библиотек Random Network Generator и RNGmotifs, разработанной в ИЦиГ СО РАН на основе модификации пакета GTrie Scanner.

Пользовательский интерфейс плагина Cytoscape обеспечивает загрузку биологической сети для анализа, формирование запроса на удаленный вычислительный сервер для реконструкции различных структурных моделей в соответствии с их спецификацией, визуализацию реконструированных структурных моделей и их сравнительный анализ в пакете Cytoscape.

С использованием разработанной системы проведен вычислительный эксперимент по реконструкции структурных моделей ряда биологических сетей, для которых удалось построить алгоритм предсказания времени расчетов структурных моделей.

Список литературы

1. *Alm E., Arkin A. P.* Biological networks // Current opinion in structural biology. 2003. Vol. 13. No. 2. P. 193–202.

2. *Gosak M., Markovič R., Dolenšek J., Slak Rupnik M., Marhl M., Stožer A., Perc M.* Network science of biological systems at different scales: A review // Physics of Life Reviews. 2018. Vol. 24. P. 118–135.

3. *Ananko E. A., Podkolodnyy N. L., Stepanenko I. L., Podkolodnaya O. A., Rasskazov D. A., Miginsky D. S., Likhoshvai V. A., Ratushny A. V., Podkolodnaya N. N., Kolchanov N. A.* GeneNet in 2005 // Nucleic Acids Res. 2005. Vol. 33. P. D425–D427.

4. *Vella D., Zoppis I., Mauri G., Mauri P., Di Silvestre D.* From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data // EURASIP Journal on Bioinformatics and Systems Biology. 2017. Vol. 6.

5. *Stuart J. M., Segal E., Koller D., Kim S. K.* A gene-coexpression network for global discovery of conserved genetic modules // Science. 2003. Vol. 302 (5643). P. 249–255.

6. *Podkolodnaya O. A., Podkolodnaya N. N., Podkolodnyy N. L.* The mammalian circadian clock: Gene regulatory network and computer analysis // Russian Journal of Genetics: Applied Research. 2015. Vol. 5. No. 4. P. 354–362.

7. *Emamjomeh A., Robat E. S., Zahiri J., Solouki M., Khosravi P.* Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data // Plant Biotechnol. Rep. 2017. Vol. 11. P. 71–86.

8. *Hu J. X., Thomas C. E., Brunak S.* Network biology concepts in complex disease comorbidities // Nature Reviews Genetics. 2016. Vol. 17. P. 615–629.

9. *Novkovic M., Onder L., Bocharov G., Ludewig B.* Graph Theory-Based Analysis of the Lymph Node Fibroblastic Reticular Cell Network // Methods Mol. Biol. 2017. Vol. 1591. P. 43–57. DOI: 10.1007/978-1-4939-6931-9_4.

10. *Grebennikov D., Loon R. van, Novkovic M., Onder L., Savinkov R., Sazonov I., Tretyakova R., Watson D. J., Bocharov G.* Critical Issues in Modeling Lymph Node Physiology // Computation. 2017. Vol. 5 (3).

11. *Ivanisenko V. A., Saik O. V., Ivanisenko N. V., Tiys E. S., Ivanisenko T. V., Demenkov P. S., Kolchanov N. A.* ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology // BMC Syst. Biol. 2015. Vol. 9. Suppl. 2:S2. DOI: 10.1186/1752-0509-9-S2-S2.

12. *Luo T., Wu S., Shen X., Li L.* Network cluster analysis of protein-protein interaction network identified biomarker for early onset colorectal cancer // Mol. Biol. Rep. 2013. Vol. 40 (12). P. 6561–6568.

13. *Yan W., Xue W., Chen J., Hu G.* Biological Networks for Cancer Candidate Biomarkers Discovery // Cancer Informatics. 2016. Vol. 15 (S3). P. 1–7. DOI: 10.4137/CIN.S39458.

14. *Albert R., Jeong H., Barabasi A.-L.* Error and attack tolerance of complex networks // Nature. 2000. Vol. 406. P. 378–382.

15. *Truong C.-D., Kwon Y.-K.* Investigation on changes of modularity and robustness by edge-removal mutations in signaling networks // BMC Systems Biology. 2017. Vol. 11 (Suppl. 7). P. 125.

16. *Erdős P., Rényi A.* On random graphs I // Publications Mathematicae Debrecen. 1959. Vol. 6. P. 290–297.

17. *Gilbert E. N.* Random graphs // Annals of Mathematical Statistics. 1959. Vol. 30. P. 1141–1144. DOI:10.1214/aoms/1177706098.

18. *Bollobas B., Riordan O. M.* Mathematical results on scale-free random graphs // Handbook of Graphs and Networks. 1st ed. Eds. S. Bornholdt, H. G. Schuster. Wiley VCH, Weinheim, 2003. P. 1–34.

19. *Orsini C., Dankulov M. M., Colomer-de-Simón P., Jamakovic A., Mahadevan P., Vahdat A., Bassler K. E., Toroczkai Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D.* Quantifying randomness in real networks // Nature Communications. 2015. Vol. 6. P. 8627.

20. *Orsini C., Mitrovic D. M., Jamakovic A., Mahadevan P., Colomer-de-Simon P., Vahdat A., Bassler K., Toroczkai Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D.* How random are complex networks // CoRR. 2015. abs/1505.07503.

21. *Newman M. E. J.* Assortative Mixing in Networks // Phys. Rev. Lett. 2002. Vol. 89. P. 208701.

22. *Nobari S. et al.* Fast random graph generation // Proc. of the 14th International conference on extending database technology. ACM, 2011. P. 331–342.

23. *Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U.* Network motifs: simple building blocks of complex networks // Science. 2002. Vol. 298. P. 824–827.

24. *Jamakovic A. et al.* How small are building blocks of complex networks // arXiv preprint arXiv:0908. 1143. – 2009.

25. *Taylor R.* Constrained switching in graphs // SIAM J. Algebraic Discrete Meth. 1982. Vol. 3 (1). P. 115–121.

26. *Simon P. C. de.* RandNetGen: a Random Network Generator. URL: http://polcolomer.github.io/RandNetGen/ (дата обращения 25.05.2018).

27. *Stanton I., Pinar A.* Constructing and sampling graphs with a prescribed joint degree distribution // ACM J. Exp. Algor. 2012. Vol. 17 (3). Article 3.5 (August 2012). 25 p.

28. *Ray J., Pinar A., Seshadhri C.* A stopping criterion for Markov chains when generating independent random graphs // Journal of Complex Networks. 2015. Vol. 3 (2). P. 204–220.

29. *Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., Ideker T.* Cytoscape: a software environment for integrated models of biomolecular interaction networks // Genome Res. 2003. Nov. Vol. 13 (11). P. 2498–2504.

30. *Killcoyne S. et al.* Cytoscape: a community-based framework for network modeling // Protein Networks and Pathway Analysis. 2009. P. 219–239.

31. *Saito R. et al.* A travel guide to Cytoscape plugins // Nature Methods. 2012. Vol. 9. No. 11. P. 1069–1076.

32. *Lotia S., Montojo J., Dong Y., Bader G. D., Pico A. R.* Cytoscape app store // Bioinformatics. 2013. Vol. 29 (10). P. 1350–1351.

33. *Доновaн А. А., Керниган Б. У.* Язык программирования Go: Пер. с англ. М.: ИД Вильямс, 2016. 432 с.

34. *Cerami E.* Web services essentials: distributed applications with XML-RPC, SOAP, UDDI & WSDL. O'Reilly Media, Inc., 2002. 304 p.

35. *Seymour K. et al.* Overview of GridRPC: A remote procedure call API for grid computing // International Workshop on Grid Computing. Springer, Berlin, Heidelberg, 2002. P. 274–278.

36. *Varda K.* Protocol buffers: Google's data interchange format // Google Open Source Blog, Jul. 2008. URL: https://opensource.googleblog.com/2008/07/protocol-buffers-googles-data.html/. (дата обращения 20.07.2018)

37. *Ribeiro P.* gtrieScanner – Quick Discovery of Network Motifs // CRACS & INESC-TEC, DCC/FCUP. URL: http://www.dcc.fc.up.pt/gtries/ (дата обращения 25.05.2018).

38. *Goh K. I. et al.* The human disease network // Proc. of the National Academy of Sciences. 2007. Vol. 104. No. 21. P. 8685–8690.

39. *Bu D. et al.* Topological structure analysis of the protein-protein interaction network in budding yeast // Nucleic acids research. 2003. Vol. 31. No. 9. P. 2443–2450.

40. *Подколодный Н. Л., Твердохлеб Н. Н., Подколодная О. А.* Анализ циркадного ритма биологических процессов в печени и почках мыши // Вавиловский журнал генетики и селекции. 2017. Т. 21, № 8. С. 903–910.

41. *Kotlyar M. et al.* Integrated Interactions Database: Tissue-specific view of the human and model organism interactomes // Nucleic Acids Res. URL: http://dcv.uhnres.utoronto.ca/iid/ (дата обращения 20.07.2018).

Материал поступил в редколлегию 13.06.2018

N. L. Podkolodnyy [1, 2], **D. A. Gavrilov** [3], **N. N. Tverdokhleb** [1, 3], **O. A. Podkolodnaya** [1]

[1] *Institute of Cytology and Genetics SB RAS*
10 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[2] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[3] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

pnl@bionet.nsc.ru

CYTOSCAPE PLUGIN FOR RECONSTRUCTION OF STRUCTURAL RANDOM GRAPH MODELS OF BIOLOGICAL NETWORKS

Modern experimental technologies in molecular biology allow reconstructing different types of biological networks, including gene and metabolic networks, networks of interatomic, gene co-expression networks, a network of diseases, etc. This article presents the program tool for reconstructing structural random graph models of biological networks, the structural regularities of which coincide with the structural regularities of the initial biological network. Such structural models can be used to test various statistical hypotheses on networks, to study the influence of structural regularities in biological networks on their function, and so on. Our tool generate the structural random graph models with the following fixed characteristics: the distribution of vertex degrees, the joint distribution of degrees of vertices, the average degree of neighboring vertices, the clustering coefficient, the clustering spectrum, the frequency of structural motifs of various sizes, etc.

The developed system is based on the client-server architecture and consists of the Cytoscape plug-application and remote computing service. The interaction between the client and the server is implemented through the gRPC framework using the Protocol Buffers (structured data serialization protocol).

The system allows to construct the structural random graph models of the given biological networks asynchronously through software Random Network Generator and GTrie Scanner. The result structural model can be loaded for visualization and analysis using the Cytoscape package. This article also presents the computational experiment for reconstruct the structural random graph models of a number of biological networks. The algorithm for estimating the time of calculations of structural models of this kind of biological networks was constructed.

Keywords: biological networks, random graphs, structural models, Cytoscape.

References

1. Alm E., Arkin A. P. Biological networks. *Current opinion in structural biology*, 2003, vol. 13, no. 2, p. 193–202.

2. Gosak M., Markovič R., Dolenšek J., Slak Rupnik M., Marhl M., Stožer A., Perc M. Network science of biological systems at different scales: A review. Physics of Life Reviews, 2018, vol. 24, p. 118–135.

3. Ananko E. A., Podkolodnyy N. L., Stepanenko I. L., Podkolodnaya O. A., Rasskazov D. A., Miginsky D. S., Likhoshvai V. A., Ratushny A. V., Podkolodnaya N. N., Kolchanov N. A. GeneNet in 2005. *Nucleic Acids Res*., 2005, vol. 33, p. D425–D427.

4. Vella D., Zoppis I., Mauri G., Mauri P., Di Silvestre D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2017, vol. 6.

5. Stuart J. M., Segal E., Koller D., Kim S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 2003, vol. 302 (5643), p. 249–255.

6. Podkolodnaya O. A., Podkolodnaya N. N., Podkolodnyy N. L. The mammalian circadian clock: Gene regulatory network and computer analysis. *Russian Journal of Genetics: Applied Research*, 2015, vol. 5, no. 4, p. 354–362.

7. Emamjomeh A., Robat E. S., Zahiri J., Solouki M., Khosravi P. Gene co-expression network reconstruction: a review on computational methods for inferring functional information from plant-based expression data. *Plant Biotechnol. Rep*., 2017, vol. 11, p. 71–86.

8. Hu J. X., Thomas C. E., Brunak S. Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 2016, vol. 17, p. 615–629.

9. Novkovic M., Onder L., Bocharov G., Ludewig B. Graph Theory-Based Analysis of the Lymph Node Fibroblastic Reticular Cell Network. *Methods Mol. Biol*., 2017, vol. 1591, p. 43–57. DOI: 10.1007/978-1-4939-6931-9_4.

10. Grebennikov D., Loon R. van, Novkovic M., Onder L., Savinkov R., Sazonov I., Tretyakova R., Watson D. J., Bocharov G. Critical Issues in Modeling Lymph Node Physiology. *Computation*, 2017, vol. 5 (3).

11. Ivanisenko V. A., Saik O. V., Ivanisenko N. V., Tiys E. S., Ivanisenko T. V., Demenkov P. S., Kolchanov N. A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst. Biol*., 2015, vol. 9, suppl. 2:S2. DOI: 10.1186/1752-0509-9-S2-S2.

12. Luo T., Wu S., Shen X., Li L. Network cluster analysis of protein-protein interaction network identified biomarker for early onset colorectal cancer. *Mol. Biol. Rep*., 2013, vol. 40 (12), p. 6561–6568.

13. Yan W., Xue W., Chen J., Hu G. Biological Networks for Cancer Candidate Biomarkers Discovery. *Cancer Informatics*, 2016, vol. 15 (S3), p. 1–7. DOI: 10.4137/CIN.S39458.

14. Albert R., Jeong H., Barabasi A.-L. Error and attack tolerance of complex networks. *Nature*, 2000, vol. 406, p. 378–382.

15. Truong C.-D., Kwon Y.-K. Investigation on changes of modularity and robustness by edge-removal mutations in signaling networks. *BMC Systems Biology*, 2017, vol. 11 (suppl. 7), p. 125.

16. Erdős P., Rényi A. On random graphs I. *Publications Mathematicae Debrecen*, 1959, vol. 6, p. 290–297.

17. Gilbert E. N. Random graphs. *Annals of Mathematical Statistics*, 1959, vol. 30, p. 1141–1144. DOI:10.1214/aoms/1177706098.

18. Bollobas B., Riordan O. M. Mathematical results on scale-free random graphs. In: Handbook of Graphs and Networks. 1st ed. Eds. S. Bornholdt, H. G. Schuster. Wiley VCH, Weinheim, 2003, p. 1–34.

19. Orsini C., Dankulov M. M., Colomer-de-Simón P., Jamakovic A., Mahadevan P., Vahdat A., Bassler K. E., Toroczkai Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D. Quantifying randomness in real networks. *Nature Communications*, 2015, vol. 6, p. 8627.

20. Orsini C., Mitrovic D. M., Jamakovic A., Mahadevan P., Colomer-de-Simon P., Vahdat A., Bassler K., Toroczkai Z., Boguñá M., Caldarelli G., Fortunato S., Krioukov D. How random are complex networks. *CoRR*, 2015, abs/1505.07503.

21. Newman M. E. J. Assortative Mixing in Networks. *Phys. Rev. Lett*., 2002, vol. 89, p. 208701.

22. Nobari S. et al. Fast random graph generation. *Proc. of the 14th International conference on extending database technology. ACM*, 2011, p. 331–342.

23. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U. Network motifs: simple building blocks of complex networks. *Science*, 2002, vol. 298, p. 824–827.

24. Jamakovic A. et al. How small are building blocks of complex networks. *arXiv preprint arXiv:0908*. 1143. – 2009.

25. Taylor R. Constrained switching in graphs. *SIAM J. Algebraic Discrete Meth*., 1982, vol. 3 (1), p. 115–121.

26. Simon P. C. de. RandNetGen: a Random Network Generator. URL: http://polcolomer.github.io/RandNetGen/ (дата обращения 25.05.2018).

27. Stanton I., Pinar A. Constructing and sampling graphs with a prescribed joint degree distribution. *ACM J. Exp. Algor*., 2012, vol. 17 (3), article 3.5 (August 2012). 25 p.

28. Ray J., Pinar A., Seshadhri C. A stopping criterion for Markov chains when generating independent random graphs. *Journal of Complex Networks*, 2015, vol. 3 (2), p. 204–220.

29. Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., 2003, Nov., vol. 13 (11), p. 2498–2504.

30. Killcoyne S. et al. Cytoscape: a community-based framework for network modeling. *Protein Networks and Pathway Analysis*, 2009, p. 219–239.

31. Saito R. et al. A travel guide to Cytoscape plugins. *Nature Methods*, 2012, vol. 9, no. 11, p. 1069–1076.

32. Lotia S., Montojo J., Dong Y., Bader G. D., Pico A. R. Cytoscape app store. *Bioinformatics*, 2013, vol. 29 (10), p. 1350–1351.

33. Donovan A. A., Kernigan B. U. Yazyk programmirovaniya Go. Transl. from Engl. Moscow, Viliyams Publ., 2016, 432 p. (in Russ.)

34. Cerami E. Web services essentials: distributed applications with XML-RPC, SOAP, UDDI & WSDL. O'Reilly Media, Inc., 2002, 304 p.

35. Seymour K. et al. Overview of GridRPC: A remote procedure call API for grid computing. *International Workshop on Grid Computing*. Springer, Berlin, Heidelberg, 2002, p. 274–278.

36. Varda K. Protocol buffers: Google's data interchange format. *Google Open Source Blog*, Jul. 2008. URL: https://opensource.googleblog.com/2008/07/protocol-buffers-googles-data.html/. (accessed 20.07.2018)

37. Ribeiro P. gtrieScanner – Quick Discovery of Network Motifs. *CRACS & INESC-TEC, DCC/FCUP*. URL: http://www.dcc.fc.up.pt/gtries/ (accessed 25.05.2018).

38. Goh K. I. et al. The human disease network. *Proc. of the National Academy of Sciences*, 2007, vol. 104, no. 21, p. 8685–8690.

39. Bu D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic acids research*, 2003, vol. 31, no. 9, p. 2443–2450.

40. Podkolodnyy N. L., Tverdokhleb N. N., Podkolodnaya O. A. Analiz tsirkadnogo ritma biologicheskikh protsessov v pecheni i pochkakh myshi. *Vavilovskiy zhurnal genetiki i selektsii*, 2017, vol. 21, no. 8, p. 903–910. (in Russ.)

41. Kotlyar M. et al. Integrated Interactions Database: Tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res*. URL: http://dcv.uhnres.utoronto.ca/iid/ (accessed 20.07.2018).

For citation:

Podkolodnyy N. L., Gavrilov D. A., Tverdokhleb N. N., Podkolodnaya O. A. Cytoscape Plugin for Reconstruction of Structural Random Graph Models of Biological Networks. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 37–50. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-37-50

УДК 004.9:57.087:577.21
DOI 10.25205/1818-7900-2018-16-3-51-63

А. В. Цуканов [1, 2], **Н. Г. Орлова** [3], **А. И. Дергилев** [2], **Ю. Л. Орлов** [1, 2]

[1] *Институт цитологии и генетики СО РАН*
пр. Академика Лаврентьева, 10, Новосибирск, 630090, Россия

[2] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[3] *Новосибирский государственный архитектурно-строительный университет* (*Сибстрин*)
ул. Ленинградская, 113, Новосибирск, 630008, Россия

ya.cukanton@yandex.ru, orlovanina2@mail.ru, arturd1993@yandex.ru, orlov@bionet.nsc.ru

ПРОГРАММЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ, КЛАСТЕРИЗАЦИИ И ВИЗУАЛИЗАЦИИ РАСПРЕДЕЛЕНИЯ САЙТОВ СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ В ГЕНОМЕ *

Исследование регуляции транскрипции генов на основе данных современных технологий высокопроизводительного секвенирования является актуальной задачей биоинформатики, требующей развития новых компьютерных средств, в том числе на основе суперкомпьютерных вычислений. Рассмотрены задачи обработки данных полногеномных профилей ChIP-seq связывания транскрипционных факторов в геномах, определения пиков профилей и поиска сайтов связывания в нуклеотидных последовательностях таких пиков. Разработаны программы для анализа положения сайтов связывания в геноме относительно районов генов, расчета кластеров таких сайтов и визуализации их положения в геноме. Рассчитаны кластеры сайтов связывания транскрипционных факторов в геноме человека по базе данных Cistrome, построены матрицы совместной встречаемости пар сайтов связывания различных транскрипционных факторов в геноме для различных типов тканей и культур клеток. Проведен вычислительный эксперимент по компьютерной генерации случайных кластеров в геноме, а также оценке встречаемости кластеров большого размера для экспериментально полученных сайтов связывания транскрипционных факторов в геноме человека. Найдены закономерности встречаемости сайтов факторов плюрипотентности в эмбриональных стволовых клетках. Разработанное программное обеспечение доступно по запросу к авторам.

Ключевые слова: геномика, секвенирование, сайты связывания транскрипционных факторов, промоторы, большие данные, статистика, визуализация.

Введение

Изучение структурно-функциональной организации генома на основе данных высокопроизводительного секвенирования ДНК продолжает оставаться магистральным направлением, развивающимся на стыке биологии и информационных технологий. Исследование организации генетической информации в геноме (на различных уровнях - ДНК, РНК и белки) требует применения современных технологических подходов высокопроизводительного секвенирования, что, в свою очередь, определяет биоинформационные задачи: первичная обработка сырых данных (картирование прочтений ДНК), анализ массивов данных (анализ дифферен-

* Работа была поддержана РФФИ и бюджетным проектом ИЦиГ СО РАН (№ 0324-2018-0017).

Авторы благодарны Ирине Вадимовне Медведевой, Владимиру Николаевичу Бабенко и Антону Геннадьевичу Богомолову за помощь в работе, и научную дискуссию.

Цуканов А. В., *Орлова Н. Г.*, *Дергилев А. И.*, *Орлов Ю. Л.* Программы статистической обработки, кластеризации и визуализации распределения сайтов связывания транскрипционных факторов в геноме // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 51–63.

© А. В. Цуканов, Н. Г. Орлова, А. И. Дергилев, Ю. Л. Орлов, 2018

циальной экспрессии генов, анализ пиков профилей ChIP-seq) и аннотация генома на основе таких прочтений (аннотирование транскриптов, определение положения сайтов связывания транскрипционных факторов) [1].

Ключевую роль в работе клетки играет регуляция транскрипции генов, которые определяют свойства клетки [2]. Изучение проблемы регуляции, развитие методов и технологий секвенирования привели к накоплению огромного количества данных, полученных различными методами секвенирования, такими как Hi-C, ChIP-seq, BS-seq, DNaseI-seq, ATAC-seq, NOMe-seq, RNA-seq [1; 3; 4]. Доступны большие объемы данных как для генома человека, так и для геномов других модельных организмов эукариот (мышь, крыса, растения). Рассмотрим несколько уровней регуляции транскрипции с точки зрения анализа данных [4].

Первый и, возможно, основной уровень – это регуляция на уровне ДНК, здесь регуляция осуществляется за счет связывания транскрипционных факторов с сайтами на ДНК, что может приводить как к активации транскрипции, так и к ее остановке [5]. Также на этом уровне выделяется влияние метилирования ДНК, которое может оказывать влияние на регуляцию за счет метилирования сайтов связывания транскрипционных факторов (ССТФ), что приводит к изменению сродства транскрипционного фактора (ТФ) к его сайту [6].

Следующий уровень – это регуляция на уровне хроматина, где в зависимости от модификаций гистонов может меняется структура хроматина, он может быть в открытом или закрытом состоянии, что влияет на транскрипцию генов. Состояние хроматина определяется по данным анализа геномных профилей секвенирования. В закрытом состоянии хроматина транскрипционные факторы и транскрипционная «машина» эукариотической клетки неспособна подобраться к сайту начала транскрипции из-за физической недоступности ДНК для этих белков [7; 8].

Третий немаловажный уровень регуляции осуществляется за счет трехмерной структуры хромосом. Трехмерная упаковка в ядре клетки может оказывать значительное влияние на регуляцию экспрессии генов за счет сближения удаленных регуляторных участков (энхансеров, сайленсеров) и промоторов генов, тем самым регулируя уровень экспрессии генов [9; 10].

Несмотря на большой массив накопленной информации, остаются пробелы в знаниях о регуляции экспрессии генов, что приводит к необходимости продолжать исследования, связанные с анализом регуляции экспрессии генов эукариот в масштабе генома, и разработку соответствующих программных инструментов. Исследование регуляции экспрессии генов эукариот в масштабе генома требует изучения сайтов связывания транскрипционных факторов (ССТФ), контролирующих транскрипцию генов, их геномной локализации, определения их генов-мишеней [3; 11]. Благодаря развитию методов высокопроизводительного секвенирования ChIP-seq, ChIP-on-chip и другим технологиям, сопряженным с иммунопреципитацией хроматина (ChIP – Chromatin ImmunoPrecipitation) [4], стал доступен огромный массив новых данных, позволяющих исследовать все сайты связывания заданного транскрипционного фактора в геноме, а также комбинации таких сайтов [3; 12].

Экспериментально установленное число сайтов в геноме может варьировать от нескольких сотен до десятков тысяч [1; 3; 4; 12]. Значительная часть ССТФ располагается в дистальных (удаленных) районах генов, что затрудняет точное определение тех генов, транскрипцию которых они регулируют. Встают задачи анализа регуляторных районов генов, в том числе дистальных, поиска закономерностей расположения в них сайтов и контекстных сигналов с помощью статистических, логических и биоинформационных методов [11; 13; 14]

Исследование влияния близко расположенных и пересекающихся по своему расположению ССТФ в промоторах и энхансерах на уровень экспрессии генов является одним из важных направлений исследований; перекрывающиеся нуклеотидными последовательностями сайты связывания изучены недостаточно [15].

С использованием данных ChIP-seq для профилей связывания ТФ в геноме мыши ранее были исследованы взаимодействия транскрипционных факторов в плане одновременного связывания различных ТФ в геномных районах (так называемые множественные локусы регуляции транскрипции) [3; 12].

Одним из важных биомедицинских приложений является построение полногеномных карт регуляторов плюрипотентности NANOG, OCT4, SOX2, KLF4 в эмбриональных стволо-

вых клетках и связанных с ними кластеров сайтов других транскрипционных факторов [3]. Накоплены экспериментальные данные о трехмерной организации геномных участков (удаленные энхансеры, пространственные домены), что служит основой для более сложных моделей регуляторных районов [1; 4].

В настоящей работе представлены скрипты для анализа сайтов по данным ChIP-seq, расчета кластеров сайтов и их визуализации в форме тепловых карт, развивающие подходы, представленные в [12] на новых данных и выполненные в другой среде программирования.

Материалы и методы

В работе использовались пики ChIP-seq для клеточной линии эмбриональных стволовых клеток человека H1. Данные в виде BED-файлов загрузили из базы данных Cistrome [1] [16], были загружены координаты пиков 38 транскрипционных факторов.

Отметим, что существует огромное количество баз данных по сайтам связывания транскрипционных факторов, которые содержат в себе как данные по ChIP-seq (Expression Atlas, Roadmap epigenomics project, ENCODE), так и данные по непосредственным координатам ССТФ и мотивам связывания ТФ (TRANSFAC, JASPAR) [17; 18], также существуют базы данных, разработанные в России, – TRRD [19; 20], GTRD [21], HOCOMOCO [22].

Для анализа данных был разработан набор скриптов на языке R, в последующем они были собраны в пакет, названый ClanChIPeaks, который можно свободно скачать из репозитория GitHub [2]. При разработке ClanChIPeaks и анализе данных также использовались сторонние пакеты из репозиториев Bioconductor (GenomicRanges, AnnotationHub, ChIPeeker и т. д.) и CRAN (ggplot2, fastclust и т. д.). Общий алгоритм анализа пиков ChIP-seq при работе в пакете ClanChIPeaks и их кластеризация представлены на рис. 1.

На первом этапе необходимо загрузить экспериментальные данные в среду R при помощи функции peaks.read(), данная функция возвращает объект GenomicRanges. Отдельно этот объект и другие широко используемые классы объектов, используемые в Bioconductor для работы с биологическими данными, представлен в статье [23].

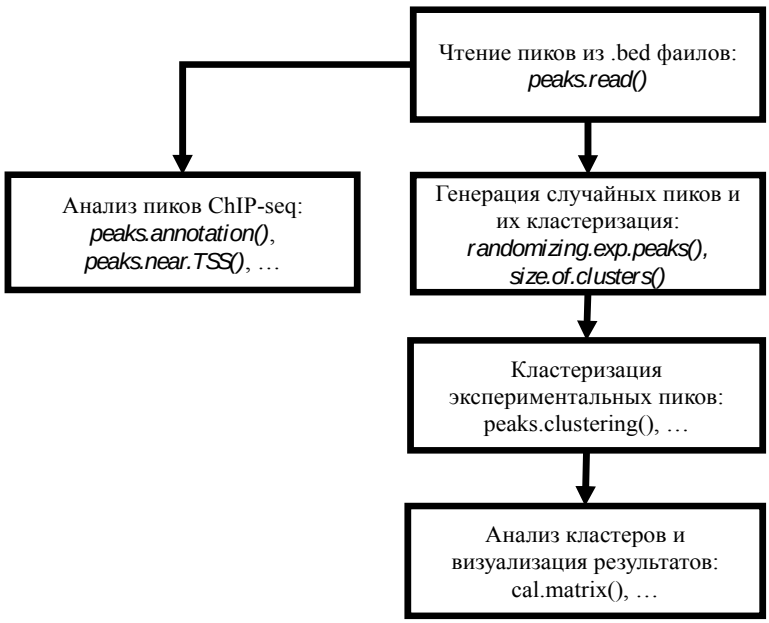


Рис. 1. Общий алгоритм анализа и кластеризации пиков ChIP-seq

[1] http://cistrome.org/db/#/.
[2] https://github.com/anton-tsukanov/ClanChiPeaks.

На втором этапе отдельно от анализа кластеров ClanChIPeaks позволяет проводить небольшой анализ пиков ChIP-seq. Так, например расчет плотности распределения пиков около начала сайтов транскрипции с использованием функции peaks.near.TSS(), а также аннотация пиков, т. е. в каком регионе находятся пики (экзон, интрон, 5'-нетранслируемая область, 3'-нетранслируемая область, межгенное пространство), для этого используется функция peaks.annotation().

На третьем этапе осуществляется генерация случайных пиков ChIP-seq, которые обладают той же общей шириной пиков, и проводится их кластеризация. Данный этап необходим для выявления размера кластера, который не будет получаться по случайным причинам.

На четвертом этапе непосредственно проводится кластеризация пиков ChIP-seq полученных экспериментально, это осуществляется при помощи функции peaks.clustering(). Данная функция возвращает стандартный объект list(), где каждый элемент list() является объектом data.frame(), содержит пики из одного кластера и информацию о них (начало пика, конец пика, ширина и др.). При помощи функции cal.matrix() можно посчитать попарную встречаемость каждого транскрипционного фактора, результат будет представлен в виде матрицы.

На последнем этапе можно проводить анализ посчитанных данных при помощи стандартных методов, встроенных в R, а также визуализировать данные при помощи пакетов ggplot2, corrplot и др.

Результаты

Общий анализ пиков ChIP-seq

На основании геномных координат (начало и конец геномного участка) пиков профилей ChIP-seq для 38 транскрипционных факторов клеточной линии H1 была рассчитана ширина каждого пика для 9 ТФ (E2F6, CREB1, MYC, ZNF143, MXI1, SP4, YY1, NRF1, NANOG); построены графики распределения ширины пиков, представленные на рис. 2.

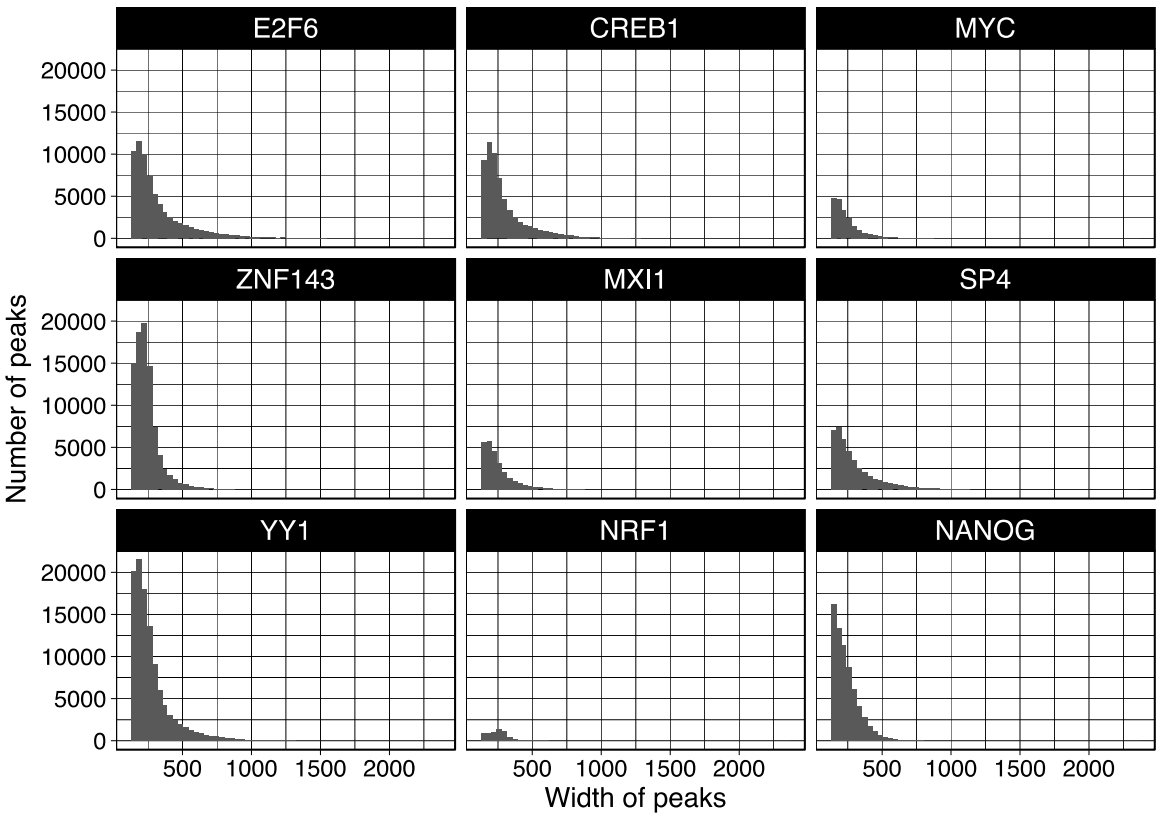


Рис. 2. Распределение размеров и ширины пиков ChIP-seq в геноме
Ось абсцисс – ширина пиков, ось ординат – количество пиков

Для представленных на рис. 2 транскрипционных факторов минимальное значение ширины пика составляет 147 п. н., медианные значения ширины пиков варьируют от 173 п. н. у MXI1 до 255 п. н. у NRF1, а максимальные значения ширины пиков – от 1074 п. н. у NANOG до 2345 п. н. у E2F6. Отметим, что сам сайт связывания ТФ имеет длину порядка 10 п. н., что делает необходимым искать точное положение сайтов связывания ТФ по их мотивам в нуклеотидных последовательностях пиков геномного профиля ChIP-seq размером в сотни нуклеотидов.

Также было посчитано распределение положения пиков ChIP-seq для 6 ТФ из общего набора в 38 ТФ клеточной линии H1 в разных участках генома относительно гена (интроны (Introns), экзоны (Exons), промоторы (Promoters), 5'-НТО (5UTR), 3'-НТО (3UTR), межгенное пространство(Intergenic)). Результаты представлены на рис. 3.

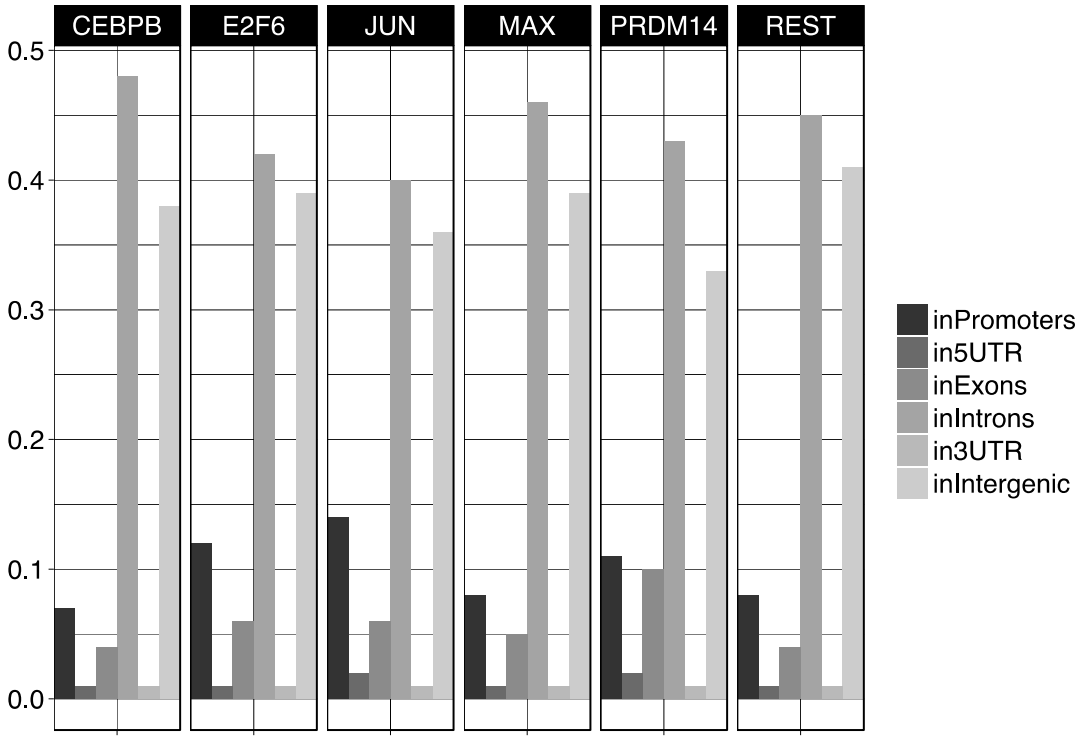


Рис. 3. Распределение пиков в геноме

Из рис. 3 видно, что большая часть сайтов находится в интронах и в межгенных районах. В то же время для ТФ JUN доля сайтов в промоторах выше, чем для ТФ CEBP и REST. Распределение сайтов в геноме дает лишь общую картину расположения сайтов относительно генов – наибольший интерес представляет определение положения сайта относительно старта транскрипции, что непосредственно влияет на транскрипцию данного гена.

Кластеризация сгенерированных пиков

Перед тем как изучать кластеризацию пиков ChIP-seq, полученных экспериментально, кластеризацию пиков изучали на сгенерированных данных, при этом суммарная ширина пиков – как экспериментальных, так и сгенерированных, совпадала. Случайные пики генерировали следующим образом. Зная общее число экспериментальных пиков на хромосоме, генерировали такое же количество псевдослучайных пиков (при помощи генератора псевдослучайных чисел) в диапазоне длины хромосомы. Далее сгенерированным пикам задавали значения ширины путем случайного выбора ширины из экспериментальных пиков.

Для кластеризации сгенерированных пиков использовали иерархический тип кластеризации и меру расстояния в Евклидовом пространстве. Кластеризацию проводили с использованием стороннего пакета fastclust из репозитория CRAN. Пики ChIP-seq кластеризовались друг с другом, если расстояние от ближайших центров пиков не превышало 25 п. н., таким образом размер кластера увеличивался до тех пор, пока расстояние между кластером и другим пиком не стало больше 25 п. н.

Сгенерированные данные ChIP-seq имели 1 485 464 пика, а точка отсечения составляла 25 п. н., для каждой новой кластеризации генерировались новые случайные пики ChIP-seq, всего было проведено 5 реплик. Результаты кластеризации сгенерированных пиков представлены в таблице.

Размер и количество кластеров,
полученных на сгенерированных пиках

Реплика	Размер кластера *			
	максимальный	количество кластеров	предмаксимальный	количество кластеров
1	4	5	3	267
2	4	6	3	263
3	4	6	3	263
4	4	4	3	294
5	4	5	3	283

* Размер кластера – количество пиков в кластере.

Из результатов, представленных в таблице, видно, что максимальный размер кластера достигает 4, а количество таких кластеров варьирует от 4 до 6. Таким образом, для экспериментальных данных ChIP-seq клеточной линии H1 (при условии, что для анализа мы используем 1 485 464 пика) мы приняли гипотезу о том, что минимальный размер кластера, который может образоваться не по случайным причинам, составляет 5 пиков ChIP-seq.

Кластеризация и изучение кластеров
на экспериментальных данных ChIP-seq

Условия кластеризации пиков ChIP-seq, полученных экспериментально, были такие же, как и для сгенерированных пиков (иерархический тип кластеризации, мера расстояния – Евклидово пространство, максимальное расстояние между пиками в одном кластере 25 п. н.). В кластеризации использовались данные по 31 транскрипционному пику ChIP-seq для клеточной линии человека H1, общее количество пиков составило 1 485 464. Сопоставление размеров кластеров, полученных на экспериментальных данных ChIP-seq и сгенерированных, представлены на рис. 4. Из графика видно, что большая часть кластеров – это одиночные сайты и пары сайтов. Логарифм значения числа кластеров уменьшается с ростом количества пиков в самом кластере, при этом для экспериментальных кластеров, полученных из экспериментальных данных, вырисовывается S-образная кривая. Стоит отметить, что логарифм от значения количества кластеров плавно уменьшается только в диапазоне 1–30, а далее значение становится нестабильным – то уменьшается, то увеличивается (что связано с малым числом таких больших кластеров). Максимальный размер кластера достигает 37 пиков, а так как количество уникальных транскрипционных факторов всего 31, следовательно, в один кластер могли попадать пики одного и того же ТФ. Для случайно сгенерированных координат не наблюдалось кластеров размером больше 4. Таким образом, кластеры сайтов размером 4 и выше не случайны, что подтверждают сделанные ранее оценки [3; 12].

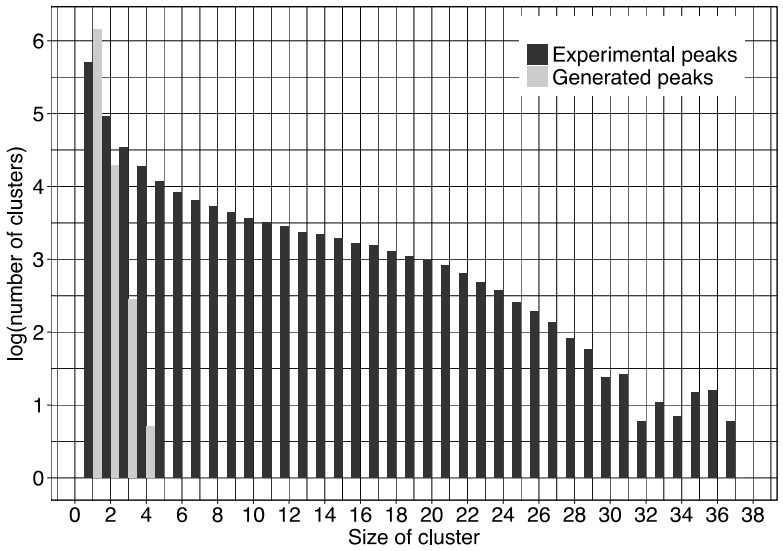


Рис. 4. Кластеры сайтов связывания ТФ (пиков ChIP-seq) по экспериментальным данным (черные столбцы) и компьютерная симуляция кластеров (генерация координат с помощью датчика случайных чисел)

Для дальнейшего анализа использовались только значимые кластеры, т. е. кластеры, размер которых составлял 5 пиков или больше. Для ТФ, входящих в такие кластеры, была посчитана матрица встречаемости, которая представляет собой квадратную симметричную матрицу чисел совместной встречаемости пары сайтов для каждых двух транскрипционных факторов из исследуемого набора. Из нее была посчитана матрица корреляций (по векторам целочисленных значений встречаемости пар сайтов). Все последующие подсчеты проводили на основе матрицы корреляций. Так, на основе матрицы корреляции был посчитан коэффициент несходства (различия), который считается как (1 – коэффициент корреляции), для каждой пары ТФ. Этот коэффициент использовался в качестве меры расстояния для построения дендрограммы (рис. 5), характеризующей взаимную встречаемость транскрипционных факторов.

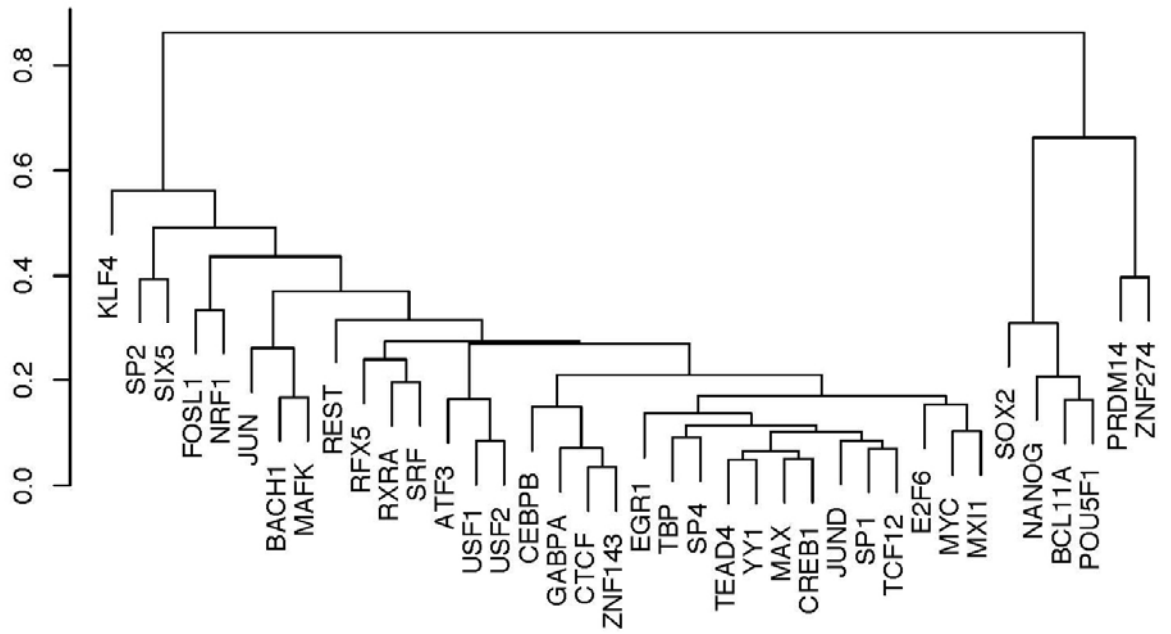


Рис. 5. Дендрограмма близости ТФ (пиков ChIP-seq) по корреляциям их взаимного расположения в кластерах

Из дендрограммы, представленной на рис. 5, можно сделать вывод, что некоторые ТФ предпочитают находиться в одном кластере. Такая, например, группа ТФ, как SOX2, NANOG, BCL11A и POU5F1, образует одну кладу, или PRDM14 и ZNF274, которые также образуют свою кладу.

Иным способом представления матрицы корреляции встречаемости ТФ является тепловая карта встречаемости ТФ (рис. 6).

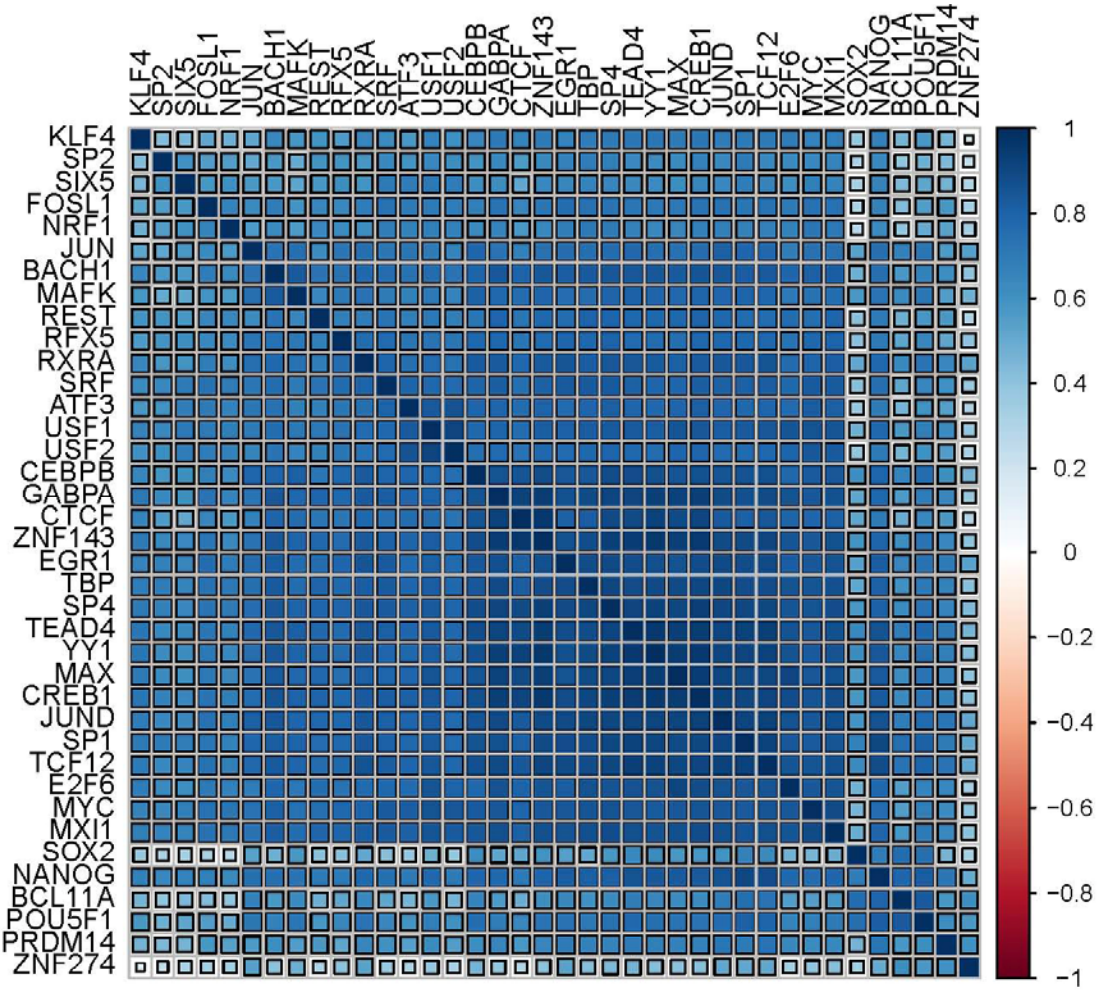


Рис. 6. Тепловая карта коэффициентов корреляции

Более темный синий цвет ячейки соответствует повышенной совместной встречаемости пары сайтов для транскрипционных факторов, представленных в соответствующих строке и столбце матрицы. Две основные ветви на дендрограмме (см. рис. 5), соответствующие факторам KLF4, SP2 и др. и факторам SOX2, NANOG и др., визуально выделяются как левый верхний и правый нижний более темные квадраты на тепловой карте (см. рис. 6). Тепловую карту построили с помощью пакета corrplot. И действительно, транскрипционные факторы SOX2, NANOG, POU5F1, PRDM14 из второго кластера являются факторами поддержания плюрипотентности, и совместная встречаемость их сайтов связывания обусловлена их общей функцией в клетке [3].

Заключение

Проведен вычислительный эксперимент по анализу кластеров сайтов связывания в геноме человека по базам данных ENCODE и Cistrome. Разработанные программы позволяют оценивать статистические параметры профилей ChIP-seq в геноме человека и модельных гено-

мах, строить распределения пиков по ширине и высоте (силе связывания с ДНК или уровнем значимости участка, представленным *p*-value), определять параметры отдельных групп сайтов. Программы позволяют рассчитывать распределение сайтов в геноме относительно старта транскрипции гена (для набора сайтов и разметки генов). Возможен статистический расчет положения сайтов в экзон-интронной структуре гена.

Программы позволяют рассчитывать колокализацию (совместную локализацию) сайтов связывания различных транскрипционных факторов, выполнять визуализацию, строить дендрограммы и тепловые карты совместной локализации сайтов.

Применение программ для анализа кластеров сайтов связывания в эмбриональных стволовых клетках человека по данным ChIP-seq из ресурса Cistrome позволило уточнить состав кластеров сайтов транскрипционных факторов, описать их функциональную роль. По частоте сигналов в исследованном наборе выделяются группы, относящиеся к NANOG, что подтверждает полученные ранее данные [3]. Среди 31 транскрипционного фактора такие группы факторов, как SOX2, NANOG, BCL11A, PRDM14 и POU5F1, имеют тенденцию встречаться совместно более часто.

Дальнейший анализ контекстных признаков в геномных последовательностях может опираться на участки низкой сложности текста (простые повторы и политракты), сайты связывания нуклеосом [24; 25]. Интеграция геномных данных позволяет решать качественно новые задачи, представляя описание полногеномной информации, такой как данные проектов ENCODE [3], FactorBook [4] в геноме человека [26]. Интересно отметить паттерны расположения участков простых нуклеотидных повторов (пониженной сложности текста) в районах однонуклеотидных полиморфизмов в геноме человека [27]. Участки простых повторов в геноме труднее картировать в геноме по коротким последовательностям прочтений ДНК при секвенировании [28]. Анализ этих сигналов вокруг кластеров сайтов связывания транскрипционных факторов позволит построить модель организации таких геномных районов [13; 14], предсказать их функцию по составу сайтов и контекстным характеристикам. В целом данное исследование кластеров сайтов связывания транскрипционных факторов в геноме развивает применение анализа регуляции экспрессии генов в эмбриональных стволовых клетках [29].

Список литературы

1. *Игнатьева Е. В., Подколодная О. А., Орлов Ю. Л., Васильев Г. В., Колчанов Н. А.* Регуляторная геномика – экспериментально-компьютерные подходы // Генетика. 2015. Т. 51 (4). С. 409–429.

2. *Levine M., Cattoglio C., Tjian R.* Looping back to leap forward: transcription enters a new era // Cell. 2014. No. 157. P. 13–25.

3. *Chen X., Xu H., Yuan P. et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells // Cell. 2008. Vol. 133. No. 6. P. 1106–1117.

4. *Кулакова Е. В., Спицина А. М., Орлова Н. Г., Дергилев А. И., Свичкарев А. В., Сафронова Н. С., Черных И. Г., Орлов Ю. Л.* Программы анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C // Программные системы: теория и приложения. 2015. Т. 6. № 2 (25). С. 129–148.

5. *Fitzgerald K. A. et al.* The role of transcription factors in prostate cancer and potential for future RNA interference therapy // Nucleic Acids Research. 2015. Vol. 43. No. 14. P. 6874–6888.

6. *Zhu H., Wang G., Qian J.* Transcription factors as readers and effectors of DNA methylation // Nature. 2016. Vol. 17. P. 551–565.

7. *Kelly T. K., Liu Y., Lay F. D., Liang G., Berman B. P., Jones P. A.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules // Genome Research. 2012. No. 22. P. 2497–2506.

8. *Hu Z., Tee W.* Enhancers and chromatin structures: regulatory hubs in gene expression and diseases // Bioscience Reports. 2017. No. 37. P. 1–14.

[3] https://genome.ucsc.edu/ENCODE/.

[4] http://www.factorbook.org.

9. *Guillaume A., Stefan M.* The three-dimensional genome: regulating gene expression during pluripotency and development // Development. 2017. Vol. 144. P. 3646–3658.

10. *Орлов Ю. Л., Тьерри О., Богомолов А. Г., Цуканов А. В., Кулакова Е. В., Галиева Э. Р., Брагин А. О., Ли Г.* Компьютерные методы анализа хромосомных контактов в ядре клетки по данным технологий секвенирования // Биомедицинская химия. 2017. № 63 (5). С. 418–422.

11. *Орлов Ю. Л., Брагин А. О., Медведева И. В., Гунбин И. В., Деменков П. С., Вишневский О. В., Левицкий В. Г., Ощепков В. Г., Подколодный Н. Л., Афонников Д. А., Гроссе И., Колчанов Н. А.* ICGenomics: программный комплекс анализа символьных последовательностей геномики // Вавиловский журнал генетики и селекции. 2012. Т. 16, № 4/1. С. 732–741.

12. *Дергилев А. И., Спицина А. М., Чадаева И. В., Свичкарев А. В., Науменко Ф. М., Кулакова Е. В., Витяев Е. Е., Чен М., Орлов Ю. Л.* Компьютерный анализ совместной локализации сайтов связывания транскрипционных факторов по данным ChIP-seq // Вавиловский журнал генетики и селекции. 2016. Т. 20 (6). С.770–778. DOI 10.18699/VJ16.194.

13. *Vityaev E. E., Orlov Yu. L., Vishnevsky O. V., Pozdnyakov M. A., Kolchanov N. A.* Computer system «Gene Discovery» for promoter structure analysis // In Silico Biology. 2002. Vol. 2. No. 3. P. 233–247.

14. *Витяев Е. Е., Орлов Ю. Л., Вишневский О. В., Беленок А. С., Колчанов Н. А.* Компьютерная система «Gene Discovery» для поиска закономерностей организации регуляторных последовательностей эукариот // Молекулярная биология. 2001. Т. 35, В 6. С. 952–960.

15. *Васькин Ю. Ю., Хомичева И. В., Игнатьева Е. В., Витяев Е. Е.* Анализ последовательностей регуляторных районов генов реляционной системой ExpertDiscovery, встроенной в пакет UGENE // Вестн. НГУ. Серия: Информационные технологии. 2012. Т. 10. № 1. С. 73–86.

16. *Mei S., Qin Q., Wu Q., Sun H., Zheng R., Zang C., Zhu M., Wu J., Shi X., Taing L., Liu T., Brown M., Meyer C. A., Liu X. S.* Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse // Nucleic Acids Res. 2017. Vol. 45. No. 4. P. 658–662.

17. *Knuppel R., Dietze P., Lehnberg W., Frech K., Wingender E.* TRANSFAC1 retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins // J. Comput. Biol. 1994. Vol. 1. P. 191–198.

18. *Mathelier A. et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles // Nucleic Acids Res. 2016. Vol. 44. P. 110–115.

19. *Кель А. Э., Колчанов Н. А., Кель О. В., Ромащенко А. Г., Ананько Е. А., Игнатьева Е. В., Меркулова Т. И., Подколодная О. А., Степаненко И. Л., Кочетов А. В., Колпаков Ф. А., Подколодный Н. Л., Наумочкин А. А.* TRRD: база данных транскрипционных регуляторных районов генов эукариот // Молекулярная биология. 1997. Т. 31, № 4. С. 636–672.

20. *Kolchanov N. A., Ignatieva E. V., Ananko E. A., Podkolodnaya O. A., Stepanenko I. L., Merkulova T. I., Pozdnyakov M. A., Podkolodny N. L., Naumochkin A. N., Romashchenko A. G.* Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. Vol. 30 (1). P. 312–7.

21. *Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F.* GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments // Nucleic Acids Res. 2017. Vol. 45 (D1). P. D61–D67.

22. *Kulakovskiy I. V., Vorontsov I. E., Yevshin I. S., Sharipov R. N., Fedorova A. D., Rumynskiy E. I., Medvedeva Y. A., Magana-Mora A., Bajic V. B., Papatsenko D. A., Kolpakov F. A., Makeev V. J.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis // Nucleic Acids Res. 2018. Vol. 46 (D1). P. D252–D259. DOI: 10.1093/nar/gkx1106.

23. *Lawrence M. et al.* Software for Computing and Annotating Genomic Ranges // PLOS Computational Biology. 2013. Vol. 8. P. 1–10.

24. *Orlov Yu. L., Potapov V. N.* Complexity: an internet resource for analysis of DNA sequence complexity // Nucleic Acids Res. 2004. Vol. 32. P. W628–W633.

25. *Орлов Ю. Л., Левицкий В. Г., Смирнова О. Г., Подколодная О. А., Хлебодарова Т. М., Колчанов Н. А.* Статистический анализ последовательностей ДНК, содержащих сайты формирования нуклеосом // Биофизика. 2006. Т. 51. С. 608–614.

26. *Спицина А. М., Орлов Ю. Л., Подколодная Н. Н., Свичкарев А. В., Дергилев А. И., Чен М., Кучин Н. В., Черных И. Г., Глинский Б. М.* Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК // Программные системы: теория и приложения. 2015. Т. 6, № 1 (23). С. 157–174.

27. *Сафронова Н. С., Пономаренко М. П., Абнизова И. И., Орлова Г. В., Чадаева И. В., Орлов Ю. Л.* Фланкирующие повторы мономеров определяют пониженную контекстную сложность сайтов однонуклеотидных полиморфизмов в геноме человека // Вавиловский журнал генетики и селекции. 2015. Т. 19 (6). С. 668–674.

28. *Naumenko F. M., Abnizova I. I., Beka N., Genaev M. A., Orlov Yu. L.* Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome // BMC Genomics. 2018. Vol. 19 (Suppl. 3). P. 92. DOI: 10.1186/s12864-018-4475-6/

29. *Дергилев А. И., Цуканов А. В., Орлов Ю. Л.* Компьютерный анализ кластеров сайтов связывания транскрипционных факторов в эмбриональных стволовых клетках // Гены и клетки. 2017. Т. 12 (3). С. 184–185.

Материал поступил в редколлегию 02.07.2018

A. V. Tsukanov [1, 2], **N. G. Orlova** [3], **A. I. Dergilev** [2], **Yu. L. Orlov** [1, 2]

[1] *Institute of Cytology and Genetics SB RAS*
10 Academician Lavrentiev Ave., *Novosibirsk*, *630090*, *Russian Federation*

[2] *Novosibirsk State University*
1 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

[3] *Novosibirsk State University of Architecture and Civil Engineering* (*Sibstrin*)
113 Leningradskaya Str., *Novosibirsk*, *630008*, *Russian Federation*

ya.cukanton@yandex.ru, *orlovanina2@mail.ru*, *arturd1993@yandex.ru*, *orlov@bionet.nsc.ru*

PROGRAMS FOR STATISTICAL ANALYSIS, CLUSTERIZATION AND VISUALIZATION OF GENOME DISTRIBUTION OF TRANSCRIPTION FACTOR BINDING SITES

The analysis of gene transcription regulation based on the data of modern technologies of high-performance sequencing is an actual task of bioinformatics. It requires the development of new computer tools including supercomputer applications. We consider the problems of processing of genome ChIP-seq profiles for detections of transcription factors binding site in a genome, determining the peaks of such profiles and search the binding sites in the nucleotide sequences of the peaks. The computer programs have been developed to analyze the location of the binding sites in the genome relative to gene regions, to calculate clusters of such sites and visualize their positions in the genome. Clusters of binding sites of transcription factors in the human genome have been calculated using the Cistrome database. We have calculated matrices of the joint occurrence of pairs of binding sites of different transcription factors in the genome for various types of tissues and cells. A computational experiment on the computer generation of random clusters in the genome was carried out, as well as an assessment of the occurrence of large clusters for experimentally obtained binding sites of transcription factors in the human genome. The patterns of occurrence of binding sites of pluripotency factors in embryonic stem cells were described. The developed software is available on request to the authors.

Keywords: genomics, sequencing, binding sites of transcription factors, promoters, big data, statistics, visualization.

References

1. Ignatieva E. V., Podkolodnaya O. A., Orlov Yu. L., Vasiliev G. V., Kolchanov N. A. Regulatory genomics: combined experimental and computational approaches. *Russian Journal of Genetics*, 2015, vol. 51 (4), p. 409–429. (in Russ.)

2. Levine M., Cattoglio C., Tjian R. Looping back to leap forward: transcription enters a new era. *Cell*, 2014, no. 157, p. 13–25.

3. Chen X., Xu H., Yuan P. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 2008, vol. 133, no. 6, p. 1106–1117.

4. Kulakova E. V., Spitsina A. M., Orlova N. G., Dergilev A. I., Svichkarev A. V., Safronova N. S., Chernykh N. S., Orlov Yu. L. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Program systems: theory and applications*, 2015, vol. 6, no. 2 (25), p. 129–148. (in Russ.)

5. Fitzgerald K. A. et al. The role of transcription factors in prostate cancer and potential for future RNA interference therapy. *Nucleic Acids Research*, 2015, vol. 43, no. 14, p. 6874–6888.

6. Zhu H., Wang G., Qian J. Transcription factors as readers and effectors of DNA methylation. *Nature*, 2016, vol. 17, p. 551–565.

7. Kelly T. K., Liu Y., Lay F. D., Liang G., Berman B. P., Jones P. A. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, 2012, no. 22, p. 2497–2506.

8. Hu Z., Tee W. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. *Bioscience Reports*, 2017, no. 37, p. 1–14.

9. Guillaume A., Stefan M. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, 2017, vol. 144, p. 3646–3658.

10. Orlov Yu. L., Thierry O., Bogomolov A. G., Tsukanov A. V., Kulakova E. V., Galieva E. R., Bragin A. O., Li G. Computer methods of analysis of chromosome contacts in the cell nucleus based on sequencing technology data. *Biomeditsinskaya Khimiya*, 2017, no. 63 (5), p. 418–422. (in Russ.)

11. Orlov Yu. L., Bragin A. O., Medvedeva I. V., Gunbin K. V., Demenkov P. S., Vishnevsky O. V., Levitsky V. G., Oshchepkov D. Y., Podkolodnyy N. L., Afonnikov D. A., Grosse I., Kolchanov N. A. ICGenomics: a program complex for analysis of symbol sequences in genomics. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*, 2012, vol. 16, no. 4/1, p. 732–741. (in Russ.)

12. Dergilev A. I., Spitsina A. M., Chadaeva I. V., Svichkarev A. V., Naumenko F. M., Kulakova E. V., Vityaev E. E., Chen M., Orlov Yu. L. Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq dat. *Russian Journal of Genetics: Applied Research*, 2016, vol. 20 (6), p.770–778. DOI 10.18699/VJ16.194. (in Russ.)

13. Vityaev E. E., Orlov Yu. L., Vishnevsky O. V., Pozdnyakov M. A., Kolchanov N. A. Computer system «Gene Discovery» for promoter structure analysis. *In Silico Biology*, 2002, vol. 2, no. 3, p. 233–247.

14. Vityaev E. E., Orlov Yu. L., Vishnevsky O. V., Belenok A. S., Kolchanov N. A. Computer system «Gene Discovery» to search for patterns in eukaryotic regulatory nucleotide sequences. *Molecular Biology*, 2001, vol. 35, B 6, p. 952–960. (in Russ.)

15. Vaskin Yu. Yu., Khomicheva I. V., Ignatyeva E. V., Vityayev E. E. Analysis of regulatory regions of genes by Expert Discovery relation system, integrated into UGENE toolkit. *Vestnik NSU. Series: Information Technologies*, 2012, vol. 10, no. 1, p. 73–86. (in Russ.)

16. Mei S., Qin Q., Wu Q., Sun H., Zheng R., Zang C., Zhu M., Wu J., Shi X., Taing L., Liu T., Brown M., Meyer C. A., Liu X. S. Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res*., 2017, vol. 45, no. 4, p. 658–662.

17. Knuppel R., Dietze P., Lehnberg W., Frech K., Wingender E. TRANSFAC1 retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol*., 1994, vol. 1, p. 191–198.

18. Mathelier A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*., 2016, vol. 44, p. 110–115.

19. Kel A. E., Kolchanov N. A., Kel O. V., Romashchenko A. G., Ananko E. A., Ignatieva E. V., Merkulova T. I., Podkolodnaya O. A., Stepanenko I. L., Kochetov A. V., Kolpakov F. A., Podkolodnyi N. L., Naumochkin A. A. TRRD: a database of transcription regulatory regions in eukaryotic genes. *Mol. Biol.*, 1997, vol. 31, no. 4, p. 636–672. (in Russ.)

20. Kolchanov N. A., Ignatieva E. V., Ananko E. A., Podkolodnaya O. A., Stepanenko I. L., Merkulova T. I., Pozdnyakov M. A., Podkolodny N. L., Naumochkin A. N., Romashchenko A. G. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*., 2002, vol. 30 (1), p. 312–7.

21. Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res*., 2017, vol. 45 (D1), p. D61–D67.

22. Kulakovskiy I. V., Vorontsov I. E., Yevshin I. S., Sharipov R. N., Fedorova A. D., Rumynskiy E. I., Medvedeva Y. A., Magana-Mora A., Bajic V. B., Papatsenko D. A., Kolpakov F. A., Makeev V. J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*., 2018, vol. 46 (D1), p. D252–D259. DOI: 10.1093/nar/gkx1106.

23. Lawrence M. et al. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 2013, vol. 8, p. 1–10.

24. Orlov Yu. L., Potapov V. N. Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic Acids Res*., 2004, vol. 32, p. W628–W633.

25. Orlov Yu. L., Levitskii V. G., Smirnova O. G., Podkolodnaya O. A., Khlebodarova T. M., Kolchanov N. A. Statistical analysis of DNA sequences containing nucleosome positioning sites. *Biophysics*, 2006, vol. 51, p. 608–614. (in Russ.)

26. Spitsina A. M., Orlov Yu. L., Podkolodnaya N. N., Svicharev A. V., Dergilev A. I., Chen M., Kuchin N. V., Chernykh I. G., Glinsky B. M. Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. *Program systems: theory and applications*, 2015, vol. 6, no. 1 (23), p. 157–174. (in Russ.)

27. Safronova N. S., Ponomarenko M. P., Abnizova I. I., Orlova G. V., Chadaeva I. V., Orlov Yu. L. Flanking monomer repeats determine decreased context complexity of single nucleotide polymorphism sites in the human genome. *Russian Journal of Genetics: Applied Research*, 2015, vol. 19 (6), p. 668–674. (in Russ.)

28. Naumenko F. M., Abnizova I. I., Beka N., Genaev M. A., Orlov Yu. L. Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. *BMC Genomics*, 2018, vol. 19 (suppl. 3), p. 92. DOI: 10.1186/s12864-018-4475-6/

29. Dergilev A. I., Tsukanov A. V., Orlov Yu. L. Computer analysis of clusters of transcription factor binding sites in embryonic stem cells. *Genes and Cells*, 2017, vol. 12 (3), p. 184–185. (in Russ.)

For citation:

Tsukanov A. V., Orlova N. G., Dergilev A. I., Orlov Yu. L. Programs for Statistical Analysis, Clusterization and Visualization of Genome Distribution of Transcription Factor Binding Sites. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 51–63. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-51-63

УДК 631.171
DOI 10.25205/1818-7900-2018-16-3-64-73

Р. М. Бандишоева

Таджикский технический университет им. М. С. Осими
пр. Академиков Раджабовых, 10, Душанбе, 734042, Таджикистан

risolatbm@mail.ru

РАЗРАБОТКА АВТОМАТИЗИРОВАННОГО МОДУЛЯ МИНЕРАЛИЗАЦИИ ПОЛИВНОЙ ВОДЫ

Статья посвящена разработке системы управления фертигацией, которая позволяет осуществлять автоматическую подачу удобрений с учетом показателей pH и EC, полученных от соответствующих датчиков. Для решения этой задачи разработан модуль минерализации, состоящий из двух подмодулей: подмодуль измерения и регулирования pH и подмодуль измерения и регулирования EC. Данные передаются микроконтроллеру для дальнейших действий по управлению системой. В качестве системы принятия решений в работе использована интеллектуальная система на основе нечеткой логики.

Ключевые слова: фертигация, нечеткий контроллер, минерализация, концентрация, раствор.

Введение

Подача удобрений к растениям через поливную воду называется фертигацией [1; 2]. Фертигация – это современная агротехнология, которая дает возможность увеличивать урожай и уменьшать загрязнение окружающей среды [3], при этом повышая эффективность использования удобрений. В фертигации контролируются время, количество и концентрация применяемых удобрений.

Состав фертигационного раствора подбирается с учетом потребностей культуры, фазы развития растения, субстрата. Важно учитывать соотношение между элементами питания на каждой стадии (преимущественно между основными микроэлементами – NPK). Для обеспечения баланса питательных веществ на средне- и тяжелосуглинистых темносероземных почвах Гиссарской долины при капельном орошении хлопчатника поливы необходимо осуществлять питательным раствором из расчета годовой нормы питательных веществ: азот – 250, фосфор – 180, калий – 60 кг/га [4]. Автоматическая система для подачи удобрений в поливную воду оснащена электрическим и гидравлическим управляемыми клапанами, а также автоматическим контролем и коррекцией pH и EC (кислотность и электропроводность). При выборе удобрения для фертигации следует рассматривать 4 основных фактора [3]:

- вид растения и стадии развития;
- почвенные условия;
- качество пресной воды;
- доступность удобрения и цена.

Разработанные модули имеют в своем составе датчики для измерения pH и EC раствора удобрения и почвы. Надо отметить, что PH и EC являются двумя важными показатели фертигации. Величина pH – это показатель кислотности раствора, т. е. соотношения кислоты и

Бандишоева Р. М. Разработка автоматизированного модуля минерализации поливной воды // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 64–73.

© Р. М. Бандишоева, 2018

щелочи (1÷14), где 1 – кислота, 14 – щелочь. Определяет способность растения усваивать питательные вещества из раствора.

Электропроводность прямо пропорциональна расстоянию между электродами L и обратно пропорциональна площади этих электродов S, которые опущены в раствор, т. е. $R = \rho(L/S)$, где ρ – удельное сопротивление раствора.

В работе создана автоматическая система управления фертигацией на основе нечеткого контроллера с учетом основных показателей датчиков и данных, полученных из базы знаний. Данные передаются микроконтроллеру для дальнейших действий по управлению системой. В модуле pH система корректирует значение pH в смесительном резервуаре с помощью подачи в раствор кислоты или щелочи, а в модуле EC система, получая данные о количестве минеральных удобрений в воде, корректирует скорость истечения поливной воды.

Заранее (по экспертным оценкам) в системе установлены значения величины pH и количество кислоты / щелочи, которое должно быть введено в смеситель. Таким образом, переменными обратной связи pH являются: расход концентрированного раствора кислоты / щелочи из кислотно-основного резервуара; значения pH питательного раствора, представляющего собой отрицательный логарифм концентрации ионов водорода в растворе; изменение pH питательного раствора представляет собой изменение отрицательного логарифма концентрации ионов водорода в питательном растворе с течением времени.

Выходом является величина расхода через кислотно-основной клапан, т. е. представляет собой количество концентрированного раствора кислоты / щелочи, которое необходимо добавить в питательный раствор для нормализации его pH. Нечеткие входные переменные контроллера следующие: вариационный pH – представляет собой изменение текущего значения pH питательного раствора во времени; разница pH – разница между заданной величиной pH и фактическим значением pH питательного раствора; кислота / щелочь – количество раствора кислоты / щелочи, который необходимо добавить.

В этой системе используются три резервуара водорастворимых удобрений N, P и K, а также емкость для смешивания удобрений (рис. 1). Концентрации C_i [кг/л] трех растворов удобрений и количества каждого компонента m_i [кг] приведены в качестве входных данных для системы управления. Рассматривая скорость потоков растворов V [л/с], система рассчитывает время T_i [с], в течение которого он должен подавать i-й раствор каждого из трех резервуаров в емкость для смешивания:

$$T_i = m_i \frac{1}{c_i V}.$$

Входным сигналом, который будет передан системе, является временной интервал между двумя процессами смешивания. Этот интервал может исчисляться в днях или часах. Учитывая это время, система будет ждать, прежде чем запускать другой процесс смешивания. Во время ожидания, система будет контролировать и поддерживать уровень влажности в почве.

Содержание влаги в почве будет контролироваться и управляться с установленной дискретностью. После того как смешивание закончено, система обеспечивает подачу этого раствора в трубопровод для поливной воды (см. рис. 1). Подача раствора выполняется насосом, который через форсунку впрыскивает раствор в трубопровод. Концентрация минерализованной воды определяется путем измерения ее электропроводности. Например, для уменьшения электропроводности необходимо увеличить скорость подачи минерализованной воды. Таблица соответствия скоростей определена экспериментальным путем.

Время наполнения смесительного бака $T_{\text{нап}}$ растворами удобрений

$$T_{\text{нап}} = \max(T_1, T_2, T_3),$$

где T_1, T_2, T_3 – соответствующие значения времени подачи из резервуара азотного, фосфорного и калийного удобрений.

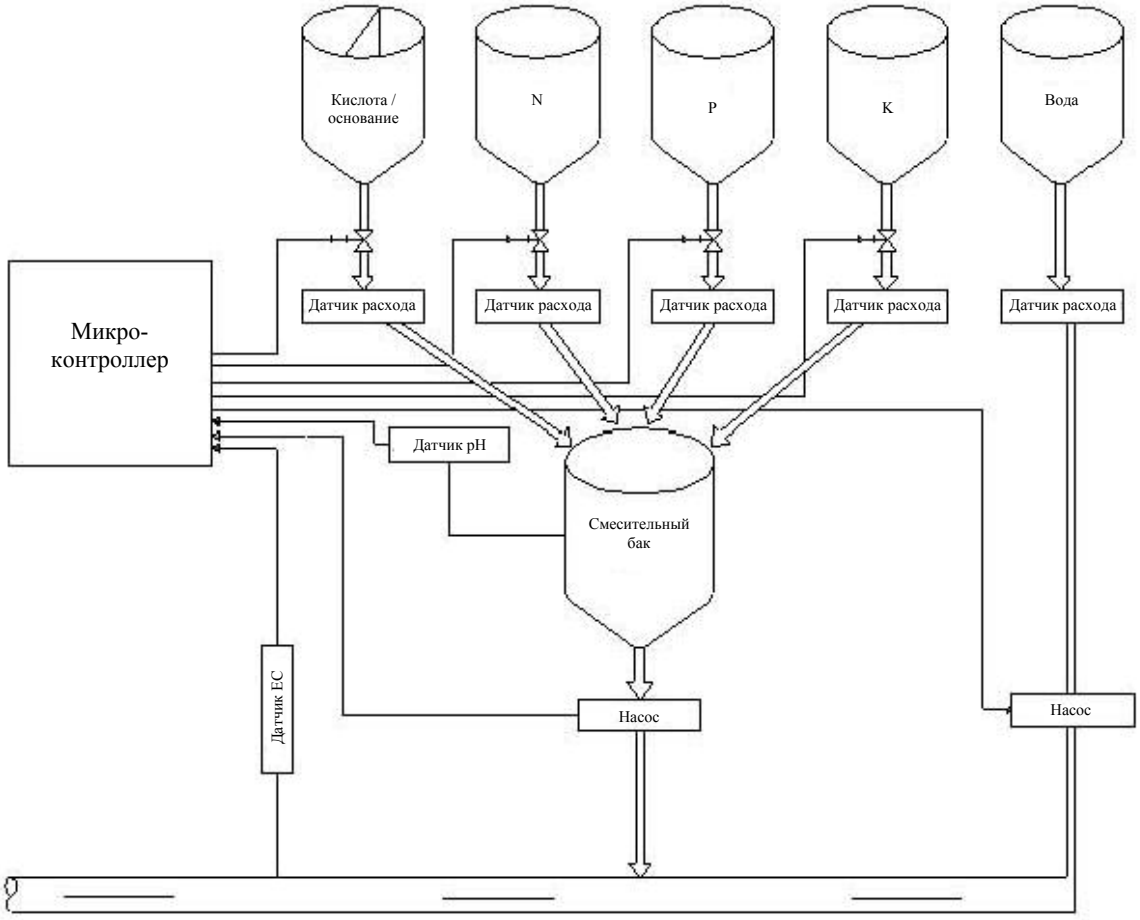


Рис. 1. Технологическая схема автоматизации добавки минеральных удобрений в поливную воду

Время, затраченное на измерение pH полученной смеси T_{pH}

$$T_{\text{pH}} = T_{\text{инд}} + T_{\text{опр}},$$

где $T_{\text{инд}}$ – время работы индикатора по определению величины pH раствора, $T_{\text{опр}}$ – время опроса датчика и передачи сигнала в модуль.

Время одного цикла $T_{\text{ц}}$ подготовки раствора минерализации массой m

$$m = \sum_{i=1}^{3} m_i + m_{\text{к/о}},$$

где m_i – масса i-го раствора минеральных удобрений в резервуаре, $m_{\text{к/о}}$ – масса подаваемого раствора кислота / основание. $T_{\text{ц}}$ определяется из выражения

$$T_{\text{ц}} = T_{\text{нап}} + T_{\text{pH}} + T_{\text{к/о}},$$

где $T_{\text{к/о}}$ – время подачи раствора кислота / основание в смеситель.

На рис. 1 показана технологическая схема подготовки и подачи удобрений совместно с капельным орошением с учетом показателей pH и EC раствора.

Разработанный нечеткий контроллер может обеспечить необходимый уровень pH раствора с учетом нелинейности поведения значений pH раствора питательных веществ. Например, переменная разности pH находится в диапазоне от –1 до 1 (данные взяты в относительных

единицах как отношение pH/pH_{max}). Благодаря известным из литературы данным, составлены следующие лингвистические значения этой переменной: «отрицательный», если значение pH находится в пределах от –1 до –0,4; «нейтральный», если pH находится в пределах от –0,2 до 0,4; «положительный», если pH находится в пределах 0,5 до 1. В системе нечеткой логики созданы входящие переменные, имеющие лингвистические значения (рис. 2).

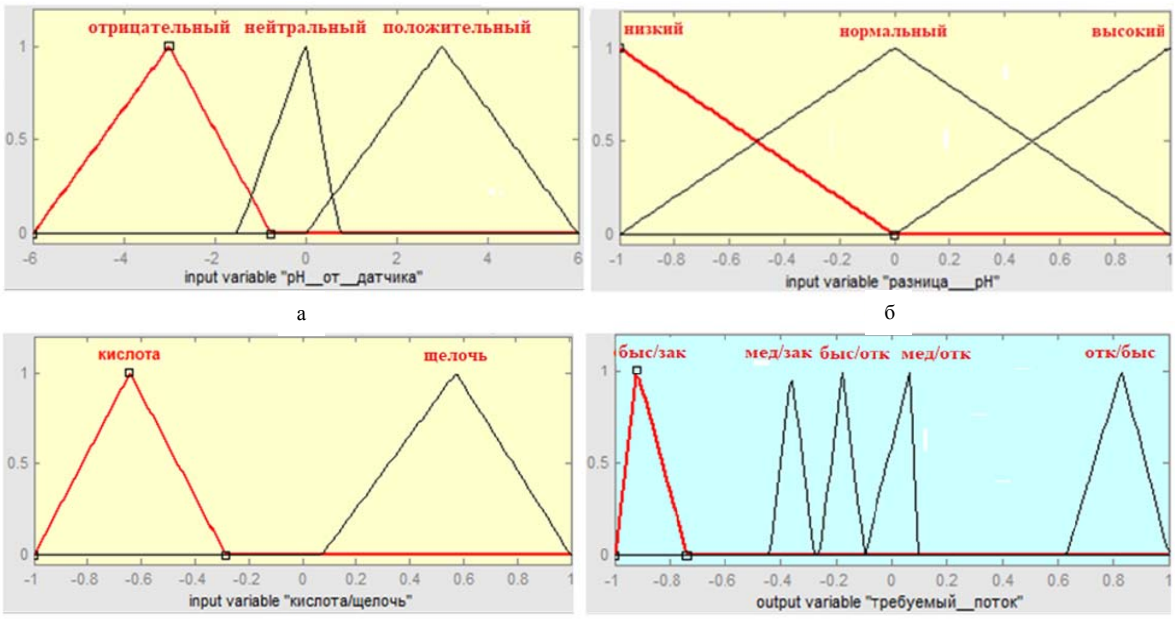


Рис. 2. Функция принадлежности:
а – для лингвистической переменной pH; *б* – для разницы измеренного и заданного значений pH;
в – для переменной подачи раствора кислота / щелочь в смеситель;
г – для выходной переменной – требуемый состав смеси

Все значения переменных и набор нечетких правил для разработки системы управления были получены на основании опроса экспертов. Чтобы точно настроить эти правила, а также функции принадлежности, мы использовали стратегию проб и ошибок, которая подразумевает настройку исполнительных органов пока набор правил не достиг удовлетворительного значения в исследуемой модели. Каждому лингвистическому входному значению в соответствии со значением функции принадлежности соответствует некоторое действие в системе. Система состоит из девяти правил, которые можно увидеть в окне редактора правил (рис. 3).

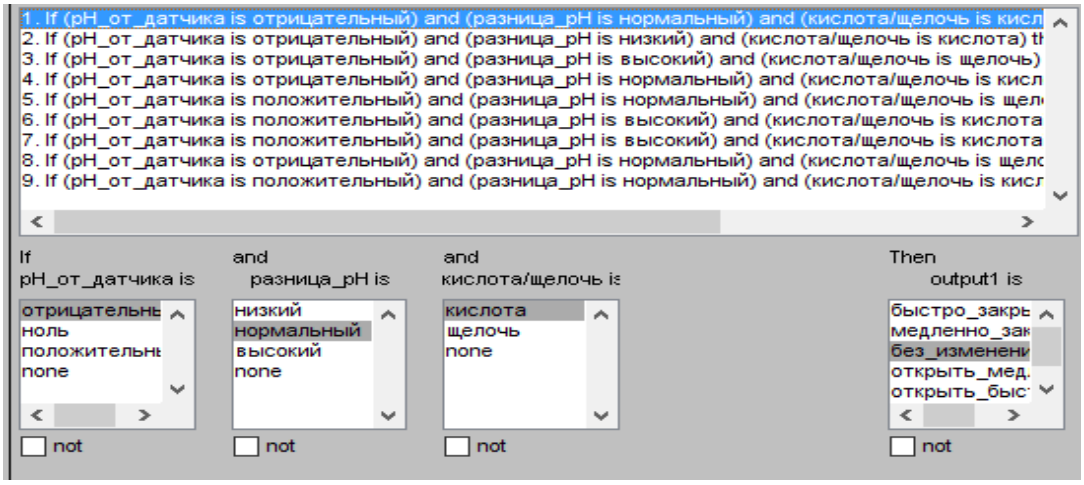


Рис. 3. Окно редактора базы знаний

Результат работы нечеткого контроллера для управления значением pH показан на рис. 4.

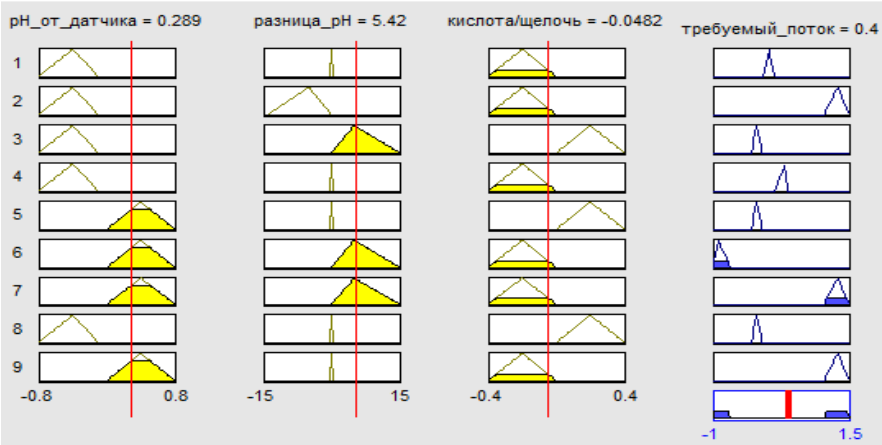


Рис. 4. Визуализация нечеткого логического вывода для управления pH

В подмодуле EC значения проводимости используются для получения информации о количестве вводимого удобрения в поток воды для полива. Высокий показатель EC раствора означает, что в составе поливной воды присутствует большое количество удобрений.

Нечеткие входные переменные контроллера: вариационный EC – представляет собой изменение текущего значения EC питательного раствора, ошибка EC – разница между заданной величиной EC и фактическим значением EC питательного раствора; фазы развития хлопчатника – три характерные фазы хлопчатника, где показатели EC имеют разные значения. Выход представляет собой поток воды от насоса. Это количество воды, требуемое для достижения необходимой концентрации поливного раствора. Основные правила, разработанные для нечеткого контроллера, приведены на рис. 5, 6.

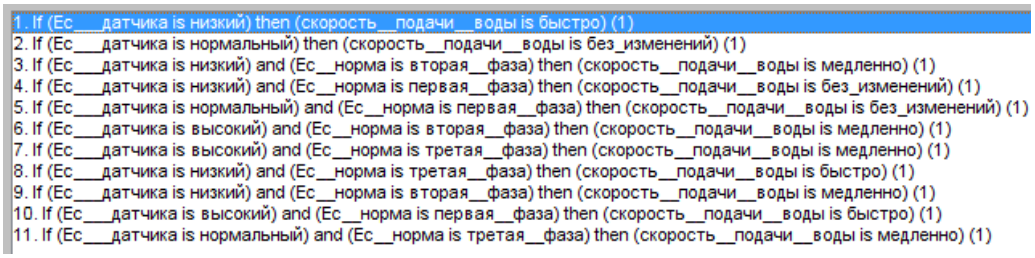


Рис. 5. Окно редактора базы знаний

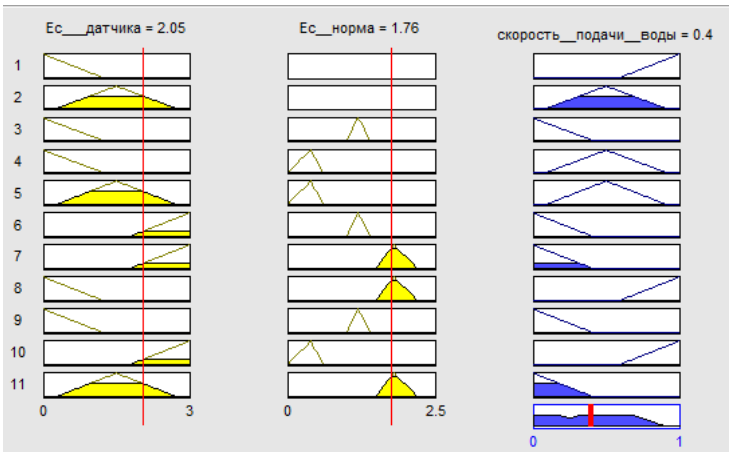


Рис. 6. Визуализация нечеткого логического вывода для управления EC

Техническая реализация решений

Блок управления состоит из соленоидных клапанов, управляемых электромагнитным реле, которое позволяет открывать / закрывать подачу удобрений в смесительный бак. Электромагнитный клапан должен быть включен / выключен в соответствии с нечеткой системой подмодулями с pH и EC. На рис. 7 приведена принципиальная электрическая схема модулей подготовки и подачи жидких удобрений для капельного орошения растений.

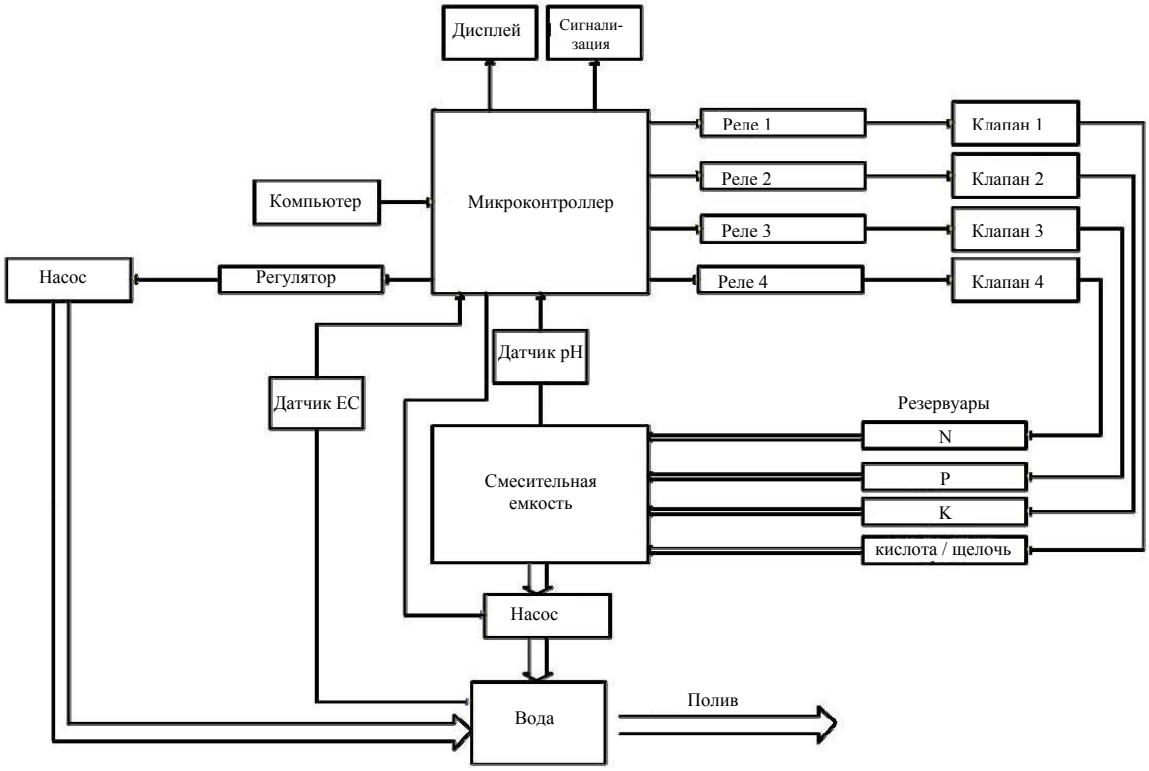
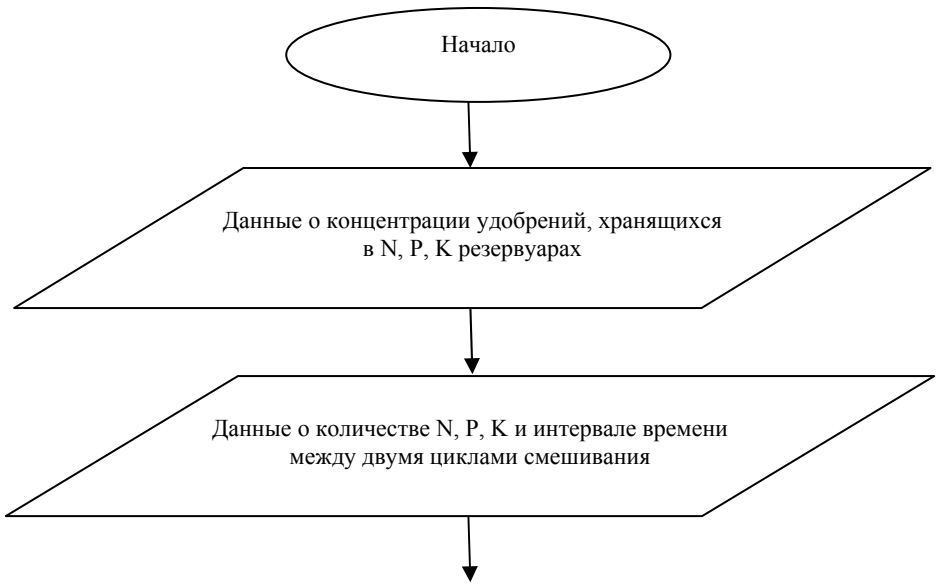
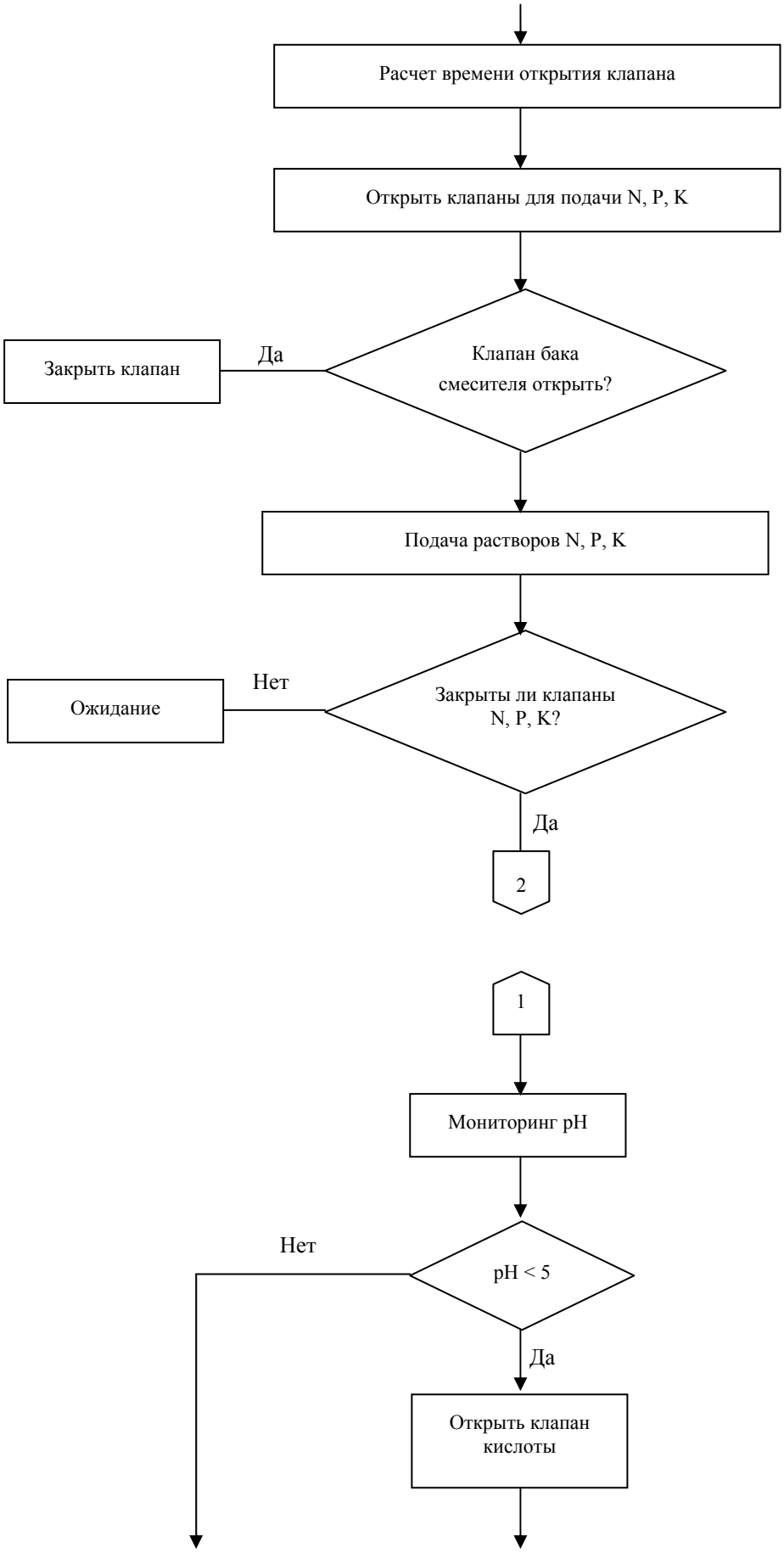
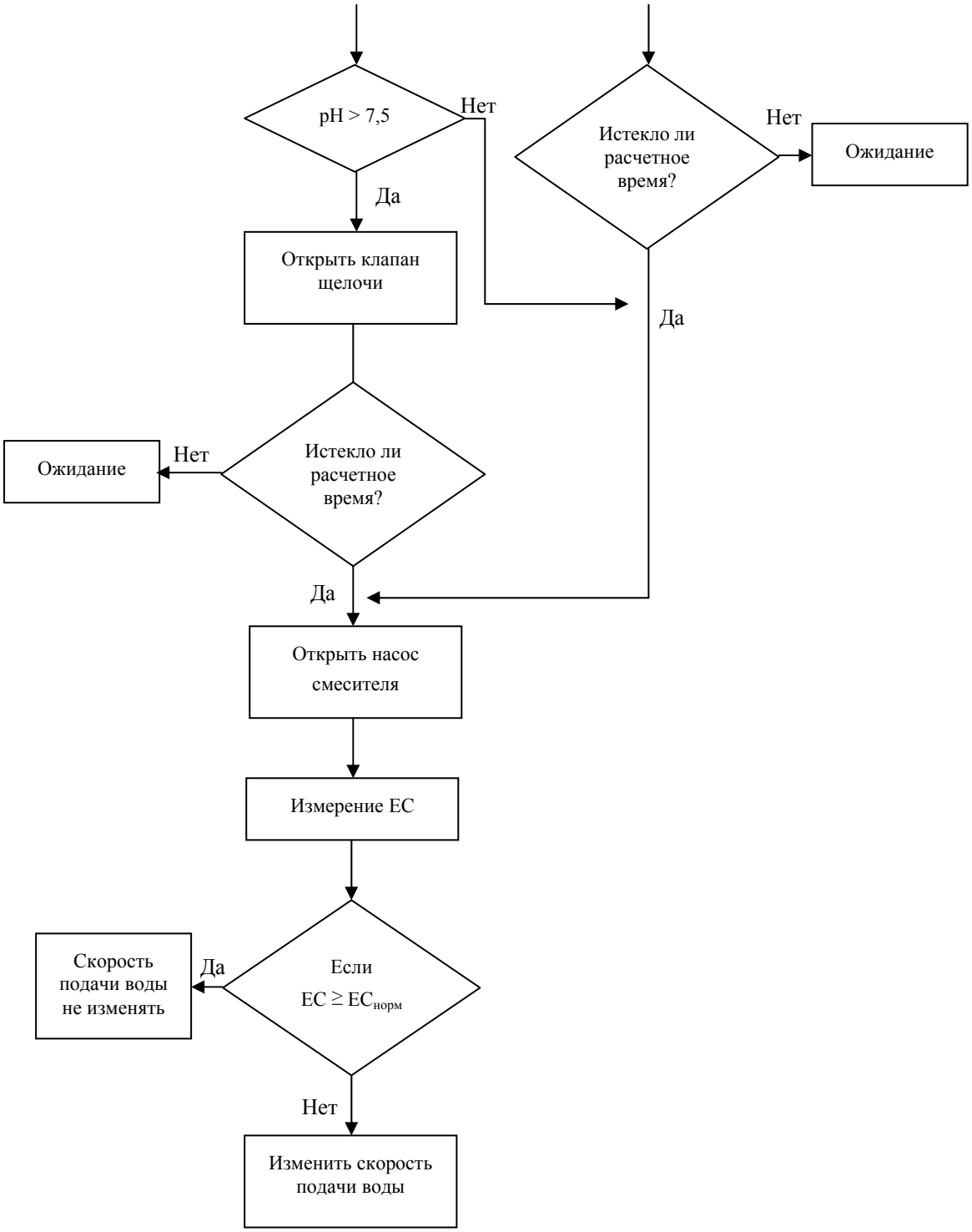


Рис. 7. Принципиальная схема модуля подготовки и подачи минеральных удобрений для капельного орошения

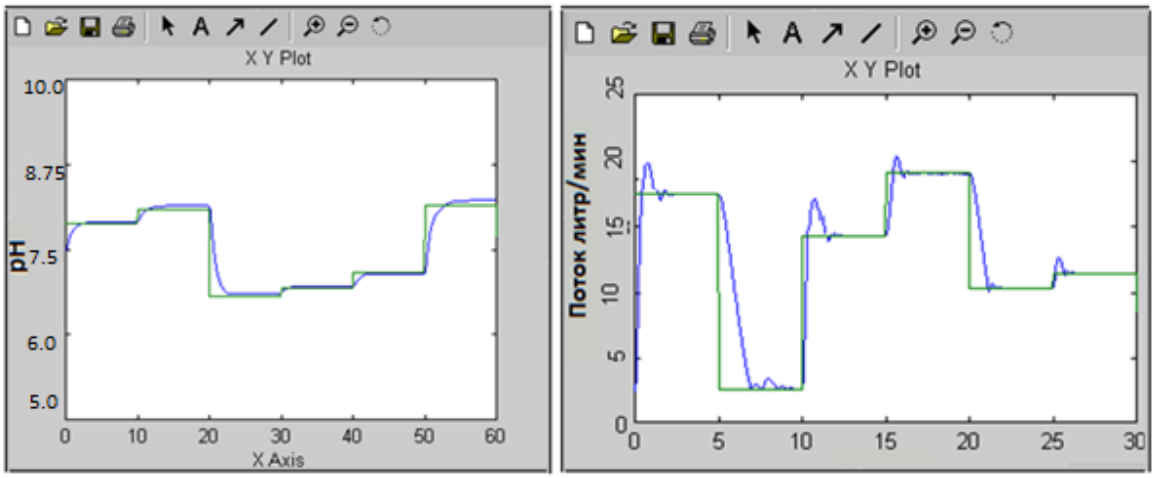
Далее мы приводим алгоритм управления системой фертигации:




Расчет времени открытия клапана
Открыть клапаны для подачи N, P, K
Клапан бака смесителя открыть?
Да
Закрыть клапан
Подача растворов N, P, K
Закрыты ли клапаны N, P, K?
Нет
Ожидание
Да
2
1
Мониторинг pH
pH < 5
Нет
Да
Открыть клапан кислоты



Чтобы проверить работоспособность системы управления, функционирующей на основе нечеткой логики, был проведен ряд экспериментов на модели в программной среде Matlab/Simulink. На рис. 9, *а* показан результат моделирования уровня pH в системе. На графике видно, что система отрабатывает изменения уровня pH (плавные линии) с достаточной точностью (< 5 %). Уровень pH для большинства сельскохозяйственных культур колеблется от 5 до 8.



а б

Рис. 9. Результаты моделирования нечетких контроллеров:
а – регулирование уровня pH; *б* – регулирование уровня EC.
Для регулирования уровня pH принят раствор кислоты с концентрацией 22 %

Для моделирования работы контроллера EC мы выбрали скорость подачи поливной воды от 4 до 25 л/мин, что соответствует диапазону концентрации питательного раствора от 10 до 62 % по сравнению со стандартным (принятым в литературе) раствором. На рис. 9, *б* показан результат моделирования подачи раствора для регулирования уровня EC раствора.

Выводы

По результатам моделирования работы системы нечеткого регулирования уровня pH и величины EC можно сделать следующие выводы:

- каждый подмодуль может работать автономно, что упрощает разработку и позволяет в дальнейшем добавлять модули по мере необходимости для улучшения и корректировки работы подмодуля;
- стабильность работы системы была продемонстрирована на примерах для производства сельскохозяйственных культур (хлопчатника);
- нечеткое управление легко адаптируется, просто и легко реализуется;
- структура нечеткого контроллера является эффективным вариантом управления подачей минерализованной воды в корневую систему растения в силу применения научно обоснованных методов управления ростом и развитием растения, которые заложены в базе знаний интеллектуальной системы.

Список литературы

1. *Bar-Yosef B.* Fertilization under drip irrigation // Fluid Fertilizer, Science and Technology / Ed. by D. A. Palgrave. New York: Marcel Dekker, 1992. P. 285–329.
2. *Садридинов С.* Инновационные подходы и факторы повышения продуктивности сельскохозяйственных культур в условиях Таджикистана. Душанбе, 2005.
3. *Kafkafi U.* Global aspects of fertigation usage // Proc. of International Symposium on Fertigation. Beijing, China, 2005. P. 8–22.
4. *Goumopoulos Ch.* An Autonomous Wireless Sensor / Actuator Network for Precision Irrigation in Greenhouses, Smart Sensing Technology for Agric. & Environ. Monitor. Berlin; Heidelberg: Springer Verlag, 2012. P. 1–20.

Материал поступил в редколлегию 25.05.2018

R. M. Bandishoeva

M. S. Osimi Tajik Technical University
10 Academicians Rajabovs Ave., Dushanbe, 734042, Tajikistan

risolatbm@mail.ru

DEVELOPMENT OF THE AUTOMATED MODULE FOR MINERALIZATION OF WATER

The article is devoted to the development of the management system of fertigation. The system allows automatic feeding of fertilizers taking into account the pH and EC values obtained from the respective sensors. To solve this problem, a mineralization module consisting of two submodules has been developed: a submodule for measuring and controlling pH and a submodule for measuring and regulating EC. The data is transferred to the microcontroller for further system management actions. As a decision-making system, the work uses an intelligent system based on fuzzy logic.

Keywords: fertigation, fuzzy controller, mineralization, concentration, solution.

References

1. Bar-Yosef B. Fertilization under drip irrigation. In: *Fluid Fertilizer, Science and Technology*. Ed. by D. A. Palgrave. New York, Marcel Dekker, 1992, p. 285–329.

2. Sadridinov S. Innovative approaches and factors of increase of productivity of agricultural crops in the conditions of Tajikistan. Dushanbe, 2005.

3. Kafkafi U. Global aspects of fertigation usage. *Proc. of International Symposium on Fertigation*. Beijing, China, 2005, p. 8–22.

4. Goumopoulos Ch. An Autonomous Wireless Sensor / Actuator Network for Precision Irrigation in Greenhouses, Smart Sensing Technology for Agric. & Environ. Monitor. Berlin, Heidelberg, Springer Verlag, 2012, p. 1–20.

For citation:

Bandishoeva R. M. Development of the Automated Module for Mineralization of Water. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 64–73. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-64-73

УДК 004.048:519.765
DOI 10.25205/1818-7900-2018-16-3-74-86

Т. В. Батура [1, 2], **А. М. Бакиева** [1]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Институт систем информатики им. А. П. Ершова СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

tatiana.v.batura@gmail.com, m_aigerim0707@mail.ru

СОЗДАНИЕ СИСТЕМЫ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ НАУЧНЫХ ТЕКСТОВ

Описан новый метод автоматического реферирования текстов. На основе предложенного метода создана система, позволяющая получать краткие аннотации научно-технических текстов и определять их темы. Процесс реферирования состоит из пяти основных шагов: предобработка, риторический анализ и преобразование текста, оценка весов, выбор предложений и сглаживание. Предлагаемый метод формирует аннотацию на основе наиболее значимых предложений исходного документа. Значимость предложений частично определяется в процессе риторического анализа, который выполняется с помощью дискурсивных маркеров и коннекторов. Также учитываются ключевые слова, многословные термины и некоторые специальные слова, которые часто встречаются в научно-технических текстах. Для извлечения ключевых слов и определения тем текста применялась аддитивная регуляризация тематических моделей.

Ключевые слова: автоматическое реферирование, теория риторических структур, дискурсивные маркеры, аддитивная регуляризация, тематические модели.

Введение

Ввиду стремительного увеличения объемов текстовой информации в Интернете активные исследования в области компьютерной лингвистики сохраняют свою актуальность. Разработка алгоритмов и создание систем автоматического реферирования, поиска и извлечения информации, классификации и кластеризации текстовых документов по-прежнему являются сложными задачами.

Подход, основанный на применении дискурсивного анализа, используется довольно широко для решения различных задач компьютерной лингвистики. Подробный обзор литературы, представленный в работе [1], показывает, что в большинстве случаев дискурсивный анализ способен улучшить качество автоматических систем на 4–44 % в зависимости от конкретной задачи.

В работе [2] теория риторических структур применяется для определения важных предложений в документе. Автор представляет входной текст в виде набора деревьев и предлагает использовать алгоритм ограничений для объединения этих деревьев. Далее применяется несколько эвристик для выбора более подходящих деревьев при формировании реферата. Автоматизированная многоязычная система реферирования текста SUMMARIST описана в [3]. Эта система сочетает в себе методы понятийного уровня знаний о мире, методы информационного поиска и статистические методы. Алгоритм состоит из трех этапов: иденти-

Батура Т. В., *Бакиева А. М.* Создание системы автоматического реферирования научных текстов // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 74–86.

© Т. В. Батура, А. М. Бакиева, 2018

фикация темы, интерпретация и генерация. SUMMARIST формирует аннотации на пяти языках: английском, японском, испанском, индонезийском и арабском.

Система автореферирования научных статей, основанная на дискурсивном анализе, описана в [4]. В ней определены семь риторических категорий. Автор работы [5] применил теорию риторических структур для создания графического представления документа. На основе структурного анализа текста вычисляются веса предложений, из которых в итоге получается краткая аннотация. В работе [6] обсуждается создание реферата, содержащего не только информацию из одного конкретного документа, но и дополнительные знания из других, похожих на него по тематике документов.

С. Митхун описывает подход, базирующийся на схемах, для формирования аннотаций на основе запросов, в которых используются структуры дискурса [7]. Этот подход выполняет четыре основные задачи, а именно: категоризация вопроса, идентификация риторических предикатов, выбор схемы и обобщение. Автор создал систему BlogSum и оценил ее производительность относительно релевантности и согласованности вопросов. Полученные результаты показывают, что предлагаемый подход решает проблему несоответствия и дискурсивной несогласованности автоматически созданных рефератов.

Исследования в этой области для английского языка достигли достаточно высокого уровня, но для текстов на русском языке данная область изучена сравнительно мало. Анализ подходов для решения проблемы автоматического формирования рефератов научно-технических текстов на русском языке проводился российскими учеными в работах [8; 9]. В исследовании [8] описаны методы и алгоритмы, учитывающие нелинейный и иерархический характер текста. С помощью риторических отношений решается проблема экстракции (извлечения фрагментов текста). С. А. Тревгода разработал систему, основанную на правилах вывода и узкоспециализированном словаре, ключевых фраз. Гибридный подход, предложенный П. Г. Осмининым [9], сочетает методы экстракции и абстракции. Этот подход был реализован автором в системе реферирования, ориентированной на автоматический перевод. Описанная система построена для текстов по теме «математическое моделирование». Были использованы не только риторические структуры, но и глаголы из предметной области «математической логики». С помощью найденных ключевых слов определяется вес предложения, затем полученная аннотация формируется в соответствии с шаблонами.

Некоторые особенности риторических отношений описаны в работах [10; 11]. Там также формулируются утверждения о свойствах этих признаков. Работа [12] описывает опыт построения корпуса на русском языке, содержащего дискурсивные маркеры. Корпус общедоступен, включает в себя тексты разных жанров, таких как научный, научно-популярный, новостной. Прежде чем использовать теорию риторических структур, ее приходится адаптировать для конкретного языка. Это связано с грамматическими особенностями. В своей статье авторы предлагают иерархию риторических отношений, которая, согласно их исследованиям, является наиболее удобной и корректной для работы с текстами на русском языке.

В нашей работе описывается подход, позволяющий получать краткие аннотации научно-технических текстов и определять их темы. Предлагаемый метод формирует аннотацию на основе наиболее значимых предложений исходного документа. Значимость предложений частично определяется в процессе риторического анализа. Для определения тем текстов применяется метод аддитивной регуляризации тематических моделей (АРТМ) [13]. Этот метод позволяет решить проблему неединственности и неустойчивости при помощи введения дополнительных ограничений на требуемое решение. В качестве регуляризаторов могут использоваться: сглаживание и разреживание распределений терминов в темах, сглаживание и разреживание распределений тем в документах и др.

Семантический анализ и особенности риторических отношений

Теория риторических структур – одна из наиболее широко используемых теорий организации текстов [14]. Согласно ей, изначально текст делится на неперекрывающиеся фрагменты, а именно на элементарные дискурсивные единицы (ЭДЕ). Кроме того, последовательные ЭДЕ связаны риторическими отношениями. Эти части известны как элементы, из которых

строятся более крупные фрагменты текстов и целые тексты. Каждый фрагмент по отношению к другим фрагментам выполняет определенную роль. Текстовая связь формируется с помощью тех отношений, которые моделируются между фрагментами в тексте.

В теории риторических структур можно определить два типа ЭДЕ. Один из них, называемый ядром, считается наиболее важной частью высказывания, другой, называемый сателлитом, поясняет ядро и считается вторичным. Ядро содержит основную информацию, тогда как сателлит содержит дополнительную информацию о ядре. Сателлит часто непонятен без ядра. Между тем выражения, в которых сателлит удален, могут быть поняты лишь в некоторой степени. Рассмотрим следующий пример:

Текст: *Дом выглядел неплохо.* ***Кроме того***, *цена была подходящая.*
Маркер: *Кроме того*
Название отношения: Elaboration

Для удобства введем следующие обозначения:

x – ядро; y – маркер; z – сателлит;
$S(x)$ – предикат для ЭДЕ, которая является ядром;
$S'(x)$ – предикат для ядра, которое начинается с прописной буквы, т. е. находится в начале предложения;
$S(z)$ – предикат для ЭДЕ, которая является сателлитом;
$S'(z)$ – предикат для сателлита, который начинается с прописной буквы;
y' – маркер с прописной буквы;
$p(\ )$ – символ пунктуации, аргументом может быть ".", ",", ":", ";".

Теперь приведенный пример может быть представлен в виде формулы исчисления предикатов: $S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.)$.

Формальное описание преобразования текста

В предлагаемом подходе риторический анализ используется на этапе построения квазиреферата. Под квазирефератом понимается перечень наиболее значимых предложений текста. Упрощенно этот этап можно описать следующим образом. Сначала необходимо найти в тексте ядерные ЭДЕ. Далее следует преобразовать высказывания, содержащие эти ЭДЕ, так, чтобы получился сокращенный текст, являющийся промежуточным между исходным текстом и готовой аннотацией. В зависимости от разных маркеров и дискурсивных отношений эти преобразования будут разными. Для формального описания действий, выполняемых системой, было принято решение использовать логику предикатов первого и второго порядков.

Предикаты первого порядка

Согласно обозначениям, введенным в предыдущем разделе, для рассмотренного примера действия, выполняемые системой на этом этапе, могут быть записаны в таком виде:

$$S'(x) \wedge p(.) \wedge y' \wedge S(z) \wedge p(.) \rightarrow S'(x) \wedge p(.) \wedge \neg(y' \wedge S(z) \wedge p(.)).$$

А именно, вначале надо найти маркер y = «*кроме того*», потом необходимо удалить его вместе с сателлитом, оставив предыдущее предложение, которое является ядерным ЭДЕ.

Для предикатов, представленных выше, мы ввели специальные *действия*, которые выполняются для создания квазиреферата. Они зависят от некоторых глаголов, существительных, маркеров и коннекторов.

Маркеры (*дискурсивные маркеры*) – это слова или фразы, которые не имеют реального лексического значения, но вместо этого обладают важной функцией формировать разговор-

ную структуру, передавая намерения говорящих. Примеры маркеров и действий, связанных с ними, приведены в табл. 1.

Коннекторы – группы слов, заменяющие маркеры и характеризующие определенные риторические отношения. Коннекторы обеспечивают связь между фразами, они показывают семантическую неполноту предложения. Например, «*в связи с этим*», «*вместе с тем*», «*тем самым*» и т. д. (табл. 2).

Таблица 1

Действия для маркеров

№	Риторические отношения	Маркеры	Действия
1	Elaboration	Кроме того	SAVE_DELETE
2	Cause-Effect	Поэтому	DELETE_SAVE
3	Contrast	Однако	SAVE_DELETE
4	Restatement	Другими словами	SAVE_DELETE
5	Elaboration	Например	SAVE_DELETE
6	Evidence	Таким образом	DELETE_SAVE

Таблица 2

Действия для коннекторов

№	Риторические отношения	Коннекторы	Действия
1	Elaboration	В связи с этим	SAVE_SAVE
2	Elaboration	Вместе с тем	DELETE_SAVE
3	Elaboration	Тем самым	SAVE_SAVE

Во время исследования мы создали словарь, состоящий из 121 маркеров и коннекторов, 120 существительных и 108 глаголов с весами, которые часто встречаются в научных и технических текстах. Всего было рассмотрено восемь действий. Ниже описаны некоторые действия.

DELETE_SAVE – это действие удаляет предыдущее предложение и сохраняет предложение с заданным маркером.

SAVE_DELETE – это действие сохраняет предыдущее предложение и удаляет предложение с заданным маркером.

SAVE_SAVE – это действие полностью сохраняет предложение с заданным маркером и предыдущим предложением.

Предикаты второго порядка

Как известно, в сложноподчиненном предложении выделяются главное и придаточное предложение. В этом случае ЭДЕ более низкого уровня вложены в ЭДЕ более высокого уровня. Для описания действий с вложенными ЭДЕ удобнее использовать предикаты второго порядка. Чтобы проиллюстрировать, как текст преобразуется в случае вложенных ЭДЕ, приведем следующий пример:

Кроме того, воздух, который поступает в морозильную камеру, уже охлаждают до 1.5 °C с помощью холодильной установки док-станции, которая составляет около 50 % тепловой нагрузки поступающего воздуха. **Таким образом**, чистым эффектом охлаждаемой док доставки является уменьшение инфильтрации нагрузки. Чистая прибыль равна разнице между уменьшением инфильтрации нагрузки морозильной камеры и холодильной нагрузки в доке судоходства. **Обратите внимание**, что док-холодильники работают при значительно более высоких температурах (1.5 вместо –23 °C) и потребляют значительно меньше энергии на ту же сумму охлаждения.

Для того чтобы записать преобразования в формальном виде, добавим следующие обозначения:

m – ядро в придаточном предложении;
n – сателлит в придаточном предложении;
$S(m)$ – предикат для ядра m;
$S(m)$ – предикат для ядра m, начинающегося с прописной буквы;
$S(n)$ – предикат для сателлита n;
$S'(n)$ – предикат для сателлита n, начинающегося с прописной буквы;
y – маркер.

$$S'(z) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(x) \wedge p(.) \wedge S'(z) \wedge p(.) \rightarrow$$
$$\neg S'(El \wedge p(,) \wedge S(m) \wedge p(.)) \wedge \neg S'(Ev \wedge p(,)) \wedge S'(S(m) \wedge p(.)) \wedge S'(x) \wedge p(.) \wedge$$
$$\neg S'(Cont \wedge S(n) \wedge p(.)),$$

где
$y = El =$ «*кроме того*»;
$y = Ev =$ «*таким образом*»;
$y = Cont =$ «*обратите внимание*».

Следует отметить, что использование формализмов логики первого и второго порядка с данной целью пока недостаточно исследовано. В будущем, возможно, придется дополнить этот формализм, чтобы учитывался порядок следования элементов в тексте.

Общее описание системы

Пусть T – текст статьи, очищенный после предварительной обработки и состоящий из предложений

$$T = \bigcup_{k=1}^{P} S_k.$$

$D = \{d_1, d_2, ..., d_N\}$ представляет собой набор дискурсивных маркеров и коннекторов, которые содержатся в этом тексте.

$V = \{v_1, v_2, ..., v_M\}$ представляет собой набор глаголов и существительных, которые часто встречаются в научных и технических текстах.

В нашем понимании задача реферирования состоит в том, чтобы найти преобразование текста T в реферат $\tilde{T}$ такое, что $\Phi: T \rightarrow \tilde{T}$, $|\tilde{T}| < |T|$. Тогда алгоритм построения реферата можно записать в виде последовательных этапов.

1. Предобработка текста. На этапе предварительной обработки из исходного текста удаляются все изображения, таблицы, предложения с формулами, информация об авторах и библиографические ссылки. Авторские аннотации были убраны и отдельно сохранены, чтобы потом можно было оценить систему, путем сравнения результата с исходной аннотацией.

2. Риторический анализ и преобразование текста. На этом шаге обнаруживаются предложения, содержащие дискурсивные маркеры и коннекторы. К этим предложениям применяются определенные действия (см. выше). В результате получается квазиреферат: $\Phi(T, D, V) = T'$.

3. Оценка весов предложений. Для формирования аннотации подсчитываются веса предложений. Опишем эту процедуру подробнее.

Пусть S'_k – произвольное предложение квазиреферата

$$T' = \bigcup_{k=1}^{P_1} S'_k.$$

При вычислении веса каждого предложения квазиреферата учитывается наличие в этом предложении ключевых слов (или многословных терминов), дискурсивных маркеров и коннекторов, а также некоторых слов, которые характерны для научных текстов. Для извлечения из текстов многословных терминов используется алгоритм Turbotopics, разработанный для определения значимых *n*-грамм в английских текстах [15]. В ходе создания системы мы адаптировали алгоритм Turbotopics для работы с текстами на русском языке.

В итоге вес каждого предложения вычисляется по следующей формуле:

$$SW = \frac{1}{L} \cdot \sum_{i=1}^{L} w_i + \frac{1}{M} \cdot \sum_{j=1}^{M} v_j + \frac{1}{N} \cdot \sum_{k=1}^{N} d_k,$$

где

$W = \{w_1, ..., w_l\}$ – множество ключевых слов и многословных выражений $(|W| = L)$;

$V = \{v_1, ..., v_M\}$ – множество значимых глаголов и существительных, которые часто встречаются в научных текстах $(|V| = M)$;

$D = (d_1, ..., d_N)$ – множество дискурсивных маркеров и коннекторов $(|D| = N)$.

4. Выбор предложений. Из полученного набора предложений (см. п. 2) для аннотации отбираются только те, вес которых (см. п. 3) превышает заданную пороговую величину β:

$$\tilde{T} = \bigcup_{k=1}^{N_1} \{S'_k : SW > \beta\},$$

где β = 0,15 является константой, которая определяется эмпирически.

5. Операция сглаживания – процедура преобразования текста, позволяющая получить связный текст из разрозненных фрагментов и при необходимости дополнительно сократить его. Например, в процессе сглаживания заменяются или удаляются некоторые слова или словосочетания и т. д. В табл. 3 приведены примеры предложений до сглаживания и после него.

Таблица 3

Примеры сглаживания

До сглаживания	После сглаживания
Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ, **так как** в некоторых ситуациях эпизоды могут быть настолько продолжительными, что задержки процесса обучения, связанные с необходимостью завершения эпизодов, будут слишком велики.	Данное преимущество TD-методов часто имеет решающее значение при использовании в ИС РВ.
Поскольку, **как уже отмечалось**, использование БСД позволяет анализировать лишь один из возможных диагнозов.	**Выявлено что**, использование БСД позволяет анализировать лишь один из возможных диагнозов.

Для сглаживания предложений используются шаблоны и индикаторы. Мы рассмотрели два типа шаблонов: для удаления фрагментов предложений (в случае когда аннотация получилась длиннее 250 слов) и для дополнения (в случаях, когда в аннотацию попал фрагмент незаконченного предложения).

Индикаторы – комбинации слов, влияющие на вес предложения. В состав индикаторов входят определенные значимые слова, которые мы включили в созданную лингвистическую базу знаний. Примеры индикаторов и действий, связанных с ними, представлены в табл. 4. В данный момент рассмотрено 95 индикаторов.

Таблица 4

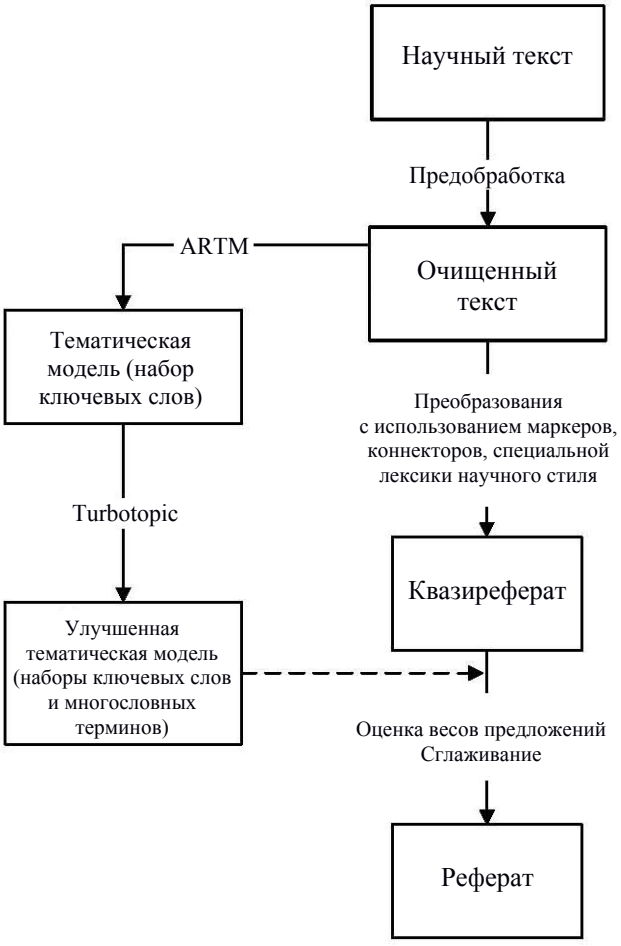
Действия для индикаторов

№	Индикатор	Действие	Результат
1	В нашей статье	REPLACE	В статье
2	Существенным является	DELETE_UNTIL_DOT	–
3	Следует подчеркнуть	DELETE_COMMA	–
4	Важным представляется	REPLACE	–

Действия, используемые при сглаживании:

REPLACE – замена индикатора на другое слово или словосочетание, или удаление данного индикатора;
DELETE_UNTIL_DOT – удаление до конца предложения, начиная с данного индикатора;
DELETE_COMMA – удаление фрагмента предложения до следующей запятой;
DELETE – удаление всего предложения.

В ходе данной работы была разработана система, блок-схема которой представлена далее:



Тематическое моделирование заключается в построении модели некоторой коллекции текстовых документов. В такой модели каждая тема представляется дискретным распределением вероятностей слов, а документы – дискретным распределением вероятностей тем.

В настоящее время существуют разные методы тематического моделирования, такие как PLSA, LDA, ARTM. Главное преимущество тематических моделей в сравнении с нейронными сетями заключается в том, что они хорошо поддаются интерпретации, пользователю понятны причины обнаружения определенных тем в тексте и структура самих тем. Кроме того, часто требуется, чтобы тематические модели учитывали разнородные данные, выявляли динамику тем во времени, автоматически разделяли темы на подтемы, использовали не только отдельные ключевые слова, но и многословные термины и т. д.

Чтобы выбрать алгоритм тематического моделирования, мы провели ряд экспериментов, результаты которых представлены в работе [16]. Было принято решение использовать алгоритм ARTM в реализации библиотеки BigARTM [17]. Благодаря своей универсальности и гибкости настройки параметров моделей ARTM позволяет комбинировать регуляризаторы, тем самым комбинируя тематические модели. Этот метод гарантирует единственность и устойчивость решения. У ARTM не наблюдается увеличение количества параметров модели с ростом числа документов, поэтому он может применяться к большим наборам данных. Кроме того, предложенная нами модификация позволяет использовать не только однословные, но и многословные выражения, что, на наш взгляд, повышает интерпретируемость модели.

Результаты

Наша система была протестирована на коллекции из 261 научной статьи на русском языке. Эта коллекция собрана на основе выложенных в открытом доступе архивов журналов «Программные продукты и системы» за 2013–2017 гг. [1] Далее приведен пример сравнения автоматически полученной и авторской аннотаций.

Автоматически полученная аннотация

В настоящей работе для решения краевой задачи для уравнения Пуассона используются различные графические ускорители и библиотека cuFFT для быстрого преобразования Фурье. В настоящей работе решение системы находится с использованием библиотеки CULA, реализующей ряд процедур пакета LAPACK на основе технологии NVIDIA CUDA. Расчеты осуществлялись на системах с различными GPU, в том числе на суперкомпьютере Ломоносов НИВЦ МГУ. Особенность использования этой библиотеки в том, что она не содержит процедур для синус-преобразований, которые в данном случае необходимы для того, чтобы полученное решение удовлетворяло нулевым граничным условиям. Процедура решения будет состоять из следующих этапов. И такие задачи могут возникать при обработке больших БД экспериментальных измерений. В зависимости от ускорителя она составляет от 44 до 150 Гфлопс, что соответствует 410 % от пиковой. Задача моделирования работы масс-спектрометров на основе ионного циклотронного резонанса и преобразования Фурье может быть решена на гибридных системах. В статье представлены результаты реализации кода Pic3D частиц в ячейке для моделирования работы масс-спектрометров на основе ионного циклотронного резонанса и преобразования Фурье на гибридных системах с CPU и GPU. Вычисления показывают, что ускорители могут быть использованы для определения кулоновского взаимодействия ионов с помощью решения первой краевой задачи для уравнения Пуассона и параллельного вычисления полей от каждой поверхности электрода через решение алгебраических систем с достаточной эффективностью.

Авторская аннотация

В работе представлено исследование эффективности параллельных программ для расчета эволюции ионов в рамках модели частиц в ячейке. Разработаны параллельные программы для гибридных вычислительных систем, содержащих устройства CPU (Central Processing Units, центральные процессоры) и GPU (Graphic Processing Units, графические ускорители). Программы применяются для прямого моделирования поведения ионов в ловушках масс-спектрометров на основе преобразования Фурье. Показана возможность использования GPU-устройств для ускорения многократного решения краевых задач для уравнения Пуассона на основе быстрого преобразования Фурье, реализованного в библиотеке cuFFT – библиотеке процедур быстрого преобразования Фурье для архитектуры CUDA (Compute Unified Device Architecture). Проведено сравнение реально достигаемой производительности и задействования полосы пропускания памяти при вычислении решения с пиковыми характеристиками GPU для разных установок. Показано, что выбранный алгоритм решения первой краевой задачи для уравнения Пуассона масштаби-

[1] Международный журнал «Программные продукты и системы». URL: http://www.swsys.ru/.

руется в соответствии с асимптотической оценкой сложности. Разработаны программы расчета полей, удерживающих ионы в ловушке в произвольной геометрии электродов для работы на гибридных системах, сочетающих в себе одновременную обработку данных на CPU и GPU. В каждом из параллельных процессов программы расчета поля решение алгебраических уравнений осуществляется на GPU через процедуры LAPACK, реализованные в составе библиотеки CULA. Результаты расчетов на суперкомпьютере «Ломоносов» показали, что эффективность параллельного использования GPU существенно зависит от выбранной схемы распределения процессов параллельной программы. Ускорители могут эффективно использоваться для определения кулоновского взаимодействия ионов с помощью решения первой краевой задачи для уравнения Пуассона. Параллельные вычисления полей на CPU от каждой поверхности электрода можно проводить совместно с решением алгебраических систем на GPU с достаточной эффективностью.

Пока не существует общепринятого эффективного способа автоматической оценки систем автореферирования [18]. Во-первых, мы пробовали оценить качество полученных аннотаций при помощи метрики ROUGE, основанной на подсчете количества совпадающих элементов в сравниваемых текстах, например, *n*-грамм или предложений [19]. В метрике ROUGE в случае подсчета совпадающих предложений текст аннотации рассматривается как последовательность предложений. Основная идея состоит в том, что чем длиннее самая длинная общая подпоследовательность *LCS* двух предложений в сравниваемых аннотациях, тем более похожими считаются эти две аннотации. Как правило, используют *F*-меру на основе *LCS* для оценки сходства между двумя величинами *X* длиной *m* и *Y* длиной *n*, считая, что *X* является образцом для сравнения, а *Y* – просматриваемый элемент. Точность, полнота и *F*-мера согласно ROUGE определяются следующим образом:

$$P_{lcs} = \frac{LCS(X,Y)}{n}, \quad R_{lcs} = \frac{LCS(X,Y)}{m},$$

$$F_{lcs} = \frac{\left(1+\beta^2\right)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}},$$

где $LCS(X, Y)$ – длина самой длинной общей подпоследовательности X и Y, а $\beta = P_{lcs} / R_{lcs}$.

Были получены следующие значения метрики ROUGE: точность 32,8 %, полнота 59,04 %, *F*-мера 34,47 %. К сожалению, в работах [8; 9], которые описывают системы обработки текстов на русском языке, не приводятся значения метрики ROUGE, поэтому нет возможности сравнить эти результаты с нашими. Также мы пришли к выводу, что некорректно сравнивать результаты работы нашей системы с результатами работы систем для английского языка, такими как, например, [20], поскольку низкие значения ROUGE могут быть связаны с особенностями языкового строя. В частности, русский язык является флективным языком с развитой морфологией, к тому же порядок слов в русском языке относительно свободный.

Во-вторых, мы воспользовались экспертной оценкой. Точность полученных аннотаций, оцененная экспертами, оказалась значительно выше. Экспертная оценка результатов реферирования показала, что 86,43 % полученных рефератов совпали с авторскими рефератами по содержанию или незначительно отличались от них (что на самом деле не всегда свидетельствует о плохом качестве реферата), и только 13,57 % представляли собой некорректно отобранные фрагменты текстов. Следует заметить, что полученная нами экспертная оценка выше, чем в работах [8; 9].

Нами было замечено, что авторы часто используют синонимы, перефразируют и меняют местами предложения. Экспертная оценка подтверждает, что порядок предложений в аннотации часто не влияет на ее общий смысл. Однако метрика ROUGE не учитывает это. Кроме того, иногда автоматически сформированная аннотация получается длиннее, чем хотелось бы (около 500 слов вместо 250). Это связано со стилем изложения самой статьи, и чаще всего означает, что в тексте имеется много содержательных предложений.

В-третьих, мы рассмотрели точность, полноту и *F*-меру, которые вычисляются способом, похожим на предложенный в работах [2; 8]. Поясним подробнее. Предположим, что автоматически полученная аннотация содержит в себе множество ключевых слов и многословных терминов W_1, множество специальных слов из научных и технических текстов V_1, множество

дискурсивных маркеров и коннекторов D_1. Объединение этих множеств обозначим N_1: $N_1 = W_1 \cup V_1 \cup D_1$. Аналогичные множества можно выделить в эталонной авторской аннотации N_2: $N_2 = W_2 \cup V_2 \cup D_2$. Тогда точность, полноту и *F*-меру будем вычислять по следующим формулам:

$$P = \frac{|N_1 \cap N_2|}{|N_1|}, \quad R = \frac{|N_1 \cap N_2|}{|N_2|},$$

$$F\text{-measure} = \frac{2 \cdot P \cdot R}{P + R}.$$

Сравнительная оценка результатов приведена в табл. 5.

Таблица 5

Оценка результатов, %

Система	Точность	Полнота	*F*-мера
Scientific Text Summarizer	**75,23**	**68,21**	**71,55**
Trevgoda 2009	67,03	64,81	66,03
Marcu 1998	73,53	67,57	70,42

Преимущество предложенных формул состоит в том, что они позволяют определить вклад каждого из признаков и разных комбинаций этих признаков в общую оценку результата. Например, можно оценить вклад только маркеров и коннекторов или только специальной научной лексики, или того и другого, но без ключевых слов и выражений и т. д. В дальнейшем мы планируем провести подобное исследование данного вопроса.

Возможное улучшение предложенного в данной статье алгоритма, по нашему мнению, состоит в том, чтобы дополнить правила удаления менее важных предложений, увеличить количество шаблонов для сглаживания, расширить списки маркеров, коннекторов и индикаторов.

Заключение

Описан подход к автоматическому построению аннотаций научно-технических текстов на русском языке. Процесс состоит из пяти шагов: предобработка, преобразование текста, оценка весов, выбор предложений и сглаживание. Преобразование предложений включает риторический анализ. На основе дискурсивных маркеров и коннекторов извлекаются наиболее значимые предложения в тексте. Также учитываются ключевые слова, многословные выражения и специальная лексика, которая часто присутствует в научных и технических текстах. Далее из полученного набора предложений для аннотации выбираются только те, вес которых превышает заранее заданную пороговую величину. Операция сглаживания позволяет получить более связный текст. В дальнейшем планируется провести эксперименты с текстами из различных научных областей на других языках.

Список литературы

1. *Ананьева М. И., Кобозева М. В.* Разработка корпуса текстов на русском языке с разметкой на основе теории риторических структур // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». М., 2016. URL: www.dialog-21.ru/media/3460/ananyeva.pdf

2. *Marcu D.* Improving summarization through rhetorical parsing tuning // VI Workshop on Very Large Corpora. 1998. P. 206–215.

3. *Hovy E., Lin Ch.-Y.* Automated text summarization and the SUMMARIST system // Proc. of the TIPSTER Text Program. 1998. P. 197–214.

4. *Teufel S., Moens M.* Summarizing scientific articles: experiments with relevance and rhetorical status // Computational Linguistics. 2002. Vol. 28 (4). P. 409–445.

5. *Bosma W.* Query-Based Summarization using Rhetorical Structure Theory // 15th Meeting of CLIN. 2005. P. 29–44.

6. *Huspi S. H.* Improving Single Document Summarization in a Multi-Document Environment. PhD Thesis. Melbourne, Australia: RMIT University, 2017. 190 p.

7. *Mithun S.* Exploiting rhetorical relations in blog summarization. PhD Thesis. Montreal, Canada: Concordia University, 2012. 230 p.

8. *Тревгода С. А.* Методы и алгоритмы автоматического реферирования текста на основе анализа функциональных отношений: Дис. … канд. техн. наук. СПб., 2009. 157 с.

9. *Осминин П. Г.* Построение модели реферирования и аннотирования научно-технических текстов, ориентированной на автоматический перевод: Дис. … канд. филол. наук. Челябинск, 2016. 239 с.

10. *Batura T. V., Bakiyeva A. M., Yerimbetova A. S. Mit'kovskaya M. V. Semenova N. A.* Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory // Bulletin of the Novosibirsk Computing Center. Series: Computer Science. 2016. Is. 40. P. 37–51.

11. *Бакиева А. М., Батура Т. В.* Исследование применимости теории риторических структур для автоматической обработки научно-технических текстов // Cloud of Science. Вып. 4, № 3. С. 450–464.

12. *Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A.* Towards building a discourse-annotated corpus of Russian // Computational Linguistics and Intellectual Technologies. 2017. Iss. 16 (23). Vol. 1. P. 194–204.

13. *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // International Conference on Analysis of Images, Social Networks and Texts (AIST). Ekaterinburg, Russia, 2015. P. 370–384.

14. *Mann W., Thompson C.* Rhetorical structure theory: Toward a functional theory of text organization // Text-Interdisciplinary Journal for the Study of Discourse. 1988. Vol. 8. No. 3. P. 243–281.

15. *Blei D. M., Lafferty J. D.* Visualizing Topics with Multi-Word Expressions // Semantic Scholar. 2009. URL: https://arxiv.org/pdf/0907.1013.pdf

16. *Батура Т. В., Стрекалова С. Е.* Подход к построению расширенных тематических моделей текстов на русском языке // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 5–18.

17. *Vorontsov K.* Welcome to BigARTM's documentation! 2015. URL: http://bigartm.readthedocs.io/en/stable/

18. *Das D., Martins A. A.* Survey on Automatic Text Summarization // Literature Survey for the Language and Statistics II course at CMU. 2007. P. 192–195.

19. *Lin Ch. Y.* ROUGE: A Package for Automatic Evaluation of Summaries // Workshop On Text Summarization Branches Out. 2004. P. 74–81.

20. *Zhang J. J., Chan H. Y., Fung P.* Improving lecture speech summarization using rhetorical information // IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). 2007. P. 195–200.

Материал поступил в редколлегию 02.07.2018

T. V. Batura [1, 2], **A. M. Bakiyeva** [1]

[1] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

[2] *A. P. Ershov Institute of Informatics Systems SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

tatiana.v.batura@gmail.com, m_aigerim0707@mail.ru

DEVELOPING THE SYSTEM FOR AUTOMATIC SUMMARIZATION OF SCIENTIFIC TEXTS

The paper describes a new method of automatic text summarization. Based on this method, a system has been created that makes it possible to obtain summaries of scientific and technical texts and to determine their topics. The summarization process consists of five main steps: preprocessing, transformation, weight evaluation, sentence selection, and smoothing. The proposed method allows receiving the summary based on important sentences of the original document. The importance of sentences is partially determined in the process of rhetorical analysis, which is performed using discursive markers and connectors. Keywords, multiword terms, and some special words that are often found in scientific and technical texts are also taken into account. We used additive regularization for topic modeling (ARTM) to extract keywords and discover the topics.

Keywords: automatic summarization, rhetorical structure theory, discourse markers, additive regularization, topic modeling.

References

1. Ananyeva M. I., Kobozeva M. V. Development the corpus of Russian texts with markup based on the Rhetorical Structure Theory. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2016"*. Moscow, 2016. URL: www.dialog-21.ru/media/3460/ananyeva.pdf/ (in Russ.)

2. Marcu D. Improving summarization through rhetorical parsing tuning. *VI Workshop on Very Large Corpora*, 1998, p. 206–215.

3. Hovy E., Lin Ch.-Y. Automated text summarization and the SUMMARIST system. *Proc. of the TIPSTER Text Program*, 1998, p. 197–214.

4. Teufel S., Moens M. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 2002, vol. 28 (4), p. 409–445.

5. Bosma W. Query-Based Summarization using Rhetorical Structure Theory. *15th Meeting of CLIN*, 2005, p. 29–44.

6. Huspi S. H. Improving Single Document Summarization in a Multi-Document Environment. PhD Thesis. Melbourne, Australia, RMIT University, 2017, 190 p.

7. Mithun S. Exploiting rhetorical relations in blog summarization. PhD Thesis. Montreal, Canada, Concordia University, 2012, 230 p.

8. Trevgoda S. A. Methods and algorithms of automatic text summarization based on the analysis of functional relations. PhD Thesis. St. Petersburg, Russia, 2009, 157 p. (in Russ.)

9. Osminin P. G. Construction of a model for abstracting and annotating scientific and technical texts focused on automatic translation. PhD Thesis. Chelyabinsk, Russia, 2016, 239 p. (in Russ.)

10. Batura T. V., Bakiyeva A. M., Yerimbetova A. S. Mit'kovskaya M. V. Semenova N. A. Methods of constructing natural language analyzers based on Link Grammar and rhetorical structure theory. *Bulletin of the Novosibirsk Computing Center. Series: Computer Science*, 2016, is. 40, p. 37–51.

11. Bakiyeva A. M., Batura T. V. Research of applicability of the rhetorical structure theory for automatic processing of scientific and technical texts. *Cloud of Science*, 2017, vol. 4, no. 3, p. 450–464. (in Russ.)

12. Pisarevskaya D., Ananyeva M., Kobozeva M., Nasedkin A., Nikiforova S., Pavlova I., Shelepov A. Towards building a discourse-annotated corpus of Russian. *Computational Linguistics and Intellectual Technologies*, 2017, iss. 16 (23), vol. 1, p. 194–204.

13. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. *International Conference on Analysis of Images, Social Networks and Texts (AIST)*. Ekaterinburg, Russia, 2015, p. 370–384.

14. Mann W., Thompson C. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 1988, vol. 8, no. 3, p. 243–281.

15. Blei D. M., Lafferty J. D. Visualizing Topics with Multi-Word Expressions. *Semantic Scholar*, 2009. URL: https://arxiv.org/pdf/0907.1013.pdf

16. Batura T. V., Strekalova S. E. An Approach to Building Extended Topic Models of Russian Texts. *Vestnik NSU. Series: Information Technologies*, vol. 16, no. 2, p. 5–18. (in Russ.)

17. Vorontsov K. Welcome to BigARTM's documentation! 2015. URL: http://bigartm.readthedocs.io/en/stable/

18. Das D., Martins A. A. Survey on Automatic Text Summarization. *Literature Survey for the Language and Statistics II course at CMU*, 2007, p. 192–195.

19. Lin Ch. Y. ROUGE: A Package for Automatic Evaluation of Summaries. *Workshop On Text Summarization Branches Out*, 2004, p. 74–81.

20. Zhang J. J., Chan H. Y., Fung P. Improving lecture speech summarization using rhetorical information. *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, p. 195–200.

For citation:

Batura T. V., Bakiyeva A. M. Developing the System for Automatic Summarization of Scientific Texts. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 74–86. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-74-86

УДК 004.6, 004.04
DOI 10.25205/1818-7900-2018-16-3-87-104

М. А. Городничев [1–3], **А. В. Комиссаров** [1, 3], **А. В. Можина** [1, 3], **П. В. Прочкин** [1, 3]
П. Д. Рудыч [1], **А. В. Юрченко** [1]

[1] *Институт вычислительных технологий СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

[2] *Институт вычислительной математики и математической геофизики СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

[3] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

maxim@ssd.sscc.ru, yurchenko@ict.nsc.ru

МОДЕЛИ И ПРОЕКТНЫЕ РЕШЕНИЯ СИСТЕМЫ ХРАНЕНИЯ И ОБРАБОТКИ ИССЛЕДОВАТЕЛЬСКИХ ДАННЫХ ECCLESIA

В ходе научных исследований порождается большое количество данных в цифровом формате, и для последующего использования этих данных (обработки, анализа, публикации) их необходимо организованно собирать и хранить. Построение информационной инфраструктуры для решения этих задач – одна из наиболее актуальных проблем в области организации работы с экспериментальными данными. Авторами настоящей статьи разрабатывается информационная система для автоматизации сбора, хранения и анализа данных, в качестве отправной точки для которой используются три задачи обработки данных из области физиологии. Рассмотрены и проанализированы возникающие в процессе разработки такой системы проблемы, а также существующие подходы и готовые решения этих и схожих задач. На основе результатов проведенного анализа предложен ряд моделей и механизмов для решения возникших проблем. Разработанные решения включают в себя модели и механизмы сбора и хранения экспериментальных данных, модель для описания и формализации сценариев обработки данных и механизмы для обработки собранных данных в распределенной вычислительной системе. В результате представлена архитектура вычислительной системы для сбора, хранения и обработки экспериментальных данных. Система предлагается в качестве инструмента для решения широкого спектра задач, возникающих при проведении научных исследований и требующих сбора, хранения и многоэтапной обработки различных типов данных.

Ключевые слова: управление научными данными, информационная система, обработка и анализ данных, данные физиологических исследований, объектное хранилище данных, функциональный язык, распределенные вычисления.

Введение

В современных исследованиях часто приходится сталкиваться с большим объемом экспериментальных данных, получаемых из различных источников. Всё чаще данные собираются и сохраняются с расчетом на то, что некоторые задачи их анализа будут поставлены в будущем. Перспективные задачи могут предполагать использование дополнительных данных, собранных независимо, в другое время другими исследователями. Нарастающие

Городничев М. А., *Комиссаров А. В.*, *Можина А. В.*, *Прочкин П. В.*, *Рудыч П. Д.*, *Юрченко А. В.* Модели и проектные решения системы хранения и обработки исследовательских данных Ecclesia // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 87–104.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2018. Том 16, № 3
© М. А. Городничев, А. В. Комиссаров, А. В. Можина, П. В. Прочкин,
П. Д. Рудыч, А. В. Юрченко, 2018

проблемы организации исследовательских данных и систематизации работы с ними стали, таким образом, самостоятельной задачей.

Управление научными (исследовательскими) данными (research data management) и «цифровое курирование» (digital curation) [1] к настоящему моменту являются устоявшимися терминами, выражающими потребности научного сообщества в инструментах и сервисах для работы с генерируемыми данными. В то время как в Европе и США эта проблематика вынесена на государственный и даже межгосударственный уровень: поддерживаются и развиваются узкоспециализированные инфраструктуры для хранения научных данных, например DataONE [1], разрабатываются и реализуются крупные программы, направленные на обоснование необходимости развития таких инфраструктур и проработку соответствующей нормативной базы, которая, в частности, будет стимулировать исследователей делиться своими данными (см. проект «Research Data e-Infrastructures: Framework for Action in H2020» [2]).

В обзоре [1] приводится «западная» оценка общей стоимости формирования инфраструктуры научных данных – 10–15 % от стоимости всей инфраструктуры науки. При этом в России на уровне федеральных властей обсуждение необходимости и путей создания инфраструктуры для науки, основанной на цифровых данных, активизировалось только в 2016–2017 гг. в связи с разработкой и принятием программы «Цифровая экономика» [2], хотя отдельные инициативы в этом направлении реализовывались и ранее [3; 4].

Несмотря на более чем 10-летнюю историю вопроса, работа с исследовательскими данными, за исключением отдельных научных направлений (например, [5]), остается слабо упорядоченной и в лучшем случае организуется путем составления и следования инструкциям, таким как «Data management: Helping MIT faculty and researchers manage, store, and share data they produce» [3]. Такое положение дел в дальнейшем будет только сдерживать развитие науки, поэтому исследователи нуждаются в создании инструментов, которые позволят им решать задачи организации сбора, хранения, обработки и анализа данных, обмена ими и их публикации.

В работе [6] предлагается подход к систематизации и соответствующей автоматизации процессов, связанных с согласованным хранением и обработкой неоднородных исследовательских данных. В общем виде рассмотрены требования к организации специализированной информационно-аналитической системы, представлено видение её архитектуры, описано состояние развития инфраструктуры Института вычислительных технологий (ИВТ) СО РАН для работы с научными данными.

В настоящей работе представлен следующий этап в рамках инкрементального подхода к созданию этой информационно-аналитической системы поддержки научных исследований. Анализируется опыт Лаборатории технологий анализа и обработки биомедицинских данных ИВТ СО РАН в решении задач сбора, хранения и анализа данных, возникших в рамках ряда физиологических исследований. На основе этого анализа формулируются минимальные требования к информационно-аналитической системе для работы с такими данными, предлагаются основные проектные решения по созданию ее прототипа – системы Ecclesia.

Одним из немногих проектов, нацеленных на целостное решение проблемы автоматизации сбора, хранения, обработки и публикации научных данных, является проект разработки платформы и сервиса для биомедицинских исследований Galaxy Project [4]. Несмотря на направленность проекта на конкретную предметную область, результаты его могут быть использованы и в других областях научных знаний. Однако сервис достаточно сложен в освоении, не поддерживает включение интерактивных сессий с пользователями в качестве этапов автоматически выполняющихся сценариев обработки данных и более приспособлен к накоплению отдельных объектов данных, воспроизводимости отдельных сценариев обработки данных, чем к систематизации накопления и автоматизации применения знаний о предметной области.

[1] https://www.dataone.org
[2] http://www.in2p3.fr/actions/informatique/media/h2020.pdf
[3] https://libraries.mit.edu/data-management/
[4] https://galaxyproject.org

Системы для хранения или обработки данных в целом можно разделить на три класса: системы общего назначения (например, Dell EMC Elastic Cloud [5], Globus [6]), специфичные для предметной области (например, SIMBioMS [7], CARMEN [8], XTENS 2 [9]) и решающие частные задачи (например, MedMining [7]). В системах общего назначения особое внимание уделяется надежности хранения, но не предоставляются инструменты для описания данных и их связей. Системы для решения частных задач, напротив, поддерживают связи между данными, но в рамках специфичной для задачи схемы, которая чаще всего не переносима на другие задачи. Среди специфичных для предметной области систем рассматривались системы хранения биомедицинских данных [8–10]. Достаточной степенью универсальности не обладает ни одна система. Так, например, они не поддерживают возможность проверки структуры загруженных данных и не позволяют вводить и тем более строить новые связи между данными.

Среди систем формирования и исполнения сценариев обработки данных также можно выделить системы общего назначения (например, Google Cloud Dataflow [10] или ActiveEon ProActive Workflows & Scheduling [11]) и системы, специфичные для предметной области (например, LabView [12] и Unipro UGENE [13] [11]). Системы общего назначения по большей части направлены на пользователей с навыками программирования. Важным примером системы, позволяющей создавать предметно-ориентированные визуальные языки программирования, является CoCoViLa [12], которая для этого требует от пользователя решать более широкую задачу, чем задание и выполнение отдельных сценариев обработки данных. Узкопредметные системы избавляют пользователей от таких сложностей, но изначально не приспособлены для решения междисциплинарных задач.

Разрешению обозначенных проблем различных систем для работы с данными и посвящена наша работа.

Анализ «пользовательских историй»

В основе требований, предъявляемых к разрабатываемой системе, лежит опыт решения задач сбора, хранения и анализа данных для ряда физиологических исследований и результаты обсуждения соответствующих проблем со специалистами-физиологами.

Задача персонализированной телемедицины:
индивидуальный подбор программы тренировки

Рассматривается задача создания велотренажера для реабилитации больных после инсульта. Врач задает программу тренировки пациента с помощью графиков требуемой скорости вращения педалей и величины сопротивления тренажера кручению педалей. Эти значения преобразуются мобильным приложением в управляющие воздействия контроллера тренажера. Необходимо индивидуально подбирать программу тренировки так, чтобы пациент получил нагрузку, адекватную его состоянию. Для этого необходимо собирать и анализировать данные, характеризующие фактическое прохождение тренировки. Собираемые данные включают в себя частоту сердечных сокращений (ЧСС) и реальную скорость кручения педалей. Мобильное приложение собирает данные и формирует пары вида «отметка времени, значение ЧСС» и «отметка времени, значение реальной скорости».

Для анализа процесса тренировки требуется сохранять и визуализировать получаемые данные, сопоставляя их с программой тренировки. Все данные являются одноканальными

[5] https://www.dellemc.com/ru-ru/storage/ecs/
[6] https://www.globus.org/
[7] https://www.simbioms.org/
[8] http://www.carmen.org.uk/
[9] https://github.com/xtens-suite/xtens-app
[10] https://cloud.google.com/dataflow/
[11] https://proactive.activeeon.com
[12] http://www.ni.com/ru-ru/shop/labview.html
[13] http://ugene.net

временными рядами (ОВР). Программу тренировки также можно представить в виде ОВР. Требуется делать выборки в заданных временных интервалах как отдельных рядов, так и их совокупностей, совмещенных по времени. Предложено ввести абстракцию ОВР на уровне решения для хранения, что позволяет унифицировано представлять данные. Помимо данных ряда предложено сохранять *атрибуты* ряда:

1) тип ряда (в данном случае некоторые идентификаторы, позволяющие отличать ряды ЧСС от рядов скорости кручения педалей и др.);
2) идентификатор устройства-источника;
3) время начала и окончания записи.

Доступ к системе хранения данных предоставляется через веб-сервис, который позволяет сохранять и извлекать ОВР. Данные ОВР хранятся в бинарных файлах, атрибуты отдельных записей со ссылкой на бинарный файл хранятся в реляционной БД. Решение позволяет унифицировано сохранять и извлекать ОВР и независимо разрабатывать различные клиентские приложения для работы с ними. В частности, реализовано сохранение данных в мобильном приложении и выгрузка данных в веб-интерфейс для визуализации.

Поскольку на уровне понятий в системе была поддержана логика работы с ОВР, пользователи получили возможность выбирать данные в веб-интерфейсе по заданным временным промежуткам записи и типу ряда, а не по именам или датам создания файлов. С пользователей также сняты типичные задачи организации данных в некоторую структуру директорий, запоминания этой структуры и ручного поиска в ней.

В ходе решения задачи проявилась необходимость обеспечивать в рамках информационно-аналитической системы реализацию понятий предметной области и поддержку работы с соответствующими объектами данных в терминах предметной области на уровне программных и пользовательских интерфейсов. Типичная система организации данных в виде файлов, собранных в структуры директорий, является в подобных ситуациях неадекватной задачам поиска и обработки данных.

Задача предобработки записей ЭЭГ

Запись электроэнцефалограммы (ЭЭГ) состоит из N каналов, в которых фиксируется динамика разности потенциалов между N электродами и электродом-референтом. ЭЭГ-записи могут использоваться для выявления схожести и различий в реакции пациентов на одинаковые стимулы [13], для восстановления позиций источников сигналов в мозге [14] и др. Однако предваряет содержательный анализ данных очистка ЭЭГ-записей от шумов [15].

В рамках этапа очистки от шумов используются запись сигналов ЭЭГ, а также информация о размещении электродов на голове пациента [16], время подачи стимулов, протокол эксперимента, в котором описаны детали проведения эксперимента. Приемы очистки записи могут применяться отдельно либо в некоторой последовательности. Наиболее распространенные из них:

1) применение частотных фильтров для удаления известных помех;
2) выявление экспертом зашумленных каналов и их удаление;
3) смена электрода-референта и соответствующий перерасчет значений ЭЭГ во всех каналах [17];
4) усиление значимого сигнала за счет суммирования выровненных по моменту предъявления стимула участков записи, так называемых эпох;
5) коррекция нулевой линии (baseline correction) [18];
6) выявление экспертом поврежденных эпох и их удаление;
7) применение анализа независимых компонент [19], выявление экспертом шумовых компонент и их удаление; каждая компонента при этом визуализируется в виде карты активности мозга.

Как правило, в физиологических лабораториях можно наблюдать, что все собираемые данные, а также данные, получаемые в ходе этапов обработки, хранятся как отдельные файлы (протокол эксперимента и вовсе может храниться в бумажном лабораторном журнале), структурирование которых в упорядоченные системы каталогов, именование, формирование наборов данных для очередных этапов обработки осуществляются вручную. Автоматизация

сбора и обработки должна освободить исследователей от этого рутинного труда и ликвидировать ошибки, обусловленные человеческим фактором. Должна быть обеспечена возможность формировать и автоматически выполнять сценарии, состоящие из набора связанных операций обработки данных, а перенос данных между операциями должен быть автоматизирован.

На практике в ходе обработки данных от пользователя требуется выбор конкретных методов. Например, существует множество алгоритмов для проведения ICA [20], нужно сделать выбор частотных фильтров и т. д. Не все эксперты в предметной области задачи являются экспертами в математической обработке данных, и для них были бы ценны советы при выборе методов обработки. Для этого целесообразно, чтобы система собирала и накапливала опыт других пользователей в составлении сценариев.

Нужно учитывать, что некоторые операции обработки данных пока (или принципиально) не могут быть автоматизированы, поэтому может потребоваться анализ и принятие решения экспертом. Таким образом, должна быть заложена возможность включать сессии взаимодействия с пользователями в качестве этапов сценариев обработки данных.

Задача сопоставления характеристик томограмм головного мозга

Задача состоит в поиске нейрональных сетей на основе анализа выявляемой с помощью функциональной магнитно-резонансной томографии (фМРТ) активности мозга пациента (например, [21–23]). С помощью фМРТ измеряется связанная с активностью нейронов гемодинамика. Каждый снимок фМРТ характеризует насыщенность крови кислородом в различных областях головного мозга. Дискретное представление насыщенности основано на разбиении куба, в который вписана голова пациента, на кубические единичные объемы – воксели. Снимок фМРТ представляет собой, таким образом, трехмерный массив чисел.

В ходе решения задачи анализируется последовательность фМРТ-снимков головы пациента. В каждом снимке выделяются группы вокселей (по выбору исследователя), и для каждой группы строится индекс активации (некоторое усреднение значений всех вокселей группы). Эти операции повторяются для всей последовательности снимков, таким образом получается временная развертка индексов активации. Представляет интерес выделение групп вокселей, демонстрирующих различную или сходную динамику индекса активации [24].

Для одного эксперимента в этой задаче нужно обработать тысячи групп вокселей во временной развертке. Каждая группа вокселей может быть рассмотрена независимо, что позволяет обрабатывать их параллельно. Таким образом, задачу целесообразно решать на параллельных вычислительных системах.

Требования к системе

На основе анализа «пользовательских историй» в совокупности с общими представлениями о назначении системы [6] сформулированы требования к минимальному жизнеспособному продукту (MVP – Minimal Viable Product). Необходимо обеспечить следующее:

1) возможность сохранять и извлекать разнородные экспериментальные данные вместе с информацией об их структуре и связях с другими данными;

2) возможность модифицировать сведения о данных и связях между ними в любой момент времени;

3) проверку соответствия данных заданной структуре с возможностью исправления выявленных несоответствий при сохранении в системе;

4) возможность задавать сложные пользовательские сценарии обработки данных;

5) обращение к данным и сохранение новых данных из сценариев;

6) выполнение сценариев на параллельных вычислительных системах, при этом пользователь должен быть по возможности избавлен от необходимости императивного параллельного программирования;

7) возможность интерактивного вмешательства пользователя в ход исполнения сценария;

8) накопление знаний о предметной области и поддержку действиям пользователя в построении сценария;

9) автоматизированный анализ и оптимизацию процессов обработки данных.

Архитектура системы и проектные решения

На основе выявленных требований к системе в ходе проектирования были выделены три ее относительно самостоятельные подсистемы:

1) хранения и управления данными;
2) формирования сценариев;
3) организации распределенной обработки.

Для возможности независимой разработки подсистем и применения различных технологий для их реализации выбрана микросервисная модель архитектуры системы, представленная на рис. 1.

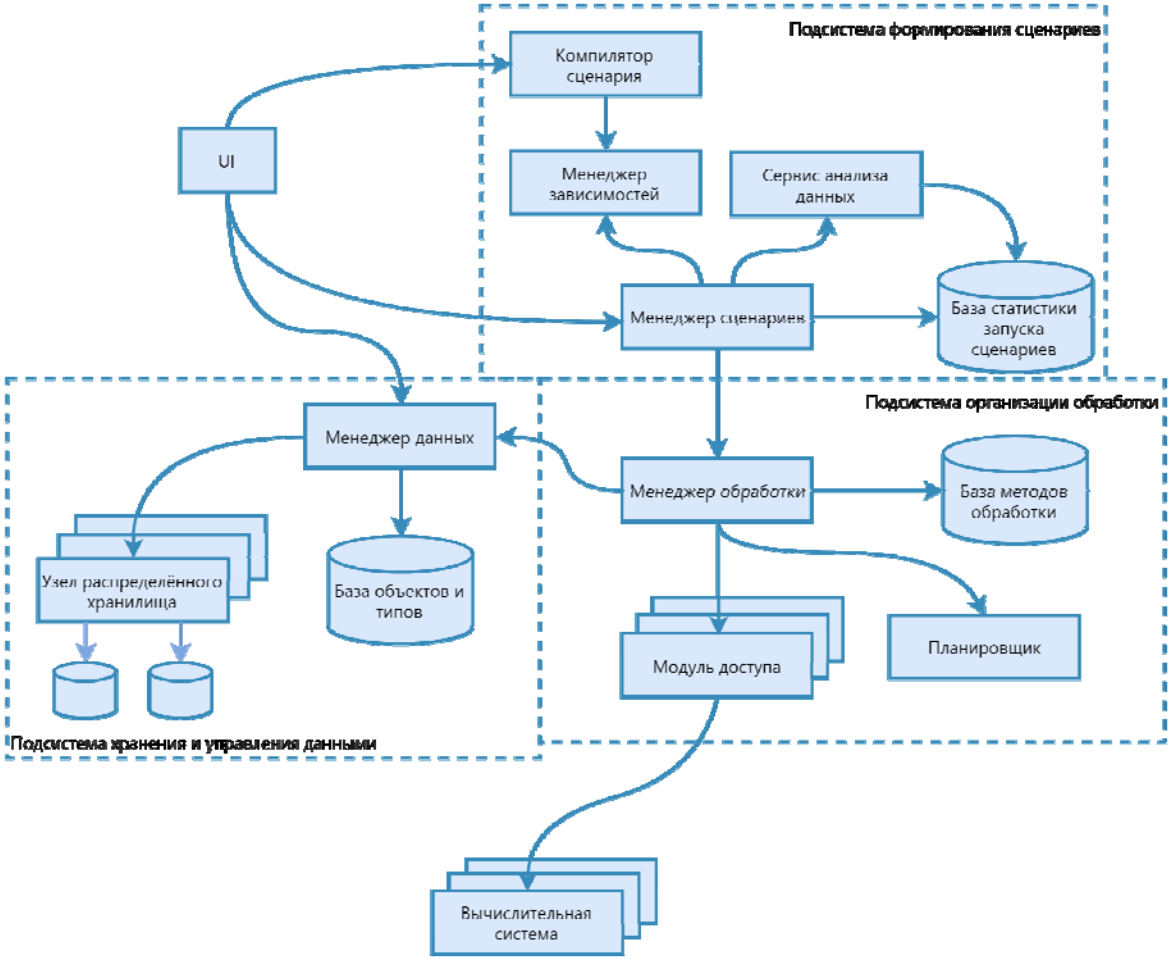


Рис. 1. Архитектура системы

Подсистема хранения и управления данными реализуется в виде надстройки над существующим распределенным высоконадежным хранилищем. Надстройка представляет собой промежуточный сервер (менеджер данных), организующий данные пользователя в соответствии с предлагаемой моделью представления данных (см. ниже) и предоставляющий программный интерфейс (API – Application Programming Interface) для работы с этими дан-

ными. На основе API менеджера данных независимо могут быть построены различные клиентские приложения. Базовым пользовательским интерфейсом является веб-интерфейс с функциями загрузки и извлечения данных, просмотра каталога данных и поиска в нем.

Модель представления данных

Для обеспечения хранения и извлечения разнородных экспериментальных данных вместе с информацией об их структуре и связях предложена следующая модель представления данных. Вводится понятие *объекта данных* как единицы хранения. Для каждого объекта при создании формируется его идентификатор. Все объекты автоматически или по указанию пользователя относятся к одному из *типов данных*, зарегистрированных в системе, в одном из *форматов данных*. Таким образом, любой файл или некоторым образом структурированный набор файлов, будучи сохраненными в качестве объекта данных, могут быть содержательно интерпретированы системой. Например, данные одного ЭЭГ-исследования на энцефалографах BrainVision сохраняются в виде тройки файлов: запись сигналов ЭЭГ в файле формата *.eeg, файл разметки ЭЭГ по времени подачи стимулов (*.vmrk) и файл, содержащий информацию о каналах энцефалографа (*.vhdr). После загрузки файлов в систему хранения эта тройка рассматривается как один объект данных типа *Запись ЭЭГ* в формате *BrainVision*. Такой подход отражает связь между файлами, позволяет обращаться к ним совместно и избегать ошибок совмещения сигналов и времени подачи стимулов от разных записей при многократном обращении.

Для обеспечения расширяемости подсистемы хранения предлагается механизм введения нового типа. Новый тип в системе задается *схемой метаданных*, списком возможных *форматов* (задаются строковыми идентификаторами) и набором *операций*, которые могут принимать на вход или вырабатывать в качестве своего результата объекты данного типа, возможно, наряду с объектами других типов (см. раздел «Подсистема организации распределенной обработки»).

Схема метаданных задает название типа, синонимы для названия и текстовое описание типа, а также специфицирует набор свойств объектов данного типа в виде множества пар (атрибут, тип). Атрибут – произвольная строка, имя некоторого свойства объекта данных, тип – строковый идентификатор типа значения атрибута, выбирается из типов, зарегистрированных в системе, например: атрибут «Дата регистрации», тип «Дата». Одним из типов является ссылка на другой объект данных по идентификатору объекта. Ссылки в метаданных на другие объекты позволяют задавать произвольные *связи* между данными. С каждым объектом данных хранятся его метаданные. Значения атрибутов заполняются пользователем или автоматически, например, некоторым клиентским программным обеспечением.

Типы, у которых задана только схема метаданных, называются *типами-стикерами* и могут быть включены в определения схем метаданных других типов для переиспользования набора атрибутов, заданных в типе-стикере. Экземпляр такого типа, стикер, – объект метаданных, содержание которого составляют только метаданные с присвоенными атрибутам значениями. Стикер может быть прикреплен к произвольному объекту данных, даже если этот стикер не включен в схему метаданных типа данного объекта.

С помощью атрибутов-ссылок и стикеров можно, в частности, сохранить связь между исходным объектом, например записью ЭЭГ, и преобразованным, например ЭЭГ, отфильтрованной от наведения электросети. Для этого нужно создать и прикрепить к полученному после фильтрации объекту стикер с названием «Без наведения» с атрибутом «Получено из» типа *Запись ЭЭГ* для сохранения ссылки на исходный объект и атрибутом «Частота наведения» типа вещественное число для сохранения параметров фильтрации (рис. 2).

Согласно требованию о проверке структуры загружаемых данных, в подсистеме вводится понятие *валидатора* – операции проверки соответствия структуры данных заданному типу и формату этих данных. Вызов этих операций при загрузке данных в хранилище позволяет автоматически выявлять несоответствие между содержанием объекта данных и указанными пользователем или автоматически определенными типом и форматом объекта данных.

Модель доступа к данным

Разработан программный интерфейс веб-сервиса доступа к данным в терминах управления ресурсами сервиса, согласно архитектурному стилю REST [25]. Типы ресурсов сервиса имеют прямое соответствие терминам модели представления данных (тип данных, формат

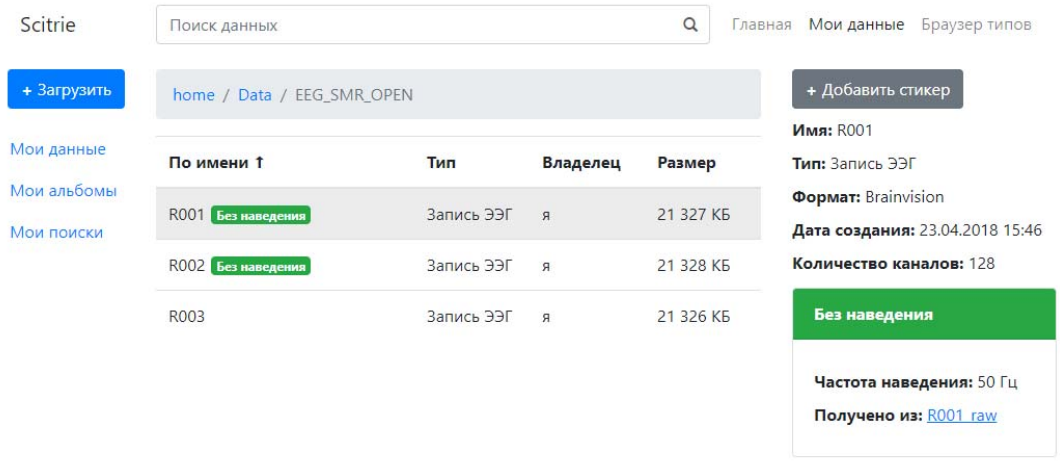


Рис. 2. Пример визуализации стикера
для преобразованной записи ЭЭГ в пользовательском интерфейсе

данных, стикер, операция, объект). Таким образом, каждый тип данных, каждый объект и т. д., а также коллекция таких сущностей, является ресурсом (в терминах REST) и получает имя (URI – Uniform Resource Identifier), посредством которого клиенты сервиса могут обращаться к ресурсу и осуществлять действия из набора CRUD [26].

Относительно способов передачи значений при запросе объектов типы разделены на два класса.

1. *Ссылочные типы*. В результате запроса объекта по его идентификатору клиенту интерфейса возвращается информация (URL – Unified Resource Locator) о том, как может быть получен доступ непосредственно к данным объекта. К этому классу относятся все сложные типы (например, запись ЭЭГ). Для получения данных (в случае записи ЭЭГ в формате BrainVision это тройка файлов, описанная выше) необходимо сделать отдельный запрос по URL. Ссылочные типы позволяют вынести в отдельный блок работ исследование и внедрение механизмов оптимизации доступа к данным, например, таких, как размещение копий данных (кэши) ближе к месту их использования.

2. *Типы значений.* Объекты таких типов возвращаются клиенту непосредственно при обращении по идентификатору объекта. К этому классу относятся примитивные типы (например, числа и строки).

Предложенная модель доступа к данным предоставляет инструменты доступа к системе хранения как из клиентских приложений (веб-интерфейс системы хранения, приложения для персональных компьютеров и мобильных устройств, программное обеспечение регистрирующего оборудования – собираемые устройствами данные могут непосредственно отправляться в хранилище, и т. д.), так и из подсистемы распределенной обработки данных.

Подсистема формирования сценариев
Модель абстрактного параллельного вычислителя

Поскольку одним из требований к системе является выполнение сценариев на различных параллельных вычислителях, то предлагается модель, в которой вычислитель может исполнять сценарии, описанные в виде частично упорядоченного множества операций. Операция характеризуется множеством входов, множеством выходов, а также именем функции, которую она применяет к входам для получения выходов. Входы и выходы – это объекты дан-

ных. Предполагается, что по имени функции может быть найдена конкретная программная реализация этой функции (программа), которую может исполнить вычислитель. Все объекты неизменяемы и либо известны до исполнения сценария (входные данные сценария), либо являются выходами операций.

Каждая операция имеет список зависимостей – список операций, которые должны быть выполнены до исполнения этой задачи. Операция может быть исполнена, если нет неисполненных операций, от которых она зависит. Таким образом, множество списков зависимостей определяет частичный порядок на множестве операций. Списки зависимостей формируются разработчиком сценария либо, как будет представлено далее, компилятором высокоуровневого языка описания сценариев. Возможность задавать произвольный порядок (помимо обусловленного зависимостями по данным) позволяет использовать операции с побочными эффектами, при этом с исполнительной системы (планировщика) снимается задача отслеживания и обработки побочных эффектов. Примерами побочных эффектов могут быть операции ввода / вывода или инициализации ресурсов вычислителя (например, буфера памяти GPU) в одной из операций, потребление в другой.

Вычислитель принимает на вход тройку множеств $<O, D, M>$, где O – множество операций, D – множество зависимостей между операциями, M – таблица сопоставления объектов данных, являющихся входными для сценария, с их именами (мнемониками), используемыми в сценарии. В таблице M указывается URI объекта, если это ссылочный тип, или непосредственно некоторая сериализация объекта в противном случае. Запись сценария в виде тройки $<O, D, M>$ называется внутренним представлением сценария.

Такая модель абстрагирует подсистему формирования сценариев от деталей реализации вычислителя, позволяет создавать варианты языков описания сценария (текстовые, визуальные) и оставить свободу выбора конкретной реализации сценария подсистеме организации распределенной обработки (см. ниже).

Накопление знаний о предметной области

Для накопления знаний предлагается использовать методы машинного обучения. Благодаря этому система сможет самостоятельно извлекать взаимосвязи между объектами предметной области исходя из действий пользователей. У подхода есть свои недостатки. Так, методы машинного обучения не дают однозначной интерпретируемости и доказуемой корректности вывода, но для целей создаваемой системы достаточно, чтобы сохранялась информация об истории происхождения того или иного утверждения.

Основная идея состоит в том, чтобы оценивать вероятность использования пользователем функции в контексте при формировании сценария на основе статистики других пользователей. Для этого адаптируется задача анализа рыночной корзины, сформулированная в [27] и получившая широкое развитие в 2006 г. во время конкурса Netflix Prize по предсказанию пользовательских оценок фильмам и рекомендациям фильмов на их основе [28]. Задача конкурса формулировалась так: есть множество пользователей U и множество фильмов F. Известна матрица R размерностью $|U| \times |F|$, содержащая пользовательские оценки фильмов. Требуется научиться предсказывать оценку пользователя и предлагать пользователю фильмы с наивысшей предсказанной оценкой. Для этой задачи эффективным решением оказались латентные модели [28].

В задаче рекомендации функций есть заметные отличия:

1) пользователь не ставит оценки функциям (это сильно отличается от основных пользовательских задач и будет отвлекать пользователя);

2) так как система рекомендаций нужна для помощи неопытным пользователям, необходимо строить рекомендации только на основе действий более опытных пользователей, т. е. информации от одних пользователей система должна доверять меньше, чем информации от других.

Для решения этих проблем предлагается следующее.

1. В качестве «оценки» функции здесь берется вероятность применения пользователем конкретной функции в сценарии с учетом контекста – набора других использованных в сценарии функций. Эта вероятность для каждого пользователя рассчитывается эмпирически

на основе анализа статистики применения функций в разработанных пользователем сценариях.

2. Вводится мера доверия системы пользователю, которая используется как вес пользователя при обучении латентной модели.

При расчете вероятности использования функций пользователем делаются два предположения. Первое заключается в том, что распределение вероятности использования функций пользователем не меняется со временем (это предположение можно смягчить, если рассматривать только сценарии не старше некоторого порогового значения), второе – вероятность использования функций в сценарии не зависит от других сценариев.

Для вычисления меры доверия пользователю используются следующие эвристики:

1) чем больше используется различных функций, тем лучше пользователь владеет методами обработки данных;

2) чем больше в среднем функций в сценариях пользователя, тем они сложнее и тем лучше он владеет методами обработки данных.

Исходя из этого мера доверия пользователю C_u вычисляется следующим образом:

$$C_u = w_0 + w_1 n(u) + w_2 \underline{n_s(u)},$$

где $n(u)$ – количество различных методов, применяемых пользователем, $\underline{n_s(u)}$ – среднее количество методов в сценариях пользователя, w_1, w_2 – веса, w_0 – константа-смещение (сейчас характеризует экспертную оценку навыка пользователя при его регистрации в системе).

Изначально веса устанавливаются вручную и при необходимости будут скорректированы по мере появления новых пользователей. В будущем при увеличении количества пользователей планируется собрать экспертные оценки их навыка и на их основе обучить модель линейной регрессии для автоматической настройки весов.

Язык формирования сценариев

В качестве базового языка описания сценариев разработан текстовый язык, за основу которого принят язык F# по следующим причинам:

1) F# является функциональным языком, что позволяет извлекать из кода информацию о функциональных зависимостях между операциями, абстрагирует от деталей исполнения сценария на конкретной вычислительной системе;

2) синтаксис F#, как и Python (который является де-факто стандартом в анализе данных), основан на отступах, в нем нет избыточных ключевых слов, поэтому код на F# краток и легко читаем;

3) F# основан на системе типов Хиндли – Милнера [29], поэтому в программах, как и в Python, не требуется явно указывать типы при определении значений и функций; при этом язык является статически типизированным, что позволяет обнаруживать большое количество ошибок до отправки сценария на исполнение;

4) компилятор F# обладает открытым исходным кодом [14], а также предоставляет программный интерфейс для анализа кода на предмет наличия ошибок, построения типизированного и нетипизированного абстрактного синтаксического дерева, а также поддержку таких функций интегрированных сред разработчика (IDE – Integrated Developer Environment), как автозавершение кода и сборка inline-документации.

Для адаптации языка к целям формирования сценариев, а также спецификации генерируемых параллельных программ изменена семантика некоторых конструкций языка:

1) let-связывание связывает объект данных (константу или выход функции) с мнемоникой, но при этом не определяет порядок вычислений и не задает других зависимостей, помимо зависимостей по данным;

2) do-связывание служит для явного обозначения побочных эффектов (например, операций ввода-вывода, если это требуется вычислителем), в случае использования do-связыва-

[14] https://github.com/fsharp/FSharp.Compiler.Service

ния с операцией все остальные операции, описанные в сценарии после нее, получают зависимость от этой операции (см. пример на рис. 3);

3) директива open используется как для подключения наборов сигнатур функций, доступных на вычислителе, так и для переиспользования других сценариев пользователя.

Набор сигнатур – это множество объявлений функций, реализации которых доступны на вычислителе в качестве исполняемых программ. Объявления оформляются в виде функций на языке F#. Они необходимы для того, чтобы произвести проверку соответствия типов в сценарии. На первом этапе наборы сигнатур добавляются в систему вручную совместно с добавлением реализаций функций в подсистему организации распределенной обработки. Впоследствии они будут генерироваться автоматически по описаниям операций.

```
open Eeg

do initConnection "ecclesia.ict.nsc.ru:5432"

let record = loadEeg "R013"

let filtered = filter 1<Hz> 40<Hz> record

let goodChannels, badChannels = splitBadChannels filtered

let epochs = extractEpochs -1000<ms> 2000<ms> goodChannels
```

Рис. 3. Пример кода на языке описания сценария (все операции должны выполняться в порядке, обусловленном зависимостями по данным, но после выполнения операции инициализации initConnection, на что указывает оператор do)

Перед запуском код сценария обрабатывается компилятором, который анализирует сценарий на наличие ошибок, выявляет зависимости между операциями и транслирует его во внутреннее представление сценария, формируя множества $<O, D, M>$.

Наборы сигнатур функций, как и сценарии, могут также подключать другие наборы и сценарии с помощью директивы open (см. рис. 3). Таким образом, исходные коды, необходимые для компиляции, образуют направленный ациклический граф. Информация о связях между исходными кодами (дугах в графе исходных кодов) выявляется компилятором и хранится совместно с исходными кодами в системе.

Компоненты подсистемы формирования сценариев

Подсистема состоит из следующих компонентов:

1) графический интерфейс пользователя (UI – User Interface) позволяет пользователю вводить сценарии обработки данных на языке описания сценариев, инициирует компиляцию сценария во внутреннее представление и передает его менеджеру сценариев;

2) компилятор сценариев анализирует сценарий на корректность и преобразует его во внутреннее представление;

3) менеджер сценариев принимает запросы от графического интерфейса, собирает информацию о сценарии в форме внутреннего представления и инициализирует его выполнение;

4) менеджер зависимостей хранит исходные коды сценариев и наборы сигнатур функций, а также связи между ними;

5) сервис анализа пользовательской статистики собирает информацию о запусках пользовательских сценариев, инициирует создание моделей машинного обучения, рассчитывает, какие функции можно предложить пользователю.

В предложенной архитектуре менеджер сценариев и компилятор никак не связаны, менеджер сценариев работает напрямую с внутренним представлением сценария. Это сделано для того, чтобы можно было подключать к системе и другие средства описания сценария, которые компилируются во внутреннее представление, например визуальные. Все компоненты представляют собой веб-сервисы и общаются друг с другом через REST API.

Подсистема организации распределенной обработки

Распределение вычислений

Обработка собранных данных зачастую является ресурсоемкой задачей из-за их большого объема или же вследствие вычислительной сложности используемых для обработки алгоритмов.

Одним из возможных решений для сокращения времени обработки является использование высокопроизводительных вычислительных систем (ВВС). Это решение позволяет существенно сократить время выполнения ресурсоемких этапов обработки, однако может быть неуместно для этапов с низкой вычислительной сложностью. Ввиду этого выполнение обработки данных исключительно на ВВС не решает проблему в полной мере, требуется оптимизация использования привлекаемых вычислительных ресурсов.

В то же время сценарии обработки данных, имеющие вид <*O*, *D*, *M*>, естественным образом выполняются в распределенной вычислительной среде: операции обработки, не зависящие друг от друга могут выполняться параллельно. Также при организации распределенной обработки собранных данных можно использовать гетерогенное множество вычислительных систем, включающее в себя как высокопроизводительные узлы для выполнения ресурсоемких этапов обработки, так и узлы для выполнения менее требовательных к вычислительным ресурсам операций. Более того, при использовании гетерогенного множества вычислителей становится возможным включать в систему специализированные узлы, более эффективно выполняющие определенные виды обработки (например, располагающие ускорителями GPU), а также вычислительные мощности, предоставляемые пользователями системы. Именно такой подход – распределенная обработка данных на гетерогенном множестве вычислительных систем – и закладывается в систему.

Модель организации распределенной обработки

На основе выбранного подхода разработана модель распределенной обработки данных в системе. Модель предполагает централизованное управление процессом обработки данных, реализуемое отдельным сервисом. Здесь с каждой операцией из внутреннего представления сценария сопоставляется реализующая ее программа, написанная на некотором языке программирования (может быть несколько таких реализаций, отличающихся нефункциональными свойствами, например, предназначенных для исполнения на различных устройствах). Эти программы обработки данных выполняются на некотором подмножестве доступных вычислительных систем, требования к системам определяются по имеющейся об операции информации (метод обработки данных, размер входных данных, нефункциональные свойства реализаций). Распределение операций по вычислительным системам происходит при составлении плана выполнения сценария.

Нужно учесть, что некоторые вычислительные системы могут и не находиться под полным контролем системы подсистемы организации распределенной обработки. Например, представляет интерес использование ресурсов суперкомпьютерных центров, однако установка системного программного обеспечения для управления обработкой данных на таких вычислителях затруднена. Решением является отказ от использования в обязательном порядке связующего программного обеспечения, разворачиваемого на используемых вычислительных системах. Вместо этого взаимодействие с вычислительными узлами решено осуществлять через штатные для используемых вычислительных систем механизмы удаленного доступа. Также если вычислительная система использует некоторую систему управления прохождением задач (СУПЗ), то задачи обработки данных Ecclesia будут ставиться в очередь СУПЗ на общих основаниях. Для обеспечения расширяемости системы новыми типами поддерживаемых вычислительных систем, механизмы взаимодействия

с вычислителями выделяются в изолированные модули внутри системы с единым внешним интерфейсом. Ограниченность контроля над вычислительными системами также требует создания виртуальной среды для всех процессов обработки данных внутри узла, для чего используется технология Docker [15].

Планирование

Для эффективного использования распределенной системы реализуется механизм оптимизирующего планирования. Используется алгоритм RDPSO (Revised Discreet Particle Swarm Optimization) поиска приближенного решения задачи в многомерном пространстве, имитирующий движение «роя частиц», адаптированный для работы с дискретным пространством решений [30]. Алгоритмом составляется план выполнения всего сценария до начала вычислений.

Не всегда можно заранее составить план выполнения всего сценария. В сценарии могут присутствовать операции, требующие взаимодействия с пользователем (например, обработки данных человеком-экспертом). Предсказать время выполнения таких операций в общем случае не представляется возможным, что делает невозможной и оптимизацию времени выполнения сценария целиком. Также к неточности оценок времени выполнения может приводить невозможность в некоторых случаях предсказать объем входных данных для некоторых операций, поскольку время выполнения операций в существенной мере зависит от объема входных данных. Эта ситуация может возникать, например, при очистке данных с различного рода датчиков от шума: соотношение объема исходных и очищенных данных неизвестно, поэтому объем очищенных данных после выполнения такой операции трудно предсказать.

Для решения этой проблемы предлагаются следующие механизмы планирования и исполнения операций обработки данных в распределенной системе.

Планирование производится в несколько проходов для подмножеств операций в сценарии. Для каждой вычислительной системы формируются очереди задач в рамках подсистемы организации распределенных вычислений. Эти очереди в контексте обсуждения алгоритма планирования следует отличать от очередей СУПЗ, которые, возможно, используются на некоторых из вычислительных систем. Для каждого прохода выбираются операции, удовлетворяющие всем следующим условиям:

- операция еще не выполняется на некоторой вычислительной системе;
- операция не зависит ни от одной незавершенной операции, требующей взаимодействия с пользователем;
- операция не зависит ни от одной операции, объем выходных данных которой не известен.

По мере выполнения сценария таким образом может быть запланировано выполнение всех операций в сценарии: операции, требующие взаимодействия с пользователем, и операции с неизвестным объемом выходных данных выполняются, необходимая для оценки времени выполнения информация становится известна.

При использовании такого механизма планирования порядок поступления операций в очереди вычислительных систем не связан ни с порядком поступления сценариев обработки данных в систему, ни с желаемыми сроками завершения выполнения сценариев. Это может привести к завершению сценариев в сроки, значительно превышающие желаемые (даже в случаях, когда нагрузка на распределенную систему и ее конфигурация позволяют завершить сценарий в срок), а также к «бесконечному ожиданию» для задач некоторых сценариев.

Чтобы избежать возникновения перечисленных проблем, предложен следующий механизм управления очередями операций для каждой вычислительной системы.

1. В процессе статического анализа сценария и оценки времени выполнения операций, проводимых перед каждым проходом планирования, для каждой операции вычисляется ее

[15] https://www.docker.com

собственный желаемый срок завершения. Эти сроки определяются на основе желаемого времени завершения для сценария и оценки относительной длительности выполнения операции.

2. На основе собственных желаемых сроков завершения операции из всех выполняемых сценариев объединяются в группы: в одну группу попадают операции, время завершения которых попадает во временное окно некоторой длительности (например, один час).

3. Очереди операций каждой вычислительной системы приоритезируются на основе распределения операций по группам: операции из групп с более ранними желаемыми сроками завершения имеют более высокий приоритет.

При оценке времени выполнения операций, помимо времени выполнения непосредственно вычислений, учитываются временные затраты на передачу данных на вычислительную систему и задержки перед началом выполнения задачи на вычислительной системе (например, вызванные наличием других задач в очереди на выполнение на этой системе).

Заключение

Поставлена проблема разработки информационной системы – инструмента для решения задачи работы с исследовательскими данными полного цикла: сбора данных, их хранения, обработки и анализа, с возможностью построения сценариев обработки и сохранения «истории» данных, формирования и обмена знаниями об обработке данных, как и самими данными на любой стадии работы с ними.

На основе трех «пользовательских историй» сформулирован ряд требований к системе для сбора, хранения и обработки исследовательских данных в ее минимальном жизнеспособном варианте (продукте, MVP – minimal viable product). В двух из этих историй авторы участвовали, разрабатывая простые программные решения для упрощения работы с данными и организации обратной связи системы сбора и обработки данных с участниками эксперимента, третья история является основой для разработки настоящего MVP.

Сформулированные требования позволили разработать следующие модели:

1) модель представления данных, позволяющую исследователям самостоятельно расширять и настраивать описания данных и их связей под нужды текущих исследований и сразу пользоваться внесенными изменениями при поиске и доступе к данным;

2) модель организации доступа к данным, позволяющую автоматизировать передачу данных между этапами обработки;

3) модель абстрактного параллельного вычислителя и язык описания сценариев, что позволяет описывать сценарии обработки данных в функциональном стиле и автоматически генерировать по этому описанию параллельные программы;

4) модель организации распределенной обработки данных на гетерогенном множестве вычислительных узлов.

Кроме того, предложены подходы к накоплению знаний о предметной области, которые будут применяться для поддержки составления новых сценариев. В частности, для этого адаптирована задача об анализе рыночной корзины.

Продолжение работы будет направлено на реализацию разработанной концепции и проектных решений, их детализацию и уточнение в ходе тестирования прототипа MVP, а именно:

1) разработку (или подбор) визуального языка описания сценария;

2) улучшение системы рекомендаций методов на основе статистики работы реальных пользователей;

3) исследование возможностей оптимизации различных этапов работы с данными, в том числе для минимизации задержек при обращении к данным;

4) исследование альтернативных алгоритмов планирования для различных классов сценариев обработки данных, в том числе на основе обратной связи с использованием машинного обучения для предсказания наиболее подходящей схемы планирования;

5) исследование возможностей построения вероятностной модели сценария в конкретной предметной области, чтобы можно было полностью генерировать новый сценарий, фиксируя пользовательские параметры.

Список литературы

1. *Земсков А. И.* Data Curation – хранение научных данных и обслуживание ими – новое направление деятельности библиотек // Научные и технические библиотеки. 2013. № 2. С. 85–101.

2. Программа «Цифровая экономика Российской Федерации»: утв. распоряжением Правительства РФ от 28 июля 2017 г. № 1632-р // СЗ РФ. 2017. № 32, ст. 5138. С. 14517–14574.

3. *Шокин Ю. И.*, *Жижимов О. Л.*, *Пестунов И. А.*, *Синявский Ю. Н.*, *Смирнов В. В.* Распределенная информационно-аналитическая система для поиска, обработки и анализа пространственных данных // Вычислительные технологии. 2007. Т. 12, № S3. С. 108–115.

4. *Новиков А. М., Пойда А. А., Поляков А. Н., Королев С. П., Сорокин А. А.* Разработка технологии и облачной информационной системы для хранения и обработки многомерных массивов научных данных // Информатика и системы управления. 2012. № 4 (34). С. 156–164.

5. *Григорьева М., Голосова М., Рябинкин Е., Климентов А.* Экзабайтное хранилище научных данных // Открытые системы. СУБД. 2015. № 04. С. 14–17.

6. *Юрченко А. В.* К концепции информационно-аналитической системы поддержки научных исследований, основанных на интенсивном использовании цифровых данных // Вычислительные технологии. 2017. Т. 22, № 4. С. 105–120.

7. *Епишев В. В., Исаев А. П., Миниахметов Р. М., Мовчан А. В., Смирнов А. С., Соколинский Л. Б., Цымблер М. Л., Эрлих В. В.* Система интеллектуального анализа данных физиологических исследований в спорте высших достижений // Вестн. ЮУрГУ. Серия «Вычислительная математика и информатика». 2013. Т. 2, № 1. С. 44–54.

8. *Krestyaninova M., Zarins A., Viksna J., Kurbatova N., Rucevskis P., Neogi S. G., Gostev M., Perheentupa T., Knuuttila J., Barrett A. et al.* A system for information management in biomedical studies SIMBioMS // Bioinform. 2009. Vol. 25. Iss. 20. P. 2768–2769.

9. *Austin J., Jackson T., Fletcher M., Jessop M., Liang B., Weeks M., Smith L., Ingram C., Watson P.* CARMEN: code analysis, repository and modeling for e-neuroscience // Procedia Comput. Sci. 2011. Vol. 4. P. 768–777.

10. *Izzo M.* Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management, Springer Theses. Cham, Switzerland: Springer, 2016. 104 p.

11. *Okonechnikov K., Golosova O., Fursov M.* Unipro UGENE: a unified bioinformatics toolkit // Bioinformatics. 2012. № 28. P. 1166–1167. DOI 10.1093/bioinformatics/bts091.

12. *Kotkas V., Ojamaa A., Grigorenko P., Maigre R., Harf M., Tyugu E.* CoCoViLa as a multifunctional simulation platform // Proc. of the 4th International ICST Conference on Simulation Tools and Techniques (SIMUTools '11). Brussels, Belgium: ICST, 2011. P. 198–205.

13. *Huettel S., McCarthy G.* What is odd about the odd-ball task? Prefrontal cortex is activated by dynamic changes in response strategy // Neuropsychologia. 2004. № 42. P. 379–386.

14. *Pascual-Marqui R. D., Lehmann D., Koukkou M., Kochi K., Anderer P., Saletu B., Tanaka H., Hirata K., John E. R., Prichep L., Biscay-Lirio R., Kinoshita T.* Assessing interactions in the brain with exact low-resolution electromagnetic tomography // Philos. Trans. A Math. Phys. Eng. Sci. 2011. № 369 (1952). P. 3768–3784.

15. *Сотников П. И.* Обзор методов обработки сигнала электроэнцефалограммы в интерфейсах мозг-компьютер // Инженерный вестник. 2014, Октябрь. № 10. С. 612–632.

16. *Oostenveld R., Praamstra P.* The five percent electrode system for high-resolution EEG and ERP measurements // Clinical Neurophysiology. 2001. Vol. 112. Iss. 4. P. 713–719. DOI 10.1016/S1388-2457(00)00527-7.

17. *Lei X., Liao K.* Understanding the Influences of EEG Reference: A Large-Scale Brain Network Perspective // Frontiers in Neuroscience. 2017. DOI 10.3389/fnins.2017.00205.

18. *Grandchamp R., Delorme A.* Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials // Frontiers in Psychology. 2011. DOI 10.3389/fpsyg.2011.00236.

19. *Hyvärinen A., Oja E.* Independent Component Analysis: Algorithms and Applications // Neural Networks. 2000. Vol. 13. P. 411–430.

20. *HongLi H., Sun Y. S. Y.* The study and test of ICA algorithms // Proc. 2005 Int. Conf. Wirel. Commun. Netw. Mob. Comput. 2005. Vol. 1. P. 602–605.

21. *Zhu H., Qiu C., Meng Y., Yuan M., Zhang Y., Ren Z., Li Y., Huang X., Gong Q., Lui S., Zhang W.* Altered Topological Properties of Brain Networks in Social Anxiety Disorder: A Resting-state Functional MRI Study // Scientific Reports. 2017. Vol. 7. DOI 10.1038/srep43089.

22. *Stillman P. E., Wilson J. D., Denny M. J., Desmarais B. A., Bhamidi S., Cranmer S. J., Lu Z.-L.* Statistical Modeling of the Default Mode Brain Network Reveals a Segregated Highway Structure // Scientific Reports. 2017. Vol. 7. DOI 10.1038/s41598-017-09896-6.

23. *Abrol A., Damaraju E., Miller R. L., Stephen J. M., Claus E. D., Mayer A. R., Calhoun V. D.* Replicability of time-varying connectivity patterns in large resting state fMRI samples // NeuroImage. 2017. Vol. 163. P. 160–176. DOI 10.1016/j.neuroimage.2017.09.020.

24. *Biswal B., Yetkin F. Z., Haughton V. M., Hyde J. S.* Functional connectivity in the motor cortex of resting human brain using echo-planar MRI // Magn. Reson. Med. 1995. № 34. P. 537–541.

25. *Fielding R. T.* Architectural styles and the design of network-based software architectures. PhD Dissertation. Irvine, US: Dept. of Information and Computer Science, University of California, 2000. P. 76–107.

26. *Martin J.* Managing the Data-base Environment. Englewood Cliffs, New Jersey: Prentice-Hall, 1983. 381 p. ISBN 0135505828.

27. *Linoff G. S., Berry M. J. A.* Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 3rd ed. John Wiley & Sons, 2004. 643 p.

28. *Koren Y., Bell R., Volinsky C.* Matrix Factorization Techniques for Recommender Systems // Computer. 2009. Vol. 42. Iss. 8. P. 30–37.

29. *Damas L., Milner R.* Principal type-schemes for functional programs // Proc. of the 9th ACM SIGPLAN-SIGACT symposium on Principles of programming languages, ACM. 1982. P. 207–212.

30. *Wu Z., Ni Z., Gu L., Liu X.* A Revised Discrete Particle Swarm Optimization for Cloud Workflow Scheduling // Computational Intelligence and Security IEEE International Conference, 2010. P. 184–188.

Материал поступил в редколлегию 01.06.2018

M. A. Gorodnichev [1–3], **A. V. Komissarov** [1, 3], **A. V. Mozhina** [1, 3], **P. V. Prochkin** [1, 3]
P. D. Rudych [1], **A. V. Yurchenko** [1]

[1] *Institute of Computational Technologies SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[2] *Institute of Computational Mathematics and Mathematical Geophysics SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[3] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

maxim@ssd.sscc.ru, yurchenko@ict.nsc.ru

INFORMATION MODELS AND PROJECT SOLUTIONS FOR THE ECCLESIA RESEARCH DATA STORING AND PROCESSING SYSTEM

Scientific research produces a lot of digital data that should be carefully gathered and stored for further usage: processing, analysis and publication. Building e-infrastructure for that is one of the most topical problems of IT (or digital) curation of science. Starting from three data-processing problems in physiology we are developing an information system for automation of gathering, stor-

ing and analyzing data. Problems encountered in development of such a system are examined and analyzed, along with existing approaches and software solutions related to these problems.

Based on results of the conducted analysis a number of models and mechanisms for solving encountered problems are proposed. Developed solutions include models and mechanisms for collecting and storing research data, a model describing and formalizing data processing scenarios and models and mechanisms for processing collected data in a distributed computer system.

As a result, an architecture for a computer system for collecting, storing and processing research data is presented. The system is proposed as a tool for solving a wide spectrum of problems in scientific research involving collecting and multi-step processing of various kinds of data.

Keywords: research data management, information system, data analysis and processing, physiology data, object storage, functional language, machine learning, distributed computing.

References

1. Zemskov A. Data Curation – a new derection in library activities. *Scientific and Technical Libraries*, 2013, vol. 2, p. 85–101. (in Russ.)

2. Program «Digital economy of the Russian Federation»: approved by the Russian Federation Government order of July 28, 2017 № 1632-p. *Collection of legislation of the Russian Federation*, 2017, vol. 32, art. 5138, p. 14517–14574. (in Russ.)

3. Shokin Y. I., Zhizhimov O. L., Pestunov I. A., Sinyavskiy Y. N., Smirnov V. V. Distributed informational-analytical system for searching, processing and analysis of spatial data. *Computational Technologies*, 2007, vol. 12, special issue 3, p. 108–115. (in Russ.)

4. Novikov A. M., Poyda A. A., Korolev S. G., Sorokin A. A. Development of Technology and Cloud Informational System for Storing and Processing of Multi-dimensional Science Data Arrays. *Information Science and Control Systems*, 2012, № 4 (34), p. 156–164. (in Russ.)

5. Grigor'eva M., Golosova M., Ryabinkin E., Klimentov A. Ekzabaitnoe hranilishe nauchnyh dannyh. *Open Systems Journal*, 2015, № 4, p. 14–17. (in Russ.)

6. Yurchenko A. V. On the concept of information-analytical system for supporting data intensive science. *Computational Technologies*, 2017, vol. 22, № 4, p. 105–120. (in Russ.)

7. Epishev V. V., Isaev A. P., Miniakhmetov R. M., Movchan A. V., Smirnov A. S., Sokolinsky L. B., Zymbler M. L., Ehrlich V. V. Physiological data mining system for elite sports. *Bulletin of the South Ural State University. Series «Computational Mathematics and Software Engineering»*, 2013, vol. 2, № 1, p. 44–54. (in Russ.)

8. Krestyaninova M., Zarins A., Viksna J., Kurbatova N., Rucevskis P., Neogi S. G., Gostev M., Perheentupa T., Knuuttila J., Barrett A. et al. A system for information management in biomedical studies SIMBioMS. *Bioinform*, 2009, vol. 25 (20), p. 2768–2769.

9. Austin J., Jackson T., Fletcher M., Jessop M., Liang B., Weeks M., Smith L., Ingram C., Watson P. CARMEN: code analysis, repository and modeling for e-neuroscience. *Procedia Comput. Sci.*, 2011, vol. 4, p. 768–777.

10. Izzo M. Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management, Springer Theses. Cham, Switzerland: Springer, 2016. 104 p.

11. Okonechnikov K., Golosova O., Fursov M., the UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, 2012, vol. 28, p. 1166–1167. DOI 10.1093/bioinformatics/bts091.

12. Kotkas V., Ojamaa A., Grigorenko P., Maigre R., Harf M., Tyugu E. CoCoViLa as a multifunctional simulation platform. Proc. of the 4th International ICST Conference on Simulation Tools and Techniques (SIMUTools '11). ICST. Brussels, Belgium, ICST, 2011, p. 198–205.

13. Huettel S., McCarthy G. What is odd about the odd-ball task? Prefrontal cortex is activated by dynamic changes in response strategy. *Neuropsychologia*, 2004, vol. 42, p. 379–386.

14. Pascual-Marqui R. D., Lehmann D., Koukkou M., Kochi K., Anderer P., Saletu B., Tanaka H., Hirata K., John E. R., Prichep L., Biscay-Lirio R., Kinoshita T. Assessing interactions in the brain with exact low-resolution electromagnetic tomography. *Philos. Trans. A Math. Phys. Eng. Sci.*, 2011, vol. 369 (1952), p. 3768–3784.

15. Sotnikov P. I. Obzor metodov obrabotki signala elektroencefalogrammy v interfeisah mozg-komputer. *Engineering Bulletin*, 2014, vol. 10, p. 612–632. (in Russ.)

16. Oostenveld R., Praamstra P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, 2001, vol. 112 (4), p. 713–719. DOI 10.1016/S1388-2457(00)00527-7.

17. Lei X., Liao K. Understanding the Influences of EEG Reference: A Large-Scale Brain Network Perspective. *Frontiers in Neuroscience*, 2017. DOI 10.3389/fnins.2017.00205.

18. Grandchamp R., Delorme A. Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials. *Frontiers in Psychology*, 2011. DOI 10.3389/fpsyg.2011.00236.

19. Hyvärinen A., Oja E. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 2000, vol. 13, p. 411–430.

20. HongLi H., Sun Y. S. Y. The study and test of ICA algorithms. *Proc. Int. Conf Wirel. Commun. Netw. Mob. Comput*., 2005, vol. 1, p. 602–605.

21. Zhu H., Qiu C., Meng Y., Yuan M., Zhang Y., Ren Z., Li Y., Huang X., Gong Q., Lui S., Zhang W. Altered Topological Properties of Brain Networks in Social Anxiety Disorder: A Resting-state Functional MRI Study. *Scientific Reports*, 2017, vol. 7. DOI 10.1038/srep43089.

22. Stillman P. E., Wilson J. D., Denny M. J., Desmarais B. A., Bhamidi S., Cranmer S. J., Lu Z.-L. Statistical Modeling of the Default Mode Brain Network Reveals a Segregated Highway Structure. *Scientific Reports*, 2017, vol. 7. DOI 10.1038/s41598-017-09896-6.

23. Abrol A., Damaraju E., Miller R. L., Stephen J. M., Claus E. D., Mayer A. R., Calhoun V. D. Replicability of time-varying connectivity patterns in large resting state fMRI samples. *NeuroImage*, 2017, vol. 163, p. 160–176. DOI 10.1016/j.neuroimage.2017.09.020.

24. Biswal B., Yetkin F. Z., Haughton V. M., Hyde J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.*, 1995, vol. 34, p. 537–541.

25. Fielding R. T. Architectural styles and the design of network-based software architectures. PhD Dissertation. Irvine, US: Dept. of Information and Computer Science, University of California, 2000, p. 76–107.

26. Martin J. Managing the Data-base Environment. Englewood Cliffs, New Jersey: Prentice-Hall, 1983, 381 p.

27. Linoff G. S., Berry M. J. A. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 3rd ed. John Wiley & Sons, 2004, 643 p.

28. Koren Y., Bell R., Volinsky C. Matrix Factorization Techniques for Recommender Systems. *Computer*, 2009, vol. 42 (8), p. 30–37.

29. Damas L., Milner R. Principal type-schemes for functional programs. *Proc. of the 9th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, 1982, p. 207–212.

30. Wu Z., Ni Z., Gu L., Liu X. A Revised Discrete Particle Swarm Optimization for Cloud Workflow Scheduling. *Computational Intelligence and Security IEEE International Conference*, 2010, p. 184–188.

For citation:

Gorodnichev M. A., Komissarov A. V., Mozhina A. V., Prochkin P. V., Rudych P. D., Yurchenko A. V. Information Models and Project Solutions for the Ecclesia Research Data Storing and Processing System. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 87–104. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-87-104

УДК 004.514
DOI 10.25205/1818-7900-2018-16-3-105-112

В. А. Зенг

Омский государственный технический университет
пр. Мира, 11, Омск, 644050, Россия

valeriyazeng@mail.ru

ФОРМИРОВАНИЕ БАЗОВОГО СЛОВАРЯ ЖЕСТОВ ДЛЯ ЕСТЕСТВЕННОГО КОМПЬЮТЕРНОГО БЕСКОНТАКТНОГО ИНТЕРФЕЙСА [*]

Исследуются возможности бесконтактных систем и интерфейсов, главные принципы работы с такими технологиями. Рассмотрена возможность применения подобных систем для упрощения взаимодействия пользователей с ограничениями возможностями здоровья с компьютерным интерфейсом. Приведены особенности и преимущества использования естественных интерфейсов и систем, основанных на жестовом управлении. Также детально рассмотрены этапы формирования базового словаря жестов для дальнейшего его применения в бесконтактном интерфейсе. В качестве дополнительного аппаратного обеспечения для получения более точных результатов распознавания таких жестов рассмотрено устройство Microsoft Kinect.

Ключевые слова: человеко-машинное взаимодействие, программное приложение, компьютерный интерфейс, прототип программы, захват движения, технологии бесконтактного взаимодействия, пользователи с ограниченными возможностями здоровья.

Введение

Современные научные работы, посвященные исследованию человеко-машинного взаимодействия, направлены в основном на создание вычислительных машин, оборудованных большим количеством различных датчиков и сенсоров, а также на изучение средств межчеловеческой коммуникации, таких как речь и сопровождающие жесты. Разрабатываемые интерфейсы ориентированы исключительно на опытных пользователей, однако почти не затрагиваются вопросы человеко-машинной коммуникации для лиц с ограниченными возможностями (инвалидов). Так, глухонемые люди не могут использовать речевые интерфейсы, а люди с проблемами мелкой моторики не способны работать с клавиатурой или жестовыми интерфейсами. Главной целью является разработка универсального бесконтактного интерфейса, пригодного для всех категорий пользователей, и реализация этого интерфейса, демонстрирующего возможности многомодальной человеко-машинной коммуникации. Такой интерфейс будет включать различные естественные для человека способы передачи и восприятия информации: речь, жесты, движения головой и телом, чтение по губам, а также комбинации этих бесконтактных модальностей. Многомодальный интерфейс сможет обрабатывать входную информацию и выводить информацию в той форме, которая доступна конкретному пользователю [1].

[*] Работа выполнена при поддержке Фонда содействия инновациям в рамках программы «УМНИК», договор № 12394ГУ/2017.

Зенг В. А. Формирование базового словаря жестов для естественного компьютерного бесконтактного интерфейса // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 105–112.

© В. А. Зенг, 2018

Возможности бесконтактных интерфейсов

Естественный интерфейс подразумевает, что основным путем взаимодействия человека с компьютером являются прикосновения, жесты, речь, а также другие виды поведения, которые практикуются в течение многих лет и/или которые являются врожденными. Прежде всего, понятие естественного интерфейса относится к процессу взаимодействия пользователя с системой – насколько оно комфортно и понятно. Управление должно быть одинаково простым, интуитивным, как для опытного пользователя системы, так и для новичка [2].

Пользователи предпочитают избавляться от барьеров, воздвигаемых традиционными интерфейсами между ними и техникой, – гораздо удобнее управлять устройствами без посредников, а жестикуляция – вполне естественный способ общения. Интерфейсы на жестикуляционном управлении предлагают привычными движениями, например, имитирующими перелистывание страницы, реализовать то же самое действие на экране. Нажатие на клавиши не относится к естественным способам общения для человека, поэтому основное преимущество жестикуляционного управления в том, что оно позволяет быстрее и проще отдавать команды устройствам. Помимо этого, подобные интерфейсы позволяют избавить экранное пространство от кнопок, клавиш и других наглядных элементов управления [3]. Кроме того, жестикуляция дает возможность сосредоточиться на экране, а не на компьютерной мыши, клавиатуре или пульте дистанционного управления.

Для реализации необходимо разработать интерфейс для работы с компьютером при помощи жестов, предварительно задаваемых пользователем. Подобный интерфейс предназначен, в основном, для помощи следующим категориям пользователей:

- людям с проблемами мелкой моторики;
- слабослышащим или глухонемым людям.

Жестикуляционные интерфейсы могут упростить инвалидам взаимодействие с электроникой. Помимо этого, существует еще несколько областей, где возможно их потенциальное применение. Одна из них – автомобили. С помощью телодвижений можно управлять развлекательной системой, очистителями стекол, фарами и другим оборудованием, не отрывая глаз от дороги. «Ford Motor» уже выпускает автомобили, у которых автоматически открывается багажник, если под ним провести ногой [1].

Жестикуляционные интерфейсы также позволят врачам и медсестрам управлять компьютерами и другими устройствами, не дотрагиваясь до них. Это очень ценная возможность в ситуациях, когда медикам нужны чистые руки, а также в случаях, когда оборудование находится вне прямой досягаемости.

Пользователь жестовой системы управления должен в первую очередь догадаться о следующем:

- что системой можно управлять с помощью жестов;
- какой набор жестов поддерживает эта система.

Эту информацию пользователь должен узнать при работе с самой системой, а не из руководства или другой документации. В целом довольно трудно ответить на вопрос, как достигнуть «affordance» (с англ. – воспринимаемая доступность) управления жестами. Термин «affordance» означает это качество системы или продукта, которое предложил Джеймс Гибсон, основатель направления в психологии, рассматривающего восприятие как процесс, не предполагающий умозаключений, промежуточных переменных. В данной ситуации, обратную связь целесообразно дополнить подсказками-предсказаниями – в процессе распознания показать, какие жесты доступны на данном этапе. Помимо конечной цели, пользователь сможет узнать, какие еще возможности поддерживает система. Также, если у некоторых жестов траектория одинаково начинается, это может стать полезным для пользователя дополнением [4].

Особенности систем жестового управления

Управление интерфейсом должно быть интуитивным и простым, он должен легко настраиваться и работать. В частности, нужно помнить, что пользователи не должны знать, как устроена система и как она работает, поэтому ошибочное распознавание и неверное исполь-

зование в идеале необходимо выявлять на стадии тестирования и игнорировать его. В случае если система неверно распознала жест, необходимо предусмотреть простой механизм отмены действия и возможность отметить, что жест распознан неправильно [5]. Чтобы персонифицировать набор жестов, необходим интерфейс, который предполагает определение персональных жестов.

Понятие доступности требует «равных прав для людей в получении доступа к информации независимо от физических и когнитивных затруднений, которые они могут испытывать в связи с временными или хроническими нарушениями и болезнями». Управление жестами должно быть доступно для пользователей с разными физическими или техническими ограничениями [6]. Также в идеале при управлении в жестовых системах не должны накладываться ограничения на окружающую среду, в которой они используются. Фактически система должна быть работоспособна в любое время, при любом освещении и при любом пространственном положении [1]. На сегодняшний день распознавание жестов в таких интерфейсах базируется преимущественно на получении данных с одной или нескольких камер, а это накладывает ограничения на среду, в которой они используются. Помимо этого, калибровка камеры может оказаться достаточно сложной задачей для среднестатистического пользователя.

При разработке системы жестового управления есть два варианта: создать универсальный набор жестов или дать возможность пользователям самостоятельно определять этот набор [5]. В первом случае достаточно сложно персонифицировать жесты, так как многое может зависеть от культуры пользователей, их личных особенностей (правша или левша) и других характеристик, которые влияют на естественные жесты пользователей. Поэтому интерфейсу, основанному на жестовом управлении, требуется либо приспосабливаться к персональным жестам, либо модифицировать универсальные жесты, либо определять персональные жесты. Возможность приспосабливаться к персональным жестам, основанная на наблюдении, предполагает обнаружение оптимальных жестов для такой задачи, однако это довольно длительный процесс, следовательно, этот вариант может не подойти для ежедневного использования [4]. Остальные варианты предпочтительнее всего для повседневных систем. Для их реализации нужно применять алгоритмы машинного обучения. При автоматической сегментации могут возникнуть проблемы алгоритма распознания, так как захватываются движения, которые не являются частью жеста. Подобные движения считаются шумом, поэтому машинные модели обучения это должны учитывать [2].

В целом список жестов для управления компьютером должен представлять собой какой-то базовый набор жестов, и при этом некоторые из них могут быть уникальными в рамках всей системы (например, жест отмены), а некоторые – не уникальны, так как один и тот же жест в зависимости от контекста может использоваться по-разному [5]. Однако при этом у системы должна быть возможность адаптации (персонализации) под каждого конкретного пользователя, а также под условия использования интерфейса.

Разработка базового словаря жестов

Формирование набора естественных жестов для бесконтактного управления имеет корни в ставших обыденными жестах при сенсорном взаимодействии, используемом в современной цифровой технике: смартфонах, планшетах, некоторых ноутбуках (laptop). Интерактивная технология «multi-touch interaction» (MT) позволяет пользователю путем прикосновений управлять графическим интерфейсом одновременно несколькими пальцами руки.

Поддерживаемые жесты: «щелчок (выбор элемента)», «сдвиг (пролистывание) вправо», «сдвиг влево», «уменьшение масштаба», «увеличение масштаба», «поворот» [2] (рис. 1). Данный набор жестов стал основой для базовых жестов бесконтактного взаимодействия

[1] ГОСТ Р ИСО 20282-1-2011 «Эргономика изделий повседневного использования. Часть 1. Требования к конструкции элементов управления с учетом условий использования и характеристик пользователя». Введен 01.12.2012. М.: Стандартинформ, 2012. 24 с.

[2] ГОСТ Р ИСО 9241-20-2014 «Эргономика взаимодействия человек-система. Часть 20. Руководство по доступности оборудования и услуг в области информационно-коммуникационных технологий». Введен 01.12.2015. М.: Стандартинформ, 2015. 44 с.

с той лишь разницей, что главными управляющими «инструментами» стали не пальцы, а кисти и руки. Исходя из этого «масштаб» управления был расширен от экрана сенсорного телефона до бесконтактного управления интерфейсом с помощью рук.

Прежде всего следует определить условные обозначения, которые будут указаны в словаре жестов как начальное, конечное положение руки и направление ее движения. Для этого используются интуитивно понятные графические примитивы, которые дополняются сопроводительными подписями (рис. 2).

Основными движениями в данном бесконтактном интерфейсе являются перелистывание вправо и влево, а также прокрутка вверх и вниз. Это базовые движения, которые на сегодняшний день интуитивно понятны любому пользователю, который хотя бы раз работал с персональным компьютером. Эти жесты осуществляются взмахом руки в нужную сторону (рис. 3). Они необходимы в ситуациях, когда информация полностью не помещается на экране монитора, разделяется на части (страницы), поэтому для просмотра новой порции информации требуется переход в следующую часть.

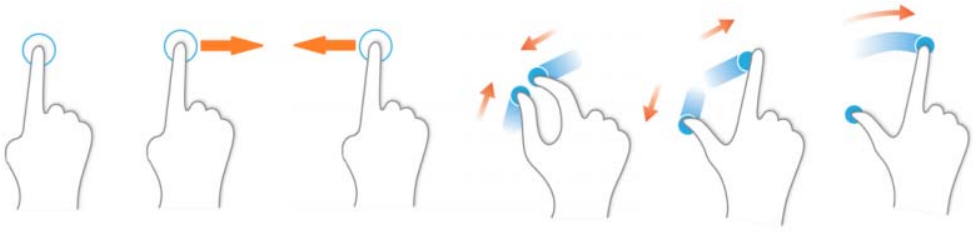


Рис. 1. Набор стандартных жестов для управления цифровым сенсорным устройством

Рис. 2. Условные обозначения словаря жестов и их расшифровка

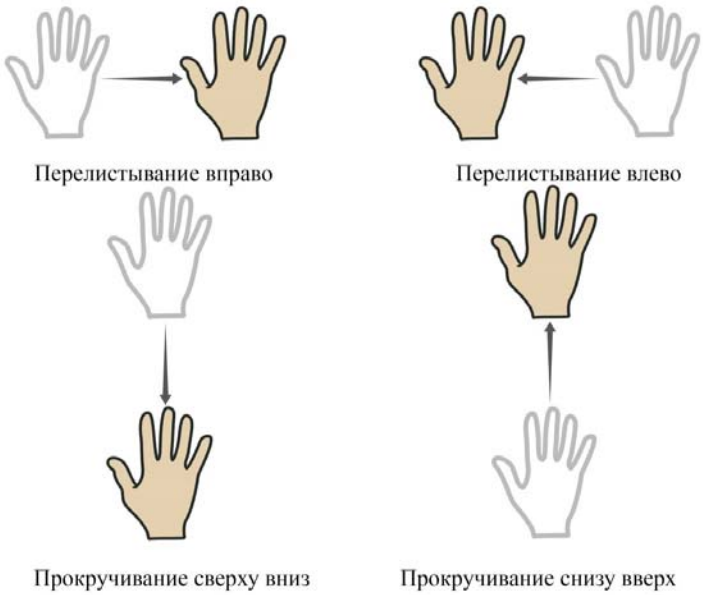


Рис. 3. Жесты пролистывания вправо или влево, прокручивания вверх и вниз

Рис. 4. Жесты предпросмотра содержимого папки и просмотра панели «Пуск»

Для облегчения взаимодействия пользователя с ограниченными возможностями здоровья (ОВЗ)с интерфейсом было разработано такое функциональное улучшение, как «предпросмотр содержимого». Зачастую пользователи, которые сортируют документы по множеству папок, не могут вспомнить точное месторасположение того или иного файла. Поэтому была создана функция «предпросмотра содержимого», которая позволяет вывести список последних открытых в определенных папках файлов, что может сэкономить много времени на поиск необходимых документов. Данная возможность реализуется путем изображения круга по часовой стрелке для раскрытия списка файлов и против часовой для их закрытия (рис. 4).

Также было принято решение создать новое движение, которое упростит работу с меню, подобным элементу «Пуск» в ОС Windows. Так как у пользователей с ограниченной моторикой возникает необходимость совершать действия в системе, затрачивая как можно меньше сил и совершая не очень активные движения, то было принято решение сделать жест максимально понятным и удобным для людей с ОВЗ [1]. Для того чтобы отобразилась панель меню, следует немного поднять ладони вверх, как при прокручивании снизу. Чтобы панель скрыть, следует сделать такое же движение, но в обратном направлении (рис. 5). Однако многие пользователи с ограничениями в движении не имеют возможности двигать две руки одновременно, к тому же в одном направлении. Поэтому интерфейс предоставляет возможность совершать данные манипуляции одной рукой.

Чтобы совершить «клик» по выбранному жестом элементу интерфейса, следует удерживать ладонь на месте 3 секунды. Если задать возможную амплитуду размаха руки в попытках удержать ее на одном месте, что не всегда удается людям с ограниченной моторикой из-за судорог или непроизвольных спазмов мышц, то можно настроить систему для восприятия дрожащей или даже колеблющейся в пределах 5–10 см руки как жест выбора элемента [3].

Последней из возможных базовых функций, адаптированных под пользователей с ОВЗ, является функция увеличения и уменьшения размеров объектов. При просмотре изображения или документа в его начальном состоянии не всегда достигаются необходимые для корректного восприятия величины объекта с самого его запуска. Поэтому функция увеличения и уменьшения размеров особенно необходима. Увеличение достигается путем раздвигания рук из центра в стороны по диагонали (см. рис. 5). Не имеет значения, какая именно рука окажется в верхней точке жеста, а какая в нижней. Система распознает оба положения как

один и тот же запрос. Соответственно уменьшение элемента может быть получено в том случае, когда руки, разведенные в стороны и находящиеся по диагонали друг от друга, собираются в центре. Причем степень увеличения и уменьшения может достигаться как поэтапно, если указать в настройках, что одно такое движение рук – это один шаг масштабирования элемента, так и плавно, что актуально для пользователей без ограничений в моторике.

Рис. 5. Жесты выбора элемента и увеличения или уменьшения объекта

Таким образом, был разработан базовый словарь жестов для естественного бесконтактного взаимодействия с компьютерным интерфейсом, который удовлетворяет основным принципам естественных интерфейсных систем.

- Быстрое погружение. Прилагая минимальные усилия и время, человек сам овладевает системой. Отсюда не следует, правда, что обучения вообще не должно быть, но оно должно быть минимальным и без посторонней помощи. Важно понимать, что мы говорим именно об интерфейсе, естественный интерфейс может быть и у промышленного робота, и у баллистической ракеты, и у шахмат, но овладение им в данных контекстах не сделает вас опытным пользователем с точки зрения системы. Недостаточно понять базовые элементы и правила их комбинирования, нужно изучить их идиоматическое использование.
- Легкое управление. Компьютер должен приспосабливаться к пользователю, а не наоборот. Человек через некоторое время перестает «думать» о том, что ему надо сделать, чтобы произошло нужное действие.
- Впечатление (игровой момент). Эффект новизны пользовательского опыта, как следствие, положительные эмоции от самой системы. Со временем и частотой использования этот эффект будет исчезать [4].

Рынок систем жестикуляционного управления относительно нов, стандартов еще практически нет, и в различных системах используются совершенно разные интерфейсы, камеры и алгоритмы. Это дает возможность выбора инструментов для использования в разработке и применении естественных интерфейсов. В разрабатываемом продукте предполагается использование устройства Microsoft Kinect в качестве дополнительного аппаратного обеспече-

ния. Оно определяет до шести человек в пространстве и 25 суставов каждого из них; распознает лица, эмоции, пульс. Аудиосистема последней модели может определить двух одновременно говорящих людей и распознать два потока речи [3]. Помимо широких возможностей и относительно низкой цены по сравнению с аналогами, устройство Microsoft Kinect может также адаптироваться под особенности пользователей, просчитывая возможные варианты жестов. Так что если человек с ОВЗ не повторит определенное движение в точности, как указано в словаре жестов, то устройство захвата движения все равно идентифицирует кисть, распознает движение и определит его тип в соответствии со словарем.

Выводы

Таким образом, создание и внедрение естественных человеко-машинных интерфейсов, основанных на автоматическом распознавании речи и жестов, предлагает пользователям-инвалидам новый способ персонифицированного бесконтактного взаимодействия с компьютером полностью без использования стандартных устройств ввода, таких как клавиатура и компьютерная мышь. Адаптивность и индивидуализация интерфейса, а также способность осуществить настройку словаря движений для управления системой в соответствии с потребностями каждой категории граждан с ОВЗ являются отличительной особенностью разработки. Еще одним важным дополнением будет возможность встраивания его в виде программного модуля для управления операционной системой MS Windows и ее системными приложениями.

Список литературы

1. *Ронжин А. Л.* Методы и многомодальные интерфейсы для бесконтактной коммуникации инвалидов с информационно-справочными системами / Российский фонд фундаментальных исследований. М., 2016. URL: http://www.rfbr.ru/rffi/portal/project_search/o_44865 (дата обращения 11.05.2018).

2. *Дроздова Ю. А.* Создание естественного пользовательского интерфейса для мобильных устройств. Обзор средств распознавания жестов руки // Студенческий научный форум: Материалы VI Междунар. студ. электрон. науч. конф. 2014. URL: http://www.scienceforum.ru/2014/527/1470 (дата обращения 11.05.2018).

3. *Зенг В. А.* Анализ подходов к созданию бесконтактных интерфейсов // Творчество молодых: дизайн, реклама, информационные технологии: Материалы XV Междунар. науч.-практ. конф. Омск, 2016. С. 107–111.

4. *Гарбер Л.* Новые пользовательские интерфейсы: жестикуляция // IEEE Computer Society, Synapses to Circuitry: Gestural Technology: Moving Interfaces in a New Direction. 2013. № 10. URL: https://www.osp.ru/os/2013/10/13039069/ (дата обращения 11.05.2018).

5. *Сорокин Д.* Упрощение в дизайне интерфейсов. URL: https://www.uplab.ru/blog/simplification-in-the-design-of-interfaces/ (дата обращения 11.05.2018).

6. *Николахина Е.* Равные права – равные возможности // Библиотека. 2015. № 10. С. 52–54.

Материал поступил в редколлегию 11.05.2018

V. A. Zeng

Omsk State Technical University
11 Mir Ave., Omsk, 644050, Russian Federation

valeriyazeng@mail.ru

GENERAL GESTURAL DICTIONARY DEVELOPMENT FOR NATURAL COMPUTER-BASED CONTACTLESS INTERFACE

This article contains the description of contactless systems and interfaces and main principles of working with these technologies. There is possibility of using such systems to simplify the interac-

tion of users with the limitations of health possibilities with a computer interface. The features and advantages of using natural interfaces and systems based on gesture control are also presented. There are stages of the formation of a basic gesture dictionary for further use in contactless interface which are described in detail as well. As additional hardware for obtaining more accurate results of recognizing such hand gestures a Microsoft Kinect device was reviewed.

Keywords: human-machine interaction, software application, computer interface, program prototype, motion capture, contactless interaction technologies, users with disabilities.

References

1. Ronzhin A. Methods and multimodal interfaces for contactless advertising. Moscow, Russian Foundation for Basic Research, 2016. URL: http://www.rfbr.ru/rffi/portal/project_search/o_44865 (accessed: 05.11.2018) (in Russ.)

2. Drozdova Yu. Creating a natural user interface for mobile devices. A review of hand gesture recognition. *Proc. of the VI International Student Electronic Scientific Conference «Student Scientific Forum»*. URL: http://www.scienceforum.ru/2014/527/1470 (accessed: 05.11.2018) (in Russ.)

3. Zeng V. Analysis of contactless interfaces creation approaches. *Young Creativity: Design, Advertising, Information Technologies: Proc. of the XV International Scientific and Practical Conference.* Omsk, 2016, p. 107–111. (in Russ.)

4. Garber L. New User Interfaces: Gesture. *IEEE Computer Society, Synapses to Circuitry: Gestural Technology: Moving Interfaces in a New Direction*, 2013, no. 10. URL: https://www.osp.ru/os/2013/10/13039069/ (accessed: 05.11.2018) (in Russ.)

5. Sorokin D. Simplification in the design of interfaces. URL: https://www.uplab.ru/blog/ simplification-in-the-design-of-interfaces/ (accessed: 05.11.2018) (in Russ.)

6. Nikolahina E. Equal rights – equal opportunities. *Library*, 2015, no. 10, p. 52–54. (in Russ.)

For citation:

Zeng V. A. General Gestural Dictionary Development for Natural Computer-Based Contactless Interface. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 105–112. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-105-112

УДК 004.04
DOI 10.25205/1818-7900-2018-16-3-113-125

И. А. Корсун [1], **Д. Е. Пальчунов** [1, 2]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Институт математики им. С. Л. Соболева СО РАН*
пр. Академика Коптюга, 4, Новосибирск, 630090, Россия

irina.korsun.nsu@gmail.com, palch@math.nsc.ru

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ОБРАБОТКИ И ИНТЕГРАЦИИ ЗНАНИЙ НА ОСНОВЕ ТЕХНОЛОГИЙ СЕМАНТИЧЕСКОЙ ПАУТИНЫ *

Статья посвящена разработке методов извлечения из текстов естественного языка определений ключевых понятий предметной области на основе теоретико-модельного подхода. Извлеченная из текстов информация путем преобразования через фрагменты атомарных диаграмм алгебраических систем представляется в виде утверждений в логике описаний (DL). Подобное представление позволяет получать тексты с большей выразительностью по сравнению с алгоритмами, где источником информации являются данные, представленные в виде баз данных или в виде выражений на формализованных языках (например, SQL).

Ключевые слова: онтология, теоретико-модельные методы, фрагменты атомарных диаграмм, определения понятий, извлечение определений, полнота определений понятий, явная определимость, неявная определимость, генерация текстов, порождение фраз естественного языка.

Введение

Современные модели представления информационных ресурсов интенсивно развиваются и внедряются в практику. Важнейшим элементом современных информационных технологий являются онтологии, которые позволяют производить автоматизированную обработку семантики информации с целью ее эффективного использования. С ростом объемов обрабатываемых данных онтологии становятся очень масштабными, что усложняет восприятие конечным пользователем хранящейся в ней информации. Так как в повседневной жизни наиболее распространенной формой представления знаний являются естественно-языковые тексты, проводится все больше исследовательских работ по извлечению описаний объектов OWL-онтологий в виде набора согласованных по смыслу предложений – определений.

Существует целый каталог систем генераций текстов [1], содержащий краткую информацию обо всех известных (более 340) разработках. Большинство из них относятся к 1963–1980 гг. и имеют алгоритмы шаблонного типа: система хранит уже готовую строку, возможно, с несколькими пропусками, которые заполняются при выдаче сообщения значениями, соответствующими характеру ошибки. Более сложные шаблонные системы дополнительно

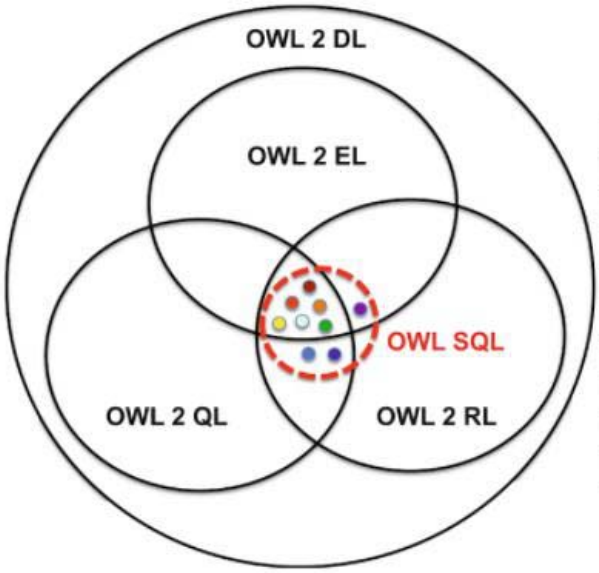
* Исследование выполнено при частичной финансовой поддержке Президиума СО РАН (проект «Инженерия интенсиональных онтологий в дедуктивных и информационных системах» Комплексной программы ФНИ СО РАН II.1).

Корсун И. А., *Пальчунов Д. Е.* Интеллектуальная система обработки и интеграции знаний на основе технологий семантической паутины // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 113–125.

© И. А. Корсун, Д. Е. Пальчунов, 2018

проводят ограниченную лингвистическую обработку генерируемого текста. Примером программ с подобной структурой могут служить следующие: диалоговая система ELIZA [2], система Employee Appraiser (Austin-Haynes) и Performance Now (KnowledgePoint). Шаблонный метод генерации, безусловно, имеет существенный недостаток. Генерируемые тексты сравнимы с естественными до тех пор, пока они имеются в единственном экземпляре. Всё, что написано на базе шаблона, похоже друг на друга за исключением тех частей, куда вставляются параметры.

Также для большинства программ источником информации являются данные, представленные в виде баз данных или в виде выражений на формализованных языках, например SQL. Примеры систем генерации из формального представления: REMIT [3], система Минока [4]. Однако подобный формат представления данных имеет существенные недостатки. Так как SQL полностью отображается на один из профилей OWL, а именно OWL SQL, то можно провести следующий анализ (рис. 1).



	EL	QL	RL
subClass	X	X	X
subProp	X	X	X
domain	X	X	X
range	X	X	X
eqClass	X	X	X
eqProp	X	X	X
inverseProp		X	X
symProp		X	X
transProp	X		X

Рис. 1. Выразительные мощности профилей OWL

Отсюда делаем вывод, что полученные определения из текстов с использованием представления информации на языке SQL достаточно тривиальны и не могут включать в себя более выразительные конструкции.

Таким образом, главная идея настоящей работы заключается в разработке системы извлечения из текстов естественного языка определений ключевых понятий предметной области на основе теоретико-модельного подхода.

Для извлечения из текстов естественного языка определений ключевых понятий предметной области мы используем результаты наших предыдущих исследований. Ранее в [5] был разработан теоретико-модельный подход к извлечению знаний из текстов естественного языка. В основе него лежит представление знаний при помощи конечных фрагментов атомарных диаграмм моделей. Были разработаны и реализованы в виде программной системы методы интерпретации различных частей речи и синтаксических связей с целью автоматического порождения сигнатуры модели [6]. В [7] были предложены алгоритмы отображения бескванторных предложений логик предикатов первого порядка сигнатуры, не содержащей функциональных символов, в логику описаний (DL).

Разрабатываемая на данный момент система сначала извлекает из утверждений на языке DL те, которые имеют отношение к классу или индивидууму, требующему описания в виде текста. Формирует фрагмент OWL онтологии при помощи алгоритмов, также представлен-

ных в [7]. Далее определяет, каким образом выбранная информация будет реализована языковыми средствами в виде предложений на естественном языке. Для этого используются так называемые текстовые планы, представляющие собой последовательность слотов, инструкции, определяющие правила их заполнения, а также лексические словари в виде OWL-онтологий.

Проблемы формального представления определений понятий: явная и неявная определимость

В данной работе мы продолжаем разработку теоретико-модельного подхода к формализации определений ключевых понятий предметной области. В частности, мы продолжаем исследование различных подходов к описанию полноты определений понятий (относительно онтологии, контекста, множества прецедентов предметной области и т. д.), явной и неявной определимости понятий, которое проводилось в [7–10].

В [8] была установлена необходимость неявных определений понятий в формальных глоссариях. В данном параграфе мы изучим взаимосвязь этих результатов с хорошо известной теоремой Бета о неявной определимости.

Необходимые определения можно взять из работ [7; 11].

В [8] было введено понятие формального глоссария.

Определение. Пусть σ – сигнатура. Последовательность предложений $<\varphi_1, \dots, \varphi_n>$ назовем *формальным глоссарием* (*определяющим понятия из* σ), если выполнено:

а) $\sigma(\varphi_1) \subset \sigma(\varphi_1 \,\&\, \varphi_2) \subset \dots \subset \sigma(\varphi_1 \,\&\, \dots \,\&\, \varphi_n) = \sigma$;

б) добавление каждого нового предложения φ_{k+1} консервативно расширяет предыдущий набор предложений $\varphi_1, \dots, \varphi_k$, т. е.

$$\mathrm{Th}(\varphi_1 \,\&\, \dots \,\&\, \varphi_k) = \mathrm{Th}(\varphi_1 \,\&\, \dots \,\&\, \varphi_n) \cap \mathrm{S}(\sigma(\varphi_1 \,\&\, \dots \,\&\, \varphi_k)).$$

Здесь $\mathrm{Th}(\varphi_1 \,\&\, \dots \,\&\, \varphi_n)$ – теория, порожденная (аксиоматизируемая) предложением $(\varphi_1 \,\&\, \dots \,\&\, \varphi_n)$. Мы говорим, что формальный глоссарий $<\varphi_1, \dots, \varphi_n>$ представляет предложение ψ, если $\mathrm{Th}(\psi) = \mathrm{Th}(\varphi_1 \,\&\, \dots \,\&\, \varphi_n)$.

Были доказаны следующие утверждения.

Теорема [8]. Существуют сигнатура σ, состоящая из имен двух понятий, и предложение ψ, определяющее смысл понятий из σ, для которых нет формального глоссария $<\varphi_1, \varphi_2>$, представляющего предложение ψ, такого, чтобы сигнатура $\sigma(\varphi_1)$ состояла из имени одного понятия, а сигнатура $\sigma(\varphi_2)$ – из имен обоих понятий.

Следствие [8]. Не всегда определения понятий могут быть представлены в виде глоссария, определяющего понятия по одному.

Следствие [8]. Не всегда определения понятий могут быть представлены в виде явного глоссария.

На первый взгляд этот результат противоречит известной теореме Бета, которая говорит, что любое неявное определение предиката можно преобразовать в явное.

Определение [11]. Пусть $\Sigma(P)$ – некоторое множество предложений сигнатуры $\sigma \cup \{P\}$, $P \notin \sigma$, т. е. $\Sigma(P) \subseteq S(\sigma)$. Будем говорить, что $\Sigma(P)$ ***неявно определяет*** отношение P, если выполнено

$$\Sigma(P) \cup \Sigma(P') \models (\forall x_1 \dots \forall x_n)\big(P(x_1,\dots,x_n) \leftrightarrow P'(x_1,\dots,x_n)\big).$$

Будем говорить, что $\Sigma(P)$ ***явно определяет*** отношение P, если существует такая формула $\varphi(x_1 \dots x_n) \in S(\sigma)$, что выполнено

$$\Sigma(P) \models (\forall x_1 \dots \forall x_n)\big(P(x_1,\dots,x_n) \leftrightarrow \varphi(x_1,\dots,x_n)\big).$$

Теорема Бета [11]. Множество $\Sigma(P)$ неявно определяет отношение P тогда и только тогда, когда $\Sigma(P)$ определяет отношение P явно.

Таким образом, теорема Бета утверждает, что понятие P, смысл которого неявно задается множеством предложений $\Sigma(P)$, может быть явно определено формулой φ.

В чем разница между понятиями неявной определимости соответственно в [8] и в [11], и почему в данном случае не возникает противоречия между указанными теоремой [8] и теоремой Бета [11]?

Дело в том, что понятие неявной определимости в смысле Бета по существу означает не определение, а задание отношения: происходит полное задание объема определяемого предиката. Формула $\varphi(x_1, ..., x_n)$ выделяет на модели множество кортежей (n), на которых предикат является истинным. Иначе говоря, это явное задание в первую очередь объема предиката (экстента в терминах анализа формальных понятий [12]), а не его содержания (интента). В то же время формальный глоссарий [8], определяющий некоторый набор понятий, ***не задает полностью*** объемы предикатов, соответствующих этим понятиям. Это, в частности, означает, что возможны обогащения одной и той же модели данным новым набором понятий (т. е. новым набором сигнатурных символов) таким, что у соответствующих предикатов будут разные объемы при разных обогащениях модели.

С другой стороны, явная определимость в смысле Бета в точности означает формульную определимость предиката P на моделях, на которых истинно множество предложений $\Sigma(P)$, а именно: в любой формуле $\psi \in F(\sigma \cup \{P\})$ мы можем заменить все вхождения предиката $P(x_1, ..., x_n)$ на формулу $\varphi(x_1, ..., x_n)$, получив в результате формулу $[\psi]^P_\varphi \in S(\sigma)$. При этом для любой модели $\mathfrak{A} \in K(\sigma)$, если $\mathfrak{A} \models \Sigma(P)$, то на модели $\mathfrak{A}$ истинность формул ψ и $[\psi]^P_\varphi$ равносильна:

$$\mathfrak{A} \models (\forall x_1 ... \forall x_k)\left(\psi(x_1,...,x_k) \leftrightarrow [\psi]^P_\varphi(x_1,...,x_k)\right).$$

Таким образом, новое понятие, описываемое предикатом P, добавленным к сигнатуре σ, и определяемое множеством предложений $\Sigma(P)$, на самом деле не дает ничего нового (с точки зрения содержания), а является своего рода «синтаксическим сахаром» – просто сокращением для формулы φ, выразимой в исходном наборе понятий σ.

Именно формульная определимость предиката P обуславливает то, что формула $\varphi(x_1, ..., x_n)$ из теоремы Бета полностью и однозначно задает объем предиката P на моделях, на которых истинно множество предложений $\Sigma(P)$.

Рассмотрим вопрос: а задает ли полностью формула φ смысл, содержание предиката P? А именно, из определения мы имеем

$$\Sigma(P) \models (\forall x_1 ... \forall x_n)\left(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\right).$$

Будет ли верно обратное, т. е. истинно ли предложение

$$(\forall x_1 ... \forall x_n)\left(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\right) \models \Sigma(P)?$$

Отрицательный ответ на этот вопрос дает следующее утверждение.

Предложение. Существуют сигнатура σ и множество предложений $\Sigma(P)$ такие, что

$$\Sigma(P) \cup \Sigma(P') \models (\forall x_1 ... \forall x_n)\left(P(x_1,...,x_n) \leftrightarrow P'(x_1,...,x_n)\right),$$

$$\Sigma(P) \models (\forall x_1 ... \forall x_n)\left(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\right),$$

но неверно

$$(\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big) \models \Sigma(P).$$

Таким образом, предложение $(\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big)$ всегда полностью задает объем предиката P, но не всегда полностью задает его смысл, содержание предиката P (если считать множество предложений $\Sigma(P)$ неявным описанием смысла понятия, обозначаемого предикатом P).

Здесь можно было бы возразить, что мы можем добавить к $\Sigma(P)$ что угодно, и при этом сохранится истинность утверждения

$$\Sigma(P) \models (\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big),$$

а утверждение

$$(\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big) \models \Sigma(P)$$

перестанет быть верным, даже если оно было верно до того. В связи с этим мы можем сформулировать вопрос: а следует ли $\Sigma(P)$ из $(\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big)$ вместе с множеством всех следствий множества предложений $\Sigma(P)$ в сигнатуре σ, т. е. вместе с множеством предложений $\mathrm{Th}\big(\Sigma(P)\big) \cap S(\sigma)$? Положительный ответ на этот вопрос дает следующее утверждение.

Предложение. Пусть $\Sigma(P) \models (\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big)$, $\varphi \in S(\sigma)$ и $\Sigma(P) \subseteq$ $\subseteq S\big(\sigma \cup \{P\}\big)$. Тогда выполнено

$$\mathrm{Th}\big(\Sigma(P)\big) \cap S(\sigma) \cup \big\{(\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big)\big\} \models \Sigma(P).$$

Тем не менее сделанное нами выше утверждение, что формула

$$(\forall x_1 ... \forall x_n)\big(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)\big)$$

полностью задает объем предиката P, но не всегда полностью задает его смысл, остается верным. Это иллюстрирует следующий пример.

Пример. Пусть $\sigma = \varnothing$, а $\Sigma(P) = \big\{(\forall x)P(x),\ (\forall x \forall y)\big(\big(P(x) \,\&\, P(y)\big) \to (x = y)\big)\big\}$. Тогда $\Sigma(P) \models (\forall x)\big(P(x) \leftrightarrow \varphi(x)\big)$, где $\varphi(x) = (x = x)$. С другой стороны, формула $\varphi(x)$ не является полным определением смысла предиката $P(x)$, задаваемого множеством предложений $\Sigma(P)$. А именно, из $\varphi(x)$ не следует, что предикат $P(x)$ выделяет ровно один элемент (т. е. предикат $P(x)$ истинен ровно на одном элементе); однако это свойство предиката P следует из его неявного определения $\Sigma(P)$. Таким образом, «явное определение» $(\forall x)\big(P(x) \leftrightarrow \varphi(x)\big)$ полностью задает объем предиката P, но не определяет его смысл.

Стало быть, из теоремы Бета не следует, что ***любое неявное определение смысла предиката можно свести к его явному определению***.

Можно было бы сказать, что ***любое неявное определение объема предиката можно свести к его явному определению***, но это также является не совсем верным. Дело в том, что у неявного определения $\Sigma(P)$ и у явного определения $(\forall x)\big(P(x) \leftrightarrow \varphi(x)\big)$ классы моделей разные. Действительно, любую модель сигнатуры $\sigma(\varphi)$ можно доопределить до модели сигнатуры $\sigma(\varphi) \cup \{P\}$ так, чтобы выполнялось утверждение $(\forall x)\big(P(x) \leftrightarrow \varphi(x)\big)$. С другой стороны, множество предложений $\Sigma(P)$ из приведенного выше примера может быть истинно только на одноэлементных моделях.

«Неявное» определение $\Sigma(P)$ объема предиката P в теореме Бета на самом деле сводится не к предложению $(\forall x_1 ... \forall x_n)(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n))$, а к множеству предложений $(\mathrm{Th}(\Sigma(P)) \cap S(\sigma)) \cup \{(\forall x_1 ... \forall x_n)(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n))\}$. Множества предложений $\Sigma(P)$ и $(\mathrm{Th}(\Sigma(P)) \cap S(\sigma)) \cup \{(\forall x_1 ... \forall x_n)(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n))\}$, как было показано выше, семантически эквивалентны, т. е. на них истинны одни и те же модели. Таким образом, «неявное» определение $\Sigma(P)$ ***объема*** предиката P сводится к «явному» определению его объема $(\forall x_1 ... \forall x_n)(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n))$, но по модулю множества предложений $(\mathrm{Th}(\Sigma(P)) \cap S(\sigma))$ сигнатуры σ. Это множество предложений задает класс моделей сигнатуры σ, на которых далее задается предикат P через его объем.

С другой стороны, является неверным утверждение, что «неявное» определение $\Sigma(P)$ ***смысла*** предиката P сводится к «явному» определению его смысла

$$(\forall x_1 ... \forall x_n)(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n)),$$

по модулю множества предложений $(\mathrm{Th}(\Sigma(P)) \cap S(\sigma))$. В рассмотренном выше примере, существенное свойство предиката P – то, что он истинен только на одном элементе (т. е. в определенном смысле задает константу – имя этого элемента), – не содержится в «явном» определении $(\forall x)(P(x) \leftrightarrow \varphi(x))$. В этом примере множество предложений $(\mathrm{Th}(\Sigma(P)) \cap S(\sigma))$ может быть истинно только на одноэлементных моделях. Поэтому из $(\mathrm{Th}(\Sigma(P)) \cap S(\sigma))$ и $(\forall x)(P(x) \leftrightarrow \varphi(x))$ дедуктивно следует предложение

$$(\forall x \forall y)((P(x) \,\&\, P(y)) \rightarrow (x = y)).$$

При этом очевидно, что в множестве предложений

$$(\mathrm{Th}(\Sigma(P)) \cap S(\sigma)) \cup \{(\forall x_1 ... \forall x_n)(P(x_1,...,x_n) \leftrightarrow \varphi(x_1,...,x_n))\}$$

свойство $(\forall x \forall y)((P(x) \,\&\, P(y)) \rightarrow (x = y))$ предиката P содержится неявно (!), в то время как в «неявном» определении $\Sigma(P) = \{(\forall x)P(x),\ (\forall x \forall y)((P(x) \,\&\, P(y)) \rightarrow (x = y))\}$ это свойство содержится явно. Получается, что в данном примере теорема Бета сводит не неявное определение смысла предиката к явному определению, а наоборот, явное определение к неявному.

Таким образом, с точки зрения смысла предиката P, а не его объема, понятия явного и неявного определения, используемые в теореме Бета, не являются полностью адекватными и требуют пересмотра.

Алгоритм формирования описания объекта из фрагмента OWL-онтологии

Общая и полная схема генерации без детализации происходящих процессов состоит из трех основных блоков:

1) планирование содержания текста – построение структуры текста;
2) микропланирование – построение планов предложений;
3) языковое оформление – реализация построенных планов предложений соответствующими грамматическими структурами.

В прикладных системах генерации к этим трем этапам часто добавляется четвертый этап – физическое представление, на котором производится форматирование текста согласно выбранному формату (PDF, HTML и др.).

Планирование состоит из следующих частей: выбор контента (информационного содержания), в котором система выбирает информацию для передачи в следующий этап конвейера, и текстовое планирование, где она планирует структуру текста, который будет сгенерирован.

Когда систему просят описать целевой объект, она прежде всего извлекает из онтологии все OWL-утверждения, связанные с объектом вниз по иерархии вплоть до уровня, указанного в настройках пользователя. Затем алгоритмы преобразуют извлеченный набор OWL операторов в триплеты (см. таблицу).

OWL-утверждения и соответствующие им формы триплетов

OWL-утверждения	Триплеты
Object – экземпляр	
ClassAssertion(Class, object)	<object, instanceOf, Class>
ClassAssertion(ObjectComplementOf(Class), object)	<object, not(instanceOf), Class>
ClassAssertion(ObjectOneOf(indiv1,indiv2) object)	<object, one of, or(indiv1, indiv2, …)>
ClassAssertion(ObjectHasValue(ojbProp, indiv) object)	<object, objProp, indiv>
ClassAssertion(ObjectHasValue(dataProp,dataValue) object)	<object, dataProp, indiv>
ClassAssertion(ObjectHasSelf(objProp) object)	<object, objProp, object>
ClassAssertion(ObjectMaxCardinality(number,prop,[Class]) object)	<object,maxCardinality(prop),number[:Class]>
ClassAssertion(ObjectMinCardinality(number,prop,[Class]) object)	<object,minCardinality(prop),number[:Class]>
ClassAssertion(ObjectExactCardinality(number,prop,[Clas s]) object)	<object,exactCardinality(prop),number[:Class]>
ClassAssertion(ObjectSomeValuesVrom(objProp,Class) object)	<object, someValuesFrom(objProp),Class>
ClassAssertion(ObjectAllValuesVrom(objProp,Class) object)	<object, allValuesFrom(objProp),Class>
ClassAssertion(ObjectIntersectionOf(C1,C2,…) object)	convert(ClassAssertion(C1, object)) convert(ClassAssertion(C2, object))…
ClassAssertion(ObjectUnionOf (C1,C2,…)object)	or(convert(ClassAssertion(C1,object)),convert (ClassAssertion(C2, object)),…)
ObjectPropertyAssertion(objProp,object,indiv)	<object, objProp, indiv>
DataPropertyAssertion(dataProp,object,dataValue)	<object, dataProp, dataValue>
NegativeObjectPropertyAssertion(objProp,object,indiv)	<object, not(objProp), indiv>
NegativeDataPropertyAssertion(dataProp,object,dataValue)	<object, not(dataProp), dataValue>
DifferentIndividuals(object,indiv)	<object, differentIndividuals, indiv>

Окончание таблицы

OWL-утверждения	Триплеты
DifferentIndividuals(indiv,object)	<object, differentIndividuals, indiv>
SameIndividual(object,indiv)	<object, sameIndividuals, indiv>
SameIndividualindiv object)	<object, sameIndividuals, indiv>
Object – класс	
EquivalentClasses(Object, Class/expr)	convert (SubClassOf (Object, Class/expr)
EquivalentClasses(Class/expr, Object)	convert (SubClassOf (Object, Class/expr)
SubClassOf(Object, Class)	<Object, isA, Class>
SubClassOf(Object, ObjectComplementOf(Class))	<Object not(isA), Class>
SubClassOf(Object,ObjectOneOf(indiv1, indiv2 …)	<Object, one of, or(indiv1, indiv2 …)>
SubClassOf(Object,ObjectHasValue(objProp, indiv))	<Object, objProp, indiv>
SubClassOf(Object,ObjectHasValue(dataProp, dataValue)	<Object, dataProp, dataValue>
SubClassOf(Object, ObjectHasSelf(objProp))	<Object, objProp, Object>
SubClassOf(Object, ObjectMaxCardinality(number,prop, [Class]))	<Object,maxCardinality(prop),number[:Class]>
SubClassOf(Object,ObjectMinCardinality(number,prop,[Class]))	<Object,minCardinality(prop),number[:Class]>
SubClassOf(Object, ObjectExactCardinality(number,prop,[Class]))	<Object,exactCardinality(prop),number[:Class]>
SubClassOf(Object, ObjectSomeValuesFrom(objProp, Class))	<Object, someValuesFrom(objProp), Class>
SubClassOf(Object, ObjectAllValuesFrom(objProp, Class))	<Object, allValuesFrom(objProp), Class>
SubClassOf(Object,ObjectIntersectionOf (C1/expr, C2/expr …)	convert(SubClassOf(C1/expr, Object)) convert(SubClassOf(C2/expr, Object)) …
SubClassOf(Object,ObjectUnionOf(C1/expr, C2/expr …)	or(convert(SubClassOf(C1/expr,Object)) convert(SubClassOf(C2/expr, Object))…)
DisjointClasses(Object, Class)	<Object, not(isA), Class>
DisjointClasses(Class, Object)	<Object, not(isA), Class>

Текстовое планирование представляет собой упорядочивание предложений по тематике с использованием дополнительных ресурсов системы.

Для работы метода группировки предложений по темам автор онтологии может определить разделы и назначить каждому из них свойства. Если разделы будут определены, то на этапе текстового планирования система распределит триплеты по соответствующим группам. Также эксперт может задать порядок самих разделов, что приводит к дополнительной итерации упорядочивания. Все секции, принадлежность свойств секциям и порядок разделов и свойств определяются в зависимых от онтологии ресурсах генерации. В нашем случае ресурсы также имеют представление OWL-онтологий.

В общем случае алгоритм сортировки имеет следующий вид:

Процедура **orderMessageTriples**

ВХОД:

t[0]: целевой объект

t[1], ..., t[n]: объекты второго уровня
L[0]: неотсортированный список триплетов, описывающих t[0]
...
L[n]: неотсортированный список триплетов, описывающих t[n]
SMap: Карта соответствий между отношениями (предикатами) и наименованием секции
SOrder: Отсортированные наименования секций
POrder: Частично отсортированный массив отношений:

ВЫХОД:
Отсортированный список триплетов

ШАГИ:
```
от  i := 0 до  n {
        orderMessageTriplesAux(L[i], SMap, SOrder, POrder)
}
от  i := 1 до  n {
        insertAfterFirst(, L[0], L[i])
}
Вернуть:  L[0]
```

Процедура **orderMessageTriplesAux**

ВХОД:
L: Неотсортированный список триплетов об одном объекте
SMap: Карта соответствий между отношениями (предикатами) и наименованием секции
SOrder: Отсортированные наименования секций
POrder: Частично отсортированный массив отношений

ЛОКАЛЬНЫЕ ПЕРЕМЕННЫЕ:
S[1], ..., S[k]: списки с триплетами по каждой секции

ВЫХОД:
Отсортированный список триплетов об одном объекте

ШАГИ:
```
<S[1], ..., S[k]> := splitInSections(L, SMap)
от  i := 1 до  k {
        S[i] := orderTriplesInSection(S[i], POrder)
}
<S[1], ..., S[k]> := reorderSections(S[1], ..., S[k], SOrder)
Вернуть: concatenate(S[1], ..., S[k])
```

Микропланирование – это блок, который позволяет от предметных знаний перейти к языковым. В нем решается, каким образом выбранная информация будет реализована языковыми средствами в виде предложений на естественном языке. В нашем случае микропланирование состоит из трех подэтапов.

1. Лексикализация концептов сообщения, т. е. выбор подходящих слов для выражения выбранного в них содержания.

Для каждого глагола, существительного или прилагательного, которые эксперт хочет использовать в планах предложений, должны быть представлены соответствующие «синтаксические формы». Большинство синтаксических форм слов русского языка можно автоматически создавать из базовых форм, используя простые правила морфологии. Расширение нашей системы подобными компонентами будет рассмотрено в дальнейших работах.

2. Агрегирование сообщений до структур, соответствующих отдельным предложениям создаваемого текста.

Системы генерации текстов часто объединяют предложения, полученные после обработки триплетов, в более длинные, чтобы улучшить читаемость. Агрегация предложения выполняется при помощи набора определенных правил, которые применяются к структурам, полученным в ходе работы текстового планировщика. Другими словами, они применяются к предложениям, уже отсортированным и сгруппированным по семантике. Действительно,

объединение не связных по смыслу предложений будет звучать неестественно. Так как агрегация происходит по четко заданным шаблонам, ее возможности полностью зависят от результата работы предыдущего этапа. Рассмотрим теперь каждое из используемых правил.

Исключение повторов одного имени существительного с несколькими прилагательными. Последовательность триплетов сообщений в форме <S, P, O1>, ..., <S, P, On> будет объединена в один триплет <S, P, и (O1, ... On)>.

Объединение последовательности триплетов, **выраженных в формах <S; M(P); O> и <S; P; O>**, **для одинаковых S и P**, причем M – это один из minCardinality, maxCardinality, exactCardinality. Применение данного правила снимает ограничение на maxMessagesPer Sentence.

Модель продается максимум в трех странах. Модель продается минимум в трех странах. Модель продается в Испании, Италии и Греции => Модель продается в трех странах: Испании, Италии и Греции.

Объединение предложений вида <S, **instanceOf**, **C> или <S**, **isA**, **C> с предложением <S**, **P**, **O>**, для того же S, где P является свойством онтологии.

Ромашка – это цветок. Ромашка растет в поле => Ромашка – это цветок, который растет в поле.

Объединение нескольких предложений, **каждое из которых является выражением триплета <S; Pi; Oi>**, **для одного и того же S**, где Pi – это свойство онтологии. В каждом из последующих или предшествующих предложений должен быть сам объект, за которым сразу следует только прилагательное. Прилагательные вписываются в результирующее предложение, сохраняя их порядок, определенный планировщиком.

Это мотоцикл. Он красный. Он дорогой. => Это красный, дорогой мотоцикл.

Объединение последовательности предложений, **связанных с одним и тем же глаголом**, **каждый из которых выражает триплет <S**, **Pi**, **Oi>**, где S одинаковый во всех предложениях, а Pi – это свойства онтологии. Результирующее предложение может быть сформировано путем однократного использования существительного и глагола. Соединительный союз «и» вставляется перед последним существительным.

Напиток имеет средний вкус. Напиток имеет умеренный аромат => Напиток имеет средний вкус и умеренный аромат.

Объединение последовательности предложений, **не связанных с одним и тем же глаголом**, **каждое из которых является представлением триплета <S**, **Pi**, **Oi>**, где S одинаковое во всех тройках, а Pi – свойства онтологии.

Вино сухое. Вино имеет средний аромат. Оно произведено в Италии. => Вино сухое, имеет средний аромат и произведено в Италии.

Архитектура системы

Для программного продукта были выработаны следующие требования:

1) форма десктопного приложения;
2) работа с уже существующей системой логического вывода;
3) работа с уже существующей системой получения фрагментов атомарных диаграмм и алгоритмами получения DL утверждений;
4) удобство и простота использования.

Под эти требования подходило большое число языков программирования, однако основной выбор пал на объектно-ориентированные. Также были сформулированы не функциональные требования:

1) быстрота и удобство разработки;
2) возможность устройства модульной архитектуры;
3) знания и умения разработки на выбранном языке программирования.

Поскольку быстрота работы программы в данном случае не является существенным требованием, был выбран кроссплатформенный язык программирования Java, так как он удовлетворил всем поставленным требованиям.

Структура программы представлена следующими директориями.

- Основные классы программы:

1) пакет, отвечающий за возможность графической работы;
2) пакет представления данных;
3) пакет классов, реализующих основную логику приложения.

- В программе используются дополнительные библиотеки:

1) Apache Commons;
2) OWLAPI – библиотека для работы с OWL-онтологиями;
3) Jena – библиотека для работы с OWL-онтологиями;
4) Hermit – машина логического вывода.

Основной модуль разработан с использованием архитектурного шаблона Pipeline. На рис. 2 представлена его use-case диаграмма, отображающая основные возможные действия пользователя.

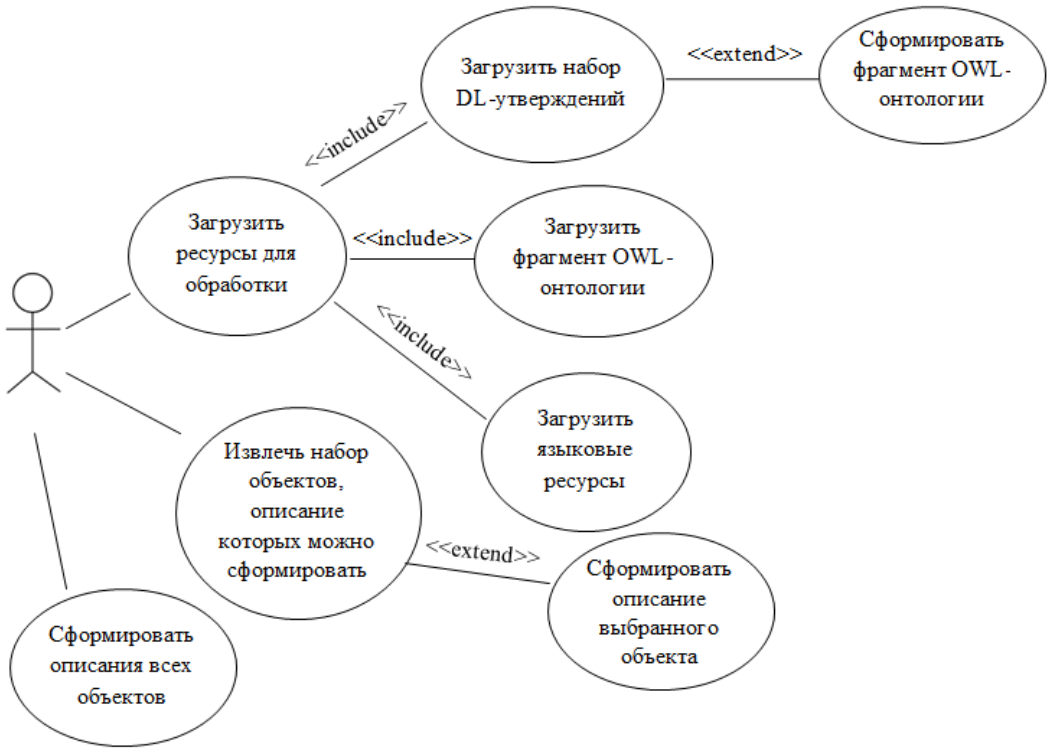


Рис. 2. Use-case диаграмма основного модуля системы

В рамках дополнительных возможностей все загружаемые элементы можно проверить машиной логического вывода Hermit.

Заключение

В работе изучены вопросы полноты определений ключевых понятий предметной области, явной и неявной определимости ключевых понятий. Исследована полнота определений понятий относительно объема и содержания соответствующих сигнатурных предикатов.

Разработаны методы интеграции фрагментов определения ключевого понятия, извлеченных из текстов естественного языка. Методы интеграции частей определения понятия основаны на представлении знаний, извлеченных из текстов естественного языка, в виде

фрагментов атомарной диаграммы алгебраической системы и на преобразовании фрагментов атомарной диаграммы в логику описаний (DL) и в OWL. Полученные OWL-спецификации погружаются в OWL-онтологию, соответствующую данной предметной области.

Разработаны методы интеграции OWL-спецификаций, представленных в OWL-онтологии, и порождения по ним фраз естественного языка.

Список литературы

1. *Bateman J., Zock M.* List of Natural Language Generation Systems. URL: http://www.fb10.-uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.html.

2. *Shah H., Warwick K., Vallverdú J., Wu D.* Can machines talk? Comparison of Eliza with modern dialogue systems // Computers in Human Behavior. 2006. No. 58. P. 278–95.

3. *Lowden B. G. T., Roeck A. N. de.* The REMIT system for paraphrasing relational query expressions into natural language // Proc. Int'l. Conf. on Very Large Data Bases. Kyoto, Japan, 1986.

4. *Minock M.* STEP A Natural Language Interface to Database. URL: http://www.cs.umu.se/~mjm/

5. *Махасоева О. Г., Пальчунов Д. Е.* Автоматизированные методы построения атомарной диаграммы модели по тексту естественного языка // Вестн. НГУ. Серия: Информационные технологии. 2014. Т. 12, № 2. С. 64–73.

6. *Махасоева О. Г., Пальчунов Д. Е.* Программная система построения атомарной диаграммы модели по тексту естественного языка. Св-во о государственной регистрации программы для ЭВМ № 2014619198, зарегистрировано 10.09.2014.

7. *Корсун И. А., Пальчунов Д. Е.* Теоретико-модельные методы извлечения знаний о смысле понятий из текстов естественного языка // Вестн. НГУ. Серия: Информационные технологии. 2016. Т. 14, № 3. С. 34–48.

8. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии. Ч. 2. Онтологии и формализация понятий // Философия науки. 2008. № 2 (37). С. 62–99.

9. *Корсун И. А., Пальчунов Д. Е.* Методы извлечения и формального представления онтологических знаний из текстов естественного языка // Знания – Онтологии – Теории (ЗОНТ-2017): Материалы Всерос. конф. с междунар. участием. 2017. Т. 2. С. 21–25.

10. *Капустина А. И., Пальчунов Д. Е.* Разработка онтологической модели тарифов и услуг сотовой связи, основанной на логически полных определениях понятий // Вестн. НГУ. Серия: Информационные технологии. 2017. Т. 15, № 2. С. 34–46.

11. *Кейслер Г., Чэн Ч. Ч.* Теория моделей. М.: Мир, 1977.

12. *Ganter B., Wille R.* Formal Concept Analysis. Mathematical foundations. Berlin, Heidelberg: Springer-Verlag, 1999.

Материал поступил в редколлегию 27.06.2018

I. A. Korsun [1], **D. E. Palchunov** [1, 2]

[1] *Novosibirsk State University*
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

[2] *S. L. Sobolev Institute of Mathematics SB RAS*
4 Academician Koptyug Ave., Novosibirsk, 630090, Russian Federation

irina.korsun.nsu@gmail.com, palch@math.nsc.ru

AN INTELLECTUAL SYSTEM FOR PROCESSING AND INTEGRATING KNOWLEDGE BASED ON SEMANTIC WEB TECHNOLOGIES

The paper is devoted to the development of model-theoretic methods of concepts definitions extraction from the natural language texts. The information extracted from texts is represented in the form of statements on the Description Logic language (DL) by transformation through fragments of atomic diagrams of algebraic systems. Such a representation allows you to get more expressive

texts, in contrast to algorithms, where information is represented in a database or by expressions in a formal language (for example, SQL).

Keywords: ontology, model-theoretic methods, fragments of atomic diagrams, concept definitions, definition extraction, completeness of definitions of concepts, explicit definability, implicit definability, text generation, natural language generation.

References

1. Bateman J., Zock M. List of Natural Language Generation Systems. URL: http://www.fb10.-uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.html.

2. Shah H., Warwick K., Vallverdú J., Wu D. Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, 2006, no. 58, p. 278–95.

3. Lowden B. G. T., Roeck A. N. de. The REMIT system for paraphrasing relational query expressions into natural language. *Proc. Int'l. Conf. on Very Large Data Bases*. Kyoto, Japan, 1986.

4. Minock M. STEP A Natural Language Interface to Database. URL: http://www.cs.umu.se/~mjm/.

5. Makhasoeva O. G., Palchunov D. E. Semi-automatic methods of a construction of the atomic diagrams from natural language texts. Vestnik NSU. Series: Information Technologies, 2014, vol. 12, no. 2, p. 64–73. ISSN 1818-7900 (in Russ.)

6. Makhasoeva O. G., Palchunov D. E. Program system for the construction of the atomic diagram of a model from natural language texts. Certificate of the State Registration of the computer program. No. 2014619198, registered 10.09.2014. (in Russ.)

7. Korsun I. A., Palchunov D. E. Model-Theoretic Methods of Extraction of Knowledge on the Meaning of Concepts from the Natural Language Texts. *Vestnik NSU. Series: Information Technologies*, 2016. vol. 14, no. 3, p. 34–48. ISSN 1818-7900 (in Russ.)

8. Palchunov D. E. Modeling of reasoning and formalization of reflection II: Ontologies and formalization of concepts. *Filosofiya nauki*, 2008, no. 2 (37), p. 62–99. (in Russ.)

9. Korsun I. A., Palchunov D. E. Methods of extraction and formal representation of ontological knowledge from natural language texts. *Proc. of the All-Russian Conference with International Participation «Knowledge-Ontology-Theory» (KONT-17)*. Novosibirsk, 2017, vol. 2, p. 21–25.

10. Kapustina A. I., Palchunov D. E. The development of ontological model of tariffs and services of mobile operator, based on logically complete definitions of concepts. *Vestnik NSU. Information Technologies*, 2017, vol. 15, no. 2, p. 34–46. ISSN 1818-7900 (in Russ.)

11. Chang C. C., Keisler H. J. Model Theory. *Studies in Logic and the Foundations of Mathematics*, 1977.

12. Ganter B., Wille R. Formal Concept Analysis. Mathematical foundations. Berlin, Heidelberg, Springer-Verlag, 1999.

For citation:

Korsun I. A., Palchunov D. E. An Intellectual System for Processing and Integrating Knowledge Based on Semantic Web Technologies. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 113–125. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-113-125

УДК 025.4.028
DOI 10.25205/1818-7900-2018-16-3-126-132

Д. С. Матусевич [1], **О. В. Изместьева** [2]

[1] *Байкальский государственный университет*
ул. Ленина, 11, Иркутск, 664003, Россия

[2] *Иркутский региональный колледж педагогического образования*
5-я Железнодорожная ул., 53, Иркутск, 664074, Россия

mds@bgu.ru, Olga_isea@mail.ru

НЕКОТОРЫЕ ПОДХОДЫ К ХРАНЕНИЮ ИНФОРМАЦИИ О КНИГООБЕСПЕЧЕННОСТИ УЧЕБНОГО ПРОЦЕССА ВЫСШЕГО УЧЕБНОГО ЗАВЕДЕНИЯ

Рассматриваются три подхода к хранению информации о книгообеспеченности учебного процесса при различных вариантах интеграции автоматизированной библиотечной информационной системы (АБИС) с информационной системой вуза: внутри АБИС в формате RUSMARC, в виде реляционной базы данных, с использованием технологии OLAP. Анализируются преимущества и недостатки каждого подхода, приводятся примеры организации хранения данных (поля RUSMARC, измерения и меры для кубов OLAP).

Ключевые слова: книгообеспеченность, вузовская библиотека, АБИС, интеграция с информационной системой вуза, OLAP.

Книгообеспеченность (КО) как понятие – это определение числа экземпляров книг, отобранных по разным критериям, в расчете на одного студента: по специальностям, направлениям и профилям обучения, по видам и формам обучения, по конкретным дисциплинам, по видам учебной литературы и т. д. В качестве основного показателя (коэффициента) книгообеспеченности предлагается показатель КО конкретной дисциплины, который определяется как частное от деления количества экземпляров учебной литературы, имеющейся в библиотеке по данной дисциплине, на число студентов, ее изучающих. Таким образом, в библиотеках учебных заведений электронная картотека книгообеспеченности (ЭКК) является активной подсистемой и реализует основные функции управления библиотечными фондами [1; 2] (рис. 1, 2).

Нормативные документы [3; 4] для высших учебных заведений определяют нормы и правила доступа обучающихся к литературе. Из [3] (через федеральные государственные образовательные стандарты (ФГОС)) следует наличие основной учебной литературы из расчета обеспечения каждого обучающегося по всем дисциплинам реализуемых образовательных программ в количестве 0,5 экземпляра на 1 студента, дополнительной учебной литературы в количестве 0,25 экземпляра на 1 студента. Также во ФГОС задано право обучающихся на «индивидуальный неограниченный доступ к электронно-библиотечной системе, содержащей издания учебной, учебно-методической и иной литературы по основным изучаемым дисциплинам», т. е. вводится применение электронных библиотечных систем (ЭБС). В [4] оценивается доля «укрупненных групп специальностей и направлений подготовки», обеспеченных литературой из ЭБС.

Матусевич Д. С., *Изместьева О. В.* Некоторые подходы к хранению информации о книгообеспеченности учебного процесса высшего учебного заведения // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 126–132.

© Д. С. Матусевич, О. В. Изместьева, 2018

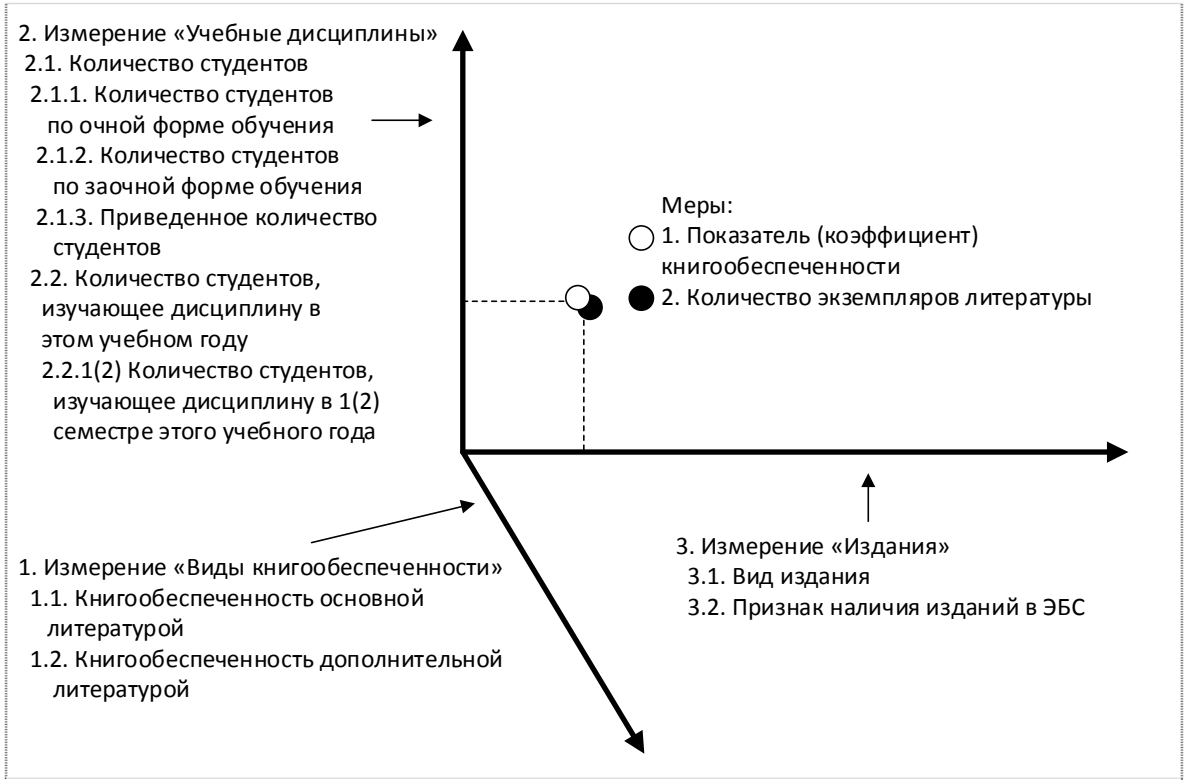


Рис. 1. Проект куба данных ЭКК с измерением «1 – Виды книгообеспеченности»

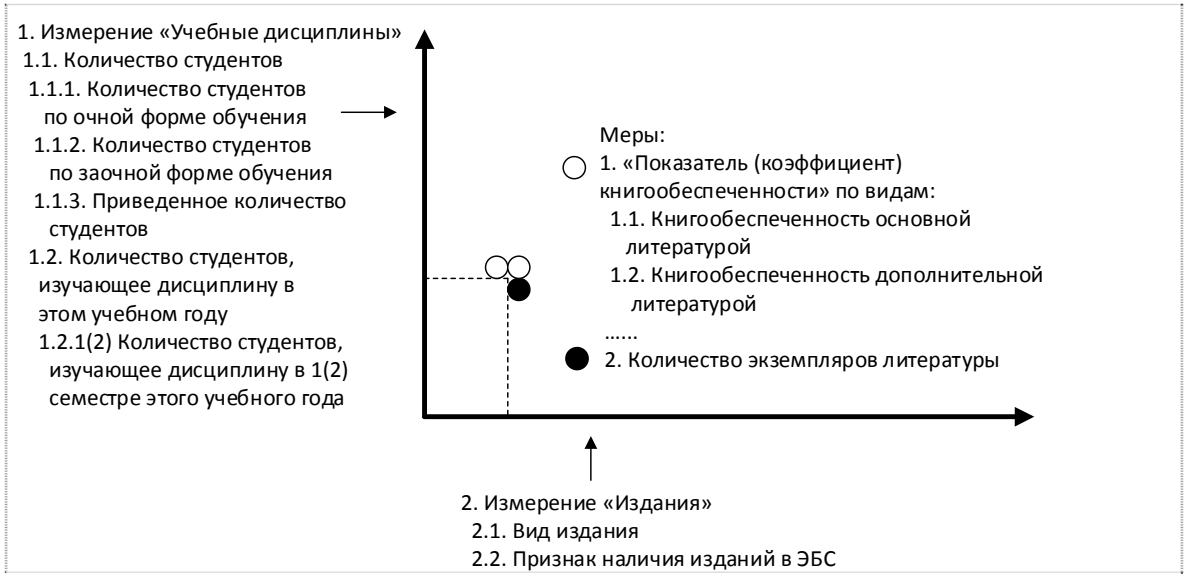


Рис. 2. Проект куба данных ЭКК с мерой «Виды книгообеспеченности»

Соответственно коэффициент КО отражает степень обеспеченности книгой того количества студентов, для которых она предназначена. Если учитывать множественность критериев для ее оценки, то становится ясно, насколько это трудоемкий процесс, требующий большого числа расчетов. Поэтому вполне закономерно, что в течение последних лет в составе многих систем автоматизации библиотек появились средства ведения ЭКК [1; 2; 5].

Основными исходными данными для функционирования ЭКК являются:

- учебные планы высшего учебного заведения;
- контингент студентов вуза;

- учебные программы дисциплин;
- издания, рекомендуемые к использованию в учебном процессе независимо от вида документа.

Таким образом, построение ЭКК ведется на стыке двух крупнейших информационных систем образовательной организации – электронного каталога библиотеки, хранимого в автоматизированной библиотечной информационной системе (АБИС), и информационной системы учебного процесса вуза (ИС ВУЗ) [1; 2].

Характер информации, хранимой в перечисленных информационных системах, определяет требования к организации системы хранения данных ЭКК:

1) совместимость с АБИС;

2) совместимость с ИС ВУЗ;

3) отсутствие необходимости в моментальном (online) обновлении данных.

Указанные требования к системе хранения данных в определенных комбинациях противоречат друг другу. Также в системе ЭКК может находиться информация, ранее накопленная, но к настоящему моменту исключенная из требований нормативных документов (например, требование к «устареваемости» фонда, циклы дисциплин и т. п.)

По результатам проведенного анализа предлагаются 3 модели хранения данных о книгообеспеченности учебного процесса в ЭКК вуза:

- модель данных на базе формата UNIMARC (RUSMARC);
- модель данных на базе реляционных баз данных (БД);
- модель данных на базе технологии OLAP [5].

Каждый метод имеет свои преимущества и недостатки в части хранения данных.

Модель данных на базе формата UNIMARC (RUSMARC)

Формат UNIMARC призван быть посредником при осуществлении обмена библиографическими записями между книгоиздателями и потребителями (библиотеками). Официальной российской версией UNIMARC является RUSMARC или российский коммуникативный формат, разработанный по заказу Министерства культуры РФ в рамках программы LIBNET под эгидой Российской библиотечной ассоциации.

Описание UNIMARC (RUSMARC) разбито на блоки с 3-значными описаниями, каждый из которых отвечает за строго определенную часть описания. Блоки с номерами больше 900 оставлены для свободного описания в так называемом «Блоке локального использования» [6]. Соответственно для ведения ЭКК предлагается использовать следующие поля из блока 900 (пример):

930 – дисциплина, которая комплектуется данным изданием;

931 – код книгообеспеченности литературы со значениями: 1 – основная литература, 2 – дополнительная литература и т. д.;

932 – наличие электронного издания со значениями: 0 – отсутствует,1 – присутствует;

933 – наличие этого издания в ЭБС, доступных в вузе со значениями: 0 – отсутствует, 1 – присутствует;

934 – название ЭБС (если присутствует);

935–939 – резерв, например для указания циклов / компонент дисциплин.

К достоинствам данной модели хранения информации следует отнести полную интеграцию в электронный каталог библиотеки. Например, библиотекарь может построить списки рекомендованной литературы по дисциплине для преподавателей и студентов. Недостатком модели является отсутствие связей с информационной системой вуза, в том числе отсутствие данных о численности студентов и, как следствие, самих коэффициентов КО. Ситуация может быть решена за счет импорта данных из ИС ВУЗ в АБИС. Предлагается использовать следующие поля:

940 – количество студентов;

941 – количество студентов по дневной форме обучения;

942 – количество студентов по заочной форме обучения;

943 – приведенное количество студентов (1 студент очной формы обучения = 10 студентов заочной формы обучения);

944 – количество студентов, изучающих дисциплину в этом учебном году;
945 – количество студентов, изучающих дисциплину в 1-м семестре в этом учебном году;
946 – количество студентов, изучающих дисциплину во 2-м семестре в этом учебном году;
947 – кафедра, ведущая данную учебную дисциплину;
948 – направление / специальность, которое изучает данную учебную дисциплину;
949 – профиль внутри направления, которое изучает данную учебную дисциплину;
950 – коэффициент книгообеспеченности основной литературой;
951 – коэффициент книгообеспеченности дополнительной литературой и т. д. [5].

Модель данных на базе реляционных баз данных

Основным назначением UNIMARC (RUSMARC) является обмен данными, т. е. данный формат изначально был оптимизирован для обмена информацией, но не для хранения и выбора. Поэтому АБИС, как правило, имеют собственный движок, основанный на принципах реляционных баз данных, либо используют тиражируемые системы управления базами данных.

Переход от формата UNIMARC (RUSMARC) к формату, поддерживающему реляционные системы управления базами данных, осуществляется по следующей схеме: каждый блок в описании RUSMARC рассматривается как отдельная сущность (см. таблицу). Перевод хранения данных ЭКК в формат реляционных БД позволяет подключать данные об учебном процессе вуза.

Следовательно, внедрение данных о книгообеспеченности в ИС ВУЗ позволит оценивать необходимость комплектования библиотеки для существующих и планируемых специальностей, проводить мониторинг состояния обеспеченности учебного процесса, оперативно информировать кафедры о книгообеспеченности отдельных дисциплин и необходимости доукомплектования литературой.

Список измерений и их иерархии

Измерения куба	Описание измерения и мер
1	Виды книгообеспеченности (см. прим. к таблице)
1.1	Книгообеспеченность основной литературой
1.2	Книгообеспеченность дополнительной литературой
2	Учебные дисциплины
2.1	Количество студентов
2.1.1.(2)	Количество студентов по дневной (заочной) форме обучения
2.1.3	Приведенное количество студентов (1 студент очной формы обучения = 10 студентов заочной формы обучения)
2.2	Количество студентов, изучающих дисциплину в этом учебном году
2.2.1.(2)	Количество студентов, изучающих дисциплину в 1-м (2-м) семестре
3	Издания
3.1	Вид издания
3.1.1(2,…)	Учебники, учебные пособия, монографии и т. д.
3.2	Признак наличия изданий в ЭБС
3.3	Количество экземпляров
3.3.1	Количество экземпляров за весь срок
3.3.2.(3)	Количество экземпляров, изданных за последние 5 (10) лет

Примечание: измерение «1 – Виды книгообеспеченности» может быть как измерением, так и мерой куба. В первом случае мерой куба является значение коэффициента КО, вычисленного как частное от деления соответствующего количества экземпляров литературы на число студентов (см. рис. 1). Во втором случае число мер куба равно числу показателей книгообеспеченности, используемых в библиотеке, например, только основной или только дополнительной литературы (см. рис. 2). Для обоих случаев коэффициент КО равен 1, если присутствует признак наличия изданий из ЭБС.

Поскольку часть информации из ИС ВУЗ может быть доступна студентам, то появляется возможность их информирования о рекомендуемой учебной литературе, распределении учебной литературы по группам, семестрам, формам обучения и т. д. [5].

Модель данных на базе технологии OLAP

При всех достоинствах применения модели хранения информации на базе реляционных БД процесс формирования показателей о КО в основном использует крайне трудоемкую операцию выбора данных из БД (т. е. команду SELECT). Таким образом, логично перейти от систем управления БД к системам управления банками данных по технологии OnLine Analytical Processing (OLAP) [5; 7].

OLAP делает мгновенный снимок реляционной БД и структурирует ее в пространственную модель для запросов. Заявленное время обработки запросов в OLAP составляет около 0,1 % от аналогичных запросов в реляционную БД.

OLAP-структура, созданная из рабочих данных, называется OLAP-куб или куб данных. Куб создается из соединения таблиц с применением схемы звезды или схемы снежинки. В центре схемы звезды находится таблица фактов, содержащая ключевые факты, по которым делаются запросы. Множественные таблицы с измерениями присоединены к таблице фактов. Эти таблицы показывают, как могут анализироваться агрегированные реляционные данные. Количество возможных агрегирований определяется количеством способов, которыми первоначальные данные могут быть иерархически отображены. Индексам массива соответствуют измерения (dimensions) или оси куба, а значениям элементов массива – меры (measures) куба [8].

Для ЭКК предлагается проект куба данных по хранению информации о книгообеспеченности учебного процесса. Для наполнения куба данных возможен экспорт данных из АБИС библиотеки (как правило, в формате RUSMARC) и данных ИС ВУЗ (как правило, из реляционных БД) [5].

Минимальное количество измерений проектируемого куба – 2: «Учебные дисциплины», описывающие перечень дисциплин, изучаемых в вузе, и «Издания», представленные количеством изданий, закрепленных за дисциплиной. Список иерархий измерений и мер куба представлен в таблице (см. выше), на рис. 1 и 2 приведены графические представления проектов кубов данных при различных вариантах измерений и мер.

Еще одной мерой куба данных является количество экземпляров литературы, закрепленных за дисциплиной, которое позволит библиотеке вуза оценить комплектование дисциплины в абсолютных показателях. Распределение экземпляров между сходными и смежными дисциплинами ведется пропорционально числу студентов, изучающих дисциплины.

Применение банков данных для хранения данных о книгообеспеченности учебных дисциплин потребует организации экспорта данных как из электронного каталога библиотеки, так и из баз данных системы ИС ВУЗ. Также необходимо разработать приложение для работы с созданным кубом данных OLAP. Однако в результате получаемая автономность от АБИС и ИС ВУЗ в сочетании с высокой скоростью работы (за счет заранее вычисленных показателей) позволяет говорить о перспективе использования технологии OLAP в плане управления книгообеспеченностью вузовских библиотек [5; 7].

Список литературы

1. *Безденежных Г. Н., Борисенко Е. С.* Электронная картотека книгообеспеченности в системе управления фондами вузовской библиотеки / Омский государственный технический университет. Омск, 2013. URL: http://lib.omgtu.ru/Data/Pages/356/Bezdenezhnih.doc (дата обращения 01.06.2018).

2. *Каллиников П. Ю.* Книгообеспеченность в «Университетской библиотеке онлайн»: рекомендательный сервис ЭБС // «Университетская библиотека онлайн»: вебинар 29.09.2017. URL: https://www.youtube.com/watch?v=dD8-zc9DKIY (дата обращения 01.06.2018).

3. Об образовании: Федер. закон от 29.12.2012 №273-ФЗ // Рос. газета. 2012. 31 дек.

4. Об утверждении показателей деятельности образовательной организации, подлежащей самообследованию: Приказ Министерства образования и науки Российской Федерации от 10 декабря 2013 г. № 1324 // Рос. газета. 2014. 19 февр.

5. Изместьева О. В. Подходы к хранению информации о книгообеспеченности учебного процесса // Культура: теория и практика: электронный научный журнал. 2016. Вып. 3 (12). С. 15. URL: theoryofculture.ru (дата обращения 01.06.2018).

6. RUSMARC в примерах: Учеб. пособие для каталогизаторов / Сост. Л. И. Беневоленская и др.; под общ. ред. О. Н. Кулиш, Б. Р. Логинова; РНБ; Нац. информ.-библ. центр ЛИБНЕТ. М.: ФАИР-ПРЕСС: Центр ЛИБНЕТ, 2005. Ч. 1: Однотомные, многотомные и сериальные издания. 999 с.

7. *Фомичева С. Г., Попкова А. А.*Современные методы анализа данных с использованием клиентских и серверных OLAP-средств: Учеб. пособие / Норильский государственный индустриальный институт. Норильск, 2016. 202 с. : цв. ил.

8. *Фомичева С. Г., Попкова А. А.* Коэффициент книгообеспеченности как критерий многомерного анализа обеспеченности учебного процесса // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2009. № 91. С. 206–220.

Материал поступил в редколлегию 05.06.2018

D. S. Matusevich [1], **O. V. Izmesteva** [2]

[1] *Baikal State University*
11 Lenin Str., Irkutsk, 664003, Russian Federation

[2] *Irkutsk Regional College of Pedagogical Education*
53 The 5th Zheleznodorozhnaya Str., Irkutsk, 664074, Russian Federation

mds@bgu.ru, Olga_isea@mail.ru

SOME APPROACHES TO STORAGE INFORMATION ABOUT TEXTBOOK PROVISION OF EDUCATION PROCESS AT UNIVERSITIES

At the article explore three approaches to storage information about textbooks provision of educational process in different types of integration Integrated Library System and University's Information System: inside the Integrated Library System in RUSMARC format, at relational database, OLAP technology. Analyzing advantages and disadvantages of each approach.

Keywords: textbooks provision, university's library, Integrated Library System, integration with University's Information System, OLAP.

References

1. Bezdenezhnykh G. N., Borisenko E. S. Electronic catalog of textbook provision at management system of the funds at university's libraries. Omsk State Technical University. URL: http://lib.omgtu.ru/Data/Pages/356/Bezdenezhnih.doc (in Russ.)

2. Kallinikov P. Yu. Textbook provision at «Online university's library». *Online university's library*: webinar at 29.09.2017. URL: https://www.youtube.com/watch?v=dD8-zc9DKIY. (in Russ.)

3. About Education: Federal Law. No. 273-FZ at 29.12.2012. (in Russ.)

4. On the approval of the performance indicators of the educational organization to self-examination: the order of the Ministry of Education and Science of the Russian Federation No. 1324 at 10.12.2013. (in Russ.)

5. Izmesteva O. V. Approaches to storage of information about textbook provision of the educational process. *Culture: theory and practice: electronic scientific journal*, 2016, iss. 3 (12), p. 15. URL: theoryofculture.ru. (in Russ.)

6. RUSMARC in the examples. Part. 1, One-volume, multi-volume and serial publications. Comp. L. I. Benevolenskaya et al.; eds. O. N. Kulish, B. R. Loginova. Moscow, FAIR PRESS, 2005. (in Russ.)

7. Fomicheva S. G., Popkova A. A. Modern methods of data analysis with the use of client and server OLAP-tools. Norilsk, NSII Press, 2016, 202 p. (in Russ.)

8. Fomicheva S. G., Popkova A. A. The coefficient of textbook provision as a criterion for multidimensional analysis of the activity of the education process. *Scientific and Technical Bulletins of the St. Petersburg State Polytechnic University. Computer science. Telecommunications. Control*, 2009, no. 91, p. 206–220. (in Russ.)

For citation:

Matusevich D. S. Izmesteva O. V. Some Approaches to Storage Information about Textbook Provision of Education Process at Universities. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 126–132. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-126-132

УДК 004.04
DOI 10.25205/1818-7900-2018-16-3-133-144

Р. С. Погодин

Новосибирский государственный университет
ул. Пирогова, 1, Новосибирск, 630090, Россия

ruspog@gmail.com

АДАПТАЦИЯ СТРУКТУРЫ МЕНЮ УСЛУГ ДЛЯ РАЗЛИЧНЫХ ТИПОВ ПОЛЬЗОВАТЕЛЕЙ С ПРИМЕНЕНИЕМ ОНТОЛОГИЧЕСКОГО МОДЕЛИРОВАНИЯ

Статья посвящена решению проблемы адаптации больших древовидных и линейных меню мобильных и интернет-услуг для различных типов пользователей на основании их интересов, социального статуса, а также иных параметров. Разработана программная система, строящая оптимальное меню услуг для классов пользователей, разделенных по социально-экономическим и физическим параметрам, с использованием модифицированного алгоритма построения оптимального графа USSD-меню. В работе используется онтологический подход для формального представления понятий данной предметной области, извлечения, представления и обработки знаний. Для адаптации интерфейсов используются модели пользователей, представляющие описания их потребностей, целей, интересов. Формализация поведения пользователей осуществляется при помощи онтологической модели мобильных и интернет-услуг. Каждого пользователя можно отнести к определенной модели на основании его физических и социальных параметров. Программа, реализующая адаптацию меню, состоит из двух модулей: модуль получения частот вызова услуг на основе запросов к онтологии и модуль оптимизации графа меню. Алгоритм оптимизации меню работает с языком описания графов DOT.

Ключевые слова: онтология, онтологическое моделирование, атомарные диаграммы, извлечение знаний, порождение знаний, иерархические меню, адаптация меню, анализ потребностей, задача оптимизации, пользовательские интерфейсы.

Введение

Статья посвящена адаптации иерархических меню для различных типов пользователей. Решается проблема адаптации больших древовидных и линейных меню мобильных и интернет-услуг на основании интересов пользователей, их социального статуса, а также иных параметров. В работе используется онтологический подход для формального представления понятий данной предметной области, извлечения, представления и обработки знаний. Оптимизация интерфейсов проводится с помощью модифицированного алгоритма построения оптимального графа USSD-меню. Формализация поведения пользователей осуществляется при помощи онтологической модели мобильных и интернет-услуг.

Информационные технологии глубоко проникли в нашу жизнь. С развитием сети Интернет и увеличением ее доступности бизнес начал двигаться в сторону автоматизации своих процессов и предоставления своих услуг в удаленном режиме: клиенту нет необходимости приходить в офис или точку продаж, достаточно открыть приложение на компьютере или телефоне, нажать несколько кнопок и радоваться совершению покупки.

Оплата ЖКХ, замена паспорта, заказ еды, бронирование отелей – всё осуществляется через глобальную паутину путем взаимодействия пользователей с меню услуг. Более того, в последнее время крупные сети кафе и ресторанов начали внедрять электронные меню: че-

Погодин Р. С. Адаптация структуры меню услуг для различных типов пользователей с применением онтологического моделирования // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 133–144.

© Р. С. Погодин, 2018

ловеку не нужно звать официанта, чтобы заказать обед – у него на столе есть планшет с установленным приложением ресторана, в котором все блюда разбиты на категории, а также есть дополнительные опции, вроде регулирования количества сахара и различных добавок. Гостю заведения необходимо просто выбрать то, что ему нравится, и ожидать, когда подадут блюда.

Очевидно, что у разных людей разные интересы. Для одного клиента нужные услуги могут быть на верхнем уровне меню, а для другого – спрятаны где-то глубоко. Соответственно, компания может потерять покупателя, который не смог найти требуемую ему услугу или товар, а покупатель, блуждая в лабиринтах программы, теряет свое время на бесполезный поиск. Аналогично с контекстной рекламой. Например, вегетарианцу нет смысла демонстрировать акцию о скидке на стейки или человеку со средним доходом – турпутевки за 2 миллиона рублей. Таким образом, чтобы избежать перечисленных проблем или свести их к минимуму, существует необходимость адаптации меню для конечного пользователя. Адаптация – это оптимизация с учетом пользовательской модели, которая представляет собой набор параметров, описывающих целевого пользователя: его возраст, бюджет, социальный статус, интересы, географическое расположение и т. п.

Области использования адаптации меню достаточно широки и разнообразны:

- USSD-сервисы;
- интернет-каталоги товаров;
- электронные меню ресторанов;
- меню порталов туристических услуг и услуг бронирования;
- меню образовательных порталов;
- диалоговые меню чат-ботов;
- интерфейсы программ.

В настоящее время уже существуют работы, которые используют онтологический подход к определению потребностей пользователей или других параметров, на основании которых можно определить наиболее релевантные предложения для каждого пользователя. Например, формульное описание диагнозов по симптомам на основании историй болезней [1] или определение предпочтений абонентов мобильных сетей по их характеристикам [2].

В данной работе предложены алгоритм получения частот использования услуг на основании модели пользователя и алгоритм оптимизации иерархических меню по полученным частотам, который является модифицированным алгоритмом оптимизации USSD-меню [3].

Алгоритм получения частот использования услуг основан на онтологическом подходе для формального представления понятий данной предметной области, извлечения, представления и обработки знаний [4–6]. Для адаптации интерфейсов используются модели пользователей, представляющие описания их потребностей, целей, интересов [10]. Формализация поведения пользователей осуществляется при помощи онтологической модели интернет-услуг [8; 9; 11].

Меню услуг

Меню услуг – это направленный ациклический граф, вершины которого представляют элементы меню, а ребра задают пути между этими элементами.

Первый уровень меню – множество вершин графа со степенью захода 0.

Разделы меню – узлы графа.

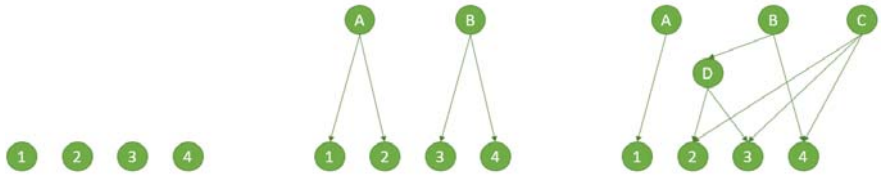
Услуги – множество вершин графа с нулевой степенью исхода.

Элемент меню – элемент множества {раздел меню, услуга}.

Переход по меню – ветвь между двумя элементами меню.

Выделяются три типа иерархических меню.

1. Линейное меню (рис. 1, *а*) – меню, в котором все услуги имеют степень захода, равную нулю. Другое название такого меню – «лента услуг». В чистом виде встречается редко, но в любом иерархическом меню можно выделить по крайней мере одно линейное меню для каждого раздела.



а б в

Рис. 1. Типы иерархических меню

2. Древовидное меню (моноиерархическое меню) (рис. 1, *б*) – меню, в котором любой элемент, не принадлежащий первому уровню, имеет степень захода, равную 1. Является частным случаем мультииерархического меню.

3. Мультииерархическое меню (рис. 1, *в*) – меню, в котором любой элемент, не принадлежащий первому уровню, имеет степень захода, равную натуральному числу. Примером такого меню является любое древовидное меню с разделом «Избранное», в котором услуги из этого раздела имеют степень захода, равную двум.

Этапы адаптации меню

Адаптация меню – это оптимизация меню для конкретного портрета пользователя на основании его параметров. Для того чтобы дать определение оптимизации меню, необходимо определить целевую функцию.

Пусть n_i – количество заказов *i*-й услуги пользователем, k – количество услуг в меню. Тогда обозначим общее количество заказов услуг пользователем: $N = \sum_{i=1}^{k} n_i$.

Пусть d_i – минимальное количество переходов в меню (нажатий клавиш или кликов мышкой) для заказа *i*-й услуги. Тогда обозначим общее количество нажатых пользователем клавиш: $D = \sum_{i=1}^{k} n_i d_i$.

Функция *D* подвергается минимизации. Но в задаче удобней работать с **вероятностью** заказа каждой услуги, а не с **количеством** заказов. Именно поэтому вводим величину $p_i = \frac{n_i}{N}$ – вероятность вызова *i*-й услуги. И вместо функции *D* проводим минимизацию функции $D_p = \sum_{i=1}^{k} p_i d_i$.

Несложно доказать, что минимизация D_p влечет за собой минимизацию *D*. Действительно, $D_p = \frac{D}{N}$, где N = const, откуда следует, что обе функции имеют экстремумы в одних и тех же точках.

Задачей оптимизации меню, предоставляющего *k* услуг, назовем минимизацию функции $D_p = \sum_{i=1}^{k} p_i d_i$., где $p_i = \frac{n_i}{N}$ – вероятность вызова *i*-й услуги, а d_i – минимальное количество нажатий клавиш, необходимое для заказа *i*-й услуги.

Для минимизации сумм $p_i d_i$ нам необходимо определить параметр p_i для каждой услуги. В данной работе применяется теоретико-модельный подход для определения вероятностей: строится онтологическая модель предметной области на основе ключевых понятий этой предметной области, ее законов и постулатов, а также множества прецедентов. Для описания прецедентов – портреты пользователей и статистика заказа услуг.

Таким образом, получаем следующие этапы адаптации меню:

1) получение информации о пользователях и составление их портретов на основе полученных характеристик;

2) построение векторов вероятностей заказа услуг для конкретного портрета пользователя;

3) оптимизация меню с помощью алгоритма оптимизации на основании вероятностей заказа услуг.

Получение информации о пользователях

Самый простой способ адаптировать меню услуг под определенного пользователя – это изучить историю его заказов этих услуг. Соответственно чем больше заказов определенной услуги, тем больший вес она будет иметь в итоге, и тем ближе к началу меню ее расположит алгоритм оптимизации.

Но здесь есть две важные проблемы. Во-первых, у пользователя такой истории может не быть (а тем более достаточно подробной, чтобы для каждой услуги можно было вычислить частоту ее использования). Во-вторых, история теряет актуальность: список услуг и портрет пользователя постоянно меняются.

Поэтому для более релевантной адаптации меню следует найти способ получения более актуальной и полной информации о пользователях и услугах. И на основании этой информации можно будет составить список характеристик пользователя и услуг и построить онтологическую модель для определения вероятностей использования определенных услуг определенными пользователями.

В наше время, наверно, почти не осталось людей, которые не имеют аккаунтов в социальных сетях ВКонтакте, Одноклассники, Facebook, к которым присоединились Twitter и Instagram. С развитием рынка смартфонов и беспроводного Интернета пользователи все больше и больше времени проводят в соцсетях. Вместе с этим соцсети оказываются отличным описанием профиля пользователей. Можно посмотреть личную информацию о человеке (пол, возраст), место жительства, интересы, количество друзей и подписчиков, его сообщества, добавленную музыку, фотографии и т. п. Самая популярная социальная сеть в России и СНГ – ВКонтакте. Именно поэтому она была выбрана в качестве целевой платформы для получения информации о пользователях. ВКонтакте предоставляет публичный программный интерфейс VK API, с помощью которого можно программно получить подробную информацию о пользователях в виде JSON-объектов (рис. 2).

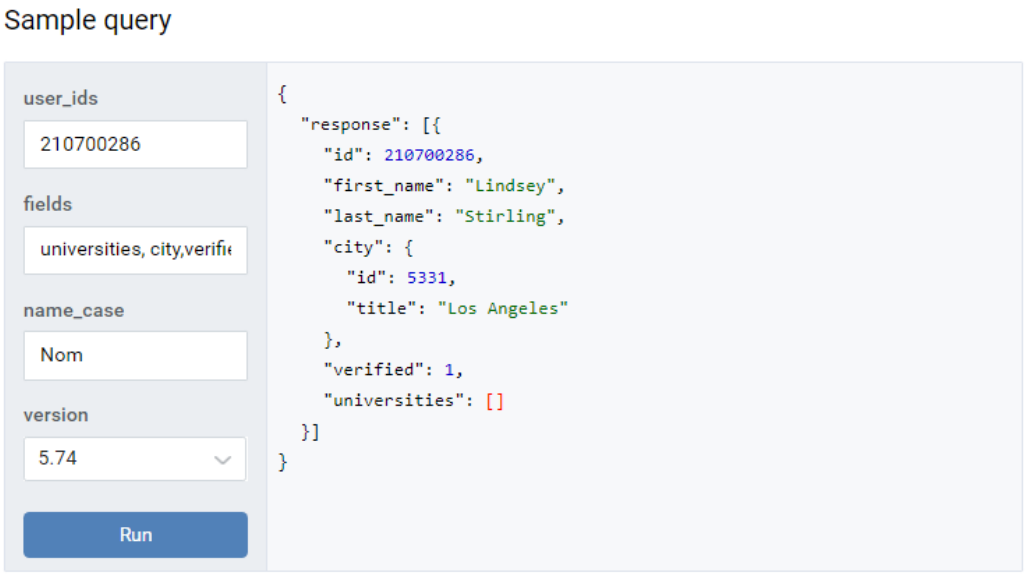


Рис. 2. Типы иерархических меню

В работе с помощью VK API были получены параметры для нескольких тысяч обезличенных пользователей, каждому из которых был приписан уникальный id для последующей идентификации. Среди параметров были: город, образование, место работы, стаж работы (по сумме лет), интересы (по ключевым словам), книги, фильмы, музыка, количество друзей, количество подписчиков, количество фотографий, сообщества.

Есть один важный параметр, который невозможно напрямую получить с помощью социальных сетей, – уровень дохода. Однако его можно приблизительно вывести с помощью информации об образовании, месте жительства, месте работы и стаже работы. Также в работе уровень дохода не исчисляется абсолютными величинами, а выражается фиксированным набором значений, аппроксимирующим определенные интервалы чисел: очень маленький, маленький, ниже среднего, средний.

Таким образом, используя социальные сети, можно построить достаточно полный портрет пользователя и список характеристик, на основании которых можно адаптировать меню под пользователя с определенным портретом.

Онтологическая модель предметной области рынка туристических услуг

В работе целевой предметной областью был выбран рынок туристических услуг. Во-первых, туризм всегда актуален, во-вторых, предложения имеют очень широкий ценовой спектр и спектр интересов, в-третьих, туристический рынок имеет очень богатую классификацию услуг: билеты на самолеты и поезда, аренда автомобилей, отели, визы, гиды, переводчики, круизы... Для построения и структурирования знаний о предметной области рынка туристических услуг была определена онтологическая модель, состоящая из четырех уровней: набор ключевых понятий предметной области, универсальные общие утверждения, эмпирические данные и оценочные знания.

Ключевые понятия предметной области

На первом уровне онтологической модели описываются понятия предметной области, а также даются определения этим понятиям. Для предметной области рынка туристических услуг описываются такие понятия, как рейс, отель, лоукостер, виза, отель, достопримечательность, культура, религия, тур (местный, внутренний, внешний), санаторий, туроператор, бронирование, сезон, попутчик, гражданство, страховка, хостел, город и т. п.

Универсальные общие утверждения

На втором уровне онтологической модели представлены знания о предметной области, т. е. строится формальное описание всех объектов, которые были выделены на первом уровне онтологической модели, а также принципов и закономерностей. Ключевым моментом является то, что все описываемые знания являются достоверными лишь на данный момент, в будущем они могут меняться. Таким образом, онтологическая модель поддерживает актуальность.

Для описания знаний выбран формат обмена данными JSON (рис. 3), так как он хорошо читается и удобно сериализуется. Каждый объект может иметь внутри себя множество свойств трех типов: характеристику, другой объект и массив объектов.

Характеристика – это такой тип свойства, который может принимать значения из конечного множества. Например, характеристика «класс обслуживания» может принимать значение из множества {Первый класс, Бизнес-класс, Премиальный-экономический класс, Экономкласс}.

Центральный объект онтологической модели предметной области рынка туристических услуг – Услуга. Он обладает четырьмя характеристиками: название, тип услуги, стоимость и дата. Характеристика «Тип услуги», как будет описано далее, требуется алгоритму подсчета вероятностей заказов услуг. Она может принимать два значения: простая услуга и контейнер.

```
Номер:
{
        Этаж;
        Количество человек;
        Тип кровати;
        Тип номера (питания);
        Период;
        Отель:
        {
                Город отеля;
                Тип отеля;
        }
}
```

```
Тур:
{
        Услуги
        [
                Страховка,
                Виза,
                Рейс,
                Номер,
                Рейс
        ]
}
```

Рис. 3. Формат описания знаний

Все остальные объекты наследуют эти характеристики и имеют свои уникальные. Особняком стоит объект «Тур», который и сам является услугой, и внутри себя содержит массив услуг, при этом его стоимость равна сумме стоимости входящих в него элементов.

Эмпирические данные

Эмпирические данные – это данные, описывающие прецеденты предметной области, иными словами, это история взаимодействия пользователей с системой, предоставляющей услуги. Формальное описание каждого прецедента задается в виде фрагмента атомарной диаграммы [7].

Атомарная диаграмма модели – это множество атомарных предложений, т. е. предложений φ вида $\varphi = (c_1 = c_2)$, $\varphi = \neg(c_1 = c_2)$, $\varphi = P(c_1,...,c_n)$, $\varphi = \neg P(c_1,...,c_n)$, где $c_1,...,c_n$ – константы, а P – предикат.

В нашем случае каждое атомарное предложение имеет следующий вид: «Человек с характеристиками [Доход: 70000, Город: Новосибирск, ...] заказал услугу с характеристиками [Название: «Рейс в Санкт-Петербург», Авиакомпания: S7, Стоимость: 10000, ...]».

Таким образом мы получаем описание всех профилей пользователей.

Вероятностные и оценочные знания

Вероятностные знания можно получить двумя способами: взять из каких-либо внешних источников или провести анализ эмпирических данных нашей онтологической модели и сопоставить их с универсальными знаниями.

Информация, получаемая в результате анализа эмпирических данных, подразделяется на два уровня: уровень данных и уровень знаний. Уровень данных – это информация, полученная полностью автоматическим способом на основе разбора ассоциативных правил и вычисления частот использования услуг по определенному алгоритму. На уровне знаний информация не может быть получена в полностью автоматическом режиме. Необходимо вмешательство эксперта, который анализирует информацию, полученную на уровне данных, и делает финальные выводы. Однако по информации, полученной на уровне данных, можно получить знания об интересах каждого пользователя, если ввести аксиому о том, что чем выше вероятность заказа услуги в прошлом, тем она выше и в будущем.

Таким образом, построив алгоритм вычисления вероятностей заказа услуг, можно будет проводить адаптацию меню на основе этой информации.

Алгоритм вычисления вероятностей заказа услуг

На втором уровне онтологической модели были описаны все туристические услуги. Каждая услуга имеет набор характеристик. Характеристики потребителей, в свою очередь, были получены путем анализа аккаунтов в соцсетях.

Итак, имеем следующее: пользователя определяют характеристики пользователя, и, в свою очередь, услугу определяют характеристики услуги. Поэтому, определив частоту, с которой каждая характеристика (с названием и значением), фигурирующая у пользователей, и каждая характеристика, принадлежащая услугам, встречались вместе в атомарном предложении, можно для каждого пользователя вычислить вероятность заказа любой услуги.

Из рассуждений, описанных выше, можно построить алгоритм, который выглядит следующим образом.

1. Для всех возможных характеристик потребителей составляется вектор весов каждой характеристики услуг по всем ее значениям на основании анализа эмпирических данных, а именно подсчета связей {Характеристика x_i пользователя – Характеристика y_j услуги} для всех x, y, i, j, где x и y – названия характеристик, а x_i и y_j – значения из множества значений. Пример: для анализа берем характеристику «Город: Новосибирск». Из всех атомарных предложений выбираются те, где пользователи имеют характеристику «Город: Новосибирск», и для каждой характеристики услуг подсчитываются ее вхождения в атомарные выражения. Таким образом, для каждого значения каждой характеристики пользователя получается список весов значений характеристик услуг:

	Компания: S7	Компания: Аэрофлот	...
Город: Новосибирск	40	40	
Город: Томск	35	32	
...			

	Стоимость: 6–10 тыс.	Стоимость: 11–15 тыс.	...
Город: Новосибирск	39	12	
Город: Томск	47	11	
...			

2. Для конкретного потребителя составляется список частот услуг следующим образом.
   A. По очереди выбирается каждая услуга и ее характеристики образуют список.
   B. Для каждой характеристики из списка подсчитывается ее вес по частотам, полученным в п. 1. Вес услуги – это среднее арифметическое весов ее характеристик.
   C. После того как посчитан вес для каждой услуги, эти веса делятся на общую сумму весов всех услуг.
   D. (Дополнительно.) Если тип услуги – контейнер (например, услуга – Тур), то шаг 2 выполняется рекурсивно для всех услуг, входящих в массив, и считается их среднее арифметическое.

В результате получаем вероятность выбора каждой услуги.

Следует отметить, что шаг 1 алгоритма выполняется всего один раз, а при каждом новом заказе услуги просто пересчитываются некоторые значения, которые относятся к характеристикам данной услуги.

Шаг 2 выполняется многократно – для каждого пользователя отдельно. Более того, с определенной периодичностью следует пересчитывать и для одного и того же пользователя, чтобы предложения услуг не теряли свою актуальность.

Таким образом, мы построили онтологическую модель предметной области рынка туристических услуг и с помощью нее для каждого пользователя вычислили вероятность заказа каждой услуги.

Оптимизация иерархических меню

Алгоритм оптимизации иерархических меню получает на вход список пар «название услуги – вероятность использования услуги». Пусть свободный узел графа – это узел графа, который не имеет «родителей». Таким образом, алгоритм оптимизации получает на вход список свободных узлов, которыми являются пары «название услуги – вероятность использования услуги». Пусть m – максимально возможное количество ветвей узла, которое задается пользователем. Если обратиться к исследованиям в области когнитивной психологии, то найдем оптимальное значение – 7 ± 2. Однако в действительности оптимальное значение зависит от задачи. Пусть n_i – количество оставшихся свободных узлов на i-м шаге итерации алгоритма.

Выбирается m свободных узлов с наименьшей вероятностью использования услуги.

Затем создается родитель этих узлов, значение частоты которого является суммой частот его прямых потомков.

Далее его потомкам присваиваются идентификаторы от 1 до m, сами потомки удаляются из списка свободных узлов, поскольку у них появился родитель, а их родитель добавляется в этот список (рис. 4). Имя родителя во время работы алгоритма задается случайным образом (например, по первым буквам имен его потомков) и в алгоритме роли не играет. В любом случае после процесса оптимизации меню для всех новых узлов название определяется человеком.

Далее все шаги повторяются до тех пор, пока количество текущих свободных узлов n_i не станет равным максимальному количеству ветвей на узле m. Эти свободные узлы и образуют корневые порталы меню.

В итоге получается оптимальная структура меню, а именно такая структура, где чем выше вероятность использования определенной услуги, тем эта услуга будет находиться ближе к корню меню. Каждая услуга в полученном меню строго определяется последовательностью переходов в графе, т. е. последовательностью кликов мышки или нажатий клавиш, которые должен сделать пользователь, чтобы вызвать ее.

В чистом виде приведенный алгоритм может в определенных ситуациях не строить оптимальное меню. Проблема заключается в том, что на последнем шаге алгоритма иногда останется менее, чем m свободных узлов. Это происходит в тех случаях, когда количество услуг k в меню будет иметь такое значение, что $k - s(m-1) < m$, где s – число шагов, за которое алгоритм построит меню. В этом случае на первом уровне меню количество ветвей будет меньше максимального (рис. 5), когда ожидаемый результат работы алгоритма – недостаток ветвей – возможен только на последнем уровне.

Для того чтобы решить эту проблему, перед первым шагом алгоритма следует оценить число свободных узлов, которое останется на последнем шаге, и, если оно окажется меньше m, на первом шаге за максимальное количество элементов на узле вместо m принять m_1, получившееся в результате оценки (рис. 6).

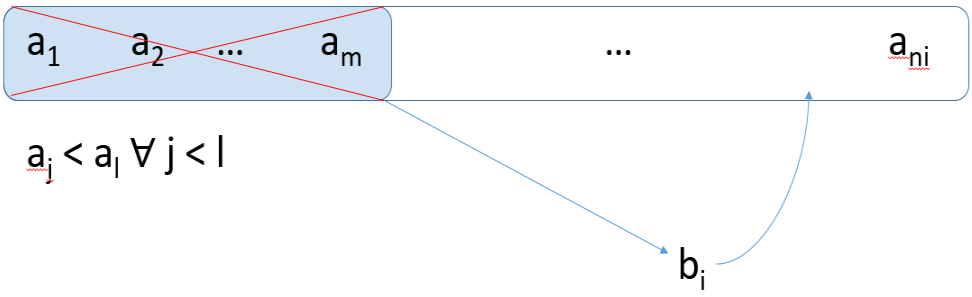


Рис. 4. Иллюстрация работы алгоритма построения оптимального графа

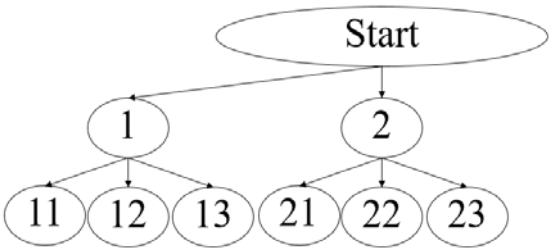


Рис. 5. Проблема недостатка ветви на первом уровне меню

```
m1 = k;
while(m1 - m > 0)
{
    m1 = m1 - m + 1;
}
```

Рис. 6. Листинг. Оценка количества вершин на последнем шаге

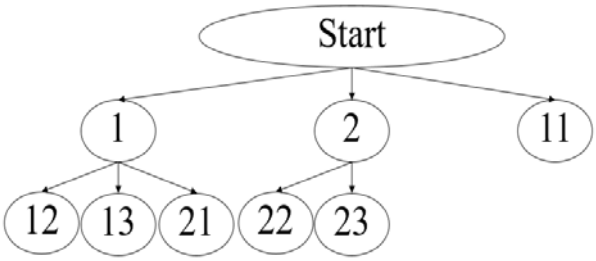


Рис. 7. Демонстрация результата оптимизации меню
с использованием предварительной оценки (для наглядности перемещения
пункты пронумерованы так же, как на рис. 5)

Таким образом, количество ветвей ниже максимального окажется именно на последнем уровне меню, что вполне допустимо, и, более того, является достаточно частой ситуацией, так как почти всегда количество услуг имеет кратность, отличную от количества подменю (рис. 7).

Тестирование и результаты

Для тестирования адаптации меню с помощью Json-генератора были созданы 913 услуг, а также построена история о 30 000 заказах услуг. Эти услуги были разбиты на следующие категории: билеты на самолет, отели, круизы, визы, сувениры, гиды, переводчики, аренда автомобилей. Каждая категория была разбита, в свою очередь, еще на несколько подкатегорий.

Тестирование проводилось на полной выборке из 30 000 заказов услуг, а также на выборках случайных 10 000 и 20 000 заказов из полной выборки. Программа тестировалась в двух режимах: режим пагинации и режим дерева. В первом режиме категории меню разбивались на страницы, а во втором режиме строилось мультииерархическое меню.

	Случайное меню	Режим пагинации	Режим дерева
10000	61198	43762	42077
20000	129614	88162	84412
30000	187309	130845	122903

Рис. 8. Результаты тестирования

Сокращение количества нажатий клавиш пользователем для выбора требуемой услуги составило от 29 до 35 % (рис. 8).

Заключение

В настоящей работе решается задача адаптации структуры меню услуг для различных типов пользователей. Разработан алгоритм составления векторов вероятностей услуг на основе построения онтологической модели предметной области.

Построена трехуровневая модель предметной области рынка туристических услуг. Характеристики пользователей были получены с помощью анализа профилей социальной сети ВКонтакте программным модулем, использующим публичный программный интерфейс VK API. Разработаны два алгоритма оптимизации меню в зависимости от его структуры: режим построения пагинации для линейного меню и режим построения дерева для древовидных меню. Если для алгоритма используется уже ранее построенное меню, в котором нельзя удалять переходы, то алгоритм поддерживает построение мультииерархического меню.

Онтологическая модель реализована программно на языке OWL в редакторе Protege с маппингом в базу данных MongoDB для более удобных операций над данными на уровне кода. В качестве языка программирования выбран C#, а в качестве платформы для приложения – ASP.NET MVC. Для работы с графами на языке C# использовалась библиотека QuickGraph, поддерживающая язык описаний DOT.

Тестирование адаптации было проведено на меню услуг, состоящим из 913 элементов, в двух режимах: пагинации и построения дерева. Среднее сокращение количества времени выбора услуги пользователем составило 29–35 %.

В дальнейшем планируется модифицировать алгоритм построения частот использования услуг.

Список литературы

1. *Пальчунов Д. Е, Яхъяева Г. Э., Ясинская О. В.* Применение теоретико-модельных методов и онтологического моделирования для автоматизации диагностирования заболеваний // Вестн. НГУ. Серия: Информационные технологии. 2015. Т. 13, № 3. С. 42–51.

2. *Долгушева Е. В., Пальчунов Д. Е.* Теоретико-модельные методы порождения знаний о предпочтениях абонентов мобильных сетей // Вестн. НГУ. Серия: Информационные технологии. 2016. Т. 14, № 2. С. 5–16.

3. *Погодин Р. С.* Адаптация структуры USSD-меню для различных типов пользователей // МНСК-2016: Информационные технологии: Материалы 54-й Междунар. науч. студ. конф. / Новосиб. гос. ун-т. Новосибирск, 2016. С. 229.

4. *Пальчунов Д. Е.* Моделирование мышления и формализация рефлексии I: Теоретико-модельная формализация онтологии и рефлексии // Философия науки. 2006. № 4 (31). С. 86–114.

5. *Пальчунов Д. Е.* Решение задачи поиска информации на основе онтологий // Бизнес-информатика. 2008. № 1. С. 3–13.

6. *Пальчунов Д. Е., Степанов П. А.* Применение теоретико-модельных методов извлечения онтологических знаний в предметной области информационной безопасности // Программная инженерия. 2013. № 11. С. 8–16.

7. *Махасоева О. Г., Пальчунов Д. Е.* Автоматизированные методы построения атомарной диаграммы модели по тексту естественного языка // Вестн. НГУ. Серия: Информационные технологии. 2014. Т. 12, № 2. С. 64–73.

8. *Palchunov D., Yakhyaeva G., Dolgusheva E.* Conceptual Methods for Identifying Needs of Mobile Network Subscribers // Proc. of the Thirteenth International Conference on Concept Lattices and Their Applications. Moscow, Russia, 2016. P. 147–160.

9. *Fensel D.* Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag, 2001. 106 p.

10. *Cadenas A., Ruiz C., Larizgoitia I., García-Castro R., Lamsfus C., Vázquez I., González M., Martín D., Poveda M.* Context Management in Mobile Environments: a Semantic Approach // Workshop on Context, Information and Ontologies (CIAO 2009). Heraklion, Greece, 2009. 8 p.

11. *Mikhailyuk A.* OWL as a Standard Model for Transdisciplinary Knowledje Representation in Semantic Web // Information Content and Processing. 2014. Vol. 1. No. 3. P. 249–261.

Материал поступил в редколлегию 31.05.2018

R. S. Pogodin

Novosibirsk State University
1 Pirogov Str., Novosibirsk, 630090, Russian Federation

ruspog@gmail.com

ADAPTATION OF SERVICE MENU STRUCTURE FOR DIFFERENT TYPES OF USERS USING THE ONTOLOGY MODELING

The purpose of the article is to solve the problem of adaptation of large arborescent and linear menus of mobile and internet services for various types of users according to their interests, social status and other parameters. There has been created the program system building an optimal menu of services for classes of users divided by physical and socio-economic parameters. When adapting the menu, a modified algorithm of construction of an optimal graph of the USSD-menu was used. An ontological approach was used during the work for formal representation of the notions of a given subject area, extraction, representation and processing of knowledge. Users' models representing descriptions of their needs, aims and interests were used for adaptation of interfaces. Formalization of users' conduct was realized with the help of an ontological model of mobile and internet services. Each user can be ascribed to a definite model based on their physical and social parameters. The program realizing adaptation of the menus contains two modules: the module of frequency of calls reception based on ontology queries and the graph of the menu optimization module. An algorithm of the optimization of the menu works with DOT graph prescription language.

Keywords: ontology, ontological modeling, atomic diagrams, knowledge acquisition, knowledge generation, hierarchical menu, menu adaptation, needs analysis, optimum problem, user interfaces.

References

1. Palchunov D., Yakhyaeva G., Yasinskaya O. The using of model-theoretic methods and ontological modeling for the automation of diagnosing diseases. *Vestnik NSU. Series*: *Information Technologies*, 2015, no. 13, p. 42–51. (in Russ.)

2. Palchunov D., Dolgusheva E. Model-theoretic methods of generating knowledge about the preferences of subscribers of mobile networks. *Vestnik NSU. Series*: *Information Technologies*, 2015, no. 14, p. 5–16. (in Russ.)

3. Pogodin R. Adaptation of USSD-menu structure for different types of users. *Proc. of the 54th International Conference MNSK-2016: Information Technologies*, 2016, p. 229. (in Russ.)

4. Palchunov D. Modeling of Thinking and Formalization of Reflection I: Theory-Model Formalization of Ontology and Reflection. *Filosofia nauki*, 2006, no. 4 (31), p. 86–114. (in Russ.)

5. Palchunov D. Solving the problem of information retrieval based on ontologies. *Biznes-Informatika*, 2008, no. 1, p. 3–13. (in Russ.)

6. Palchunov D., Stepanov P. The using of model-theoretical methods of ontological knowledge extraction in the information security domain. *Programmnaya inzheneria*, 2013, no. 11, p. 8–16. (in Russ.)

7. Makhasoeva O., Palchunov D. Automated methods for constructing an atomic diagram of a model from the text of a natural language. *Vestnik NSU. Series*: *Information Technologies*, 2014, no. 12, p. 64–73. (in Russ.)

8. Palchunov D., Yakhyaeva G., Dolgusheva E. Conceptual Methods for Identifying Needs of Mobile Network Subscribers. *Proc. of the Thirteenth International Conference on Concept Lattices and Their Applications*. Moscow, Russia, 2016, p. 147–160.

9. Fensel D. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag, 2001, 106 p.

10. Cadenas A., Ruiz C., Larizgoitia I., García-Castro R., Lamsfus C., Vázquez I., González M., Martín D., Poveda M. Context Management in Mobile Environments: a Semantic Approach. *Workshop on Context, Information and Ontologies (CIAO 2009)*. Heraklion, Greece, 2009, 8 p.

11. Mikhailyuk A. OWL as a Standard Model for Transdisciplinary Knowledje Representation in Semantic Web. *Information Content and Processing*, 2014, vol. 1, no. 3, p. 249–261.

For citation:

Pogodin R. S. Adaptation of Service Menu Structure for Different Types of Users Using the Ontology Modeling. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 133–144. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-133-144

УДК 519.246.8
DOI 10.25205/1818-7900-2018-16-3-145-158

К. С. Чирихин [1, 2], **Б. Я. Рябко** [1, 3]

[1] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

[2] *Сибирский государственный университет телекоммуникаций и информатики*
ул. Кирова, 86, Новосибирск, 630102, Россия

[3] *Институт вычислительных технологий СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

chirihin@gmail.com

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ ТОЧНОСТИ МЕТОДОВ ПРОГНОЗА, БАЗИРУЮЩИХСЯ НА АРХИВАТОРАХ

В теории информации известно, что методы сжатия данных могут быть использованы для прогнозирования стационарных процессов. В данной работе предложен базирующийся на архиваторах алгоритм прогнозирования временных рядов и проведено экспериментальное исследование его эффективности. В процессе работы описанного алгоритма могут быть использованы произвольные методы сжатия данных, причем прогнозные значения от разных методов комбинируются, и наибольшее влияние на конечный результат оказывает метод, способный сильнее других сжать временной ряд. Данный алгоритм может быть использован для прогнозирования рядов с дискретными и непрерывными алфавитами. Для повышения точности прогноза возможно применение существующих методов предварительной обработки данных. Экспериментальное исследование эффективности предложенного алгоритма проводилось на временных рядах из M3 Competition и ряде Т-индекса, при этом были использованы хорошо известные архиваторы. Результаты вычислений показали, что полученный метод обладает сравнительно высокой точностью и быстродействием.

Ключевые слова: универсальное кодирование, прогнозирование временных рядов.

Введение

Задача прогнозирования привлекает внимание многих исследователей в силу ее большой практической значимости. Например, существует большое количество приложений в экономике (можно построить прогноз для уровня безработицы, объемов промышленного производства и т. д.), геофизике (прогнозирование числа солнечных пятен, уровня моря) и во многих других областях человеческой деятельности. Математически временной ряд может быть описан как случайный процесс с дискретным временем [1], значения которого, как правило, находятся на равном расстоянии друг от друга. При прогнозировании временного ряда требуется оценить значения процесса в нескольких будущих моментах времени на основании имеющейся в наличии его предыстории.

Для решения описанной задачи разработано большое количество разнообразных методов – как статистических, так и из области машинного обучения. К наиболее распространенным можно отнести экспоненциальное сглаживание, модель авторегрессии – скользящего

Чирихин К. С., *Рябко Б. Я.* Экспериментальное исследование точности методов прогноза, базирующихся на архиваторах // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 3. С. 145–158.

© К. С. Чирихин, Б. Я. Рябко, 2018

среднего и различные ее модификации, а также экспертные системы и нейронные сети [2; 3]. Тем не менее задача построения высокоточного прогноза еще далека от разрешения.

Поскольку между сжимаемостью последовательности и ее вероятностью существует тесная связь, одним из возможных подходов к решению данной задачи является использование методов сжатия данных. В статье [4] было показано, что любой архиватор может быть использован для прогнозирования. Важно отметить, что в архиваторах, помимо теоретических результатов, используются различные эвристики, повышающие их способность улавливать закономерности во встречающихся на практике данных и улучшающие степень сжатия.

В данной статье разрабатывается и исследуется метод прогнозирования временных рядов, основанный на распространенных архиваторах и библиотеках для сжатия данных (таких, как библиотека zlib [1], лежащая в основе архиватора gzip, библиотека ppmd [5], используемая среди прочих алгоритмов в 7-z, и др.). Для экспериментальной оценки эффективности метода были проведены расчеты для временных рядов из известного исследования M3-Competition [3] (далее M3C), основной целью которого было сравнение точности различных методов прогнозирования на реальных данных преимущественно из социально-экономической сферы. Проведено сравнение предлагаемого метода с методами, участвовавшими в этом исследовании. Еще одним рядом, результаты вычислений для которого приводятся в данной статье, является временной ряд T-индекса [2], тесно связанный с солнечными пятнами.

Результаты экспериментального исследования показывают, что описываемый метод обладает точностью, сравнимой с другими методами прогнозирования, и представляет практический интерес.

Прогнозирование с помощью методов сжатия данных для случая конечного алфавита

Сначала покажем, как связаны сжатие данных и прогнозирование. Пусть $X = x_1, x_2, ..., x_t$ – временной ряд (или в терминах теории информации – передаваемое сообщение), порожденный некоторым вероятностным источником. Все члены временного ряда x_i принадлежат конечному множеству A, называемому *алфавитом*. В теории информации хорошо известно, что для любого разделимого кода выполняется неравенство Крафта – Макмиллана [6]:

$$\sum_{X \in A^t} 2^{-|\phi(X)|} \le 1, \tag{1}$$

где A^t – множество всевозможных последовательностей длиной t над алфавитом A, $|\phi(X)|$ – длина кодового слова для сообщения X при кодировании по методу ϕ.

Величину $2^{-|\phi(X)|}$ далее для краткости будем называть *кодовой вероятностью* сообщения X.

В работе [4] было предложено использовать неравенство (1) для задания распределения вероятностей на множестве кодируемых сообщений:

$$P_\phi(X) = \frac{2^{-|\phi(X)|}}{\sum_{Y \in A^t} 2^{-|\phi(X)|}}.$$

Условная вероятность того, что следующие h символов $x_{t+1}, x_{t+2}, ..., x_{t+h}$ будут равны соответственно $a_1, a_2, ..., a_h$, $a_i \in A$, может быть найдена с использованием имеющейся предыстории по формуле

$$P_\phi\left(x_{t+1} = a_1, x_{t+2} = a_2, ..., x_{t+h} = a_h \middle| x_1, x_2, ..., x_t\right) =$$

[1] Zlib Home Site URL: https://zlib.net (дата обращения 23.03.2018).

[2] T Index FAQ. Australian Government/Bureau of Meteorology. URL: http://www.sws.bom.gov.au/Educational/5/2/1 (дата обращения 23.03.2018).

$$= \frac{2^{-|\phi(x_1,x_2,...,x_t,a_1,a_2,...,a_h)|}}{\sum_{Y\in A^h} 2^{-|\phi(x_1,x_2,...,x_t,y_1,y_2,...,y_h)|}}. \qquad (2)$$

Пример 1. Приведем пример вычислений по формуле (2). Построим прогноз на два шага вперед для последовательности 0110011001 над алфавитом $\{0,1\}$. Будем поочередно дописывать в конец данной последовательности различные комбинации длиной 2 из нулей и единиц и сжимать ее каким-либо архиватором. Например, воспользуемся библиотекой zlib версии 1.2.11, в частности, ее функцией compress2 с флагом Z_BEST_COMPRESSION. Получившиеся длины кодовых слов и дальнейшие вычисления приведены в табл. 1. Жирным шрифтом в столбце 1 выделены два «дописанных» символа. Последовательности в программе на языке C++ были представлены как массивы типа unsigned char, на каждый символ последовательностей отводился 1 байт (единица была представлена байтом со значением 1_{10}, а нуль – со значением 0_{10}).

Таблица 1

Пример построения распределения вероятностей для следующих двух элементов временного ряда

Последовательность	Размер сжатого представления, бит	Кодовая вероятность	Вероятность
011001100**100**	128	2.939E-39	1.520E-5
011001100**101**	128	2.939E-39	1.520E-5
011001100**110**	112	1.926E-34	9.961E-1
011001100**111**	120	7.523E-37	3.891E-3
Сумма	–	1.933E-34	1.000

Как видно из табл. 1, согласно данному методу, следующими двумя символами будут 10 с вероятностью 0.996. Таким образом, архиватор смог успешно выявить закономерность даже на короткой последовательности.

Более подробное изложение метода, описанного в данном и следующем разделах, а также обоснование приведенных в них формул можно найти в книге [7].

Прогнозирование временных рядов с непрерывным алфавитом

Описанный в предыдущем разделе метод применим только для временных рядов с конечным алфавитом, но на практике чаще всего встречаются ряды, у которых алфавитом является некоторый отрезок вещественной прямой. В таких случаях необходимо перейти от подобного отрезка к конечному алфавиту. Это можно сделать путем разбиения отрезка на k непересекающихся равных интервалов с номерами $A = \{0, 1, ..., k-1\}$ (обозначим интервал с номером i через q_i), последующего преобразования временного ряда в ряд номеров интервалов и прогнозирования номеров для следующих значений. Предположим, что требуется построить прогноз на h шагов вперед. Тогда вероятность того, что в момент времени i, $1 \le i \le h$, значение процесса попадет в интервал с номером $j \in A$, может быть получена из маргинального распределения вероятностей по совместному распределению номеров, вычисленному по формуле (2):

$$P_\phi\left(x_{t+i} \in q_j\right) = \sum_{a_1,...,a_{i-1},a_{i+1},...,a_h \in A} P_\phi\left(a_1,...,a_{i-1},j,a_{i+1},...,a_h\right). \qquad (3)$$

В качестве точечного прогноза можно использовать математическое ожидание случайной величины, значениями которой являются середины интервалов, и их вероятности задаются по формуле (3).

Рассмотрим далее вопрос о выборе количества интервалов *k*. Проблема заключается в том, что при малых *k* точность прогноза может оказаться низкой из-за грубого округления, а при больших – из-за слишком короткой предыстории процесса и присутствия шумов в данных. Кроме того, с увеличением *k* экспоненциально возрастает время вычислений. В данной работе был использован подход, описанный в статье [8]. Отрезок, из которого принимают значения члены временного ряда, разбивается последовательно на 2^i интервалов, $i = 1, 2, ..., n$, $k = 2^n$. Для каждого *i* прогноз строится независимо, и затем прогнозы взвешиваются. Пусть x_i – член исходного временного ряда, а $x_i^{[j]}$ – соответствующий ему номер интервала при разбиении исходного отрезка на 2^j интервалов. Взвешивание можно провести по следующей формуле:

$$P_\phi\left(x_1^{[n]}, x_2^{[n]}, ..., x_t^{[n]}\right) = \frac{\sum_{i=1}^{n} \omega_i 2^{-\left|\phi\left(x_1^{[n]}, x_2^{[n]}, ..., x_t^{[n]}\right)\right| + t(n-i)}}{\sum_{i=1}^{n} \sum_{Y \in N_i^t} \omega_i 2^{-|\phi(Y)| + t(n-i)}}, \quad (4)$$

где

$N_i = \left\{0, 1, ..., 2^i - 1\right\}$ – алфавит, состоящий из номеров интервалов;

$\omega_i \geq 0$, $\sum_{i=1}^{n} \omega_i = 1$ – весовые коэффициенты, в данной работе вычислялись по формуле

$$\omega_i = \begin{cases} \frac{1}{i} - \frac{1}{i+1}, i \neq n, \\ \frac{1}{n+1}, i = n. \end{cases}$$

Поясним, зачем прибавлять величину $t(n-1)$ к длине кодового слова в формуле (4). Без нее нельзя сравнивать длины кодовых слов для сообщений из разных алфавитов. Предположим, что область значений временного ряда поочередно разбивалась на 2 и 4 интервала (разбиения 1 и 2 соответственно). В данном случае $n = \log_2 4 = 2$. Заметим, что каждому интервалу разбиения 1 соответствует 2 интервала разбиения 2 (нулю соответствуют нуль и единица, единице – два и три). Для того чтобы сообщить, в какой из двух возможных интервалов разбиения 2 попадает элемент из ряда с разбиением 1, требуется 1 бит. Поскольку всего элементов *t*, требуется добавить $t(2-1) = t$ бит к ряду с разбиением 1.

Для того чтобы избежать больших отрицательных степеней при расчетах по формуле (4), после того, как длины всех кодовых слов будут известны, можно вычесть наименьшую из них из каждого кодового слова. Данная идея будет проиллюстрирована в примере 2.

Далее покажем, как можно взвесить прогнозы, полученные по различным архиваторам. Пусть имеется *n* методов сжатия данных (архиваторов) $\phi_1, \phi_2, ..., \phi_n$. Можно скомбинировать прогнозы, полученные от каждого архиватора в отдельности, в общий прогноз по следующей формуле:

$$P_{\phi_1, \phi_2, ..., \phi_n}\left(X \middle| Y\right) = \frac{\sum_{i=1}^{n} \gamma_i 2^{-|\phi_i(YX)|}}{\sum_{Z \in N^h} \sum_{i=1}^{n} \gamma_i 2^{-|\phi_i(YX)|}}, \quad (5)$$

где

$\gamma_i \geq 0$, $\sum_{i=1}^{n} \gamma_i = 1$ – весовые коэффициенты;

$Y = y_1, y_2, ..., y_t$ – данный временной ряд;

$X = x_1, x_2, ..., x_h$ – одно из возможных продолжений ряда;

YX – ряд, полученный путем дописывания ряда X в конец ряда Y.

В данной работе при вычислениях по формуле (5) были использованы равные веса: $\gamma_i = \frac{1}{n}$.

Отметим, что формулы (4) и (5) можно использовать совместно. В вычислениях, приводимых в данной статье, сначала оценки вероятностей, полученные при различных мощностях разбиений, взвешивались по формуле (4), и тем самым получалось единственное распределение интервалов (содержащее номера интервалов из наибольшего рассматриваемого разбиения) для каждого используемого архиватора. Затем распределения от разных архиваторов объединялись в одно по формуле (5).

Пример 2. Рассмотрим, как построить прогноз для временного ряда с непрерывным алфавитом при помощи двух архиваторов. Пусть дан следующий временной ряд:

3.4	0.1	3.9	4.8	1.5	1.8	2.0	4.9	5.1	2.1

Требуется построить по нему прогноз на два шага вперед. Для этого воспользуемся библиотекой zlib, реализацией алгоритма сжатия Re-Pair [9] и формулой (5). Сначала перейдем к конечному алфавиту. Разобьем интервал, в который попадают все значения временного ряда, на 2 и 4 части одинаковой длины, построим прогнозы независимо для этих разбиений, а затем скомбинируем их по формуле (5).

Наименьшее значение в рассматриваемом временном ряде $m = 0.1$, наибольшее $M = 5.1$. Хотя все значения ряда попадают в отрезок $[0.1; 5.1]$, отступим на некоторую величину от крайних значений (поскольку следующее значение может быть больше (меньше) предыдущего максимума (минимума)). При расчетах, результаты которых приводятся в данной статье далее, отступ составлял 10 % от ширины интервала, поэтому в данном примере поступим также. Получим $m' = m - 0.1(M - m) = -0.4$, $M' = M + 0.1(M - m) = 5.6$. При разбиении на два интервала одинаковой длины в случае, если значение временного ряда меньше чем 2.6, будем считать, что оно попало в интервал с номером 0, иначе в интервал с номером 1. Получим следующий временной ряд:

3.4	0.1	3.9	4.8	1.5	1.8	2.0	4.9	5.1	2.1
1	0	1	1	0	0	0	1	1	0

Будем последовательно дописывать к ряду всевозможные комбинации из символов $\{0,1\}$ и сжимать получающиеся сообщения. Размеры сжатых файлов и соответствующие вычисления вероятностей приведены в табл. 2. При «склеивании» кодовых вероятностей, полученных от zlib и Re-Pair, были использованы веса $\gamma_1 = \gamma_2 = 0.5$.

Проведем аналогичные вычисления для разбиения области значений ряда на 4 интервала. Получим следующий временной ряд:

3.4	0.1	3.9	4.8	1.5	1.8	2.0	4.9	5.1	2.1
2	0	2	3	1	1	1	3	3	1

Интервалу с номером 0 при разбиении на 2 интервала соответствуют интервалы 0, а при разбиении на 4 интервала – 1, соответственно интервалу с номером 1 – интервалы 2 и 3. Поэтому, например, прогнозному значению 13 при разбиении на 4 интервала соответствует прогнозное значение 01 при разбиении на два интервала.

К значениям в пятом и шестом столбцах табл. 2 было прибавлено 12 бит, поскольку длина t каждой из последовательностей равна 12, и $\log_2 4 - \log_2 2 = 1$. В данной таблице наименьшая длина сжатого сообщения составила 100 бит, для уменьшения риска переполнения при вычислениях с плавающей точкой эту величину можно вычесть из всех значений в столбцах 2, 3, 5, 6 табл. 2.

Таблица 2

Результаты вычислений при разбиении на два и четыре интервала

Сообщение (4 интервала)	Размер сжатого сообщения, бит		Сообщение (2 интервала)	Размер сжатого сообщения, бит	
	zlib	Re-Pair		zlib	Re-Pair
202311331**00**	160	112	101100011**00**	120 + 12	104 + 12
202311331**01**	160	112			
202311331**10**	144	120			
202311331**11**	136	104			
202311331**02**	160	120	101100011**01**	128 + 12	88 + 12
202311331**03**	160	112			
202311331**12**	144	120			
202311331**13**	144	120			
202311331**20**	160	120	101100011**10**	136 + 12	88 + 12
202311331**21**	160	112			
202311331**30**	160	112			
202311331**31**	160	112			
202311331**22**	160	112	101100011**11**	136 + 12	88 + 12
202311331**23**	160	120			
202311331**32**	160	112			
202311331**33**	144	112			

Далее нужно перейти к кодовым вероятностям, а затем взвесить и нормировать их. Окончание расчетов приведено в табл. 3.

Теперь построим прогноз на первый и второй шаги. Просуммировав вероятности в строках таблицы, в которых первый дописанный символ равен 0, получим вероятность того, что следующее значение временного ряда попадет в интервал с номером 0. Аналогичным образом вычисляются вероятности для интервалов с номерами 1, 2 и 3. Распределения вероятностей номеров интервалов, а также середины соответствующих интервалов, приведены в табл. 4. В качестве прогнозных значений используются математические ожидания, значения которых также приведены в таблице.

Уменьшение трудоемкости алгоритма

Обозначим через $|A|$ мощность алфавита источника (максимальная мощность разбиения в случае непрерывного алфавита). Если требуется построить прогноз на h шагов, то потребуется сжать $|A|^h$ последовательностей. В то же время в МЗС для ежемесячных данных требовалось вычислять прогноз на 18 шагов вперед, что приводит к необходимости сжатия $2^{18} = 262144$ последовательностей при минимально возможной мощности алфавита $|A| = 2$. Сократить время вычислений позволяет следующий подход. Предположим, что h является четным. Удалим из ряда все члены с четными номерами и построим прогноз на $h/2$ шагов вперед с нечетными номерами. Затем удалим из исходного ряда все элементы с нечетными номерами и построим прогноз для недостающих $h/2$ элементов с четными номерами. В результате вместо $|A|^h$ возможных продолжений ряда получим $2|A|^{h/2}$ вариантов. Для повышения точности можно построить прогноз на первые $h/2$ шагов по полному ряду. Данный метод легко можно обобщить. К примеру, если h кратно 6, то на первом этапе нужно оставить только члены ряда с номерами i, для которых $i \bmod 6 = 0,$ и вычислить прогноз на $h/6$ шагов вперед и т. д.

Таблица 3

Окончание расчетов из примера 2

Сообщение (4 интервала)	Кодовая вероятность (4 интервала)		Сообщение (2 интервала)	Кодовая вероятность (2 интервала)		Кодовая вероятность после «склеивания»	Вероятность после «склеивания»
	zlib	Re-Pair		zlib	Re-Pair		
202311133100	2.939E-39	2.441E-04	10110001100	2.328E-10	1.562E-05	6.485E-05	2.150E-05
202311133101	2.939E-39	2.441E-04				6.485E-05	2.150E-05
202311133110	1.926E-34	9.537E-07				4.053E-06	1.344E-06
202311133111	4.930E-32	0.063				0.016	0.005
202311133102	2.939E-39	9.537E-07	10110001101	9.095E-13	1.000	0.250	0.083
202311133103	2.939E-39	2.441E-04				0.250	0.083
202311133112	1.926E-34	9.537E-07				0.250	0.083
202311133113	1.926E-34	9.537E-07				0.250	0.083
202311133120	2.939E-39	9.537E-07	10110001110	3.553E-15	1.000	0.250	0.083
202311133121	2.939E-39	2.441E-04				0.250	0.083
202311133130	2.939E-39	2.441E-04				0.250	0.083
202311133131	2.939E-39	2.441E-04				0.250	0.083
202311133122	2.939E-39	2.441E-04	10110001111	3.553E-15	1.000	0.250	0.083
202311133123	2.939E-39	9.537E-07				0.250	0.083
202311133132	2.939E-39	2.441E-04				0.250	0.083
202311133133	1.926E-34	2.441E-04				0.250	0.083
Сумма	–		–	–		3.016	1.000

Таблица 4

Вычисление прогнозных значений на два шага вперед для ряда из примера 2

Номер интервала	Маргинальная вероятность		Середина интервала
	шаг 1	шаг 2	
0	0.166	0.166	0.35
1	0.171	0.171	1.85
2	0.332	0.332	3.35
3	0.332	0.332	4.85
Математическое ожидание	3.093	3.093	–

Используемые архиваторы

При проведении экспериментальных расчетов, результаты которых приведены в данной статье, были использованы библиотеки zlib, ppmd [3] и реализация алгоритма Re-Pair [9]. В их основе лежат разные идеи. В zlib реализована схема Deflate, в которой используется алгоритм LZ-77 совместно с кодом Хаффмана. Библиотека ppmd является реализацией алгоритма PPM (Prediction by Partial Matching, предсказание по частичному совпадению). Re-Pair – алгоритм сжатия данных, основанный на грамматической модели. В процессе работы этого алгоритма выполняется построение контекстно-свободной грамматики, задающей язык из единственной цепочки – сжимаемого файла. При этом сжатие достигается за счет замены повторяющихся фрагментов во входной последовательности на нетерминальные символы.

Экспериментальное исследование

В рамках данной работы была выполнена программная реализация описанного метода и проведены вычисления для набора временных рядов из исследования M3-Competition [3], а также временного ряда Т-индекса. Соревнование M3-Competition проводилось в 2000 г., и основной его целью являлось сравнение точности методов прогнозирования на реальных данных. В нем были представлены 3 003 временных ряда, длина которых составляла от 14 до 144 наблюдений. Присутствовали ежегодные, ежемесячные и ежеквартальные данные, а также небольшое количество рядов, не относящихся ни к одной из вышеперечисленных категорий (категория «другие»). Требовалось построить прогноз на 6 шагов для ежегодных данных, на 8 шагов для ежеквартальных данных и данных из категории «другие», на 18 шагов для ежемесячных данных. На веб-сайте Международного института прогнозистов [4] размещены временные ряды из M3C и реально зафиксированные значения прогнозируемых величин, поэтому имеется возможность провести аналогичные вычисления и оценить качество построенного прогноза.

В данной статье используется один из способов оценки точности в M3C – симметричная средняя абсолютная процентная ошибка (symmetric mean absolute percentage error, sMAPE), определяемая по следующей формуле:

$$\text{sMAPE}\left(\hat{X}, X\right) = \sum_{i=1}^{h} \frac{\left|\hat{x}_i - x_i\right|}{\left(\hat{x}_i + x_i\right)/2} \cdot 100, \tag{6}$$

где

$\hat{X} = \hat{x}_1, \hat{x}_2, ..., \hat{x}_h$ – ряд прогнозных значений;

$X = x_1, x_2, ..., x_h$ – ряд зафиксированных значений.

Поскольку ни один временной ряд из M3C не содержит отрицательных значений, величина, вычисляемая по формуле (6), не может быть отрицательной.

При вычислениях по описываемому в данной статье алгоритму использовалась предварительная обработка данных. В частности, для ежемесячных и ежеквартальных данных использовалось выделение сезонной компоненты временного ряда (была использована реализация [5] метода STL [10]). Затем выполнялось построение прогноза для тренда и случайной составляющей, а сезонная компонента принималась постоянной. Кроме того, для всех категорий временных рядов осуществлялся переход к первой разности (первой разностью ряда $x_1, x_2, ..., x_t$ называется ряд $x_2 - x_1, x_3 - x_2, ..., x_t - x_{t-1}$) и применялось сглаживание по формуле

[3] ppmd_sh URL: https://github.com/Shelwien/ppmd_sh (дата обращения 23.03.2018).

[4] M3-Competition/International Institute of Forecasters URL: https://forecasters.org/resources/time-series-data/m3-competition (дата обращения 23.03.2018).

[5] URL: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/stl.html (дата обращения 23.03.2018).

$x_i = \left(2x_i + x_{i-1} + x_{i-2}\right)/4$. В процессе экспериментального исследования была предпринята попытка выбора такого порядка разности, при котором среднеквадратичное отклонение для ряда будет наименьшим. За исключением одного случая (ежемесячные данные) такой подход приводил к небольшому ухудшению точности прогноза по сравнению с постоянным выбором первой разности.

Поскольку при вычислениях возникает потребность в работе с очень маленькими числами с плавающей точкой, в программной реализации описанного метода была использована библиотека для арифметики с высокой точностью [6].

При построении всех прогнозов применялась ранее описанная процедура оптимизации вычислений при помощи прореживания ряда. При прогнозировании ежегодных, ежеквартальных и других данных оставлялся каждый второй элемент временного ряда, при прогнозировании ежемесячных – каждый шестой. Выбор таких значений был обусловлен соображениями трудоемкости: при прогнозировании ежегодных данных для каждого ряда дважды строился прогноз на 3 шага вперед, ежемесячных данных – шесть раз на 3 шага вперед и т. д.

Результаты вычислений приведены в табл. 5–8. Формат таблиц близок к таблицам из [3]. В таблицах приводятся данные, полученные по различным архиваторам и некоторым их комбинациям. В строках «Лучший М3С» и «Худший М3С» содержатся соответственно наименьшие и наибольшие ошибки на каждый шаг среди всех участвовавших в М3С архиваторов.

В двух из четырех таблиц метод прогнозирования на основе архиваторов показал точность, сравнимую с другими методами. На ежеквартальных и ежемесячных данных его точность оказалась сравнительно низкой, но, тем не менее, при прогнозировании на первые четыре шага сопоставимой с методами из М3С. При этом увеличение количества интервалов с 16 до 32 ни в одном случае не привело к существенному повышению точности прогноза.

Далее приведем результаты вычислений для ряда T-индекса, который рассчитывается метеорологическим бюро Австралии и тесно связан с солнечными пятнами. Временной ряд ежемесячных значений T-индекса можно найти на веб-сайте http://listserver.ips.gov.au/mailman/listinfo/ips-tindex-predictions. На этом же сайте размещается прогноз T-индекса (на данный момент доступен прогноз на 44 шага вперед), а также архивные данные. В рамках нашей работы для каждого из месяцев с апреля 2011 по апрель 2016 г. был построен прогноз на 18 шагов вперед (в этом временном интервале отсутствуют пропущенные значения), и проведено сравнение его точности с прогнозом метеорологического бюро.

В данном ряде, в отличие от рядов из М3С, встречаются отрицательные значения, поэтому точность прогноза оценивалась по средней абсолютной ошибке (mean absolute error, MAE), вычисляемой по следующей формуле:

$$\mathrm{MAE}\left(\hat{X}, X\right) = \frac{1}{h}\sum_{i=1}^{h}\left|\hat{x}_i - x_i\right|, \tag{7}$$

где

$\hat{X} = \hat{x}_1, \hat{x}_2, ..., \hat{x}_h$ – ряд прогнозных значений;

$X = x_1, x_2, ..., x_h$ – ряд зафиксированных значений.

Вычисления были проведены следующим образом. На основании данных из файла за некоторый месяц (например, январь 2014 г.) строился прогноз на 18 шагов вперед. Затем по более позднему файлу на 18 месяцев по формуле (7) рассчитывалась ошибка вычисленного прогноза и прогноза, содержащегося в исходном файле (за январь 2014 г. в данном примере, в каждом файле помимо зафиксированных значений приводятся прогнозные значения). Результаты расчетов приведены в табл. 9. Как видно из этой таблицы, при прогнозировании на 1 шаг метод на основе архиваторов оказался точнее метода метеорологической службы, в остальных случаях точность прогноза метеорологической службы оказалась выше.

[6] Bignum C++ library URL: https://www.ttmath.org/ (дата обращения 23.03.2018)

Таблица 5

Результаты прогнозирования ежегодных данных из M3C

Метод	Ошибка sMAPE для номера шага *h*						Среднее		Количество рядов
	1	2	3	4	5	6	1–4	1–6	
Лучший M3C	7.6	12.1	16.1	18.2	20.8	22.7	13.65	16.42	645
Худший M3C	10.7	15.2	20.8	24.1	28.1	31.2	17.57	21.59	645
zlib (16 интервалов)	9.9	14.9	20.8	22.9	28.0	28.2	17.13	20.79	645
ppmd (16 интервалов)	10.1	14.5	20.6	22.4	27.3	27.2	16.76	20.25	645
rp (16 интервалов)	10.5	14.9	19.5	21.9	26.2	27.4	16.70	20.06	645
zlib + ppmd + rp (16 интервалов)	10.4	14.5	19.4	21.8	26.5	27.3	16.51	19.97	645
zlib + ppmd + rp (32 интервала)	10.3	14.5	19.3	21.8	26.4	27.3	16.49	19.95	645

Таблица 6

Результаты прогнозирования ежеквартальных данных из M3C

Метод	Ошибка sMAPE для номера шага *h*							Среднее			Количество рядов
	1	2	3	4	5	6	8	1–4	1–6	1–8	
Лучший M3C	4.8	6.6	7.4	8.8	9.4	10.9	12	7	8.04	8.96	756
Худший M3C	7.7	8.9	9.1	10.7	11.8	13.7	15.7	8.86	9.6	10.96	756
zlib (16 интервалов)	5.8	7.6	9.0	11.5	14.0	14.4	17.0	8.47	10.39	12.02	756
ppmd (16 интервалов)	5.8	7.5	8.8	10.6	12.7	13.3	15.4	8.16	9.77	11.12	756
rp (16 интервалов)	6.5	9.0	10.0	12.0	13.5	13.6	15.2	9.35	10.75	11.96	756
zlib + ppmd + rp (16 интервалов)	5.8	7.5	8.8	10.5	13.0	13.2	15.0	8.16	9.80	11.11	756
zlib + ppmd + rp (32 интервала)	5.8	7.5	8.8	10.5	13.0	13.2	14.9	8.16	9.81	11.11	756

Таблица 7

Результаты прогнозирования ежемесячных данных из M3C

Метод	Ошибка sMAPE для номера шага *h*										Среднее			Количество рядов
	1	2	3	4	5	6	8	12	15	18	1–4	1–8	1–18	
Лучший M3C	11.2	10.7	11.7	12.4	11.8	12.2	12.6	13.2	16.2	18.2	11.54	12.06	13.85	1428
Худший M3C	15.3	13.8	15.7	17.0	15.3	15.6	17.4	17.5	22.2	24.3	15.39	15.89	18.4	1428
zlib (16 интервалов)	14.4	14.9	17.3	19.4	20.9	17.9	20.4	19.0	25.6	26.5	16.51	18.42	21.23	1428
ppmd (16 интервалов)	14.0	14.1	15.7	20.2	16.8	20.6	17.7	18.0	24.6	25.5	17.23	18.76	19.95	1428
rp (16 интервалов)	15.6	15.7	17.9	19.7	21.2	18.5	19.1	20.7	26.4	26.9	15.44	17.26	21.55	1428
zlib + ppmd + rp (16 интервалов)	14.0	14.1	15.8	19.6	21.2	18.3	19.1	20.6	25.9	26.7	15.86	18.02	21.13	1428
zlib + ppmd + rp (32 интервала)	14.0	14.1	15.8	19.6	21.2	18.4	19.1	20.6	26.0	26.7	15.86	18.02	21.13	1428
ppmd (16 интервалов, подбор порядка разности)	13.4	13.2	14.7	17.5	19.1	16.4	17.5	17.6	22.3	24.4	14.70	16.52	18.87	1428

Таблица 8

Результаты прогнозирования прочих (other) данных из М3С

Метод	Ошибка sMAPE для номера шага *h*							Среднее			Количество рядов
	1	2	3	4	5	6	8	1–4	1–6	1–8	
Лучший М3С	1.6	2.7	3.8	4.3	5.3	5.1	6.0	3.17	3.86	4.38	174
Худший М3С	2.7	3.8	5.4	6.3	7.8	7.6	9.2	4.38	5.49	6.3	174
zlib (16 интервалов)	2.5	3.4	4.9	6.3	8.3	7.7	8.9	4.25	5.49	6.33	174
ppmd (16 интервалов)	2.4	3.2	4.7	5.1	7.1	6.9	8.2	3.85	4.90	5.68	174
rp (16 интервалов)	2.7	3.9	5.8	6.9	8.0	8.5	10.3	4.84	5.97	6.87	174
zlib + ppmd + rp (16 интервалов)	2.4	3.2	4.7	5.1	7.3	6.8	8.1	3.85	4.91	5.70	174
zlib + ppmd + rp (32 интервала)	2.4	3.2	4.7	5.0	7.2	6.8	8.1	3.84	4.90	5.69	174

Таблица 9

Результаты расчетов для временного ряда Т-индекса

Метод	Ошибка MAE для номера шага *h*										Среднее			Количество рядов
	1	2	3	4	5	6	8	12	15	18	1–4	1–8	1–18	
Прогноз метеорологического бюро	13.0	14.2	14.8	15.7	16.5	18.0	21.2	23.5	25.6	27.0	14.43	16.69	21.09	61
zlib (16 интервалов)	12.3	16.0	18.3	20.4	19.1	21.5	24.9	24.5	28.8	26.5	16.75	19.70	23.68	61
ppmd (16 интервалов)	11.9	15.2	17.4	20.9	21.5	22.7	18.8	25.7	30.6	25.5	16.34	18.56	22.87	61
rp (16 интервалов)	12.5	18.1	18.6	18.3	24.9	24.5	22.6	33.4	38.3	41.9	16.86	21.10	28.2	61
zlib + ppmd + rp (16 интервалов)	11.9	15.2	17.4	20.8	21.5	22.0	18.8	24.9	29.4	24.2	16.31	18.46	22.57	61
zlib + ppmd + rp (32 интервала)	11.8	15.2	17.4	21.2	21.5	22.0	18.8	25.4	28.7	26.1	16.39	18.51	22.67	61

Выводы

В рамках данной работы был разработан и экспериментально исследован метод прогнозирования временных рядов, базирующийся на широко известных архиваторах. Показано, что точность данного метода во многих случаях сравнима с точностью других методов прогнозирования и он представляет практический интерес. Возможно совмещение предлагаемого метода с распространенными методами предварительной обработки данных для повышения точности прогноза.

Список литературы

1. *Kendall M. G.*, *A. Stuart.* The Advanced Theory of Statistics: Design and analysis, and time-series. The Advanced Theory of Statistics. Hafner, 1976.

2. *Hyndman R. J.*, *Athanasopoulos G.* Forecasting: principles and practice. OTexts, 2014.

3. *Makridakis S.*, *Hibon M.* The M3-Competition: results, conclusions and implications // International journal of forecasting. 2000. Vol. 16. No. 4. P. 451–476.

4. *Рябко Б. Я.* Прогноз случайных последовательностей и универсальное кодирование // Проблемы передачи информации. 1988. Т. 24, №. 2. С. 3–14.

5. *Shkarin D.* PPM: One step to practicality // Proc. Data Compression Conference. IEEE, 2002. P. 202–211.

6. *Cover T. M.*, *Thomas J. A.* Elements of information theory. John Wiley & Sons, 2012.

7. *Ryabko B.*, *Astola J.*, *Malyutov M.* Compression-based methods of statistical analysis and prediction of time series. Switzerland: Springer International Publishing, 2016.

8. *Ryabko B.* Compression-based methods for nonparametric prediction and estimation of some characteristics of time series // IEEE Transactions on Information Theory. 2009. Vol. 55. No. 9. P. 4309–4315.

9. *Bille P.*, *Gørtz I. L.*, *Prezza N.* Space-Efficient Re-Pair Compression // Data Compression Conference. IEEE, 2017. P. 171–180.

10. *Cleveland R. B.*, *Cleveland W. S.*, *Terpenning I.* STL: A seasonal-trend decomposition procedure based on loess // Journal of Official Statistics. 1990. Vol. 6. No. 1. P. 3.

Материал поступил в редколлегию 10.05.2018

K. S. Chirikhin [1, 2], **B. Ya. Ryabko** [1, 3]

[1] *Novosibirsk State University*
1 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

[2] *Siberian State University of Telecommunications and Information Sciences*
86 Kirov Str., *Novosibirsk*, *630102*, *Russian Federation*

[3] *Institute of Computational Technologies SB RAS*
6 Academician Lavrentiev Ave., *Novosibirsk*, *630090*, *Russian Federation*

chirihin@gmail.com

EXPERIMENTAL STUDY OF THE ACCURACY OF COMPRESSION-BASED FORECASTING METHODS

In information theory it is known that methods of data compression can be used for forecasting of stationary processes. In this paper an compression-based algorithm for time series forecasting was proposed and empirical study of its accuracy was carried out. The algorithm can operate with arbitrary methods of data compression. During the steps of the algorithm predicted values from different methods are combined, and the greatest impact on the end result is exerted by the method with the best compression ratio for the series. The algorithm can be used for forecasting of time series with discrete and continuous alphabets. To improve the accuracy of the forecast existing meth-

ods of time series preprocessing can be used. The empirical study of the efficiency of the proposed algorithm was conducted on time series from the M3 Competition and the T-index series. To generate forecasts well-known archivers were used. The results of the calculations showed that the obtained method has a relatively high accuracy and speed.

Keywords: universal coding, time series forecasting.

References

1. Kendall M. G., Stuart A. The Advanced Theory of Statistics: Design and analysis, and time-series. The Advanced Theory of Statistics. Hafner, 1976.

2. Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice. OTexts, 2014.

3. Makridakis S., Hibon M. The M3-Competition: results, conclusions and implications. *International journal of forecasting*, 2000, vol. 16, no. 4, p. 451–476.

4. Ryabko B. Ya. Prediction of random sequences and universal coding. *Problems of information transmission*, 1988, vol. 24, no. 2, p. 87–96. (in Russ.)

5. Shkarin D. PPM: One step to practicality. *Proc. Data Compression Conference*. IEEE, 2002, p. 202–211.

6. Cover T. M., Thomas J. A. Elements of information theory. John Wiley & Sons, 2012.

7. Ryabko B., Astola J., Malyutov M. Compression-based methods of statistical analysis and prediction of time series. Switzerland, Springer International Publishing, 2016.

8. Ryabko B. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Transactions on Information Theory*, 2009, vol. 55, no. 9, p. 4309–4315.

9. Bille P., Gørtz I. L., Prezza N. Space-Efficient Re-Pair Compression. *Data Compression Conference*. IEEE, 2017, p. 171–180.

10. Cleveland R. B., Cleveland W. S., Terpenning I. STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 1990, vol. 6, no. 1, p. 3.

For citation:

Chirikhin K. S., Ryabko B. Ya. Experimental Study of the Accuracy of Compression-Based Forecasting Methods. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 3, p. 145–158. (in Russ.)

DOI 10.25205/1818-7900-2018-16-3-145-158

СВЕДЕНИЯ ОБ АВТОРАХ

Бабенко Владимир Николаевич – кандидат биологических наук, старший научный сотрудник Института цитологии и генетики СО РАН (Новосибирск), старший научный сотрудник лаборатории компьютерной геномики факультета естественных наук Новосибирского государственного университета

Бабенко Роман Олегович – магистрант механико-математического факультета Новосибирского государственного университета

Бакиева Айгерим Муратовна – аспирантка факультета информационных технологий Новосибирского государственного университета

Бакулина Анастасия Юрьевна – кандидат биологических наук, заведующая лабораторией структурной биоинформатики и молекулярного моделирования факультета естественных наук Новосибирского государственного университета

Бандишоева Рисолат Мирзошоевна – старший преподаватель кафедры автоматизированных систем обработки информации и управления Таджикского технического университета им. акад. М. С. Осими (Душанбе, Таджикистан)

Батура Татьяна Викторовна – кандидат физико-математических наук, доцент, старший научный сотрудник Института систем информатики им. А. П. Ершова СО РАН (Новосибирск)

Богомолов Антон Геннадьевич – инженер Новосибирского государственного университета, младший научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Брагин Анатолий Олегович – кандидат биологических наук, научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Гаврилов Денис Андреевич – магистрант факультета информационных технологий Новосибирского государственного университета

Галиева Айя Галиевна – студентка факультета информационных технологий Новосибирского государственного университета

Городничев Максим Александрович – младший научный сотрудник Института вычислительных технологий СО РАН, младший научный сотрудник Института вычислительной математики и математической геофизики СО РАН, старший преподаватель Новосибирского государственного университета

Губанова Наталья Владимировна – кандидат биологических наук, старший научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Дергилев Артур Игоревич – магистрант механико-математического факультета Новосибирского государственного университета

Зенг Валерия Андреевна – аспирантка Омского государственного технического университета

Изместьева Ольга Владимировна – библиотекарь Иркутского регионального колледжа педагогического образования

Ковалев Сергей Сергеевич – студент Новосибирского государственного медицинского университета, лаборант Института цитологии и генетики СО РАН (Новосибирск)

Комиссаров Александр Владимирович – магистрант факультета информационных технологий Новосибирского государственного университета, инженер-программист Института вычислительных технологий СО РАН (Новосибирск)

Корсун Ирина Андреевна – магистрант факультета информационных технологий Новосибирского государственного университета

Леберфарб Елена Юрьевна – кандидат медицинских наук, научный сотрудник Института цитологии и генетики СО РАН, доцент Новосибирского государственного медицинского университета

Матусевич Дмитрий Сергеевич – старший преподаватель кафедры информатики и кибернетики Байкальского государственного университета (Иркутск)

Медведева Ирина Вадимовна – кандидат биологических наук, научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Можина Алина Викторовна – магистрант факультета информационных технологий Новосибирского государственного университета, инженер-программист Института вычислительных технологий СО РАН (Новосибирск)

Орлов Юрий Львович – доктор биологических наук, профессор РАН, старший научный сотрудник Института цитологии и генетики СО РАН, заведующий лабораторией компьютерной геномики факультета естественных наук Новосибирского государственного университета, главный научный сотрудник Института морских биологических исследований РАН им. А. О. Ковалевского (Севастополь)

Орлова Нина Геннадьевна – кандидат физико-математических наук, доцент кафедры высшей математики Новосибирского государственного архитектурно-строительного университета – НГАСУ (Сибстрин)

Пальчунов Дмитрий Евгеньевич – доктор физико-математических наук, ведущий научный сотрудник Института математики СО РАН (Новосибирск)

Погодин Руслан Сергеевич – магистрант факультета информационных технологий Новосибирского государственного университета, программист ООО «Вероника» (Новосибирск)

Подколодная Ольга Александровна – кандидат биологических наук, старший научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Подколодный Николай Леонтьевич – старший научный сотрудник Института вычислительной математики и математической геофизики СО РАН, ведущий аналитик Института цитологии и генетики СО РАН (Новосибирск)

Прочкин Павел Вячеславович – магистрант факультета информационных технологий Новосибирского государственного университета, инженер-программист Института вычислительных технологий СО РАН (Новосибирск)

Рудыч Павел Дмитриевич – инженер-программист Института вычислительных технологий СО РАН (Новосибирск)

Рябко Борис Яковлевич – доктор технических наук, главный научный сотрудник Института вычислительных технологий СО РАН, профессор кафедры компьютерных систем факультета информационных технологий Новосибирского государственного университета

Табанюхов Кирилл Александрович – студент Новосибирского государственного аграрного университета

Твердохлеб Наталья Николаевна – младший научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Цуканов Антон Витальевич – аспирант Института цитологии и генетики СО РАН, инженер лаборатории компьютерной геномики факультета естественных наук Новосибирского государственного университета

Чадаева Ирина Витальевна – младший научный сотрудник Института цитологии и генетики СО РАН (Новосибирск)

Чирихин Константин Сергеевич – аспирант факультета информационных технологий Новосибирского государственного университета, ассистент кафедры прикладной математики и кибернетики Сибирского государственного университета телекоммуникаций и информатики (Новосибирск)

Юрченко Андрей Васильевич – кандидат физико-математических наук, первый заместитель директора Института вычислительных технологий СО РАН (Новосибирск)

Авторы представляют статьи на русском или английском языке объемом от 0,5 авторского листа (20 тыс. знаков) до 1 авторского листа (40 тыс. знаков), включая иллюстрации (1 иллюстрация форматом 190 × 270 мм = 1/6 авторского листа, или 6,7 тыс. знаков).

Публикации, превышающие указанный объем, допускаются к рассмотрению только после индивидуального согласования с редакцией журнала.

Требования к оформлению основного текста и иллюстративных материалов

Текст рукописи должен быть представлен в редколлегию в виде файла MS Word. Гарнитура Times New Roman, размер шрифта 14, межстрочный интервал 1,5, размеры полей – стандартные значения текстового процессора. Форматирование – выравнивание по ширине страницы, переносы слов отключены, красной строки нет. Не допускается ручное форматирование абзацев (пробелами, лишними переводами строк, разрывами страниц).

Подзаголовки набираются полужирным шрифтом. Для выделения частей текста и отдельных слов и словосочетаний допускается использование курсива или полужирнго шрифта. Подчеркивание, разрядка, изменение основного кегля и выделение цветом и т. п. не используются.

Все иллюстрации к рукописи статьи должны быть приложены в отдельных файлах. При этом в тексте может содержаться как включенное изображение с указанием имени файла, так и только указание. Все иллюстрации, содержащие схемы, графики, алгоритмы, скриншоты и другие изображения должны быть представлены в максимально высоком качестве, без каких-либо потерь и искажений (.jpg, .tif).

Все иллюстрации должны иметь подрисуночную подпись.

Нумерация последовательная и неразрывная от начала статьи. Не допускается использование других наименований, кроме «рис.» и усложнение нумерации (например, «рис. 3.2.»). Ссылка на иллюстрацию в тексте должна быть приведена в круглых скобках: (рис. 1).

Формулы должны быть набраны с использованием редактора **MathType**. Кегль основных символов – 11, греческие символы набираются прямым шрифтом, латинские – курсивом. Нумеруются только те формулы, на которые автор ссылается в тексте.

Ссылки на литературу указываются цифрами в квадратных скобках. Список литературы нумеруется в порядке цитирования и оформляется в соответствии с ГОСТом на библиографическое описание. Ссылки на неопубликованные работы, а также на интернет-ресурсы (кроме электронных изданий, поддающихся библиографическому описанию) оформляются в виде сноски.

Дополнительные материалы

К статье обязательно прилагаются следующие материалы.

- Информация об авторе / авторах на русском и английском языках:
  - ФИО полностью,
  - сведения об ученой степени и ученом звании,
  - должность и место работы, почтовый адрес,
  - электронный адрес,
  - контактный телефон.
- Индекс УДК (Универсальной десятичной классификации).
- Название статьи на русском языке и его авторский перевод на английский язык.
- Аннотация статьи на русском и английском языках.
- Ключевые слова (не более 10–15), на русском и английском (Keywords) языках.
- Список литературы на русском и английском (References) языках.

Доставка материалов

Материалы предоставляются в редакцию только по электронной почте:

inftech@vestnik.nsu.ru