Нейроинформатика, её приложения и анализ данных

XXX Всероссийский семинар

Красноярск
2022

Министерство науки и высшего образования РФ
Сибирское отделение РАН
Российская ассоциация нейроинформатики
Институт вычислительного моделирования СО РАН

НЕЙРОИНФОРМАТИКА, ЕЁ ПРИЛОЖЕНИЯ И АНАЛИЗ ДАННЫХ

МАТЕРИАЛЫ
ТРИДЦАТОГО ВСЕРОССИЙСКОГО СЕМИНАРА
30 сентября – 2 октября 2022 года

Красноярск 2022

УДК 007.52 (042.3)
Н 43

Нейроинформатика, её приложения и анализ данных: Материалы XXX Всероссийского семинара, 30 сентября – 2 октября 2022 года / Под ред. М.Г. Садовского, отв. за вып. М.Ю. Сенашова; – Красноярск: Институт вычислительного моделирования СО РАН, 2022. – 189 с.

В сборнике представлены материалы XXX Всероссийского семинара «Нейроинформатика, её приложения и анализ данных», проходившей в городе Красноярске 30 сентября – 2 октября 2022 года.

Основной задачей конференции является всесторонний и высоко квалифицированный обмен новейшими достижениями в различных областях нейроинформатики (как теории нейросетей, так и в области различных практических приложений), а также в области нелинейного статистического анализа многомерных данных, обладающих нетривиальными структурами.

Большое внимание уделено анализу областей применимости и точности методов обработки многомерных данных, анализу устойчивости различных новых (нелинейных) методов кластеризации, разбор большого числа конкретных случаев, иллюстрирующих эти проблемы и достижения.

Материалы предназначены для научных работников, преподавателей, студентов и аспирантов соответствующих специальностей.

Конференция проводится при поддержке Красноярского математического центра, финансируемого Минобрнауки РФ в рамках мероприятий по созданию и развитию региональных НОМЦ (Соглашение № 075-02-2022-873).

ISBN 978-5-6047078-2-1
9 785604 707821

ПЕРВЫЕ РЕЗУЛЬТАТЫ ИЗУЧЕНИЯ МИКРОБИОТЫ У БОЛЬНЫХ РАССЕЯННЫМ СКЛЕРОЗОМ

В.Г.Абрамов[3], А.А.Молявко[1,2], М.Е.Туник[3], А.А.Тетерлева[1], [4]А.В.Моргун, [4]И.А.Ларионова, К.О.Туценко[3,4], Д.В.Похабов[3], М.Г.Садовский[2,3,4]

[1]Сибирский федеральный университет, ИФБиБТ, ***tenth_smith@mail.ru***

[2]Институт вычислительного моделирования СО РАН, ***msad@icm.krasn.ru***

[3]ФГБУ ФСНКЦ ФМБА России

[4]ФГБОУ ВО КрасГМУ им. проф. В.Ф.Войно-Ясенецкого Минздрава России

Микробиота человека — это совокупность микроорганизмов, обитающих непосредственно на человеке. Всю микробиоту человека естественно разделять на группы по «среде обитания»: выделяют микробиоту кожных покровов, микробиоту слизистых и микробиоту, живущую «внутри» человека — например, в кишечнике или в органах дыхания. Микробиота человека играет важнейшую роль в жизни человека [1 – 6]. Связь состава микробиоты и развитием болезней широко изучалась ранее.

Анализ микробиоты человека требует развития и применения инструментов быстрого определения её видового состава и количественных показателей. Один из эффективных инструментов здесь — секвенирование 16S РНК бактерий с последующей идентификацией их таксономического положения [7 – 9]. Здесь встаёт задача эффективного (в смысле вычислительных затрат) и надёжного (в смысле точности определения таксономического положения) определения видовой принадлежности секвенированной последовательности. Традиционным методом сравнения последовательностей является т.н. выравнивание, основывающееся на идее редакционного расстояния [10]. Несмотря на широкую распространённость, выравнивание обладает рядом существенных недостатков, которые не могут быть преодолены. К таким недостаткам относятся:

– расходимость метода. Укладка двух последовательностей может продолжаться сколь угодно долго, если не предпринимать специальных — и никак не обусловленных самим методом — усилий;

– основная идея редакционного расстояния, лежащего в основе метода выравнивания, заключается в выборе минимального числа т.н. элементарных преобразований, позволяющих сопоставить две последовательности однозначно. Выбор укладки, обеспечивающей минимум этих операций, не однозначен и ведущей идеей для понижения этого произвола был выбор т.н. штрафных функций. Иными словами, разные элементарные преобразования «стоили» по-разному. И минимум укладки находился не по числу элементарных операций, а по их общей «стоимости» (score); выбор системы весов для элементарных преобразований носит полностью произвольный характер и самим методом никак не обусловлен;
– произволен и выбор иных параметров;
– наконец, укажем ещё одну, возможно, ключевую, трудность метода выравнивания — очень большие вычислительные сложности (вычислительные затраты) и низкая эффективность при работе с такими важными в биологии типами мутаций, как вставки и выпадения.

Совсем недавно [11 – 13] был предложен новый метод, полностью свободный от указанных выше недостатков; так, этот метод вообще не имеет свободных параметров. Другой важной особенностью метода является его высокая эффективность в поиске замен типа «вставка/удаление». Именно этот метод был применён для анализа генетического материала, необходимого для изучения состава микробиоты человека. Опишем кратко этот метод.

Пусть имеются две (нуклеотидных) последовательности T_1 и T_2 длины N_1 и N_2, соответственно ($N_1 \neq N_2$). Будем называть **наложением** последовательности T_1 на T_2 такое их взаимное соответствие, когда i-му символу в последовательности T_1 поставлен в соответствие первый символ T_2 и так далее. Если длина T_2 меньше длины оставшейся части последовательности T_1, то тогда наложение означает, что все элементы второй последовательности целиком располагаются против элементов T_1. Если же вторая последовательность длиннее, чем остаток первой, то наложение оканчивается там, где находится последний элемент T_1. Соответственно, самой первой задачей сравнения двух последовательностей T_1 и

T_2 является определение числа точно совпавших символов в текущем наложении; и так для всех наложений.

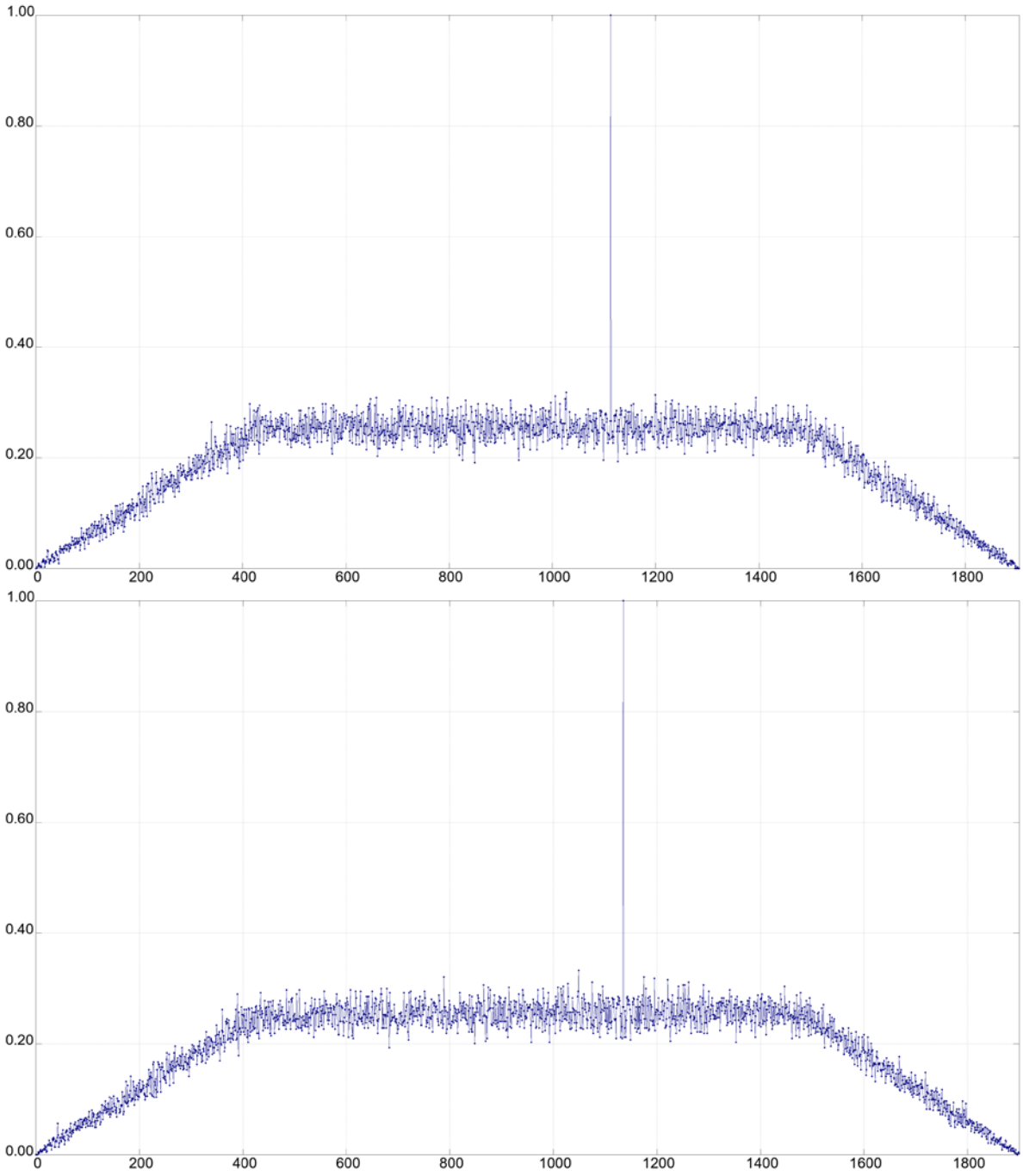


Рис.1. Пример свёртки, вычисленной для сиквенса № 105 и гена *Bacteroides*, идентификатор AE015928 и сиквенса № 107 и гена *Phocaeicola*) идентификатор AB5107031; видовые названия не приведены. По горизонтали отложен номер наложения, по вертикали — значение свёртки.

Ключевая идея подсчёта числа таких совпадений состоит в том, чтобы вычислить значение свёртки для двух последовательностей. Для этого каждая из последовательностей преобразуется в четыре $(0,1)$ последовательностей по следующему правилу: в первой из четырёх копии последовательности все символы A заменяются на 1, а остальные — на нули; аналогичная замена делается для оставшихся трёх копий последовательности, в которых заменяется один символ, а остальные «обнуляются». Пусть $T = a_1, a_2, \dots, a_N$ и $U = b_1, b_2, \dots, b_M$, где N и M — длины последовательностей T и U, соответственно. Тогда свёртка этих двух последовательностей есть последовательность $\mathfrak{S}$ длины $N + M$, для получения которой нужно одну из последовательностей (U для определённости) инвертировать, переписав в обратном порядке, и последовательно перебрать все наложения, подсчитывая в каждом из их число точно совпавших нуклеотидов, определяемых в пределах данного наложения. Полученные значения свёртки (для каждого из наложений) и будут давать число точно совпавших нуклеотидов в наложении.

На первый взгляд данная задача представляется вычислительно трудоёмкой, однако на самом деле вычисление свёртки может быть произведено весьма быстро. Для этого используется специальная процедура — быстрое преобразование Фурье. Кроме того, данный метод позволяет применить крупномасштабное распараллеливание, что существенно (на порядки) ускоряет расчёты и делает их вообще выполнимыми. Все детали этого подхода изложены в [ш1 – ш3].

Материалы и методы

Для целей нашего исследования использовались гены 16 S РНК бактерий. Генетический материал брался из открытого источника данных RDP (http://rdp.cme.msu.edu/) [14]. Эта база данных использовалась для идентификации сиквенсов 5 региона 16S РНК бактерий, полученных от трёх доноров биологического материала; все доноры имеют подтверждённый диагноз рассеянный склероз; сиквенсы были получены М.Р. Кибиловым (ЦКП «Геномика» СО РАН, г.Новосибирск). Всего в нашем распоряжении было 512 сиквенсов, полученных

от трёх доноров; следует подчеркнуть, что это число является сводным: не все доноры содержали каждый сиквенс. Основной задачей работы было изучение особенностей видового состава пристеночной микробиоты кишечника человека с тем или иным заболеванием.

Для сравнения полученных от ЦКП сиквенсов с образцами из упомянутой базы данных генов 16 S РНК бактерий (21195 записей) использовались свёрточные функции, а не традиционные методы сравнения, основывающиеся на выравнивании. При вычислении свёрток использовалась исходная база, проверка на однородность её содержимого не проводилась. В последствии выяснилось, что упомянутая база содержала несколько записей генов 16 S РНК эукариот, а также гены цианобактерий. Заметим, что эта особенность исходной базы генов 16 S РНК сыграла важную положительную роль при анализе сиквенсов, полученных от доноров.

Результаты

Для определения видового состава микробиоты кишечника доноров использовалась база данных RDP, а само по себе сравнение проводилось путём вычисления каждой из 21195 генов, представленных в базе, с каждым сиквенсом, полученным из ЦКП. Тем самым, были подсчитаны свёртки для 10851840 пар последовательностей. Все вычисления проводились на ПК с восемью ядрами и 32 Гб оперативной памяти; время счёта составило порядка 100 часов.

На Рисунке 1 представлены графики свёрток, вычисленные для двух сиквенсов (№№ 105 и 107) и двух генов 16 S РНК бактерий, идентификаторы приведены в подписи к рисунку. Поясним смысл этих графиков. Во-первых, это не графики в строгом смысле: на самом деле, свёртка представляет собой (конечную) числовую последовательность, а отрезки между точками, собственно, и указывающими на значения свёртки в данной точке, проведены для облегчения восприятия. Во-вторых, как и видно из рисунка, общая длина свёртки немного не превышает 1900, что и составляет сумму длин того гена 16 S РНК бактерии, для которого подсчитывалась данная свёртка. В-третьих, хорошо видно, что в доста-

точно грубом приближении свёртка выглядит как трапеция; это естественно для случая вычисления свёртки для двух последовательностей разной длины. Как и было сказано выше, каждое значение свёртки (в каждой точке, соответствующей текущему наложению) равно числу точно совпавших — в данном наложении — нуклеотидов. При этом важно, что положение этих совпавших нуклеотидов вдоль по наложению может меняться: два разных наложения могут иметь одинаковые значения свёртки, но при этом только по значению свёртки ничего нельзя сказать о том, какие именно нуклеотиды совпали и где они располагаются вдоль по наложению.

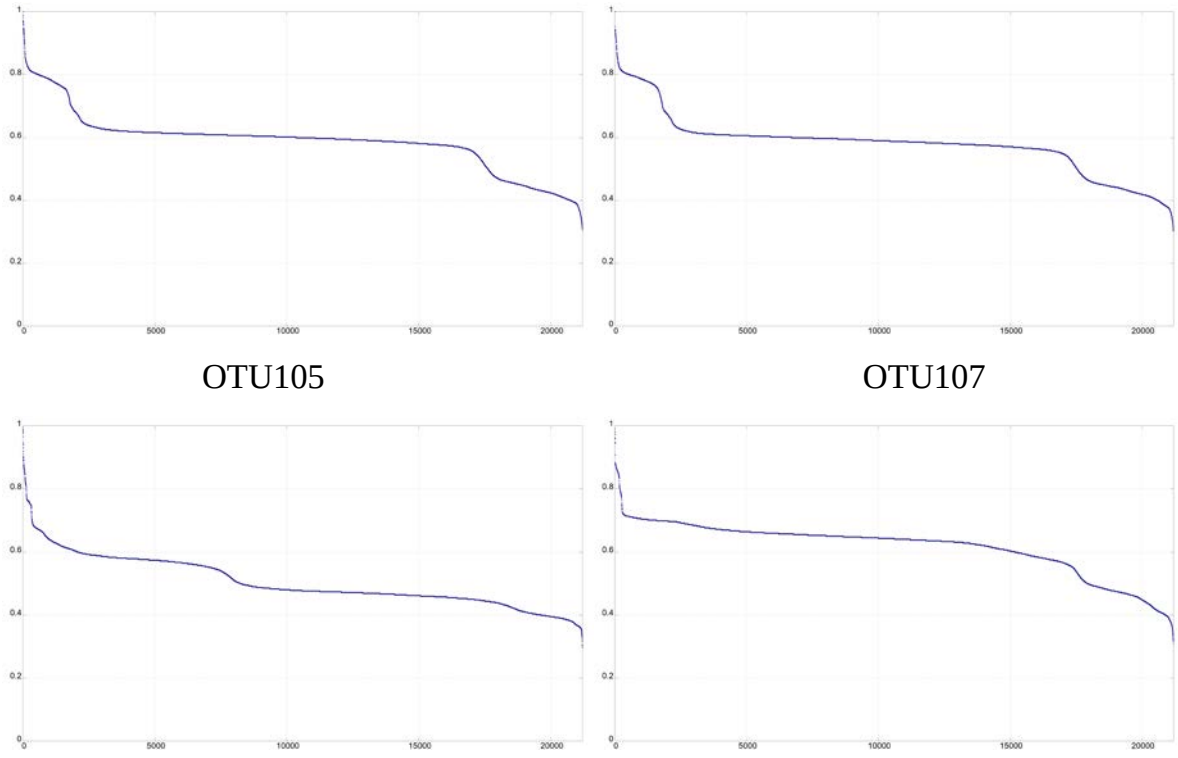


Рис.2. Кривые убывания максимумов свёрток, полученных для четырёх сиквенсов пациентов с РС, для всех 21195 генов 16 S РНК бактерий. По горизонтали отложены номера генов 16 S РНК бактерий, для которых было получено данное значение свёртки, по вертикали — значение свёртки.

Хорошо видно, что в целом свёртка очень похожа на кусочно-гладкую ломаную (трапецию), зашумленную каким-то случайным либо сложно организо-

ванным процессом. Такой характер поведения свёртки объясняется случайными совпадениями, встречающимися во всяком наложении; вопрос о том, как можно понизить уровень такого шума, обсуждается ниже.

Обратим теперь внимание на два высоких пика (единичной длины), расположенных в районе 1150-го наложения. Поскольку исходный смысл свёртки — это число совпавших нуклеотидов, постольку её абсолютное значение зависит от длины меньшей последовательности. Для того, чтобы можно было сравнивать свёртки, получаемые для сиквенсов разной длины, мы ввели нормировку на длину более короткой последовательности — сиквенса в нашем случае. Пики единичной высоты, показанные на Рис.1, соответствуют таким наложениям показанных там сиквенсов на соответствующие гены 16 S РНК бактерии, при которых имеет место полное совпадение всех нуклеотидов из сиквенса с нуклеотидами в гене 16 S РНК бактерии, по всей длине сиквенса. Понятно, что такое хорошее совпадение встречается далеко не всегда; действительно, во всех 10851840 парах свёрток, которые были нами подсчитаны, единичные значения достигались только в 106 парах.

Оценить характерные значения свёрток, получаемых для одного сиквенса, можно по Рис. 2. На этом рисунке приведены четыре примера сиквенсов, для которых максимальное значение свёртки, определяемое по всем 21195 генам 16 S РНК бактерий, упорядочено по убыванию. Иными словами, каждая точка на этом рисунке — это максимальное значение свёртки, полученное при сравнении с тем или иным (из 211195) геном 16 S РНК бактерий.

На начальном этапе анализировались только те пары «*сиквенс – ген*», для которых максимальное значение по всем парам было равно единице. Это означает, что в этих парах нашёлся ген 16 S РНК бактерии, для которого данный сиквенс был полным вложением: то есть, указанный сиквенс является подпоследовательностью в этом гене. Понятно, что такое хорошее совпадение по типу вложенности не является обязательным. Однако следует сказать, что тип «кривой», показывающей убывание максимумов, фактически во всех парах одинаков; единственное отличие состоит в том, что с единицы он начинается только у 106

пар.

Обратим теперь внимание на правую часть «графиков», представленных на Рис. 2. Там чётко выделяется (также немногочисленная) группа генов, являющихся «антивложениями»: это те гены, для которых значения свёртки составляют меньше 0,4, что означает, что лишь 40 % нуклеотидов в самом оптимальном наложении сиквенса на ген совпадают. Заметим, что это значение свёртки является оценкой сверху для числа совпадений: кроме точно совпавших олигонуклеотидов, свой положительный вклад в значение свёртки вносят и случайные совпадения.

Для всех четырёх примеров, представленных на Рис. 2, характерным является поведение значений свёрток по типу кластеризации: это означает, что можно сформировать весьма значительные по числу наборы генов 16 S РНК бактерий, которые дают очень близкие по величине значения свёрток с данным сиквенсом. Так, на Рис. 2 (OTU105) это гены, для которых значение свёртки лежит в (узкой) окрестности значения 0,6; аналогичное поведение можно увидеть и на Рис. 2 (OTU107). Для сиквенса OTU958 это типичное значение даже превосходит 0,6 и лежит на уровне 0,65 ÷ 0,70.

Конечно, такие грубые оценки сходства двух нуклеотидных последовательностей едва ли могут служить надёжным основанием для атрибуции сиквенса. Здесь требуются дальнейшие инструментальные исследования, направленные на построение системы сравнения результатов традиционного выравнивания и по свёрточным функциям, однако можно с высокой уверенностью утверждать, что значение свёртки на уровне 0,6 может служить веским основанием для того, чтобы отвергнуть гипотезу о подобии данного сиквенса и тех генов, для которых значения свёртки находятся в окрестности 0,6: действительно, это означает, что реальный уровень совпадений (за исключением случайных, которые только увеличивают значение свёртки) будет меньше 0,5. Такой низкий уровень гомологии на уровне мононуклеотидного состава не может считаться приемлемым для того, чтобы такие две нуклеотидные последовательности считать подобными либо близкими. Следует подчеркнуть, что использованный нами метод сравнения

нуклеотидных последовательностей на основе свёрточных функций обладает важным достоинством: он легко и эффективно находит мутации типа вставки/выпадения.

Обсуждение

Основной задачей данной работы является иллюстрация нового метода сравнения символьных (конкретно, нуклеотидных) последовательностей — метода Шайдурова, который основан на вычислении свёртки двух (0,1) последовательностей. Данный метод имеет три существенных преимущества перед традиционным подходом (выравниванием), основанном на определении редакционного расстояния:

- отсутствие расходимости, характерной для выравнивания;
- полное отсутствие свободных параметров, выбор которых для случая выравнивания полностью лежит за пределами самого метода, и
- высокая эффективность в обработке такого типа мутаций, как вставки/выпадения, доставляющие максимальную трудность для выравнивания.

Высокая вычислительная эффективность, скорость работы, возможность распараллеливания, тем не менее, «не проходят даром»: непосредственное вычисление свёрток ещё не даёт полной картины того, насколько близки две последовательности. Перечислим проблемы, которые требуют дополнительных исследований.

1. Проблема шума. Вычисление свёрток для посимвольного совпадения даёт очень зашумленный сигнал, порождённый большим числом случайных и бессодержательных совпадений, наблюдающихся в наложении — особенно для длинных последовательностей в случае, когда их длины близки.
2. Проблема локализации. Вычисление свёртки позволяет за один проход вычислить значения для всех мыслимых наложений; максимальные значения свёртки соответствуют таким наложениям, для которых наблюдается максимальное же посимвольное совпадение. Пусть теперь у исследователя есть все основания ожидать, что совпадающий фрагмент обладает очень высокой

степенью подобия, однако его длина весьма и весьма мала по сравнению с длиной сравниваемых последовательностей. Тогда локализовать — указать точное положение такого короткого фрагмента с очень высокой степенью подобия весьма затруднительно. К счастью, чем сильнее различие длин сравниваемых последовательностей, тем проще локализовать совпадающий участок. Если сравниваются две последовательности существенно разной длины и для какого-то наложения наблюдается ярко выраженный пик, то это значит, что в более длинной последовательности содержится фрагмент очень высокого уровня подобия более короткому, и положение этого фрагмента в длинной последовательности локализуется следующим образом: он находится в интервале длиной 2Δ, где Δ — длина короткой последовательности с центром в точке максимума свёртки. Для случая, показанного на Рис. 1 локализовать начало участка, имеющего полное совпадение с той OTU, с которой проводилось сравнение, можно с точностью до 400 нуклеотидов.

Ещё одна содержательная проблема относится не столько к самому методу Шайдурова, сколько к содержательности тех задач, решение которых требует сравнения нуклеотидных (либо символьных) последовательностей. Метод Шайдурова позволяет с высочайшей вычислительной эффективностью вычислить значения свёртки для всех вообще наложений двух последовательностей. Так, в качестве эксперимента мы вычислили свёртку между первой и второй хромосомами человека; это заняло на персональном компьютере с 8 ядрами и 32 Гб оперативной памяти 22 минуты. Однако возникает естественный вопрос: а для какого рода содержательных задач может потребоваться столь объёмные сравнения — по фрагментам длиной в сотни миллионов символов?

Очевидно, что в настоящее время гораздо больший содержательный интерес вызывают задачи поиска всех фрагментов сравнительно малой длины (например, до десяти тысяч символов), располагающихся внутри очень протяжённой (сотни миллионов) последовательности и имеющих уровень подобия не ниже заданного. Отметим, что метод Шайдурова позволяет очень эффективное распараллеливание для решения подобного рода задач; более того, формализа-

ция такого поиска может существенно способствовать решению задачи локализации. Сравнивая две очень длинные последовательности, можно одну из них разделить на набор (быть может, пересекающихся) подпоследовательностей существенно меньшей длины и вычислять свёртки между одной очень длинной последовательностью и одной очень короткой; это сильно повысит точность локализации и уменьшит влияние шума.

Отметим ещё одно важное обобщение этого метода. В настоящей работе все свёртки считались для односимвольного совпадения; для этого исходные (нуклеотидные) последовательности преобразовывались в четыре бинарных. Однако можно расширить алфавит бинаризации и обрабатывать не четыре, а 16 $(0,1)$ последовательностей, что даст в качестве значений свёртки не число совпавших символов в наложении, а число совпавших пар символов, непосредственно примыкающих друг к другу. Очевидно, что доля случайно совпавших пар будет заметно меньше, чем доля случайно совпавших нуклеотидов, что понизит уровень шума. Ничто не мешает увеличивать алфавит бинаризации и подсчитывать свёртки для k-плетов, что, вообще говоря, может понизить уровень шума до любой заранее заданной величины. Отметим, что этот подход реализован в настоящее время для $1 \leq k \leq 7$, однако детальное обсуждение этих результатов выходит за рамки настоящей работы.

Укажем в заключение ещё одну важную задачу, возникающую в связи с применением метода Шайдурова для сравнения генетических последовательностей. К настоящему времени выравнивание занимает лидирующее положение вреди методов сравнения символьных последовательностей. Многие и многие исследователи, особенно молодые, считают данный метод единственно возможным и активно используют различные программы, реализующие выравнивание; при этом использование таких программ далеко не всегда даёт пользователю возможность понять «кухню» метода и в результате исследователь становится заложником используемого программного обеспечения. Но несмотря на это, выравнивание стало фактически стандартом по умолчанию в сравнении последовательностей. На практике это означает, что для успешного и осознанного приме-

нения метода Шайдурова для сравнения нуклеотидных последовательностей требуется провести своего рода инструментальное исследование: сопоставить результаты сравнения, полученные на тех или иных эталонных наборах последовательностей этими двумя разными методами — выравниванием и свёрточными функциями. Сделать это надо не для сравнения самих методов, а для построения системы интерпретации результатов, получаемых разными методами. Эта работа должна быть проделана в ближайшее время.

Список литературы

1. Аверина, О. В., В. Н. Даниленко. Микробиота кишечника человека: роль в становлении и функционировании нервной системы. // Микробиология 86.1 (2017): 5-24.
2. Бухарин, О. В., Перунова, Н. Б. Роль микробиоты в регуляции гомеостаза организма человека при инфекции // Журнал микробиологии, эпидемиологии и иммунобиологии, 2020. (5), 458-467.
3. Wang B., Yao M., Lv L., Ling Z., Li L. The human microbiota in health and disease // Engineering. 2017 Feb 1;3(1):71-82.
4. Martínez J.E., Vargas A., Pérez-Sánchez T., Encío I.J., Cabello-Olmo M., Barajas M. Human microbiota network: unveiling potential crosstalk between the different microbiota ecosystems and their role in health and disease // Nutrients. 2021 Aug 24;13(9):2905.
5. Innao V., Allegra A.G., Musolino C., Allegra A. New frontiers about the role of human microbiota in immunotherapy: the immune checkpoint inhibitors and CAR T-cell therapy era // International Journal of Molecular Sciences. 2020 Nov 24;21(23):8902.
6. Ewald DR, Sumner SC. Human microbiota, blood group antigens, and disease // Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2018 May;10(3):e1413.
7. Tang, Yi-Wei, Nicole M. Ellis, Marlene K. Hopkins, Douglas H. Smith, Deborah E. Dodge, and David H. Persing. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. // Journal of

clinical microbiology 36, no. 12 (1998): 3674-3679.

8. Clarridge III, Jill E., Silvia M. Attorri, Qing Zhang, and John Bartell. 16S ribosomal DNA sequence analysis distinguishes biotypes of *Streptococcus bovis*: *Streptococcus bovis* biotype II/2 is a separate genospecies and the predominant clinical isolate in adult males // Journal of clinical microbiology 39, no. 4 (2001): 1549-1552.
9. Clarridge III, Jill E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases // Clinical microbiology reviews 17, no. 4 (2004): 840-862.
10. Левенштейн В.И. Двоичные коды с исправлением выпадений и вставок символа // 1. Проблемы передачи информации, 1965. 1(1), pp.12-25.
11. Molyavko, A., Shaidurov, V., Karepova, E. and Sadovsky, M., 2020, May. Highly parallel convolution method to compare DNA sequences with enforced in/del and mutation tolerance. In International Work-Conference on Bioinformatics and Biomedical Engineering (pp. 472-481). Springer, Cham.
12. А.А. Молявко, Е.Д. Карепова, М.Г. Садовский Сравнение последовательностей методом Шайдурова с применением технологии MPI для распараллеливания // Информатизация и связь, 2022, – в печати
13. Molyavko A.A., Karepova E.D., Borovikov I.A., Mutovina O.A., Sadovsky, M.G. Comparison of search efficiency in symbol sequences with mismatches between alignment and Shaidurov's method // CEUR Workshop Proceedings, Vol., 3047, 2021, SibData, pp.93 – 97.
14. Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic acids research, 42(Database issue), D633–D642. https://doi.org/10.1093/nar/gkt1244

ИДЕНТИФИКАЦИЯ ВРЕМЕННЫХ РЯДОВ СТИМУЛОВ, ПОЛУЧЕННЫХ ИСКУССТВЕННОЙ НЕЙРОННОЙ СЕТЬЮ, ПО ПАТТЕРНУ НЕЙРОННОЙ АКТИВНОСТИ[1]

С.И. Барцев[1,2], Г.М. Маркова[2]

[1] Институт биофизики СО РАН –

обособленное подразделение ФИЦ КНЦ СО РАН, ***bartsev@yandex.ru***

[2] Сибирский федеральный университет, Красноярск, Россия

Одна из задач современной когнитивной нейронауки – реконструкция информации, которая была получена и обработана мозгом. В рамках концепции нейронных коррелятов [1], в качестве исходных данных для реконструкции используются паттерны нейронной активности. Известно, что данные паттерны, соответствующие тому или иному полученному стимулу, индивидуальны и динамичны [2-5], что затрудняет реконструкцию. Эти особенности также исследуются и воспроизводятся на нейросетевых моделях. В работах [6,7] показано, что информация о стимуле, полученном искусственной рекуррентной нейронной сетью (РНС) при прохождении теста отложенного сравнения с образцом (ОСО), представлена в нейронной активности в динамичной форме. Несмотря на это, для данной задачи продемонстрирована возможность идентификации стимула по паттерну нейронной активности с помощью метода центроидов и нейросетевого декодирования [7], при этом точность метода центроидов достигает не более 75-80%, а метод нейросетевого декодирования позволяет безошибочно идентифицировать полученный стимул по паттерну нейронной активности отдельно взятой сети.

В ходе теста ОСО испытуемый (РНС, другая система или живое существо) получает первый стимул, спустя некоторое время – второй, после чего требуется решить, различались эти стимулы или совпадали. Следует отметить, что такая задача значительно упрощена по сравнению с реальными задачами, которые вынуждены решать живые организмы, а также реальные или виртуальные

[1] Исследование выполнено при финансовой поддержке РФФИ, Правительства Красноярского края и Красноярского краевого фонда науки в рамках научного проекта № 20-41-240003.

роботы-аниматы. Хотя в тесте ОСО конкретная продолжительность паузы между стимулами может варьироваться, испытуемый имеет дело с однократными событиями, четко разделенными во времени, а не с непрерывным потоком. Одна из задач, где испытуемый вынужден принимать решение в непрерывном потоке событий – рефлексивные игры. Однако даже простейшая рефлексивная игра «Чет-нечет» при условии, что игроки потенциально способны использовать внутренние репрезентации поведения противника (например, РНС [8]), порождает огромное множество возможных цепочек событий, что исключает воспроизводимость экспериментальных ситуаций. Чтобы убедиться, что идентификация характерного временного ряда событий, с которым РНС сталкивается в процессе игры, принципиально реализуема, представляется удобным использовать комбинации фиксированных последовательностей ходов (ряды) в качестве «квази-противника» для РНС.

Цель настоящей работы – оценить возможность идентификации фиксированной временной последовательности стимулов, получаемых простой РНС в ходе рефлексивной игры «Чет-нечет», по паттерну нейронной активности.

Использовались простейшие РНС малого размера (15 нейронов), функционирование которых описывается формулами:

$$\alpha_i^{n+1} = \frac{\rho_i^n}{a + \left|\rho_i^n\right|}, \tag{1}$$

$$\rho_i^n = \sum_j w_{ij}\alpha_j^n + A_i^n, \tag{2}$$

где α_i^n – выходной сигнал i-того нейрона на n-ном такте; w_{ij} – матрица весовых коэффициентов; A_i^n – входной сигнал, поступивший на i-тый нейрон на n-ном такте; a – константа, определяющая крутизну переходной характеристики нейрона. Информация о ходе противника на предыдущем такте игры поступала в РНС по двум входам: 01, если противник выбрал «0», и 10, если «1». Аналогично соотношение сигналов двух выходных нейронов определяло ход самой РНС.

Использовалась квадратичная функция потерь:

$$C = \frac{1}{2}\sum_{i=1}^{2}(\alpha_i^n - \delta_i^n)^2, \quad (3)$$

где α_i^n и δ_i^n – имеющийся и требуемый сигналы на выходном нейроне РНС в момент времени n. Требуемый сигнал определялся в соответствии с тем, играла РНС за «Чет» или «Нечет»: в первом случае от сети требовалось сделать тот же ход, что противник, во втором – противоположный. В данной работе РНС играли за «Чет».

Обучение РНС проводилось по алгоритму обратного распространения ошибки с глубиной 5 ходов. Выбор данного алгоритма основан на его повсеместном использовании и относительной простоте реализации.

Комбинации фиксированных последовательностей стимулов (ряды), которые использовались как «квази-противники» РНС, приведены в следующей таблице.

Таблица

Фиксированные последовательности стимулов

Ряд 1	110011001100
Ряд 2	101100101100
Ряд 3	010011010011
Ряд 4	111000111000

Динамика нейронной активности РНС визуализировалась как траектория в пространстве нейронной активности, где каждая точка – состояние РНС на текущем шаге игры, а ее координаты – уровни возбуждения нейронов с соответствующим порядковым номером. Визуализация траекторий и обработка данных с помощью метода главных компонент проводилась в пакете Scilab (https://www.scilab.org/). Для генерации и функционирования РНС использовалась среда разработки Lazarus (https://www.lazarus-ide.org/).

РНС обучались игре «Чет-нечет» в режиме непрерывной подачи стимулов. Подаваемый на входы РНС ряд менялся через каждые 60 тактов, общая продолжительность обучения 2000 тактов. Опытным путем было определено, что

такого количества тактов достаточно, чтобы РНС описанной конфигурации обучались распознавать получаемый ряд и реагировать правильно (согласно игре за позицию «Чет» – делать тот же ход, что противник).

Обученные РНС способны решать вышеописанную задачу, следовательно, в паттернах нейронной активности РНС формируется представление сущности получаемого ряда, притом инвариантное к конкретному такту предъявления. Успешная реконструкция конкретного ряда, который получает РНС, по ее нейронной активности позволит доказать существование данного динамического инварианта.

Для идентификации получаемых РНС рядов применялся метод нейросетевого декодирования [7]. В качестве декодера использовались слоистые нейронные сети (ДН), имевшие 15 входов (по числу нейронов РНС, нейронная активность которых использовалась), 4 выходных нейрона с линейной функцией активации (4) и 6-14 нейронов скрытого слоя с сигмоидальной функцией активации (5):

$$f_o(x)=\begin{cases}0, & if\ x\le 0,\\ x, & if\ x>0\ \&\ x<1,\\ 1, & if\ x\ge 1.\end{cases} \tag{4}$$

$$f_h(x)=\frac{1}{2}\left(\frac{x}{a+|x|}+1\right). \tag{5}$$

Функция потерь ДН также была квадратичной (3). Для обучения использовался алгоритм обратного распространения ошибки. Отклик ДН определялся по номеру выходного нейрона с наибольшим сигналом.

Входными данными для ДН служили уровни возбуждения нейронов обученных РНС при обработке фиксированных временных рядов стимулов. Нейронная активность РНС записывалась построчно; каждая строка состояла из значений активности 15 нейронов РНС в данном цикле, всего 6 строк для 6 последовательных тактов. Далее каждая строка ставилась в соответствие с номером того ряда, который в тот момент подавался на вход РНС. Для обучения ДН использовалась нейронная активность РНС, зарегистрированная в стабильном ре-

жиме работы, т.е. после первых 10-12 тактов подачи текущего ряда. Всего для каждой обученной РНС было записано 96 строк. Строки подавались на входы ДН в случайном порядке. Поскольку нейронная активность каждой нейронной сети имеет индивидуальные особенности, обучение декодеров также проводилось индивидуально.

Оценка качества функционирования обученных ДН проводилась следующим образом. Нейронная активность РНС записывалась при получении рядов, по 50 тактов на каждый, затем последовательно подавалась на вход ДН. В таком режиме представлялось удобным сопоставлять результат декодирования, или реконструкции, с настоящим номером ряда. Порядок подачи данных на входы обученной ДН не влияет на качество реконструкции, поскольку из-за отсутствия обратных связей ДН неспособна сохранять информацию о предыдущих полученных данных в паттернах возбуждения.

В отличие от теста ОСО [7], ДН без скрытого слоя не справились с идентификацией рядов. Добавление скрытого слоя, т.е. преобразование линейного декодера в нелинейный, позволило обучить ДН с приемлемой точностью до 10^{-10}. Минимальное количество нейронов скрытого слоя при этом оказалось равным 6. ДН со скрытым слоем, состоящим из 6-14 нейронов, в режиме оценки качества функционирования реконструировали правильно от 60 до 80% данных. Для сравнения, ДН, идентифицировавшие стимул из теста ОСО, обучались до нулевой ошибки и демонстрировали качество функционирования 100%.

Отмечено, что дальнейшее увеличение числа скрытых слоев ДН не приводит к заметному изменению качества функционирования. Вопрос о том, как увеличить качество, предполагается рассмотреть в дальнейших исследованиях. В настоящей работе цель состояла в демонстрации самой возможности идентификации полученного ряда по нейронной активности РНС.

Неспособность линейных ДН идентифицировать ряд говорит о том, что информации, которую получают выходные нейроны в этом случае, недостаточно, и требуются дополнительные преобразования. Выходные нейроны РНС, получая при обработке рядов сигналы от остальных нейронов, решают задачу, ана-

логичную декодированию. Следовательно, они также нуждаются в дополнительной обработке полученных сигналов, чтобы сформировать отклик. Тогда существенным представляется не только текущий обрабатываемый стимул, но и предыдущие, информация о которых хранится в остаточном возбуждении нейронов РНС. Данный результат демонстрирует еще одно качественное отличие задачи (рефлексивной игры и ее моделирования при взаимодействии РНС с фиксированными рядами) от теста ОСО.

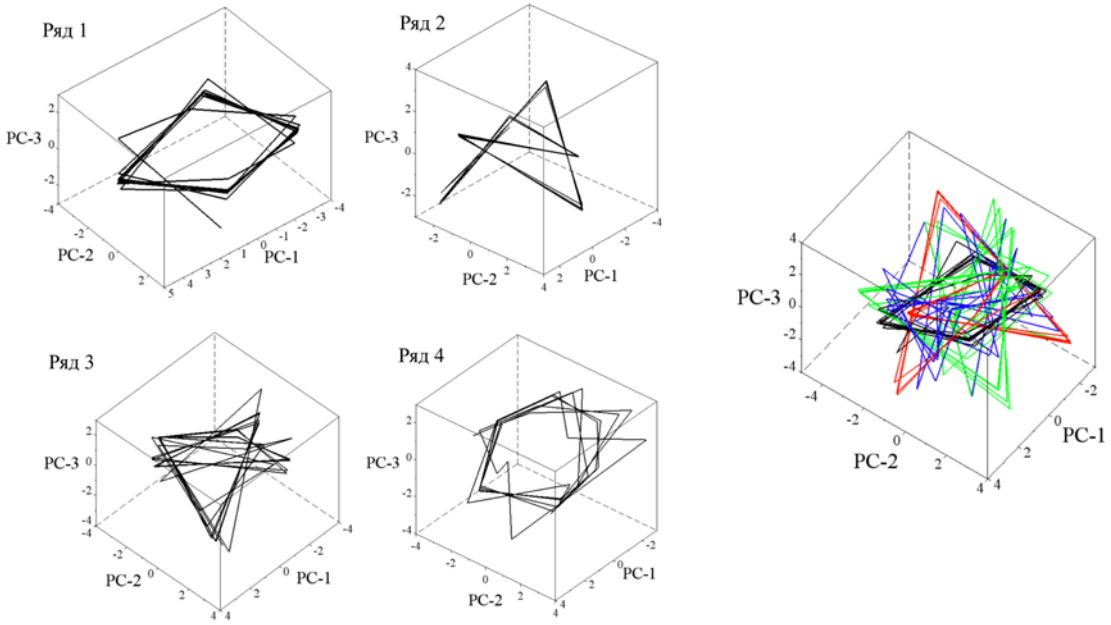


Рис. Примеры траекторий нейронной активности РНС. Слева – траектории, соответствующие каждому ряду по отдельности; справа – соответствующие четырем рядам на одном графике. Данные обработаны методом главных компонент.

В режиме оценки качества также было обнаружено, что обученные ДН не справляются с правильной реконструкцией в течение первых тактов предъявления нового ряда после его смены. Данный результат говорит о временной нестабильности паттерна нейронной активности РНС, возникающей при попытке «распознавания» РНС нового ряда. После того, как РНС удается «определить» ряд, паттерн стабилизируется и выходит к динамическому инварианту. В терминах нелинейной динамики данный процесс можно описать переходом от одного аттрактора нейронной активности (соответствующего прежнему получаемому

ряду) к другому. Визуализация нейронной активности РНС позволяет оценить вид этих аттракторов (см. Рис.).

Как видно из рисунка, аттракторы нейронной активности РНС являются сложными циклами. При этом имеются области, в которых траектории, соответствующие разным рядам, проходят достаточно близко друг к другу (см. Рис., справа). Существование таких областей – одно из возможных объяснений сравнительно низкой точности идентификации рядов по сравнению со стимулами ОСО. Наличие близких или пересекающихся участков траекторий нейронной активности может обуславливаться повторяющимися фрагментами сигналов в разных рядах: например, фрагмент 1100 встречается в рядах 1, 2, 4 (см. Таблицу). Проверка данной гипотезы будет проведена в дальнейших исследованиях.

Наконец, во входных данных, подаваемых на входы ДН в режиме оценки качества функционирования, треть значений была заменена на симулированные данные – случайные числа в диапазоне, соответствующем реальной нейронной активности. ДН реконструировали эти данные с высокой степенью уверенности (т.е. отклик на выходном нейроне близок или равен 1) как один из рядов. При этом у разных ДН, обученных на нейронной активности одной и той же РНС, этот ряд различался. Такой результат позволяет предположить, что в процессе обучения ДН на самом деле учится распознавать не четыре, а три класса данных; при этом в четвертый попадают все те входные данные, которые не подошли ни к одному из трех предыдущих. Аналогичный результат был также получен на ДН, обученных реконструировать стимул в ходе теста ОСО.

В работе показано, что фиксированная временная последовательность стимулов, получаемых простой рекуррентной нейронной сетью, может быть идентифицирована по паттерну нейронной активности с помощью нейросетевого декодера. Получена точность идентификации до 80%. Для данной задачи требуется нелинейный декодер (с одним и более скрытым слоем), что говорит о существенной роли остаточных уровней возбуждения, в которых кодируется информация о предыдущих полученных сетью стимулах. Траектория нейронной активности при получении сетью определенной последовательности стимулов имеет

циклический вид. Полученные результаты позволяют заключить, что при обработке фиксированного временного ряда стимулов формируется динамический инвариант – репрезентация данного ряда в нейронной активности сети, что в свою очередь говорит о наличии аттракторов нейронной активности, соответствующих каждой фиксированной последовательности событий.

Список литературы

1. Crick F. A framework for consciousness / F.Crick, C.Koch // Nature Neuroscience. – 2003. – Vol.6. – № 2. – P.119 – 126.
2. Meyers E.M. Dynamic population coding and its relationship to working memory / E.M.Meyers // Journal of Neurophysiology. – 2018. – Vol.120. – №. 5. – P.2260 – 2268.
3. Meyers E.M. Dynamic population coding of category information in inferior temporal and prefrontal cortex / E.M.Meyers, D.J.Freedman, G.Kreiman, E.K.Miller, T.Poggio // Journal of neurophysiology. – 2008. – Vol.100. – № 3. – P.1407 – 1419.
4. Barak O. Neuronal population coding of parametric working memory / O.Barak, M.Tsodyks, R.Romo // Journal of Neuroscience. – 2010. – Vol.30. – P.9424 – 9430.
5. Stokes M.G. Dynamic coding for cognitive control in prefrontal cortex / M.G.Stokes, M.Kusunoki, N.Sigala, H.Nili, D.Gaffan, J.Duncan // Neuron. – 2013. – Vol.78. – № 2. – P.364 – 375.
6. Miconi T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks / T.Miconi // Elife. – 2017. – Vol.6. – P.e20899.
7. Барцев С.И. Нейросетевое декодирование информации о внешнем стимуле по паттерну нейронной активности рекуррентной нейронной сети / С.И.Барцев, П.М.Батурина, Г.М.Маркова // Доклады Российской академии наук. Науки о жизни. – 2022. – Т.502. – №1. – С.48 – 53.
8. Bartsev S. Recurrent and multi-layer neural networks playing "Even-Odd": reflection against regression / S.Bartsev, G.Markova // IOP Conf. Series: Materials Science and Engineering. – 2020. – Vol.734. - № 1. – P. 012109.

КОНСТРУКТОР ПРОДУКЦИОННЫХ ЭКСПЕРТНЫХ СИСТЕМ С ЭЛЕМЕНТАМИ НЕЧЁТКОЙ ЛОГИКИ FLM_BUILDER И ИНТЕГРАЦИЯ ЕГО МОДЕЛЕЙ В ПОЛЬЗОВАТЕЛЬСКИЕ ПРОЕКТЫ

Н.А. Болсуновский, А.Д. Пронин, В.А. Углев

Сибирский федеральный университет, ***uglev-v@yandex.ru***

Изучение систем искусственного интеллекта стала для IT-профиля в современных ВУЗах нормой, даже если программа подготовки не специализируется на разработке программного обеспечения. Если, например, для постижения восходящих методов искусственного интеллекта (искусственных нейронных сетей, генетических алгоритмов и пр.) в учебных заведениях присутствуют достаточно развитые системы (тот же IBM Watson Studio Neural Network Modeler), позволяющие конструировать модели принятия решений и интегрировать их в свои проекты, то для нисходящих методов их значительно меньше. Обучение же должно поддерживаться практическими работами для актуальных для учащихся задачах [1]. Разработка экспертных систем (ЭС) с доской объявлений или на прецедентах практически всегда требует программирования даже на уровне реализации прототипа. Немного проще дело обстоит для продукционных ЭС и метода нечёткой логики (НЛ): существует ряд программ-конструкторов, позволяющих познакомиться с логикой создания таких систем без углубления в программирование (например, пакеты FuziCalc, iThink, Knowledge Craft, Kappa-PC, FLEX, ILOG RULES и даже MATLAB [2]). Но оптимизм по поводу массовости применения [3] таких инструментов до сих пор не оправдался: даже разработанные в таких мощных средах как CLIPS или Prolog ЭС проблематично интегрировать в пользовательские проекты. На этапе обучения основам инженерии знаний это создаёт дополнительные проблемы, поэтому опишем наш опыт её решения от настольных версий к сетевым (в виде сервиса).

Для проектирования продукционной ЭС требуется сформировать базу знаний и дать возможность разработчику через интерфейсы её обработать с помощью решателя. Так как инженер по знаниям и программист в процессе создания

ЭС должны являться разными людьми [4], то обучение широкого круга специалистов (включая аналитиков и системщиков) не должно делать упор на написание сложного программного кода. Важными требованиями является быстрота прототипирования, логичность и простота инструментария, наличие механизма НЛ, а также простота интеграции в приложения, не требующие наличия специализированной среды разработки ЭС. Для ЭС с продукционными правилами также важно контролировать полноту базы знаний, с которой учащийся должен работать в естественно-языковой форме. Учитывая приведенную специфику, покажем решение на примере программы FLM_Builder.

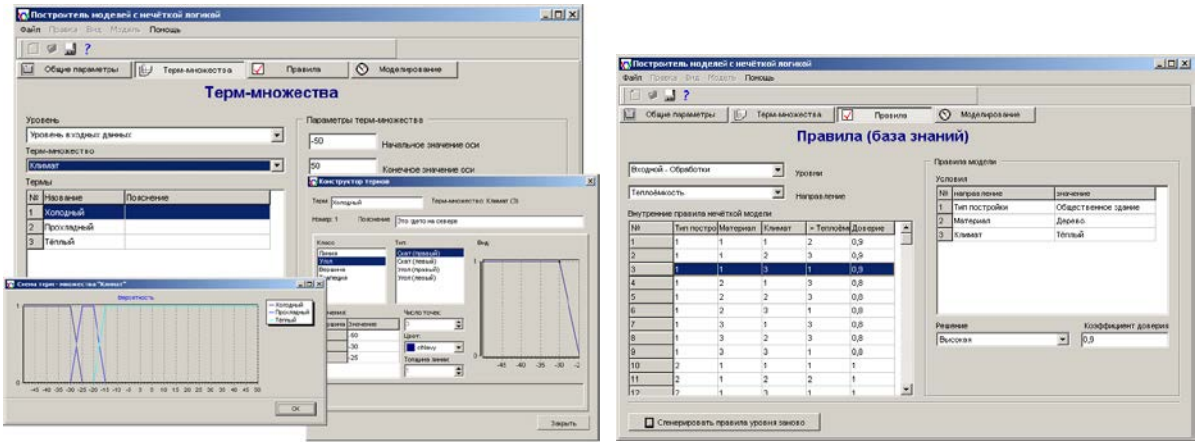


Рис. 1. Примеры экранных форм второго и третьего этапов описания модели (задача подбора типа утеплителя)

Программа Fuzzy Logic Model Builder была разработана в 2005 году [5] для обучения студентов Красноярского государственного технического университета (с 2007 г. Сибирского федерального университета) специальностей информационного профиля, не специализирующихся на разработке программного обеспечения. В её основе был положен принцип композиционного вывода [6], входы обрабатывались методом НЛ [7] (использован критерий минимума). Модуль работы с моделями был реализован в виде модуля на языке программирования Object Pascal (FLM_modul.pas), а интерфейсная часть в виде оконного приложения в среде Delphi. Создание ЭС происходило в режиме конструктора и производилось за 3 этапа (рис. 1): формирование архитектуры процесса рассуждений (формы дерева принятия решений); описание этапов принятия решений (включая кон-

струирование характеристических функций для фазификации); заполнение базы знаний (набора продукционных правил). Результирующий файл с расширением *.flm сохранялся в виде текстового файла (нотация конфигурационных фалов ini).

Созданную в приложении FLM_Builder модель ЭС пользователь мог интегрировать в произвольные приложения в среде Delphi: для этого ему было необходимо скопировать файл flm в каталог нового проекта и туда же разместить FLM_modul.pas, после чего загрузить модель и обратиться к функции её просчёта (вектор входных значений подавался в виде аргументов, см. левую часть рис. 2). Таким образом, весь объём работ по интеграции flm-модели ограничивался 5-7 строками типового кода [8]. Так как изначально все студенты обучались основам программирования в Delphi, то сложностей в организации простейших интерфейсов при выполнении практических работ по данной теме у них не возникало. Пример кода, необходимого для интеграции flm-модели приведен ниже (три фактора на входе, два ответа на выходе для модели MyESModel.flm):

```
uses FLM_modul;
…
procedure(…)
var
        MyES : TFuzzyModel;
    ESIn : RealArray;
    ESOut: StringArray;
begin
    …
    SetLength(ESIn,3);
    MIn[0] := StrToFloat(Edit1.Text);
    MIn[1] := StrToFloat(Edit1.Text);
    MIn[2] := StrToFloat(Edit1.Text);
    MyES := TFuzzyModel.Create;
    MyES := LoadModelFile('MyESModel.flm');
```

```
    SetLength(MOut,2);
    ESOut := CalcModel(MyES, ESIn);
    …
    MyES.Free;
end;
```

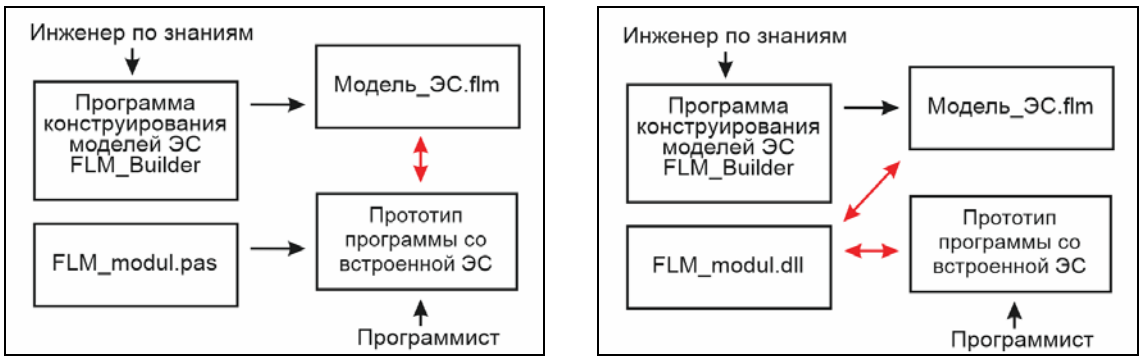


Рис. 2. Схема интеграции flm-файлов в пользовательские проекты на базе модуля FLM_modul.pas (слева) и его dll версии (справа)

Дальнейшее расширение номенклатуры языков программирования и специальностей студентов привело к тому, что потребовалось интегрировать flm-модели в код приложений за рамками учебного процесса (включая использование в проектах на языке C++). Для этого библиотека FLM_modul.pas была откомпилирована в виде dll-библиотеки (2009), что немного изменило технологию (см. правую часть рис. 2).

Необходимость интеграции flm-моделей в проекты на платформе 1C (2010) позволило отработать технологию совмещения dll с подходом «обертывания» (API Wrapper, см. левую часть рис. 3 [9]). А в 2015 г. выходит отдельная версия модуля для языка C++ (рис. 3 [10]). Объём и содержание кода, требующиеся для интеграции в проекты на C++, практически не изменились. Всё это позволило применять программу FLM-Builder для обеспечения практической части подготовки студентов различных уровней (специалисты, бакалавры, магистранты) в области инженерии знаний.

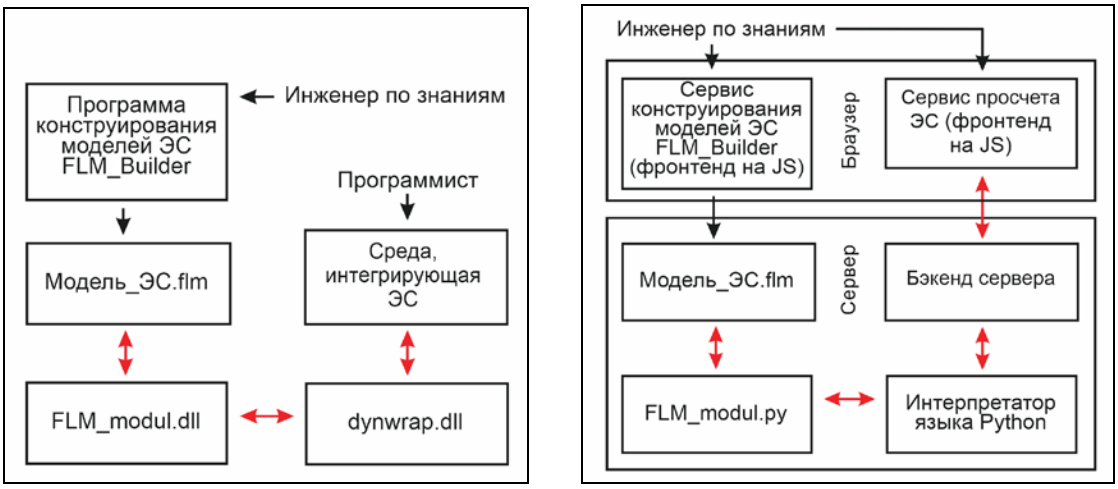


Рис. 3. Схема интеграции flm-файлов в пользовательские проекты на базе технологии Wrapper (слева) и для серверного приложения с интерпретатором языка Python (справа)

За 17 лет применения программы с её помощью было выполнено более 200 практических работ, подготовлено несколько десятков дипломных работ и магистерских диссертаций, а также реализован ряд научных проектов. В качестве примеров можно привести исследования в области информационной безопасности [11], автоматизации работы программно-аппаратных комплексов [12-13], повышении качества функционирования автоматизированных обучающих систем [14-17].

В процессе эксплуатации программы FLM_Builder и модулей интеграции основные подходы к разработке систем и технологии существенно изменились. Настольные и сетевые приложения начали выходить из «моды», уступая сервисному подходу (через интернет-формы). Первая попытка реализации такого функционала была успешно предпринята ещё в 2008 г. [11], но она была реализована не в сервисной архитектуре и нуждалась в полной переработке. Среди недостатков программы можно отметить следующие: ограничение числа слоев (этапов) логического вывода (нельзя сделать более двух); коллекция продукционных правил при переходе между слоями формируется как комбинация состояний всех анализируемых факторов (нельзя указать частичную связность); вызов пользовательских функций в процессе просчета не предусмотрен; просчет выво-

дит только итоговое решение ЭС без протоколирования промежуточных значений; сложности интеграции в код языка Python и веб-формы. Всё это потребовало поставить задачу перепроектирования FLM_Builder как сервиса (см. табл.).

Таблица.

Характеристика различных версий программы FLM_Builder

Версия, год	Модуль интеграции	Среды интеграции	Число слоев логич. вывода	Максимальное число сотояний гипотезы	Тип формата flm	Полносвязность
v1, 2005	FLM_modul.pas	Delphi	2	10	ini	Да
v2, 2010	FLM_modul.pas	Delphi	2	15	ini	Да
v2.5, 2010	FLM_ES_Calc.dll	1C и пр.	2	15	ini	Да
v3, 2015	FLM_ES_CCalc.cpp	C++	2	15	ini	Да
v4.b, 2022	FLM_modul.py	Python и JS	произвольное		xml	Нет

Новая версия программы реализована в виде модуля FLM_modul.py, для которого был написан визуальный конструктор на языке Java Script (JS) в виде web-сервиса. flm-модели стали сохраняться в нотации XML, который подгружается и просчитывается интерпретатором языка Python на стороне сервера. При проектировании ЭС инженер по знаниям взаимодействует с экранными формами на JS, а при необходимости просчета используется связка JS – серверный язык (например, PHP) – интерпретатор языка Python - FLM_modul.py – Модель_ЭС.flm (см. правую схему с рис. 3).

Специфика интерфейсной части программы-конструктора ЭС заключается в том, что последовательность логического вывода описывается в виде ориентированного графа методом drag&drop (см. рис 4 и 5, правая часть окна), а описание параметров узлов дерева принятия решений ЭС (рис. 4 слева) и самой базы знаний (рис. 5 слева) осуществляется через форму с настройками. Каждый узел

графа соответствует терм-множеству состояний анализируемого фактора. Не только входные терм-множества (зеленые кружки) могут включать наборы характеристических функций, но и промежуточные (розовые) и выходные (красные). За счет возможности связывать входные факторы с любым из последующих слоев логического вывода (см., например, рис. 6) достигается сокращение объёма базы знаний, состоящих из продукционных правил (см. примеры экранных форм на рис. 4 и 5). Характеристические функции для метода нечёткой логики также формируются в виде таблицы значений и в перспективе будут поддерживать загрузку из внешних файлов с первичными мнениями экспертов

На данный момент программная реализация FLM_Builder v4b находится в процессе опытной эксплуатации и доработки. Её развертывание с открытым доступом предполагается на базе учебного ресурса aesu.ru (функции конструирования, сохранения и открытия моделей из flm-файлов, просчет моделей). Загрузка модуля FLM_modul.py позволит использовать продукционные ЭС с элементами НЛ в новом формате *.flm как в программах на языке Python (по аналогии с правой схемой с рис. 2), так и в интернет-сервисах по схеме, приведенной на правой схеме с рис. 3.

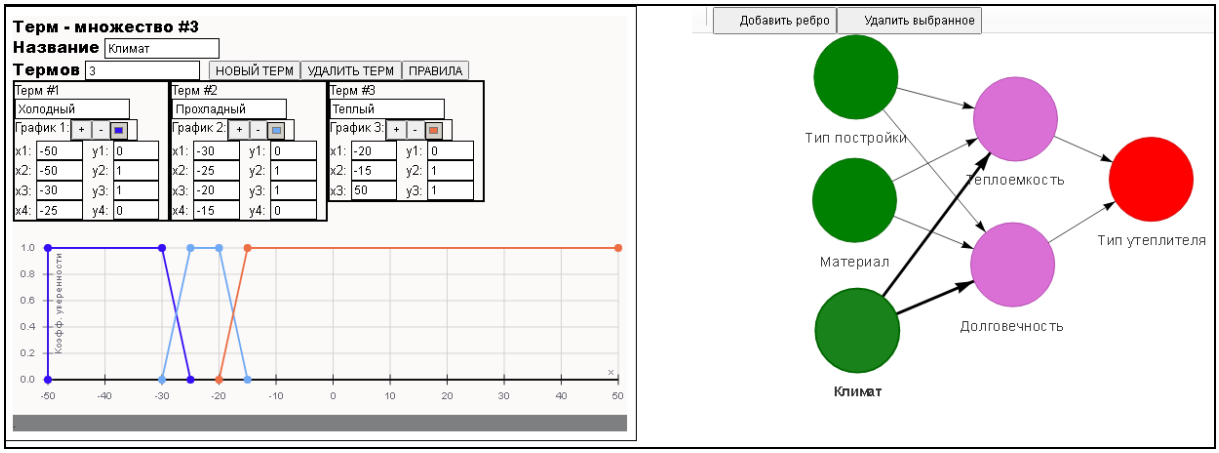


Рис. 4. Пример описания терм-множества (слева) в составе дерева принятия решений (справа) в окне браузера (задача подбора типа утеплителя)

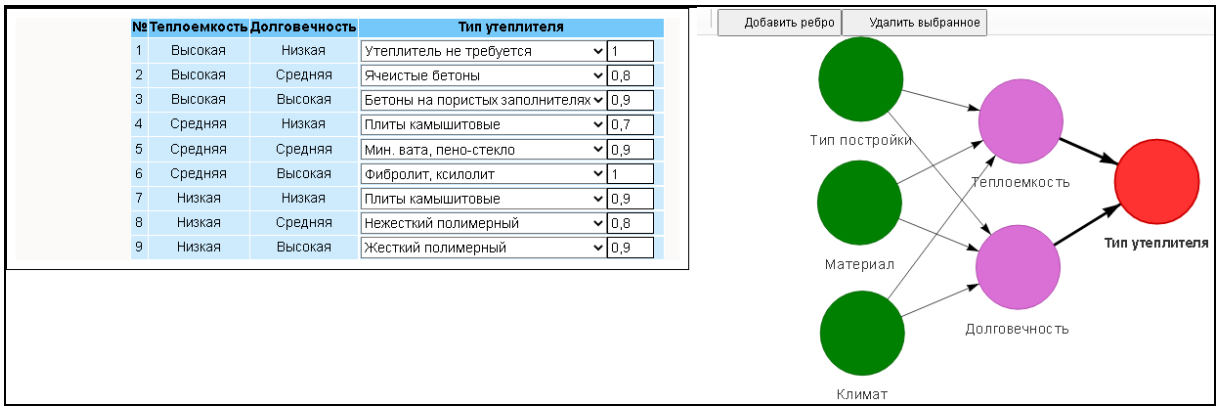


Рис. 5. Пример формирования набора продукционных правил для перехода от промежуточному к выходному уровню вывода в окне браузера (задача подбора типа утеплителя)

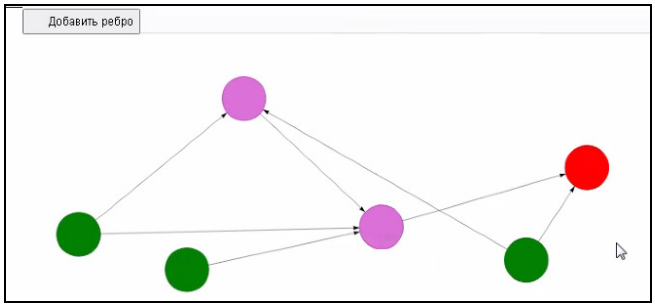


Рис. 6. Пример формирования дерева принятия решений произвольной конфигурации в окне сервиса конструирования ЭС FLM_Builder

Список литературы

1. Миркес Е.М., Углев В.А. Обучение методам искусственного интеллекта студентов информационного профиля // Интеллект и наука: Материалы XI Международной конференции. - Красноярск: Центр информации, 2011. – С. 112-114.
2. Садыкова А.М. обзор программ для создания экспертных систем // Студенческий научный форум: Материалы IX Международной конференции, 2017. https://scienceforum.ru/2017/article/2017037705.
3. Масалович А. И. Этот нечеткий, нечеткий, нечеткий мир //PC Week/RE. – 1995. – №. 16. – С. 35-36.
4. Гаврилова, Т. А. Базы знаний интеллектуальных систем / Гаврилова, Т. А. Хо-

рошевский, В. Ф. – СПб.: Питер, 2001. – 384 с.

5. Углев В.А. Модуль создания экспертных систем с нечеткой логикой для среды программирования Delphi «FLM_modul.pas» // Компьютерные учебные программы и инновации. Каталог. – 2006. – №12. – С. 9-10. http://viperson.ru/data/200902/ kupi122006.pdf
6. Джексон, П. Введение в экспертные системы: Пер. с англ. – М.: Вильямс, 2001. 624 с.
7. Zadeh L. Fuzzy Sets. Information and Control, 8(3), June 1965. – pp. 338-353.
8. Углев В.А., Добронец Б.С. Разработка экспертных систем с применением внешних модулей // Молодёжь и наука: начало XXI века: материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых учёных: в 3 ч. Ч. 1. - Красноярск: ИПЦ КГТУ, 2006. – С. 305 - 306.
9. Углева Е.В., Углев В.А. Интеллектуализация работы программ на платформе 1С // Интеллект и наука: Материалы X Международной научно-практической конференции. - Красноярск: ИПК СФУ, 2010. – С. 288-289.
10. Углев В.А., Брюханов С.И. Динамически подключаемая библиотека "FLM_ES_CCalc". – М.: Роспатент, 2015. – №2015610514 от 13.01.2015.
11. Сухинин Д.И., Углев В.А. Использование экспертных систем для классификации пользователей высоконагруженных информационных систем с распределенной архитектурой на примере тематических ресурсов // Нейроинформатика, ее приложения и анализ данных: XVI Всероссийский семинар. - Красноярск, 2008. – С. 142-145.
12. Самонов С.С., Углев В.А., Князькин Ю.М. Интеллектуальная бортовая система поддержки принятия решений транспортирования космических аппаратов // Исследования наукограда. – №2 (12). – 2015. – С.37-43.
13. Яковлева М.О., Углев В.А. Методическое и алгоритмическое обеспечение интеллектуальной системы управления в технологии «умный дом» // Современные инновационные технологии подготовки инженерных кадров для горной промышленности и транспорта 2016: Материалы международной конференции. - Днепропетровск: НГУ, 2016. – С. 252-258.

14. Углев В.А. Подход к интеграции модульных экспертных систем с пользовательскими приложениями на примере создания проекта для адаптивного тестирования // Современные техника и технологии: Материалы XIII международной научной конференции. В 3 т. Т. 2. - Томск: ТПУ, 2007. – С. 454-456.
15. Савицкая А.А., Углев В.А. Методика автоматизированного синтеза перечня литературы для научных исследований с помощью систем поддержки принятия решений // Инновационное развитие науки и образования: Материалы II Международной научно-практической конференции. Ч. 1. - Пенза: Наука и просвещение, 2018. – С. 91-94.
16. Дудкина М.В., Углев В.А. Результаты эксперимента по индивидуализации маршрута изучения дисциплины // Евразийская педагогическая конференция: Материалы II Международной конференции. - Пенза: Наука и просвещение, 2018. – С. 26-29.
17. Фоминых Н. А. Результаты эксперимента по индивидуализации состава электронного учебного курса «Моделирование систем» с использованием механизма рефлексии //Робототехника и искусственный интеллект: Материалы XII Всероссийской конференции. – Красноярск: ЛИТЕРА-принт, 2020. – С. 240-247.

АНАЛИЗ МЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ МОДЕЛИ РЕАНАЛИЗА NCEP GFS ДЛЯ АТМОСФЕРЫ Г. КРАСНОЯРСКА [2]

О.С.Володько[1], Н.А.Буряк[2], А.В.Дергунов[3]

[1]Институт вычислительного моделирования СО РАН, ***osv@krasn.ru***

[2]Сибирский федеральный университет, Институт математики и фундаментальной информатики

[3]ФИЦ КНЦ СО РАН, Красноярск, Россия, ***alexdergunov@icm.krasn.ru***

По сведениям Минприроды РФ, Красноярск считается одним из нескольких населенных пунктов РФ с наиболее грязным воздухом, концентрация загрязняющих веществ в атмосфере города нередко выше допустимых норм [1].

Загрязнение атмосферного воздуха по данным многочисленных исследований оказывает заметное воздействие на здоровье населения – может существенно ухудшать функцию легких и вызывать обострения бронхиальной астмы, повышать риски развития рака легких, острых сердечно-сосудистых осложнений, развития заболеваний коронарных артерий, и проч. [2]

Анализ метеорологических данных может внести заметный вклад в исследование механизмов формирования и распространения загрязнений в атмосфере города [3].

Для исследования и прогнозирования периодов повышенной концентрации загрязняющих веществ в пограничном слое атмосферы Красноярска предполагается использование метеорологических данных реанализа глобальной модели атмосферы GFS (Global Forecast System), разработанной NCEP (National Centers for Environmental Prediction, США) [4]. В составе этих данных – несколько сотен слоев с характеристиками атмосферы на различных вертикальных уровнях, которые вычисляются на регулярной сетке с пространственным разрешением 0.25° (~25 км) с периодичностью 4 раза в сутки. Модель NCEP GFS считается одной из лучших моделей атмосферы, широко используется многими исследователями [5, 6].

Для сокращения размерности большого объема метеорологических дан-

[2] Исследование выполнено за счет средств гранта РНФ и Красноярского краевого фонда науки № 22-21-20117

ных, полученных с модели реанализа NCEP GFS, были проведены корреляционный анализ и анализ методом главных компонент. Для исследования были взяты наборы метеорологических данных на разных вертикальных уровнях – от поверхности земли до высоты 3000 м, всего получилось 157 метеорологических параметров (значения температуры, влажности воздуха, давления, ветра, количества осадков, облачности, показатели влажности почвы, концентрации озона в атмосфере и др.). Данные были получены с июня 2019 г. по март 2022 г.

Корреляционный анализ был проведен между одинаковыми метеорологическими параметрами на разных слоях, что позволило убрать факторы с высокой корреляцией и избавиться от мультиколлинеарности. Для расчета корреляции между признаками использовался коэффициент корреляции Пирсона [7]. На Рис.1, в качестве примера, представлена матрица корреляций влажности воздуха на различных слоях.

Для определения существования линейной зависимости между коррелируемыми данными были построены диаграммы рассеяния между одинаковыми метеорологическими параметрами на разных слоях. Пример таких диаграмм для влажности воздуха представлен на рис.2-3.

Влажность воздуха 1000 мбар	0,50	0,35	0,21	0,11	0,04
Влажность воздуха 975 мбар	0,58	0,42	0,28	0,17	0,09
Влажность воздуха 950 мбар	0,78	0,60	0,45	0,34	0,24
Влажность воздуха 925 мбар	0,95	0,79	0,63	0,49	0,39
Влажность воздуха 900 мбар	1,00	0,90	0,74	0,58	0,46
Влажность воздуха 850 мбар	0,90	1,00	0,89	0,71	0,56
Влажность воздуха 800 мбар	0,74	0,89	1,00	0,90	0,73
Влажность воздуха 750 мбар	0,58	0,71	0,90	1,00	0,90
Влажность воздуха 700 мбар	0,46	0,56	0,73	0,90	1,00
Влажность воздуха 1000 мбар	0,50	0,35	0,21	0,11	0,04
Влажность воздуха 975 мбар	0,58	0,42	0,28	0,17	0,09
Влажность воздуха 950 мбар	0,78	0,60	0,45	0,34	0,24
Влажность воздуха 925 мбар	0,95	0,79	0,63	0,49	0,39
Влажность воздуха 900 мбар	1,00	0,90	0,74	0,58	0,46
Влажность воздуха 850 мбар	0,90	1,00	0,89	0,71	0,56
Влажность воздуха 800 мбар	0,74	0,89	1,00	0,90	0,73
Влажность воздуха 750 мбар	0,58	0,71	0,90	1,00	0,90
Влажность воздуха 700 мбар	0,46	0,56	0,73	0,90	1,00
	Влажность воздуха 900 мбар	Влажность воздуха 850 мбар	Влажность воздуха 800 мбар	Влажность воздуха 750 мбар	Влажность воздуха 700 мбар

Рис.1. Матрица корреляций значений влажности воздуха на различных слоях.

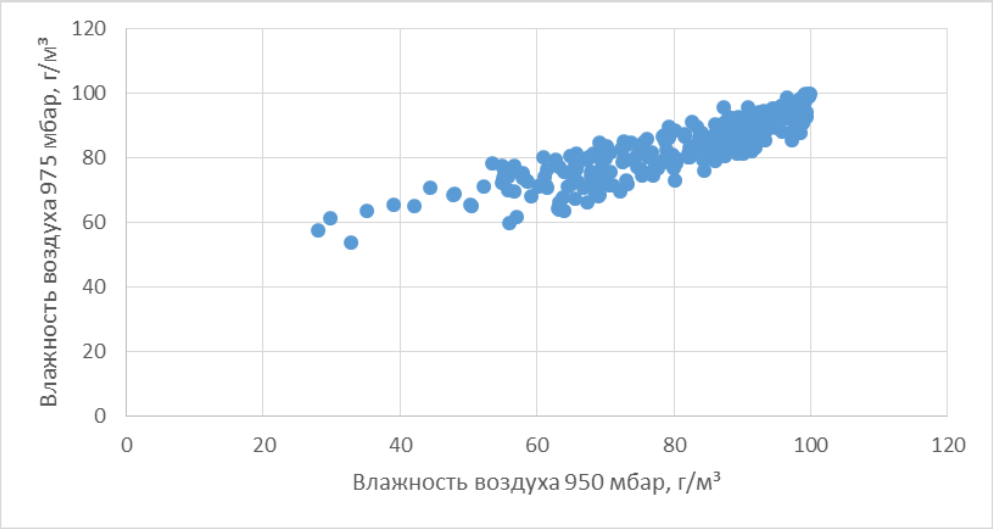


Рис.2. Диаграмма рассеяния между влажностью воздуха на слое 1000 мбар и влажностью воздуха на слое 950 мбар.

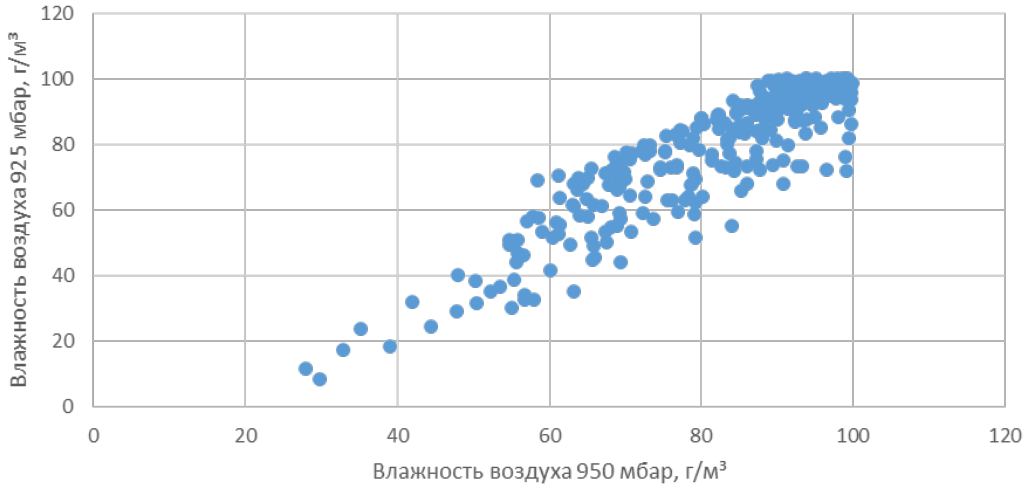


Рис.3. Диаграмма рассеяния между влажностью воздуха на слое 925 мбар и влажностью воздуха на слое 950 мбар.

Для дальнейшего сокращения размерности данных, к оставшимся после корреляционного анализа 58 метеорологическим параметрам, был применен метод главных компонент, который позволяет уменьшить размерность данных, сохранив при этом максимум информации [8].

В таблице показано, что в результате применения метода главных компонент первые 18 главных компонент содержат 90 % дисперсии.

Таблица

Распределение дисперсии по главным компонентам

Главная компонента	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Процент дисперсии	29,84%	13,37%	11,21%	7,45%	4,77%	4,27%	3,11%	2,85%	2,24%	1,78%	1,59%	1,43%	1,35%	1,25%	1,12%	0,91%	0,84%	0,80%

Таким образом, с помощью корреляционного анализа были отобраны 58 метеорологических параметров из 157 исходных. Метод главных компонент позволил сократить размерность данных с 58 метеорологических параметров до 18 главных компонент.

Полученные главные компоненты могут быть использованы для построения регрессии главных компонент [9] с целью предсказания периодов, способствующих накоплению вредных (загрязняющих) веществ в приземном слое атмосферного воздуха.

Список литературы

1. Eremkin A.I. Regulation of emissions, pollutants into the atmosphere / A.I.Eremkin, I.M.Kvashin, U.I.Unkerov. —Assoc. of Constr. Univers, Moscow, 2001. — 176 p.
2. Kaufman Y.J. A satellite view of aerosols in the climate system / Y.J.Kaufman, D.Tanré, O.Boucher // Nature. — 2002. — V. 419. — №. 6903. — P. 215 – 223.
3. Матвеев Л.Т. Общая метеорология. Физика атмосферы / Л.Т.Матвеев. — Ленинград: Гидрометеоиздат. — 1984. — 753 с.
4. The Global Forecast System (GFS). Documentation Retrieved from: https://www.emc.ncep.noaa.gov/emc/pages/numericalforecast_systems/gfs.php
5. Durai V.R. Prediction of Indian summer monsoon in short to medium range time scale with high resolution global forecast system (GFS) T574 and T382 / V.R. Durai, S.K. Roy Bhowmik // Climate dynamics. — 2014. — V. 42. — №. 5. — P. 1527 – 1551.

6. Shin U. Predictability of PM2.5 in Seoul based on atmospheric blocking forecasts using the NCEP global forecast system / S.H.Park, J.S.Park, J.H.Koo, C.Yoo, S.Kim, J.B.Lee // Atmospheric Environment. — 2021. — V. 246. — p. 118141.
7. Chicco D. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation / D.Chicco, M.J.Warrens, G. Jurman // PeerJ Comput. Sci. — 2021. — V. 7. — p. e623.
8. Jolliffe I.T. Principal Component Analysis. / I.T.Jolliffe. — Springer, New York, 2002 — 487 p.
9. Zhou Q. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network / H.Jiang, J.Wang, J.Zhou. // Science of the Total Environment. — 2014. — V. 496. — P. 264-274.

ОБРАБОТКА ДАННЫХ ДАТЧИКА MSU-GS КОСМИЧЕСКОГО АППАРАТА «АРКТИКА-М1» С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

О.А. Дубровская[1], Е.В. Пермяков[2], Д.А. Можаров[2], И.А. Галочкин[3], А.Г. Окунев[4], В.Ю. Кудинов[5]

[1]Федеральный исследовательский центр информационных и вычислительных технологий, ***dubrovskaya_oa@list.ru***

[2]ФГБУ «Научно-исследовательский центр космической гидрометеорологии «Планета» Сибирский центр, ***mozharov.daniil.a@gmail.com, permyakovyegor1204@gmail.com***

[3]Группа Компаний «Карбис», ***galochkinia@mer.ci.nsu.ru***

[4]Высший Колледж Информатики Новосибирского Государственного Университета, ***okunev73@mail.ru***

[5]Новосибирский государственный университет, ***v.kudinov@g.nsu.ru***

В связи с запуском в феврале 2021 года отечественного космического аппарата (КА) «Арктика-М1» возникла необходимость создания технологий обработки данных спутника. На данном спутнике установлено многозональное сканирующее устройство гидрометеорологического обеспечения MSU-GS для получения мультиспектральных изображений облачности и поверхности Земли в видимом и ИК диапазонах.

Существует множество методов и алгоритмов выделения объектов по цифровым данным [1]. Для обработки спутниковых снимков, часто прибегают к показательным индексам и ансамблевым алгоритмам кластеризации по спектральным и пространственным характеристикам. Распознавание объектов и кластеризация изображений датчика MSU-GS КА «Арктика-М1» проводилось с использованием нейронных сетей. Трудности классификации данных связаны с изменчивостью признаков – отражательная способность меняется в зависимости от времени суток, сезона, полосы захвата спектрометров, положения Солнца и так далее. Часто классификация бывает неопределенной, поскольку элементы

растра могут принадлежать сразу к нескольким классам – это так называемые «смешанные элементы».

Целью данной работы является разработка технологии выделения различных объектов таких как вода, суша, облачные структуры, снег/лед с использованием нейронных сетей по данным КА «Арктика-М1».

Для выделения водных объектов были сформированы маски с использованием индексов NDWI (Normalized Difference Water Index) в двух вариациях [2]:

- С использованием ближнего инфракрасного канала

$$NDWI = \frac{(X_{nir} - X_{swir})}{(X_{nir} + X_{swir})}$$

- С использованием зеленого канала

$$NDWI = \frac{(X_{green} - X_{nir})}{(X_{green} + X_{nir})}$$

Здесь для КА Landsat X_{nir} - канал с длиной волны 0,845—0,885 мкм , X_{swir} - 1,560 -1,660 мкм, X_{green} - 0,525—0,600 мкм.

В первом варианте данный индекс используется для определения содержания воды в растениях, во втором для определения водных тел.

Формирование масок для обучающей выборки с использованием NDWI позволило выделить класс суша/вода, так как яркостная температура отличается от температуры облаков и снега. При выделении водных объектов и суши на снимках остаются классы облачность, снег/лед.

В работе был использован метод transductive transfer learning [3], который позволяет обучить нейронную сеть на данных MODIS для кластеризации данных MSU-GS. Для сегментации изображения и кластеризации объектов была выбрана архитектура сверточной нейронной сети U-net, стандартная для решения таких задач.

Для обучения сети при формировании датасета были использованы снимки, полученные инструментом MODIS в инфракрасных каналах, диапазоны

длин волн которого близки к диапазонам в каналов MSU-GS (Таблица). Обучающая выборка строилась на данных в инфракрасных диапазонах, так как значительная часть зоны видимости КА Арктика может находиться в области тени для видимого участка спектра.

Таблица

Каналы и диапазоны волн MSU-GS и MODIS

Каналы MSU-GS	Длины волн (мкм)	Каналы MODIS	Длины волны(мкм)
4	3.50 - 4.00	20	3.660 - 3.840
5	5.70 - 7.00	27	6.535 - 6.895
6	7.50 - 8.50	28	7.175 - 7.475
7	8.20 - 9.20	29	8.400 - 8.700
8	9.20 - 10.2	30	9.580 - 9.880
9	10.2 - 11.2	31	10.780 - 11.280

Обучение нейронной сети проходило в 4 этапа для получения наибольшей точности предсказаний. На первом этапе использовались 200 файлов, сформированные из снимков MODIS, содержащие данные за сутки, без какой-либо фильтрации по времени суток. На втором этапе был увеличен объем обучающей выборки 2300 файлов. На третьем этапе из обучающей коллекции были исключены снимки, относящиеся к ночному времени суток. На четвертом этапе оставлены только дневные снимки, исключены снимки переходного периода (восход/закат), так как при этом происходит некорректный расчет NDWI. В итоге обучающая выборка составила 1824 снимка, из них 365 отобрано для валидации.

В представленной работе нейронная сеть обучена на основе коррекции ошибок, пройдя, так называемое обучение с учителем, в замкнутой системе с обратной связью, без включения окружающей среды.

Четвертый этап обучения нейронной сети проходил на кластере с 1Тб оперативной памяти и видеокартой Tesla V100. Процесс обучения занял 16 часов и 139 эпох, точность при обучении достигла 82%.

Таким образом, предложена технология для выделения класса вода/суша при инвертировании которого получается класс облачность/снег/лед, на спутниковых снимках КА «Арктика-М1». Применение transductive transfer learning при использовании сверточной нейронной сети U-net позволило выделить данные объекты. На рисунке 1 справа представлена маска класса вода/суша при обработке снимка по предсказанию нейронной сети, слева изображение этого же снимка в RGB каналах.

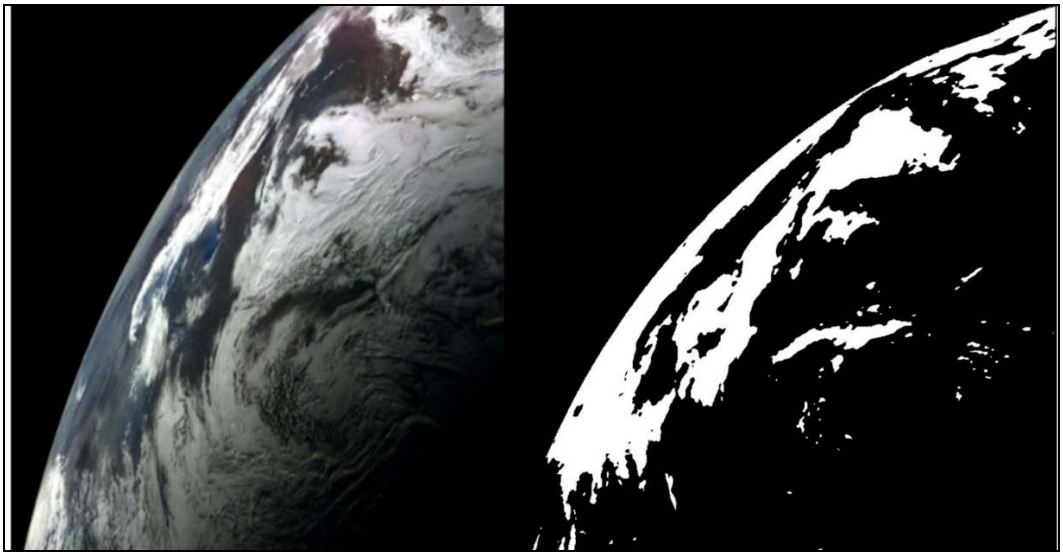


Рис.1. Изображения снимка КА «Арктика-М1» в RGB диапазонах слева, бинарная маска на основе предсказания нейронной сети (белое - класс вода+суша) справа.

Для валидации рассчитанных масок было проведено попиксельное сравнение с продуктами, полученными по различным алгоритмам на тот же момент времени (Рис. 2). Сравнение проводилось по данным КА Metop-B и Meteosat.

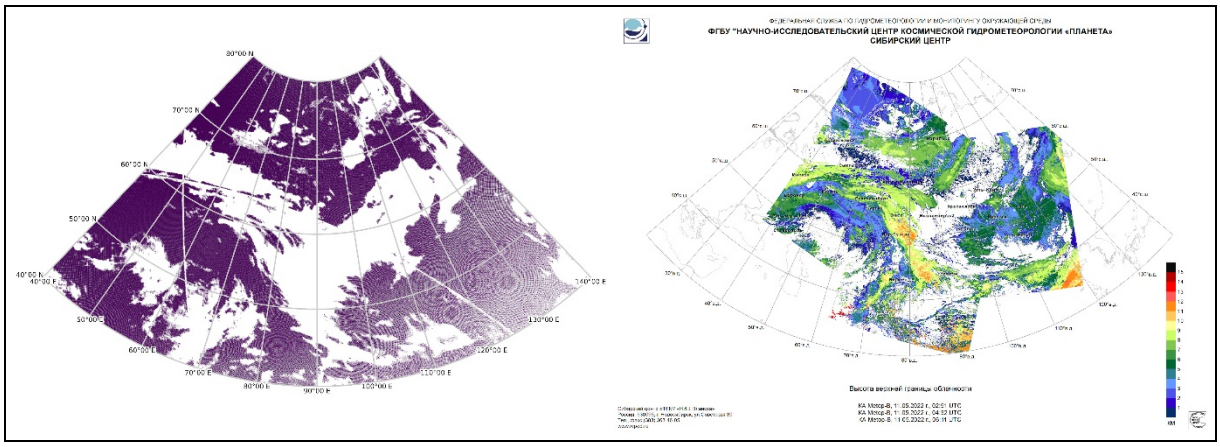


Рис.2. Маска облачности КА «Арктика-М1» полученная по НС слева, высота верхней границы облачности по данным КА Metop-B справа

Таким образом, предложена технология для выделения различных классов на спутниковых снимках КА «Арктика-М1», которые представляют интерес для дальнейшей обработки и получения различных продуктов в области метеорологии.

Список литературы

1. Sinyavskiy Yu.N. Experimental evaluation of nonparametric clustering algorithms for image segmentation / Yu.N. Sinyavskiy, S.A. Rylov, I.A. Pestunov. — E3S Web of Conferences, 2020. — Vol.223. — Art.02008. ISSN 2267-1242. https://doi.org/10.1051/e3sconf/202022302008

2. Bo-cai Gao NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. / Bo-cai Gao. — Remote Sensing of Environment, 1996. — P. 257-266. ISSN 0034-4257, https://doi.org/10.1016/S0034-4257(96)00067-3.

3. Arnold A. A Comparative Study of Methods for Transductive Transfer Learning / A. Arnold, R. Nallapati and W. W. Cohen. — Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), 2007. — P. 77-82. https://doi.org/10.1109/ICDMW.2007.109

АЛГОРИТМ ОБУЧЕНИЯ СЕГМЕНТНОЙ СПАЙКОВОЙ МОДЕЛИ НЕЙРОНА ДЛЯ РЕАЛИЗАЦИИ ИНКРЕМЕНТНОГО ОБУЧЕНИЯ

Е.А.Еременко[1], А.М.Корсаков[1], А.В.Бахшиев[2]

[1]Центральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики (ЦНИИ РТК), ***a.korsakov@rtc.ru, elizaveta.yeremenko@gmail.com***

[2]Санкт-Петербургский политехнический университет Петра Великого (СПбПУ), ***palexab@gmail.com***

Введение

В рамках нейроморфного подхода разрабатываются системы, которые строятся на принципах работы биологических нейронных сетей. Примером реализации нейроморфного подхода являются спайковые нейронные сети, которые воспроизводят импульсную природу биологических нейронных сетей. Как правило, существующие нейронные сети, в том числе спайковые, адаптируются к новым данным только за счёт изменения числа нейронов и связей между ними в сети [1], [2], в то время как в биологических нейронных сетях может происходить как изменение топологии нейронных структур, так и дендритных деревьев отдельных нейронов.

Сегментная спайковая модель нейрона (ССМН) [3] была разработана для систем управления и обработки информации. Устройство модели нейрона позволяет производить обучение нейронных сетей, построенных на ней, как за счёт изменения топологии сети, так и на уровне структур отдельных нейронов. Разработаны архитектуры нейронных сетей, построенные на ССМН, позволяющие решать задачу классификации на малой обучающей выборке [4]. Однако существующий алгоритм структурного обучения ССМН [5] позволяет использовать модель только в оффлайн сценарии, так как алгоритм предполагает изменение структуры нейрона только путём увеличения числа сегментов, а само обучение предполагается только один раз за всё время использования модели. Данная работа предлагает новый алгоритм обучения ССМН, позволяющий использовать модель нейрона в сценариях инкрементного обучения [6].

Сегментная спайковая модель нейрона

ССМН допускает большое разнообразие структур, получаемых из стандартных сегментов. В данной работе используется структура модели нейрона, представленная на рисунке 1. Используемая структура не предполагает ветвления дендритов и исходит из того, что входные сигналы будут поступать только на возбуждающие синапсы входных сегментов дендритов.

Модель нейрона состоит из дендритов, тела нейрона и генераторной зоны, где при превышении суммарным потенциалом на теле нейрона определенного порога происходит генерация выходного импульса. Дендриты и тело нейрона состоят из сегментов. Сегменты дендритов и тела нейрона также являются составными элементами: каждый сегмент состоит из минимум одного тормозного и одного возбуждающего синапсов. Модели тормозных и возбуждающих синапсов идентичны по устройству и обработке сигнала, но имеют разное воздействие на нейрон. Возбуждающие синапсы оказывают гиперполяризующее, положительное воздействие на нейрон, а тормозные – деполяризующее, отрицательное воздействие. На рисунке 1 возбуждающие синапсы намерено нарисованы вне входных сегментов дендритов.

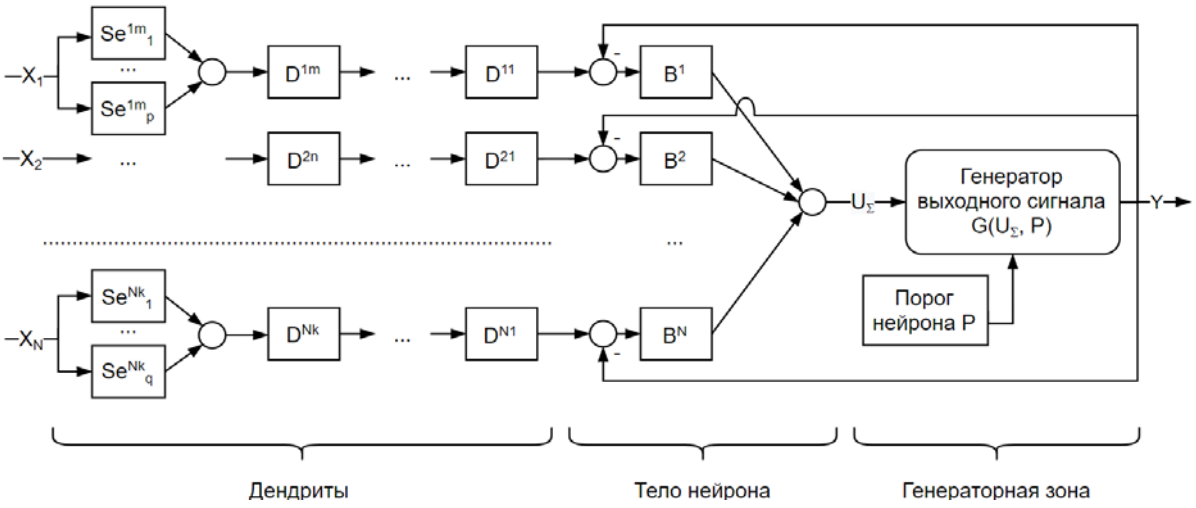


Рис.1. Структура сегментной спайковой модели нейрона. Se_l^{ij} – входной синапс l, связанный с сегментом дендрита D^{ij}; D^{ij} – j-й сегмент i-о дендрита; B^k – k-й сегмент тела нейрона.

ССМН используется в нейронных сетях или других моделях, где после обучения ССМН заданному паттерну импульсов модель нейрона способна с некоторой погрешностью описывать класс паттернов, к которому принадлежал заданный паттерн. Паттерны импульсов получаются путём увеличения размерности на 1 и последующего кодирования исходных данных методом, описанным в работе [4].

Обучение ССМН паттерну импульсов состоит в изменении структуры модели нейрона таким образом, чтобы для данного паттерна импульсов максимизировать значение суммарного потенциала на входе в генераторную зону нейрона. Для достижения этого необходимо, чтобы сигналы с тела нейрона поступали на вход генераторной зоны одновременно. Это достигается задержкой всех импульсов кроме последнего поступившего на тело нейрона. Для изменения времени поступления на вход генераторной зоны импульсов изменяется число сегментов в дендритах. Добавление дополнительного сегмента в дендрит приводит к задержке поступления сигнала на тело нейрона и уменьшению максимального значения сигнала. Для компенсации потери максимального значения сигнала на входные сегменты дендритов добавляются дополнительные возбуждающие синапсы. Так происходит изменение всех дендритов, кроме дендрита, на который приходил импульс, поступающий на тело нейрона последним. Этот дендрит называется калибровочным. Кодирование исходных данных проводится так, что калибровочным всегда является последний дендрит – дендрит с индексом N. Весь процесс обучения состоит из последовательности итераций. Одна итерация обучения представляет собой получение на вход нейрона паттерна для обучения, обработки нейроном входного паттерна и изменение структуры модели нейрона для лучшего соответствия паттерну.

Существующий алгоритм принятия решения о том, как изменять структуру, исходит из предположения, что нейрон перед началом обучения будет находиться в исходном состоянии: тело нейрона состоит из N сегментов; нейрон имеет N дендритов, состоящих из одного сегмента; сегменты тела нейрона и дендритов имеют только по одному тормозному и одному возбуждающему си-

напсу. Это не позволяет обучать уже обученную чему-либо модель нейрона другим примерам, или получать структуры нейрона путём последовательного неполного изучения нескольких примеров. Для реализации инкрементного обучения ССМН необходимо исходить из следующих положений: обучение модели может происходить из произвольного состояния модели; обучение может начаться и закончиться в любой момент, в том числе до того, как модель полностью изучит некоторый паттерн; в процессе обучения может происходить изменение паттерна, которому должна обучаться модель нейрона.

Предлагаемый алгоритм обучения модели нейрона

Перед началом процесса обучения ССМН находится в некотором произвольном состоянии: модель состоит из тела нейрона, генераторной зоны, N дендритов, каждый из которых состоит из некоторого числа сегментов, и на входных сегментах дендритов имеется некоторое произвольное число синапсов.

Пусть на итерации t на вход нейрона поступает некоторый паттерн эталонных импульсов, имеющих задержки d_1^t, d_2^t, …, d_N^t. На каждой последующей итерации обучения для каждого дендрита происходит увеличение или уменьшение на один, или не изменение количества сегментов в дендрите и синапсов на входном сегменте дендрита, то есть процессы синхронизации и нормализации существующего алгоритма [5] в предложенном алгоритме происходят одновременно, а не последовательно. Для определения того, как изменять структуру, на входе в каждый сегмент тела нейрона фиксируется значение максимума внутриклеточного потенциала и время его фиксации. Для нахождения оптимального числа сегментов дендрита j целью является минимизация модуля разности времени фиксации максимума значения внутриклеточного потенциала τ_N^t на сегменте тела нейрона, связанном с калибровочным дендритом, и времени фиксации максимума внутриклеточного потенциала τ_j^t на j-м сегменте тела нейрона:

$$\left|\tau_N^t - \tau_j^t\right| \longrightarrow \min. \qquad (1)$$

Для нахождения оптимального числа синапсов на входном сегменте дендрита j для данного числа сегментов дендритов целью является минимизация

разности в значениях текущего максимума внутриклеточного потенциала h_j^t и его значения h_N^t для сегмента тела нейрона, связанного с калибровочным дендритом:

$$\left|h_N^t - h_j^t\right| \longrightarrow \min. \qquad (2)$$

Процессы принятия решения об изменении числа сегментов дендритов и синапсов и фактического изменения их разделены: принятие решения об изменении происходит в конце текущей итерации по результатам измерений значения сигналов на текущей итерации, а реализация принятых решений в начале следующей итерации.

Алгоритмы принятия решения о том, как необходимо производить изменение числа сегментов и синапсов в дендрите, представлены в форме псевдокода с использованием математических обозначений на рисунках 2.А и 2.Б соответственно.

Для сохранения информации о принятых решениях, как изменять структуру нейрона, до момента их реализации алгоритм использует дополнительные переменные DSt_j и $SeSt_j$, показывающие насколько в дендрите j необходимо изменить число сегментов и синапсов соответственно. В первых строках обоих алгоритмов (рис. 2.А и рис. 2.Б) вычисляются значения переменных w_j^t и b_j^t, значения и знаки которых затем используются в алгоритме. Функция sgn(f) возвращает знак f. Знак переменных позволяет определить, в каком направлении, увеличения или уменьшения, необходимо изменять число сегментов в дендритах или синапсов на входных сегментах дендритов.

Настройки ССМН могут быть разными, поэтому алгоритм должен быть независимым от значений параметров сегментов дендритов и синапсов и не может при окончании процесса обучения опираться только на абсолютные значения. Однако алгоритм позволяет закончить обучение раньше, используя следующие значения, задаваемые пользователем: e_2 – неотрицательное число, при не превышении которого считается, что число сегментов в дендрите оптимально; e_3 – неотрицательное число, при не превышении которого считается, что число

синапсов оптимально.

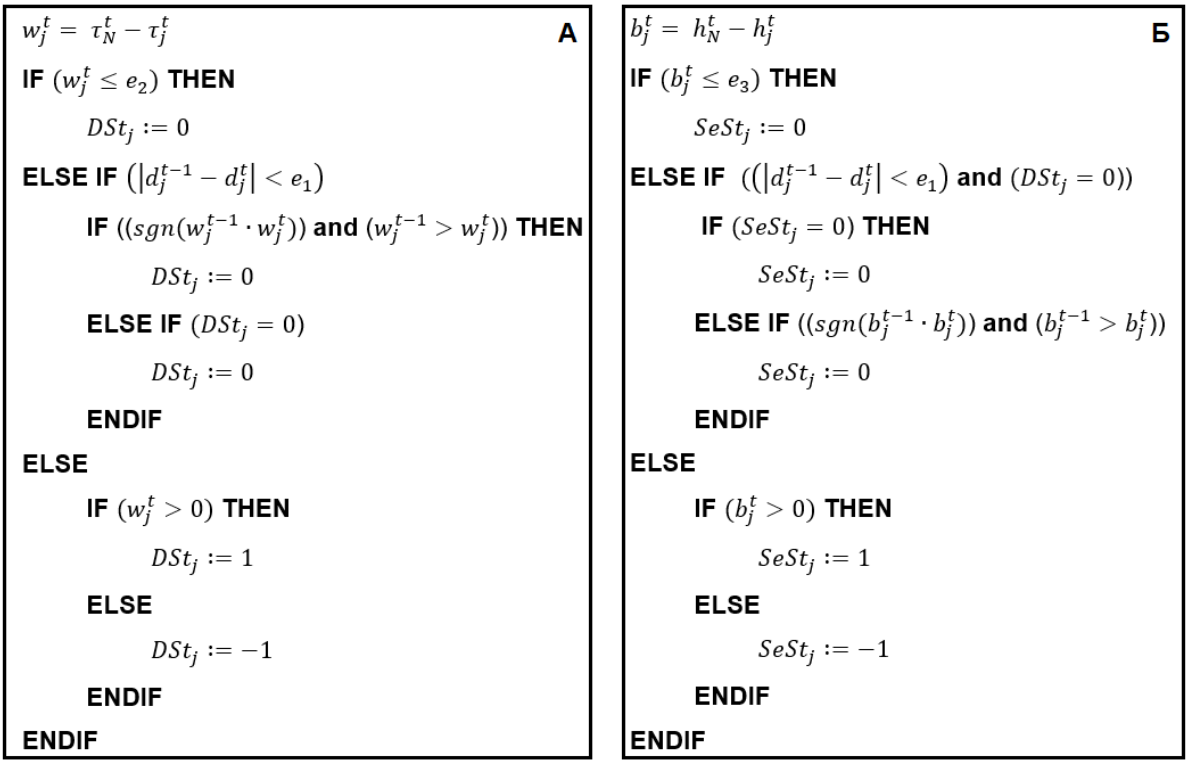


Рис.2. Алгоритмы принятия решения о том, как необходимо производить изменение числа сегментов (А) и синапсов (Б) в дендрите.

Нельзя также полагаться только на динамику изменения параметров w_j^t и b_j^t, поскольку между итерациями могут изменяться входные паттерны, что может создавать видимость нахождения оптимального количества сегментов дендритов или синапсов. Во избежание этого алгоритм проверяет соответствие поданных задержек паттерна, используя e_1 – допустимую погрешность разности временных задержек паттернов на итерациях $t-1$ и t, при которой паттерны считаются одинаковыми. В наиболее простом случае значения переменных e_1, e_2 и e_3 можно считать равными нулю.

Число синапсов зависит от числа дендритов. Это значит, что может быть найдено оптимальное число синапсов по условию (2) для данного числа сегментов дендрита и входного паттерна, однако на следующей итерации может измениться число сегментов дендрита, и потребуется новое определение оптимального числа синапсов, поэтому важно не допустить преждевременных решений об

окончании изменения числа синапсов. Для этого алгоритм принятия решения об изменении числа синапсов для дендрита j проверяет значение переменной DSt_j.

Экспериментальное исследование и результаты

Для демонстрации новых возможностей модели был проведён эксперимент на обучение одной ССМН двум паттернам из одного класса. Сначала модель обучалась одному паттерну, затем на вход модели подавался другой паттерн. Целью эксперимента являлось получение промежуточных структур между структурами, которые были бы получены при изучении паттернов по отдельности, и исследование этих структур для подтверждения или опровержения гипотезы, что промежуточные структуры могут позволить объединить знания двух паттернов и лучше отличать изученный класс от остальных.

В ходе эксперимента после полного изучения первого паттерна модель изучает второй паттерн в течение нескольких итераций так, что длины дендритов изменяются не более чем на 1, а число синапсов при прочих равных изменяется до максимально возможного соответствия текущих максимумов потенциалов на теле нейрона уровню сегмента тела нейрона, связанного с калибровочным дендритом. После этого производится решение задачи бинарной классификации датасета полученной структурой нейрона. Далее вновь производится обучение, пока не будет полностью изучен второй пример.

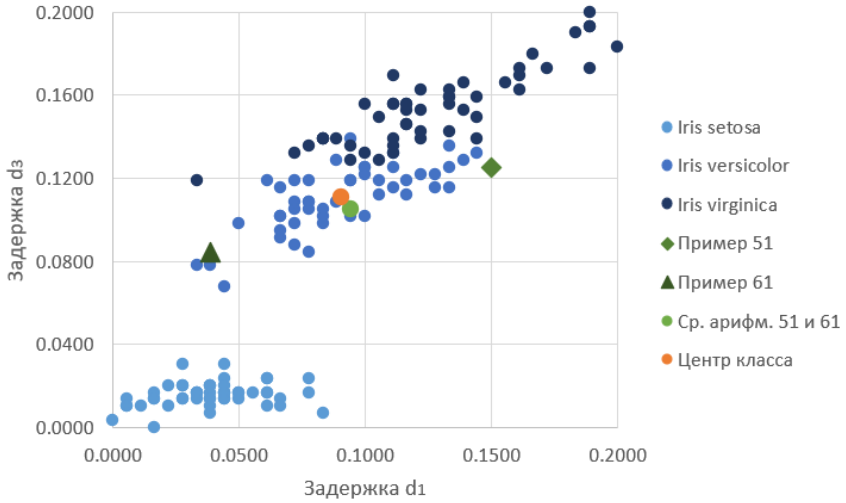


Рис.3. Проекция датасета Iris, преобразованного в паттерны, на плоскость задержек d_1 и d_3

Эксперимент проводился на датасете Iris [7]. Для эксперимента были выбраны примеры 51 и 61, расположенные в противоположных частях второго класса так, что среднее арифметическое примеров располагается рядом с центром класса, получаемым как среднее арифметическое всех примеров класса. Уровень знаний нейрона оценивался по количеству неверных ответов при распознавании всего датасета.

Результаты эксперимента представлены в таблице. Динамика количества неверных ответов модели показывает, что до 6-й итерации обучения включительно модель расширяет объём знаний о втором классе за счёт изучения 61 примера. Однако после 6-й итерации структура модели становится всё более непохожей на структуру после изучения 51 примера, поэтому происходит забывание знаний 51 примера о втором классе, результаты распознавания ухудшаются и постепенно принимают уровень примера 61.

Таблица

Количество неверных ответов при распознавании всего датасета

Итерация обучения примеру 61														Среднее арифметическое примеров
1	2	3	4	5	6	7	8	9	10	11	12	13	14	
28	20	13	12	12	8	10	18	26	37	43	49	50	43	20

Эксперимент показывает, что возможно объединение знаний двух примеров и даже получение лучших результатов, чем могли бы быть получены при изучении среднего арифметического исходных паттернов. Однако результаты обучения нескольким примерам напрямую зависят от структуры датасета и выбора примеров в нём, поскольку обучение позволяет объединить знания выбранных примеров и является зависимым от способности выбранных примеров описывать свой класс.

Заключение

Предложенный алгоритм изменил организацию процесса обучения моде-

ли, объединив последовательные процессы синхронизации и нормализации существующего алгоритма, что привело к уменьшению времени, затрачиваемого на обучение модели паттерну, в 1,5 раза при прочих равных.

Предложенный алгоритм позволяет преодолеть такие недостатки исходного алгоритма, как возможность обучения только один раз, невозможность обучения нескольким примерам, невозможность прервать или начать обучение по необходимости. Устранение данных недостатков исходного алгоритма позволяет использовать ССМН в сценариях инкрементного обучения, где ей может потребоваться расширить знания или изменить представление о данных за счёт поступления новых примеров в произвольный момент времени.

Предложенный алгоритм является только начальным этапом на пути к реализации инкрементного обучения нейронный сетей, построенных на ССМН. Необходимы дальнейшие исследования методов обучения ССМН нескольким примерам, методов борьбы с катастрофическим забыванием для ССМН и методов реализации инкрементного обучения нейронных сетей, построенных на ССМН, как на уровне сети, так и на уровне отдельных нейронов. Отметим также, что поскольку ССМН применяется не только для решения задач машинного обучения, возможно применение предложенного алгоритма в других направлениях исследований ССМН.

Благодарности

Работа проводилась в рамках выполнения государственного задания Минобрнауки России за 2022 год № 075-01623-22-00 «Исследование и разработка биоподобной системы управления поведением мобильных роботов на базе энергоэффективных программно-аппаратных нейроморфных средств».

Список литературы

1. Yoon, J. Lifelong Learning with Dynamically Expandable Networks / J. Yoon, E. Yang, J. Lee, S.J. Hwang. E-print: https://arxiv.org/abs/1708.01547
2. Lobo, J.L. Evolving Spiking Neural Networks for online learning over drifting data streams [Text] / J.L. Lobo, I. Lana, J. Del Ser, M.N. Bilbao, N. Kasabov // Neural

Networks. — 2018. — Vol.108. — P.1 – 19.

3. Бахшиев, А.В. Сегментная спайковая модель нейрона CSMN / А.В. Бахшиев, А.А. Демчева // Известия вузов. ПНД. — 2022. — Т.30. — № 3. — С. 299 – 310.
4. Корсаков, А.М. Применение сегментной спайковой модели нейрона со структурной адаптацией для решения задачи классификации / А.М. Корсаков, Л.А. Астапова, А.В. Бахшиев // Информатика и автоматизация. — 2022. — Т.21. — № 3. — С.493 – 520.
5. Бахшиев, А.В. Структурная адаптация сегментной спайковой модели нейрона / А.В. Бахшиев, А.М. Корсаков, Л.А. Астапова, Л.А. Станкевич // Труды VII Всероссийской конференции. Нижний Новгород, 20–24 сентября 2021 г. — Н. Новгород: Институт прикладной физики Российской академии наук, 2021. — С. 30 – 33.
6. Gepperth, A. Incremental learning algorithms and applications [Text] / A. Gepperth, B. Hammer // ESANN. — 2016.
7. UCI Machine Learning Repository: Iris Data Set. — URL: https://archive.ics.uci.edu/ml/datasets/iris (дата обращения: 15.05.2022).

УЛУЧШЕНИЕ ВИДИМОСТИ ТКАНЕЙ ГОЛОВНОГО МОЗГА В УСЛОВИЯХ МАССИВНОГО КРОВОТЕЧЕНИЯ ПО ДАННЫМ NIR-КАМЕРЫ И ШИАРЛЕТ-ПРЕОБРАЗОВАНИЯ ИЗОБРАЖЕНИЙ

А.В.Медиевский[1], А.Г.Зотин[2], К.В.Симонов[3], А.С. Кругляков[3]

[1]Красноярский государственный медицинский университет
имени профессора В.Ф. Войно-Ясенецкого. ***amedievsky@yandex.ru***,

[2]Сибирский государственный университет науки и технологий
имени академика М.Ф. Решетнева. ***zotin@sibsau.ru***

[3]Институт вычислительного моделирования СО РАН. ***simonovkv50@gmail.com***

Введение

В настоящее время наиболее часто отдают предпочтения малоинвазивным операциям. Они отличают от обычных тем, что для их проведения достаточно небольшого разреза. У данного вида операций есть свои преимущества и недостатки. Так из-за малого разреза невозможно самостоятельно увидеть внутренние ткани. В связи с этим необходимо использовать дополнительное оборудование, наиболее часто используемым является эндоскопическая камера. Это единственное устройство, позволяющее визуализировать ткани внутри пациента. Однако в случае развития интраоперационного кровотечения вся полость заполняется кровью, в связи с чем возможна полная потеря видимости объектов интереса.

На данный момент не существует видеоэндоскопического оборудования, которое могло бы визуализировать интересующие ткани сквозь массивный объём излившейся крови. В таких условиях бывает крайне тяжело осуществить гемостаз, так как все манипуляции хирург проводит вслепую. Для решения данной проблемы предлагается применение особого эндоскопического оборудования со встроенным сенсором ближнего инфракрасного диапазона. А с целью улучшения восприятия исходных снимков предлагается использовать набор алгоритмов для шумоподавления и улучшения яркостных характеристик, а

также шиарлет-преобразования с применением цветового кодирования объектов интереса.

1. Осуществляемые мероприятия при возникновении кровотечения

Популярность эндоскопических и лапароскопических операций способствовала разработке и внедрению новых средств для достижения гемостаза там, где традиционные методы оказались неприменимы[1]. Они имеют свои особенности и при бесконтрольном использовании некоторые из них могут вызывать ряд осложнений [2]. В момент развития кровотечения, помимо самих способов гемостаза, хирургу могут потребоваться способы восстановления видимости. И чем лучше видимость, тем эффективнее будет проведёт гемостаз.

Для улучшения визуализации применяются такие методики, как dry-field maneuver [3 – 4] и small-chamber irrigation technique (SCIT) [5]. Метод dry-field maneuver основан на аспирации ликвора из желудочков, что способствует выявлению и коагуляции источника кровотечения. В методе SCIT используется дополнительная жидкость для орошения операционного поля. Раствор вымывает эритроциты непосредственно перед камерой, что очищает поле зрения эндоскопа. Но использование существующих методов визуализации возможно только при работе с незначительными кровотечениями. При массивном кровотечении хирургу остаётся осуществлять гемостаз вслепую.

2. Взаимодействие ближнего инфракрасного диапазона с кровью

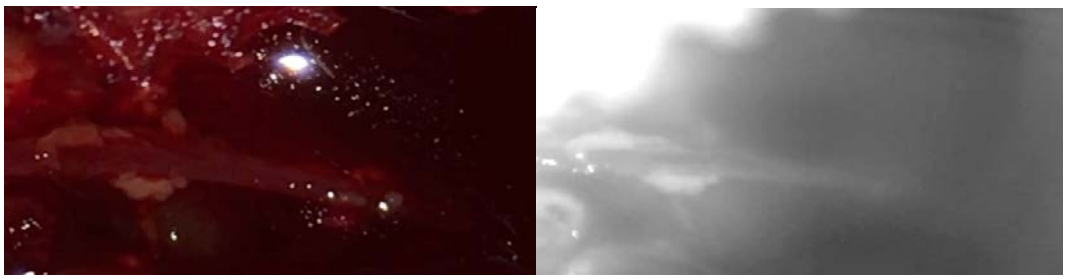
Для возобновления видимости во время кровотечения был выбран эндоскоп, улавливающий только лучи ближнего инфракрасного спектра (NIR). Выбор обуславливается тем, что лучи NIR способны проникать в ткани глубже, чем видимый свет [6], также NIR в большей степени поглощается гемоглобином, чем окружающими тканями [7]. Кровь имеет значительно выше коэффициент поглощения лучей NIR в диапазоне от 800 нм до 1050 нм в сравнении с коэффициентом поглощения того же спектра тканями, задействованными во время операции.

На основе данного свойства работает веновизаторы, позволяющие визуализировать периферические кровеносные сосуды в тех ситуациях, когда их

не способен обнаружить человеческий глаз для выполнения медицинской манипуляции [8]. С учетом таких особенностей NIR лучей была выдвинута гипотеза о том, что данный спектр позволит визуализировать интересующие ткани в условиях массивного интраоперационного кровотечения.

3. Получение исходного изображения

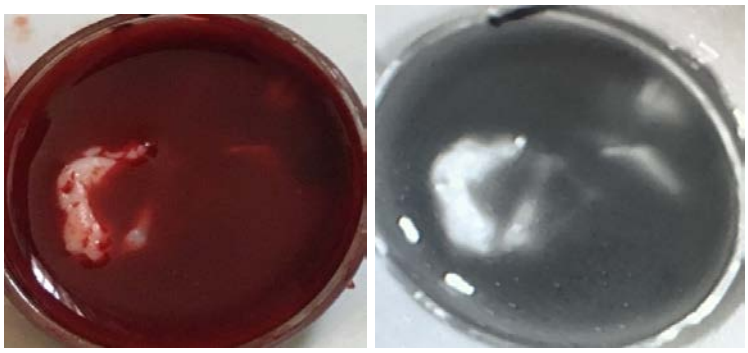
Для проверки достоверности методики визуализации биологических тканей подготовлена модель операционного поля. С этой целью использовался гистологический препарат человеческого головного мозга, помещённый в пластиковую прозрачную чашку. Затем была оценена структура среза без крови. Это позволило создать эталонное изображение для проверки разрабатываемой методики. Проверка осуществлялась при помощи сравнения одного и того же препарата (рис. 1) до добавления крови, заснятом в видимом спектре (рис. 2а), и после добавления крови в ближнем инфракрасном диапазоне (рис. 2б).



а б

Рис. 1. Эталонное изображение: (а) операционное поле до добавления крови, заснятом в видимом спектре; (б) операционное поле с кровью в спектре NIR.

На рисунке 2 продемонстрированы отличия при съёмке интересующих структур в диапазоне видимого излучения и в ближнем инфракрасном спектре.



(а) (б)

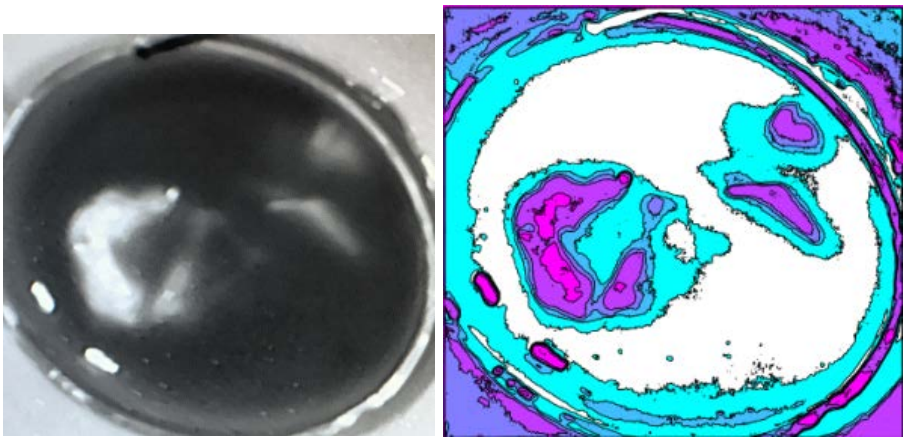
Рис. 2. Модели операционного поля при кровотечении: (а) съёмка в диапазоне видимого излучения; (б) съёмка в ближнем инфракрасном диапазоне.

Применение камер NIR (рис. 2b) позволяет рассмотреть структуры, которые располагаются более глубоко, по сравнению с обычной съёмкой (рис. 2a). Полагаем, что у хирургов, несмотря на массивное кровотечение, появляется возможность в режиме реального времени получать изображение операционного поля и осуществить необходимые манипуляции.

В результате проведенного поиска концепций, нацеленных на решение той же проблемы, найдена работа, позволяющая различать кровеносные сосуды и кровотечение [9], но данная работа не может использоваться в экстренном порядке для нахождения разорвавшегося сосуда, так как определение сосуда занимает продолжительное количество времени и данный кровеносный сосуд должен находиться на поверхности.

4. Обработка NIR снимков

Для оперативной адекватной оценки хирургом видимой операционной области разработана вычислительная методика обработки снимков (рис. 3). В методике заложены процедуры, призванные скорректировать яркостные характеристики, произвести шумоподавление, а также улучшить контрастность и детализацию тканей объектов интереса. Оценка эффективности работы предлагаемых алгоритмов осуществлялось посредством экспертного анализа.



(а) (б)

Рис. 3. Результаты применения методов обработки: (а) контрастирование исходного изображения BCET mean100; (б) шиарлет преобразование и цветовое кодирование.

На рисунке 3а представлен результат контрастирования исходного снимка, все визуализируемые объекты становятся пропорционально ярче. Повышается яркость исследуемой ткани мозга. Сохраняются все детали, и в большей степени затемняется фон. Поэтому объект исследования воспринимается лучше. Далее применялись методики шиарлет-преобразования и цветового кодирования. Результат представлен на рисунке 3б. Данный вариант совместил все необходимые характеристики: более полный градиент, точность морфологии и хорошая контрастность с фоном.

Заключение

Разработана вычислительная методика для визуализации операционного поля во время массивного кровотечения, которая способна функционировать в режиме реального времени. Для ее апробации планируется в следующих работах применение данного метода в ходе полноценной нейрохирургической операции. Это технически и «юридически» возможно благодаря тому, что разработанная методика визуализации предлагается в виде дополнения к имеющимся способам визуализации. Особенностью предлагаемого подхода является то, что предлагаемый метод работает обособленно и обрабатывает только входящее

изображение, поэтому не требуется от нейрохирургов вводить дополнительное оборудование в операционную рану (пространство).

Список литературы

1. Yao H.H. Haemostasis in neurosurgery: what is the evidence for gelatin-thrombin matrix sealant? / H.H.Yao, M.K.Hong, K.J.Drummond // Journal of clinical neuroscience: official journal of the Neurosurgical Society of Australasia. — 2013. — V.20(3). — P.349 – 356.
2. Das J.M. Bone Wax in Neurosurgery: A Review / J.M.Das // World neurosurgery. — 2018. — V.116. — P.72 – 76.
3. Turhan T. Dry-field maneuver for controlling the massive intraventricular bleeding during neuroendoscopic procedures. Child's nervous system / T.Turhan // Journal of the International Society for Pediatric Neurosurgery. — 2018. — V.34(3). — P.541 – 545.
4. Oertel J. Management of severe intraoperative hemorrhage during intraventricular neuroendoscopic procedures: the dry field technique / J.Oertel, S.Linsler, A.Csokonay, H.Schroeder., S.Senger // Journal of neurosurgery. — 2018. — V.131(3). — P. 931 – 935.
5. Manwaring J.C. The small-chamber irrigation technique (SCIT): a simple maneuver for managing intraoperative hemorrhage during endoscopic intraventricular surgery / J.C.Manwaring, A.El Damaty, J.Baldauf., H.W.Schroeder // Neurosurgery. — 2014. — V.10. — Suppl. 3. — P. 375 – 379.
6. Shourav M.K. Visualization of superficial vein dynamics in dorsal hand by near-infrared imaging in response to elevated local temperature / M.K.Shourav, J.Choi, J.K.Kim // J. Biomed Opt. — 2021. — V.26(2). — 026003.
7. Wang L. Infrared imaging of hand vein patterns for biometric purposes / L.Wang, G.Leedham, S.Gho // IET Computut. Vis. — 2007. — V.1. — №3. — P.113 – 122.
8. 3D Near Infrared and Ultrasound Imaging of Peripheral Blood Vessels for Real-Time Localization and Needle Guidance / A.I.Chen, M.L.Balter, T.J.Maguire, M. L.Yarmush // Medical image computing and computer-assisted intervention:

MICCAI. International Conference on Medical Image Computing and Computer-Assisted Intervention. — 2016. — V.9902. — P.388 – 396.

9. Laurence A. Multispectral diffuse reflectance can discriminate blood vessels and bleeding during neurosurgery based on low-frequency hemodynamics / A.Laurence, A.Bouthillier, M.Robert, D.K.Nguyen, F.Leblond // Journal of biomedical optics. — 2020. — V.25(11). — 116003.

АМИНОКИСЛОТЫ, КОДИРУЕМЫЕ ГЕНАМИ ТРАНСПОРТНЫХ РНК БАКТЕРИЙ, ЯВЛЯЮТСЯ ВЕДУЩИМ ФАКТОРОМ КЛАСТЕРИЗАЦИИ ЭТИХ ГЕНОВ ПО ТРИПЛЕТНЫМ ПРОФИЛЯМ

О.А.Мутовина[1], М.Г.Садовский[2,3,4]

[1]Сибирский федеральный университет, ИФБиБТ, ***mutovina.ole4ka@mail.ru***

[2]Институт вычислительного моделирования СО РАН, ***msad@icm.krasn.ru***

[3]Федеральный Сибирский научно-клинический центр ФМБА России,

[4]Красноярский государственный медицинский университет МЗ РФ

Заметный прогресс масштабного автоматизированного секвенировании ДНК и распространения компьютерных алгоритмов для быстрой автоматической идентификации местоположения генов транспортных РНК в секвенированных геномах привели к экспоненциальному увеличению доступной информации о последовательности тРНК. На сегодняшний день полностью секвенировано более сотни геномов эукариот, бактерий и архей, вирусов, и секвенирование многих геномов продолжается. Подробная информация о последовательностях, структуре и встречаемости генов тРНК в различных геномах позволяет проводить надежный систематический анализ филогенетических зависимостей, использования антикодонов, характерных структурных особенностей тРНК и другого [1–4]. Большой объём и разнообразие информации в данной области делает возможным развитие алгоритмов для структурирования и классификации этих данных. Это позволяет проводить анализ данных, содержащихся в генетическом материале, с разных сторон и в разных аспектах. Наиболее интереснымии информативными являются анализ связи структуры, функции и таксономии носителя. При этом такого рода исследование можно проводить напредмете анализ любой из пар составляющей, так и анализ с учетом сочетаний всех трех сторон, что представляет больший интерес. Настоящая работа посвящена анализу всех трёх аспектов.

Актуальность работы связана с большим количеством работ, посвященных изучению различных генетических данных и структур. Большой объём данных требует систематизации и классификации, подкрепленных не

только биологическими данными, но совокупностью таких данных и некоторых расчетов. Гены транспортных РНК бактерий являются достаточно древними генами [5] и присутствуют у всех организмов, за исключением геномов вирусов, что позволяет получать более точные результатыпри изучении филогении. Гены тРНК могут быть полезны в изучении распределения бактерий по группам, определяемым таксономией или функциональной ролью.

Целью данной работы является анализ связи триплетного состава генов транспортных РНК бактерий с их таксономией и функциональной ролью этих генов. Для этого была сформирована база генов транспортных РНК, проведена их кластеризация методом упругих карт и методом динамических ядер, проанализировано распределение генов по кластерам с точки зрения таксономического состава и функциональной роли (переносимой аминокислоты).

В бактериальных клетках набор РНК в основном состоит из рРНКи тРНК (до 20 %). В геноме бактерий гены тРНК группируются в опероны. Например, *Bacillus subtilis* имеет 86 генов тРНК, кодирующих 35 различных видов тРНК. Эти гены организованы в 21 оперон. Похожие ситуации есть и у других бактерий [7]. Геномы не кодируют изоакцепторные гены тРНК для всех кодонов. У бактерий сканирование более чем 100геномных последовательностей показало, что у 20 кодонов постоянно отсутствовал ген тРНК с комплементарным антикодоном у всех исследованных видов эубактерий и архебактерий. Репертуар генов тРНК может быть ещё меньше у отдельных видов, например, Ɣ-протеобактерия *Pseudomonas aeruginosa* имеет всего 37 изоакцепторных генов тРНК в одной, двух или трех копиях [6, 8, 9]. Гены тРНК, как правило, присутствуют в большом количестве копий в геномах большинства организмов, от прокариот до эукариот, но количество копий генов для каждого вида тРНК (имеется ввидутРНК с одним и тем же антикодом), широко варьирует от вида к виду. Для любой активно делящейся клетки эффективность трансляции одного определённого кодона определяется количество тРНК в клетке. Количество копий гена определённой тРНК в геноме определяет концентрацию этой тРНК в клетке [10, 11]. Таким образом, содержание

гена тРНК определяет наличие относительного разнообразия изоакцепторов тРНК, которое, в свою очередь, определяет эффективность трансляции кодонов. В работах [4, 11, 12] показано, что число генов, кодирующих каждую тРНК, не сохраняется между царствами. Изменчивость числа генов тРНК может довольно сильно различаться: одни виды тРНК могут отсутствовать у целой филогенетической ветви, а другие, наоборот, преобладать. Для отдельных видов изучались факторы и причины, влияющие на такое распределение количества генов тРНК [11, 13]. Однако неизвестны принципы, которые лежат в основе эволюции популяции генов тРНК [11].

Разные бактерии имеют различную структуру и организацию генов транспортных РНК. В данном случае под структурой понимается способ составления последовательность гена тРНК. Такая структура имеет несколько особенностей: кодирование 3′-конца CCA, наличие интрона и сохранение определённых элементов, таких как нуклеотиды, расположенные выше или ниже исходной последовательности тРНК. Так же к особенностям относится длина дополнительных 5′ и 3′-выступающих последовательностей и спейсеров между двумя генами тРНК в мультимерном транскрипте, наличие процессинговых сигналов и вторичных структур. Кроме того, нельзя не отнести к особенности положение генов тРНК в том или ином участке хромосомы. Последняя особенность больше относится к организации генов тРНК. Как и относится к организации то, как гены тРНК организованы в опероны вместе с другими генами тРНК или генами, кодирующими другие типы РНК [14].

Как и для других генов, транскрипция последовательности гена одной какой-то тРНК имеет начало и конец. Среди генов тРНК существуют 5′-лидерные последовательности, которые не имеют каких-либо других генов между местом начала транскрипции и самой последовательностью гена. Если же другие гены, разделяющие точку начала и саму последовательность, существуют, то такие гены являются спейсерной последовательностью. Таким же образом будет и с 3′-концом. Если перед терминатором закодирован другой ген, то он определяется как спейсер. Если же такой ген отсутствует, то после-

довательность, расположенная ниже кодируемой тРНК и будет 3′-концом. Наличие хромосомно кодируемого 3′-конца CCA является сильно различающейся особенностью для разных бактерий. У некоторых бактерий, таких как *E. coli*, 3′-CCA-конец кодируется во всех генах тРНК. Однако существуют бактерий, у которых 3′-CCA-конец отсутствует либо в некоторых генах тРНК, либо во всех. В таких случаях 3′-CCA-конец добавляется к тРНК посттранскрипционно, как у эукариот. Большая часть бактериальных генов тРНК не имеют интронов (за некоторым исключением), то есть их последовательности тРНК непрерывны, в отличие от эукариот и архей, имеющих от 4 % до 25 % генов тРНК с интронами. Бактериальные интроны тРНК относятся к типу интронов первой группы самосплайсинга, тогда как интронам тРНК эукариот и архей требуется сплайсинговая эндонуклеаза для вырезания. Другой особенностью бактериальных геномов является идентичность нуклеотидов, находящихся рядом с началом кодирующей последовательности тРНК. В таком случае первым нуклеотидом зрелой тРНК будет нуклеотид, расположенный непосредственно ниже этого сайта. Большинство бактериальных тРНК имеют в этом положении букву G. Если же нуклеотид располагается непосредственно перед кодирующей последовательностью тРНК, то в большинстве случаев это будет буква U. Такая идентичность важна в отношении процессинга 5′-конца тРНК [14–23].

Гены бактериальных тРНК организованы в опероны и транскрибируются как полицистронные транскрипты с одного общего промотора. Этообъясняет часто встречаемое близкое расположение многих генов тРНК на одной и той же цепи бактериальной хромосомы. Однако существует достаточное количество генов тРНК, транскрибируемых как моноцистронные предшественники. Такое распределение генов тРНК на моно- и полицистронные транскрипты варьирует у разных организмов. Бывает, что большинство генов тРНК закодировано в полицистронных оперонах, а бывает так, что почти все гены тРНК организованы как моноцистронные единицы транскрипции. Так же варьирует среднее и максимальное число генов тРНК в бактериальном опе-

роне. Так, например, *Streptomyces coelicolor* имеет 65 генов тРНК, из которых 71 % транскрибируется в виде моноцистронных транскриптов или полицистронных транскриптов, в которых отсутствуют другие гены тРНК. В качестве примера для сравнения можно привести *Listeria monocytogenes*, у которой из 67 генов тРНК только 12 % имеют подобную организацию. Ещё одной общей чертой для бактериальных генов тРНК является то, что данные гены достаточно часто располагаются в оперонах рибосомальных РНК. Здесь так же есть различные вариации. Может быть так, что оперон рРНК не содержит ни одного гена тРНК, а может быть так, что все опероны рРНК содержат гены тРНК. Помимо генов рРНК, бактериальные опероны с генами тРНК могут содержать последовательности, кодирующие белок. В основном это белки, принимающие участие в трансляции, то есть рибосомальные субъединичные белки [4, 7, 14, 24–26].

У бактерий существует два или более изоакцептора тРНК всего для восьми аминокислот; остальные 12 аминокислот аминоацилированы только одним видом тРНК. Например, у бактерий есть только одна тРНК для пар кодонов для аспартата, аспарагина, гистидина и тирозина, каждая из которых оканчивается либо на C, либо на U. В то время как G в положении колебания антикодона может образовывать пару оснований как с C, так и с U [6, 27].

Материалы и методы

Гены транспортных РНК брались из открытой базы данных генов тРНК GtRNAdb: Genomic tRNA Database. Данная база содержит прогнозы генов тРНК, сделанные с помощью tRNAscan-SE для полных или почти полных геномов. К настоящему времени количество генов тРНК в базе составляет 241 975, без учёта псевдогенов. Таксономический состав базы достаточно разнообразен и представлен 32 типами бактерий, включающих различные таксоны: класс, порядок, семейство, род, вид. В начале работы общее число генов в скачанной базе составляло 246028.

Так как исходная база генов имеет большой разброс по количеству объектов в разных таксономических группах, проводилась индексация. Большое количество данных в группе генов одного таксона в 63-мерном пространстве может образовать кластер, притягивающий к себе соседние точки и таким образом вносить искажения, которые могут значительно повлиять на результаты. Поэтому с целью увеличения точности результатов проводилась обработка материала и отбор генов. За наименьшую таксономическую единицу были взяты штаммы, которые соответствуют идентификаторам. Задан диапазон, при котором штамм с количеством генов меньше 50 исключается из базы, а штамм с количеством больше 1000 индексируется (прореживается) случайным образом. Также были исключены объекты, не имеющие видовой принадлежности и имеющие нерасшифрованные триплеты для аминокислоты. В результате индексирования количество генов транспортных РНК в исследуемой базе сократилось до 171143.

Проблема связи между структурой нуклеотидной последовательности определенного организма с его таксономическим положением довольно часто встречается при изучении процессов эволюции и их молекулярного уровня. Что понимается под термином *структура*? Это триплетный состав изучаемых генов транспортных РНК, то есть частотный словарь триплетов: список всех триплетов, встречающихся в последовательности, с указанием их частот.

Частотные словари представляют собой точки в многомерном пространстве, имеющем размерность 64. В работе частотные словари рассматривались как точки в 63-мерном пространстве, так как для лучшей кластеризации из анализа исключается один триплет, имеющий минимальное стандартное отклонение, определяемое по всей базе изучаемых последовательностей. Это объясняется тем, что сумма всех частот триплетов равна единице и исключаемый триплет вносит наименьший вклад в различимость объектов. Для работы с данным пространством частотных словарей необходимо ввести метрику. В данной работе использовалась Евклидова метрика.

Кластерный анализ позволяет выделить максимально близкие и одно-

родные по составу группы и выявить структурированность полученных групп. Такие группы представляют собой непересекающиеся подмножества набора данных, обладающие тем свойством, что данные, принадлежащие разным кластерам, различаются между собой гораздо больше, чем данные, принадлежащие одному кластеру. В данной работе использовалось два методы кластеризации: метод упругих карт, относящийся к нелинейной кластеризации, и метод динамических ядер, относящийся к линейной кластеризации. Данные методы выбирались исходя из поставленных задач: обнаружения скопления групп генов и выявление признака, по которому те или иные гены попадают в определенный кластер.

Алгоритм построения упругой карты состоит в следующем. Первым этапом является определение первой и второй главных компонент. На втором этапе строится плоскость на первых двух главных компонентах как на осях и на эту плоскость проектируются все точки, находящиеся в пространстве. Затем определяется минимальный квадрат, содержащий все проекции точек. Далее каждая точка соединяется математической пружиной со своей проекцией на плоскости. Такая пружина имеет бесконечную растяжимость и не меняет свои свойства по мере растяжения. Третий этап заключается в замене жёсткого квадрата гибкой мембраной, обладающей растяжимость и эластичностью. После этого вся система отпускается, в результате чего мембрана достигает формы, которая соответствует минимуму общей энергии деформации. Последним шагом является переопределение точек в результате построения ортогональных проекций для каждой точки. Затем систему освобождают от пружин и она релаксирует до тех пор, пока исходный квадрат снова не станет плоским; он и будет являться упругой картой во внутренних координатах.

Метод кластеризации, описанный в данной работе, подразумевает определение локальной плотности точек. Плотность точек здесь определяется как среднее число точек на небольшом участке упругой карты. Дляэтого каждая точка снабжается какой-либо куполообразной функцией с максимумом в этой точке, например, функцией Гаусса:

$$f_j(r) = \exp\left\{-\frac{(r - r_j)^2}{\mu^2}\right\}$$

Здесь r_j — координата j–ой точки, μ — полуширина этой функции, полностью аналогичная стандартному отклонению для случая нормального распределения случайной величины — подгоночный параметр, определяющий контрастность картины локальной плотности. Далее вычисляется функция, которая является суммой всех определенных выше функций

$$F(r) = \sum_{i=1}^{N} f_i(r)$$

Здесь N — количество точек.

Для реализации методов кластеризации, применяемых в работе, использовалось программное обеспечение *ViDaExpert* — свободно распространяемое программное обеспечение, реализующее ряд простых и расширенных статистических методов и удобный графический интерфейс для применения этих методов к таблице данных, которая может содержать как числовые значения характеристик, так и определенные метки для объектов и функций [28–30].

Кластеризация генов тРНК методом упругих карт

Кластеризация изучалась на упругих картах 16 × 16; при построении карты использовались следующие параметры: исключался триплет ТАТ, имеющий наименьшее стандартное отклонение, радиус корреляции при построении локальной плотности устанавливался равным 0,15, значение остальных параметров задавалось по умолчанию. Начнём с распределения генов тРНК на кластеры в разрезе принадлежности аминокислот к разным кластерам. Наблюдаемые кластеры довольно однородны по составу генов транспортных РНК, несущих одну и ту же аминокислоту — гены, кодирующие одну и ту же аминокислоту, в большинстве случаев концентрируются в одном кластере. Тем не менее, для каждой аминокислоты находятся такие точки, которые при разбиении на кластеры оказываются не в своём кластере.

На Рисунке 1 показана общая картина распределения генов транспортных РНК бактерий по упругой карте; на этом рисунке каждая аминокислота

показана своим цветом. Хорошо видно, что картина распределения упомянутых генов в целом достаточно упорядочена: как правило, гены транспортных РНК, кодирующие соответствующие переносчики, группируются в весьма плотные кластеры, а эти последние отделены друг от друга.

Следует подчеркнуть, что даже индексированная база включает в себя более 150000 генов; это означает, что на Рисунке 1 представлено такое большое число точек. Понятно, что анализ распределения точек по упругой карте и даже их визуальное восприятие здесь весьма затруднены. Для того, чтобы облегчить восприятие и анализ распределения точек, мы построили серию индивидуализованных распределений. Для этого на упругих картах отображались только те точки, которые соответствуют генам тРНК, переносящим только одну аминокислоту. Картина локальной плотности при этом строилась по всем точкам.

Индивидуальные распределения генов тРНК, кодирующих одну и ту же аминокислоту, показаны на Рисунке 2. Анализ распределения генов, кодирующих одну аминокислоту, начнём с рассмотрения двух из них: триптофана и метионина. Это две единственные аминокислоты, не имеющие синонимических кодонов. Анализ распределений, показанных на Рисунке 2, показывает, что соответствующие гены разбились на три кластера, для каждого из генов. При этом картины распределения только для Cys, Gln, Glu, Lys и Gly могут рассматриваться (с известным приближением) как двухкластерные.

При этом большинство генов образуют несколько кластеров, однако заметная часть точек — по крайней мере, такая, которую нельзя считать случайной — разбросана по нескольким отдельным кластерам. Таким образом, хорошо видно, что, например, цистеин и глутамин образуют два кластера и имеют небольшое число выпадающих точек. Гистидин и лейцин, напротив, образуют только один кластер, но при этом большое количество точек разбросано по кластерам других аминокислот.

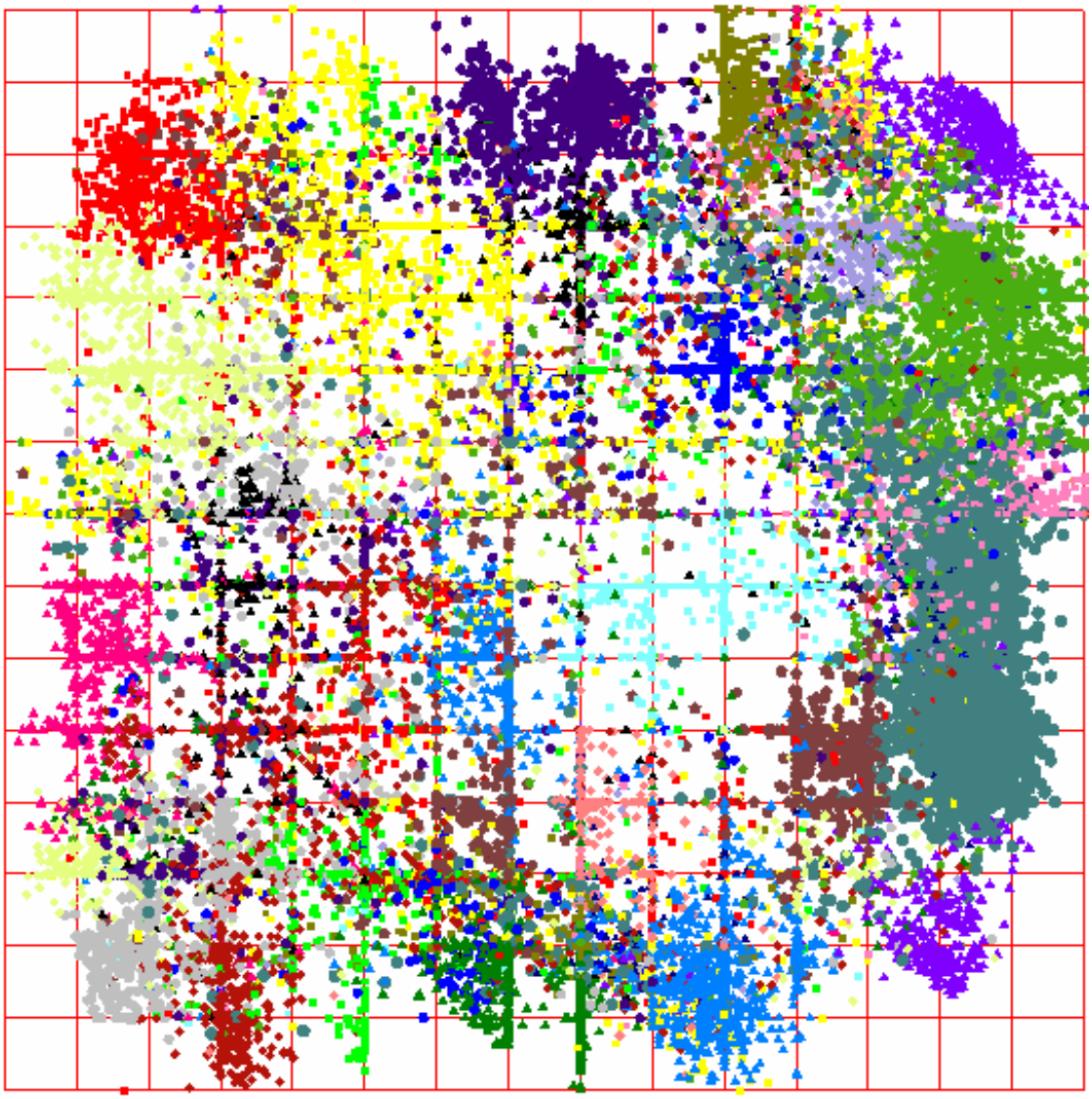


Рис. 1. Распределение генов тРНК всех 20 аминокислот по упругой карте 16 × 16. Разными цветами обозначены гены, переносящие соответствующие аминокислоты; специфическое распределение синонимических кодов не показано.

Здесь следует обратить внимание на важный факт: всё разнообразие генов транспортных РНК определяется несколькими факторами. Во-первых, следует выделить различия, обусловленные синонимией кодонов, кодирующих одну и ту же аминокислоту. Тем самым, говоря о кластерной структуре генов транспортных РНК, необходимо учитывать эту синонимию. Рис. 2 очень хорошо иллюстрирует это влияние: хорошо видно, что гены тРНК ряда аминокислот, кодируемых синонимами, образуют отдельные кластеры.

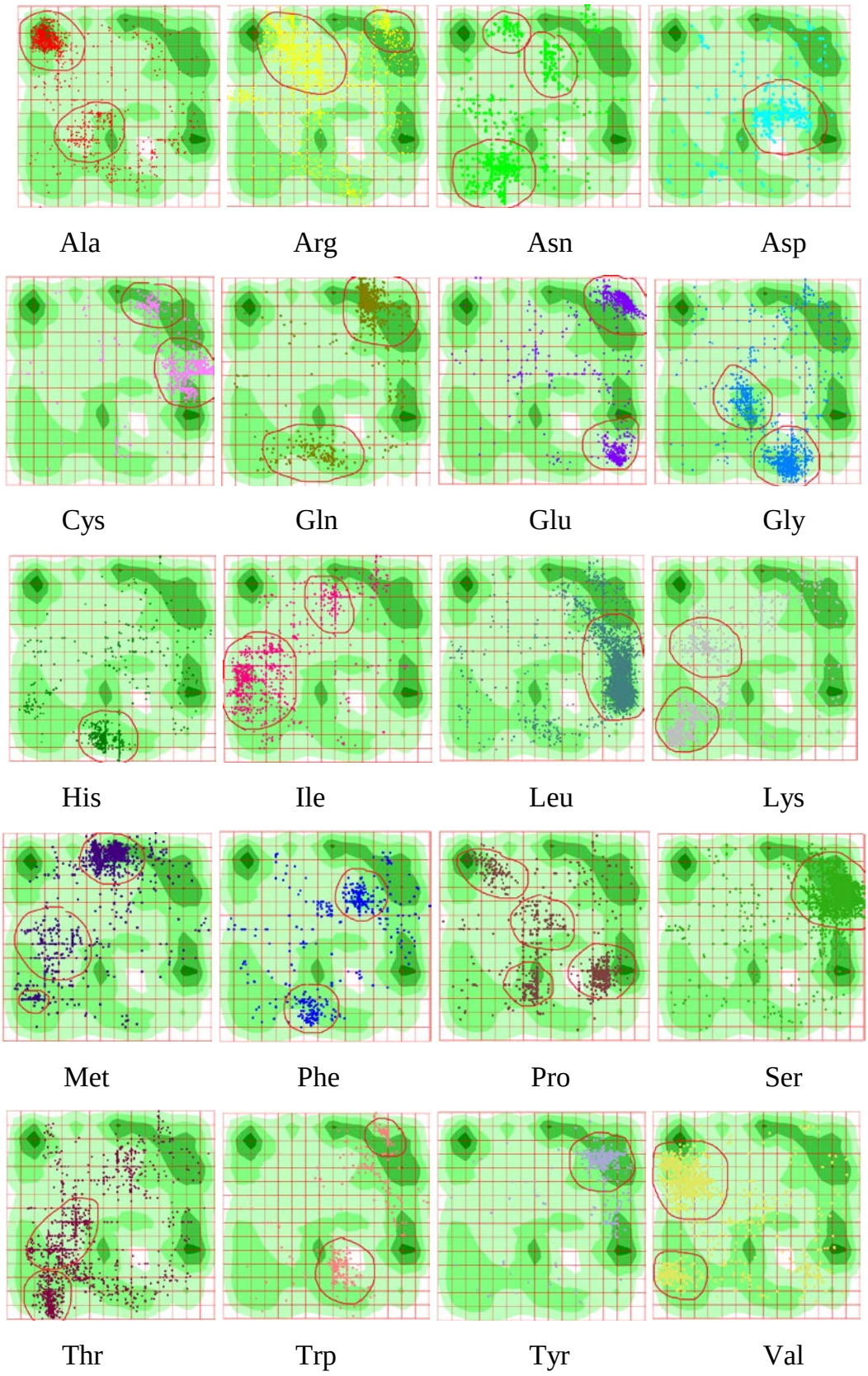


Рис. 2. Карты индивидуального распределения генов тРНК по упругой карте 16 × 16; локальная плотность определялась по всем генам.

Во-вторых, имеется и «внутрикодонная» вариабельность в последовательности нуклеотидов в генах транспортных РНК, описанная выше. Всё это приводит к тому, что гены тРНК бактерий, имея незначительную вариацию в последовательности нуклеотидов, формируют достаточно плотные кластеры. Следует сказать, что распределение генов тРНК по кластерам (классам) проверялось также и одним из классических линейных методов классификации — методом динамических ядер (известным в англоязычной литературе как *k*-means). Наши исследования показали, что никакой сколько-нибудь устойчивой классификации этим методом получить нельзя, если число классов меньше (при $2 \leq k \leq 10$) числа аминокислот. В связи с этим мы не приводим здесь результатов построения таких классификаций.

Помимо связи структуры (триплетного словаря) и функции (специфической переносимой аминокислоты), мы проверяли и связь полученных кластеров с таксономией. Для этого анализировался таксономический состав кластеров, образуемых в пределах одной аминокислоты; иными словами, для анализа использовалась только те гены, которые кодируют перенос одной и той же аминокислоты. Таким образом, собиралась база для каждой отдельной аминокислоты и уже на основании такой базы строились упругие карты. Установлено, что в пределах одной аминокислоты образуются кластеры, некоторые из которых являются достаточно плотными. Анализ таксономического состава каждого такого кластера показал, что кластеры никак таксономически не обусловлены. В одном кластере находятся штаммы, различающиеся вплоть до типа. Кластеры же представляют собой скопления точек, соответствующих определенным изодекодерам. Данные результаты, полученные методом нелинейного анализа, говорят о преобладании функции над таксономией для генов транспортных РНК.

Список литературы

1. Sprinzl, Mathias. Compilation of tRNA sequences and sequences of tRNA genes / Mathias Sprinzl, Konstantin S Vassilenko // Nucleic acids research. — 2005. — Vol. 33, no. suppl_1. — Pp. D139–D140.

2. Ardell, David H. Computational analysis of tRNA identity / David H Ardell // FEBS letters. — 2010. — Vol. 584, no. 2. — Pp. 325–333.

3. Fujishima, Kosuke. tRNA gene diversity in the three domains of life / Kosuke Fujishima, Akio Kanai // Frontiers in genetics. — 2014. — Vol. 5. — P. 142.

4. Marck, Christian. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features / Christian Marck, Henri Grosjean // Rna. — 2002. — Vol. 8, no. 10. — Pp. 1189–1232.

5. Eigen, Manfred. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis / Manfred Eigen, Ruthild Winkler Oswatitsch, Andreas Dress // Proceedings of the National Academy of Sciences. — 1988. — Vol. 85, no. 16. — Pp. 5913–5917.

6. Direct RNA nanopore sequencing of *Pseudomonas aeruginosa* clone C transcriptomes / Marie-Madlen Pust, Colin Francis Davenport, Lutz Wiehlmann, Burkhard Tu¨mmler // Journal of Bacteriology. — 2021. — Pp. JB–00418.

7 13. Exploring the regulation of tRNA distribution on the genomic scale / Kimberly A Dittmar, Evelyn M Mobley, Agnes Jancso Radek, Tao Pan // Journal of molecular biology. — 2004. — Vol. 337, no. 1. — Pp. 31–47.

8 Pust, Marie-Madlen. Bacterial tRNA landscape revisited / MarieMadlen Pust, Kenneth N Timmis, Burkhard Tu¨mmler // Environmental Microbiology. — 2022.

9 On the Track of the Missing tRNA Genes: A Source of Non-Canonical Functions? / Ricardo Ehrlich, Marcos Davyt, Ignacio L´opez et al. // Frontiers in molecular biosciences. — 2021. — Vol. 8. — P. 84.

10. An evolutionarily conserved mechanism for controlling the efficiency of protein translation / Tamir Tuller, Asaf Carmi, Kalin Vestsigian et al. // Cell. — 2010. — Vol. 141, no. 2. — Pp. 344–354.

11. A Role for tRNA Modifications in Genome Structure and Codon Usage / Eva Maria Novoa, Mariana Pavon-Eternod, Tao Pan, Lluıs Ribas de Pouplana // Cell. — 2012. — Vol. 149, no. 1. — Pp. 202–213. https://www. sciencedirect.com/science/article/pii/S0092867412002127.

12. Gerber, Andre P. RNA editing by base deamination: more enzymes, more targets,

new mysteries / Andre P Gerber, Walter Keller // Trends in biochemical sciences. — 2001. — Vol. 26, no. 6. — Pp. 376–384.

13. Withers, Mike. Archaeology and evolution of transfer RNA genes in the Escherichia coli genome / Mike Withers, Lorenz Wernisch, Mario Dos Reis // RNA. — 2006. — Vol. 12, no. 6. — Pp. 933–942.
14. Pettersson, BM. tRNA Gene Structures in Bacteria: Ph.D. thesis / Acta Universitatis Upsaliensis. — 2009.
15. The making of tRNAs and more–RNase P and tRNase Z / Roland K Hartmann, Markus Goessringer, Bettina Spaeth et al. // Progress in molecular biology and translational science. — 2009. — Vol. 85. — Pp. 319–368.
16. The complete genome sequence of Escherichia coli K-12 / Frederick R Blattner, Guy Plunkett III, Craig A Bloch et al. // science. — 1997. — Vol. 277, no. 5331. — Pp. 1453–1462.
17. Li, Hong. Complexes of tRNA and maturation enzymes: shaping up for translation / Hong Li // Current opinion in structural biology. — 2007. — Vol. 17, no. 3. — Pp. 293–301.
18. Muralitharan, Gangatharan. Evidence on the Presence of tRNA{fMet} Group I Intron in the Marine Cyanobacterium Synechococcus elongatus / Gangatharan Muralitharan, Nooruddin Thajuddin // Journal of microbiology and biotechnology. — 2008. — Vol. 18, no. 1. — Pp. 23–27.
19. Hopper, Anita K. tRNA transfers to the limelight / Anita K Hopper, Eric M Phizicky // Genes & development. — 2003. — Vol. 17, no. 2. — Pp. 162–180.
20. Trotta, Christopher R. tRNA splicing: an RNA world add-on or an ancient reaction? / Christopher R Trotta, John Abelson // Cold Spring Harbor monograph series. — 1999. — Vol. 37. — Pp. 561–584.
21. tRNAdb 2009: compilation of tRNA sequences and tRNA genes / Frank Juhling, Mario M¨orl, Roland K Hartmann et al. // Nucleic acids research. — 2009. — Vol. 37, no. suppl_1. — Pp. D159–D162.
22. Evidence that substrate-specific effects of C5 protein lead to uniformity in binding and catalysis by RNase P / Lei Sun, Frank E Campbell, Nathan H Zahler, Michael E Harris // The EMBO journal. — 2006. — Vol. 25, no. 17. — Pp. 3998–4007.

23. Brannvall, Mathias. Importance of the+ 73/294 interaction in Escherichia coli RNase P RNA substrate complexes for cleavage and metal ion coordination / Mathias Brannvall, BM Fredrik Pettersson, Leif A Kirsebom // Journal of molecular biology. — 2003. — Vol. 325, no. 4. — Pp. 697–709.
24. Inokuchi, Hachiro. Structure and expression of prokaryotic tRNA genes / Hachiro Inokuchi, Fumiaki Yamao // tRNA: Structure, Biosynthesis, and function. — 1994. — Pp. 17–30.
25. Structure and transcription of eukaryotic tRNA gene / Stephen Jefferson Sharp, Jerone Schaack, Lyan Cooley et al. // Critical Reviews in Biochemistry. — 1985. — Vol. 19, no. 2. — Pp. 107–144.
26. Paule, Marvin R. Survey and summary transcription by RNA polymerases I and III / Marvin R Paule, Robert J White // Nucleic acids research. — 2000. — Vol. 28, no. 6. — Pp. 1283–1298.
27. Morris, Rana C. The effect of queuosine on tRNA structure and function / Rana C Morris, Kenneth G Brown, Mark S Elliott // Journal of Biomolecular Structure and Dynamics. — 1999. — Vol. 16, no. 4. — Pp. 757–774.
28. Gorban, Alexander N. Principal manifolds and graphs in practice: From molecular biology to dynamical systems / Alexander N. Gorban, Andrei Zinovyev // International Journal of Neural Systems. — 2010. — Vol. 20, no. 03. — Pp. 219–232. — PMID: 20556849. https://www. worldscientific.com/doi/abs/10.1142/S0129065710002383.
29. Gorban, A. N. Principal Manifolds for Data Visualisation and Dimension Reduction / A. N. Gorban, A. Yu. Zinovyev // Lecture Notes in Computational Science and Engineering / Ed. by A N Gorban, B Kegl, D Wunsch, A Yu Zinovyev. — Berlin – Heidelberg – New York: Springer, 2007. — Vol. 58. — Pp. 153–176.
30. Gorban, Alexander N. Fast and user-friendly non-linear principal manifold learning by method of elastic maps / Alexander N. Gorban, Andrei Yu. Zinovyev // 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015. — 2015. — Pp. 1–9. https://doi.org/10.1109/DSAA. 2015.7344818.

О СВЯЗИ ТРИПЛЕТНОЙ СТРУКТУРЫ ГЕНОВ ТРАНСПОРТНЫХ РНК ЧЕЛОВЕКА С ПЕРЕНОСИМОЙ АМИНОКИСЛОТОЙ

Я.В.Недорез[1], М.Г.Садовский[2,3,4]

[1]Новосибирский государственный университет, ***y.nedorez@g.nsu.ru***

[2]Институт вычислительного моделирования СО РАН, ***msad@icm.krasn.ru***

[3]Федеральный Сибирский научно-клинический центр ФМБА России,

[4]Красноярский государственный медицинский университет МЗ РФ

Введение

Задача изучения связи функции нуклеотидных последовательностей (в том числе и кодируемой ими), их структуры и таксономии носителей соответствующего генетического материала является ключевой в исследованиях в области молекулярной биологии, биоинформатики, биофизики. Понятно, что такое изучение требует выбора подходящего генетического материала, для которого три указанных выше понятия достаточно строго определены. Одним из интересных объектов здесь являются гены транспортных РНК. Во-первых, эти гены кодируют вполне определённый класс молекул РНК, осуществляющих перенос аминокислотных остатков к «месту сборки белка» — рибосоме. С этой точки зрения функцию данных генов можно считать фиксированной, а малые функциональные различия — вид переносимой аминокислоты — строго (насколько это вообще возможно в биологических исследованиях) определёнными.

Современные методы секвенирования и анализа генетического материала также не оставляют желать лучшего в определении таксономического положения того организма, для которого ведётся такое исследование. В рамках настоящей работы мы рассматривали гены тРНК человека, что, строго говоря, исключает из триады *структура – функция – таксономия* таксономическую составляющую. Итак, целью данной работы является изучение связи между структурой генов тРНК человека, понимаемой как профиль триплетного состава, и видом той аминокислоты, которая переносится геном.

Транспортные РНК представляет собой краткие нуклеотидные последова-

тельности, в состав которых входит порядка 70 – 100 оснований. Ключевая функция тРНК состоит в ее участии в процессе синтеза белков: тРНК каждого типа ковалентно связывается с остатком соответствующей ей аминокислоты, после чего доставляет последнюю в рибосому для дальнейшего этапа биосинтеза, комплементарно соединяясь антикодоном с кодоном матричной РНК. Транспортную РНК можно считать одну из наиболее классических молекул РНК: она присутствует абсолютно во всех живых организмах и играет важнейшую роль в обеспечении работы биологической системы.

Вторичная структура тРНК широко известна: как правило, её представляют в виде клеверного листа с четырьмя отходящими плечами, каждое из которых, кроме акцепторного, заканчивается петлей. Наиболее важным для нашей работы является антикодоновое плечо, содержащее антикодон — ключевой триплет, комплементарно связывающийся с соответствующим кодоном матричной РНК в процессе трансляции.

Разнообразие молекул тРНК принято классифицировать по изоакцепторам. Изоакцепторами называют группу всех тРНК организма, несущих одну и ту же аминокислоту [1]. Помимо этого, внутри группы изоакцепторов тРНК делят на изодекодеры — тРНК, имеющие один и тот же антикодон, но различающиеся по первичной структуре (последовательности нуклеотидов, не входящих в состав антикодона). Примечательно, что различия в первичной структуре изоакцепторов не являются результатом посттрансляционной модификации РНК, а обусловлены большим количеством различающихся генов изодекодеров. Так, в геноме человека среди порядка 500 генов тРНК более 270 кодируют изодекодеры [2]. Проведённые ранее исследования [2, 3] показывают, что различия первичной структуры изодекодеров влияют на эффективность их функционирования. Это, а также факт существования разных генов изодекодеров в геноме одного организма даёт основание предполагать наличие взаимосвязи между первичной структурой генов изодекодеров и функциями, выполняемыми уже зрелыми изодекодерами и изоакцепторами.

Структуры как самих тРНК, так и их генов сильно различаются как между

разными доменами жизни, так и между видами, несмотря на малый размер последовательности. Не так давно было обнаружено множество неканонических генов тРНК и их структур. К таковым относятся, например, перестроенные, фрагментированные или расщеплённые гены тРНК [4]. Существует предположение, что такого рода разнообразие является следствием коэволюции структуры тРНК и фермента её процессинга — эндонуклеазы сплайсинга. Для эукариот число генов тРНК колеблется в пределах от 170 до 570, а количество типов изоакцепторных тРНК колеблется от 41 до 55. Значительно варьируется и количество генов изоакцепторов — от 10 до 246. Из 446 известных генов тРНК человека транслируется порядка 274 тРНК различной структуры.

Следует также отметить, что клетке требуется менее 61 разновидностей изодекодеров для взаимодействия между антикодонами тРНК и всеми смысловыми кодонами. В эукариотических организмах такая экономия реализована стратегией «истощения» тРНК, содержащих аргинин или гуанин в первой позиции антикодона. Например, каждая из аминокислот Phe, Tyr, His, Asn, Asp, Cys кодируется парой синонимичных кодонов, различающихся только последним основанием, U или C (UUC и UUU для Phe, ACC и ACU для Tyr и т. д.). В процессе синтеза белка тРНК, содержащие G в первом основании антикодона, считывают как кодоны NNC, так и NNU вышеозначенных аминокислот. То же верно для тех из четырех кодонов Ser, в которых последним основанием является пиримидин (также U или C). Наоборот, тРНК, содержащие A в «неоднозначной» позиции, используются для декодирования пиримидина при чтении синонимичных кодонов Val, Pro, Thr, Ala и некоторых других. Полный список существующих в организме человека типов тРНК (и их количество, проанализированное в ходе исследования) приведен в таблице.

При этом гены тРНК эукариот более сложны, чем предполагалось ранее. Среди всех генов оказалось большое число генов изодекодеров (более 50 % от числа всех генов тРНК для человека и шимпанзе). Наблюдается филогенетическое родство генов изодекодеров у эукариот. Это указывает на то, что некоторые изодекодеры могут выполнять уникальные функции в организмах видов, при-

надлежащих к одной филогенетической ветви. Существует предположение, что такое количество генов изодекодеров связано с увеличением генома позвоночных в процессе эволюции. С одной стороны, возможно, что эти последовательности возникли в результате нейтрального дрейфа генов и не выполняют отдельной функции. Однако возможно и обратное: новые вариации генов тРНК могут исполнять неканонические функции, которые ещё не определены исследователями [5].

Таблица 1

Распределение изодекодеров генов тРНК.

Прочерком обозначены кодоны без соответствующего изодекодера.

	U		C		A		G		
U	Phe	—	Ser	9	Tyr	1	Cys	—	U
	Phe	10	Ser	—	Tyr	13	Cys	29	C
	Leu	4	Ser	4	Stop/SeC	1	Stop	—	A
	Leu	7	Ser	4	Stop	—	Trp	7	G
C	Leu	9	Pro	9	His	—	Arg	7	U
	Leu	—	Pro	—	His	10	Arg	—	C
	Leu	3	Pro	7	Gln	6	Arg	6	A
	Leu	9	Pro	4	Gln	13	Arg	4	G
A	Ile	14	Thr	9	Asn	—	Ser	—	U
	Ile	3	Thr	—	Asn	22	Ser	8	C
	Ile	5	Thr	6	Lys	12	Arg	6	A
	Met	10	Thr	5	Lys	15	Arg	5	G
G	Val	10	Ala	22	Asp	—	Gly	—	U
	Val	—	Ala	—	Asp	22	Gly	14	C

Val	5	Ala	8	Glu	7	Gly	9	A
Val	11	Ala	4	Glu	8	Gly	5	G

Эталонный геном человека hg19, использованный в нашем исследовании, всего содержит 613 генов тРНК [6]. Однако лишь около 400 из них кодируют тРНК, которые могут складываться в правильную структуру клеверного листа с нулевым или максимум одним несоответствием в стебле. Весь репертуар генов тРНК человека включает в себя порядка 60 – 100 пар оснований — это составляет всего 0,0019% от объёма всей последовательностей генома. Референсный геном человека содержит неожиданно большое количество разнообразных последовательностей генов тРНК [7, 8, 9], что само по себе является благодатной почвой для исследований, вызывая вопросы: зачем организму столько разнообразных генов тРНК, есть ли закономерности в структуре этих последовательностей и, наконец, могут ли РНК, транслируемые их этих генов, выполнять не канонические функции?

Кратко опишем интроны в генах тРНК и их распределение среди разных организмов. Интроны в составе транспортной РНК были обнаружены во организмах представителей всех царств жизни, и их точечное вырезание играет решающую роль для функции тРНК. Особый интерес представляют интронные кольцевые РНК, которые образуются в процессе сплайсинга тРНК в организмах многоклеточных животных и образуют очень стабильные интроны. Известно, что в клетках животных эти стабильные интроны образуют новый класс некодирующих РНК [10].

Опишем постановку задачи. Имеется ансамбль из 416 генов тРНК человека. Требуется проверить гипотезу о том, что близкие по структуре гены кодируют тРНК, переносящие одну и ту же аминокислоту и наоборот, верно ли, что тРНК, переносящие одну и ту же аминокислоту, имеют близкую структуру. Ответ на эти два вопроса требует определения того, что будет пониматься под структурой. Разнообразие структур, наблюдаемых в нуклеотидных последовательностях, весьма велико [11] и зачастую зависит от фантазии исследователя. Для целей

нашего исследования необходимо определить структуру таким образом, чтобы она допускала применение арсенала средств строго анализа и сравнения, а также была сравнительно простой, в первую очередь, чтобы исключить из рассмотрения различные специальные случаи. Наиболее подходящим объектом на роль структуры нуклеотидной последовательности является её частотный словарь, в нашем случае — частотный словарь триплетов.

Частотный словарь триплетов W_j — это список всех подряд стоящих троек нуклеотидов $\omega = \nu_1\nu_2\nu_3$ вместе с указанием той частоты, с которой эта тройка встречается в изучаемой последовательности. Частота определяется как отношение числа копий этой тройки, обнаруживаемых в последовательности, к общему числу всех последовательностей; в нашем случае общее число всех троек равно длине последовательности. Такое преобразование отображает последовательность (гена тРНК в нашем случае) в точку в 63-мерном пространстве: поскольку сумма частот всех троек от AAA до TTT равна единице, то один триплет можно исключить из рассмотрения. Естественно исключать тот триплет, для которого стандартное отклонение его частоты, определяемое по всей базе исследования, минимально; действительно, такой триплет вносит наименьший вклад в различимость генов.

Основная идея ответа на поставленный вопрос заключается в следующем: требуется проверить, образуют ли изучаемые гены, преобразованные в частотные словари триплетов, кластеры, и если да, какова структура этих кластеров. Иными словами, верно ли, что в один кластер группируются (по преимуществу) гены, ответственные за перенос вполне определённой аминокислоты и исключения являются редкими? Мы пользовались двумя методами кластеризации: методом упругих карт [12] и методом динамических ядер [13].

Методы кластеризации

Метод динамических ядер (МДЯ), известный также как k-means или k-средних, является одним из самых популярных и широко известных методов классификации данных. Суть метода заключается в группировке элементов век-

торного пространства на заранее заданное количество классов. Метод является методом классификации без учителя, линейным. Это означает, что при построении классификации не используется никакой априорной информации, а линейность означает, что делителями множества точек на классы являются гиперплоскости в соответствующем пространстве (в пространстве частот триплетов в нашем случае).

На первом этапе применения данного метода требуется определить количество классов K, на которые следует разделить всё множество исследуемых точек. Первое разделение по K классам происходит случайным образом: среди множества точек инициализируют K случайных главных точек (иначе называемых динамическими ядрами класса). Далее для каждого из случайно выбранных динамических ядер необходимо найти минимальное расстояние от каждой точки до каждого ядра и разделить точки по классам в соответствии с минимальным расстоянием (под расстоянием в данном случае подразумевается Евклидово расстояние). Так, если расстояние от точки i до ядра M меньше, чем до остальных ядер, то точка попадает в класс M.

На следующем этапе для каждого класса необходимо вычислить центр масс (среднее арифметическое всех точек, попавших в данный класс). Теперь найденные центры являются новыми динамическими ядрами классов. После этого необходимо повторить итерацию с нахождением минимального расстояния от каждой точки до каждого центра и вычислением нового центра масс. Итерации необходимо проводить до тех пор, пока система не достигнет равновесия, и изменения внутрикластерного расстояния более не будут происходить. Зацикливание данного алгоритма практически невозможно, поскольку при каждой итерации суммарное квадратичное отклонение уменьшается.

К недостаткам алгоритма k-means можно отнести неоднозначность финальной конфигурации, что требует дополнительных усилий для выявления кластерной структуры в данных. Иногда выбор случайных ядер на первом этапе приводит к неадекватным кластеризациям. Лучшим вариантом является выбор в качестве ядер максимально удалённых друг от друга точек; при этом первые два

ядра инициализируют по максимальному значению всех попарных расстояний, а другие ядра выбираются таким образом, чтобы расстояние от нового ядра до уже выбранного было максимально. Кроме того, кластеризация также может оказаться плохой, если неправильно задать число кластеров. В этом случая для решения проблемы рекомендуется проводить разбиения на разные количества классов K и выбрать то количество, при котором наблюдается наилучший показатель кластеризации [14].

Относительно новый метод упругих приближает многомерные (63-мерные в нашем случае) данные многообразиями малой размерности; в нашем случае мы использовали многообразия размерности 2. Напомним, что наши данные представляют собой набор записей (точек), где каждая запись соответствует тому или иному гену тРНК человека и характеризуется набором из 64 чисел, являющихся частотами соответствующих триплетов. Перед проведением анализа триплет с наименьшим стандартным отклонением (ТАС) исключался из рассмотрения для избавления набора данных от линейности.

Метод наименьших квадратов позволяет аппроксимировать многомерные данные кривой либо поверхностью, однако тип этой кривой (поверхности) строго задан заранее. Метод упругих карт свободен от этого ограничения. Он также аппроксимирует многомерные данные, минимизируя суммарную энергию деформации (см. ниже), однако вместо фиксации типа кривой либо поверхности используются существенно менее сильные топологические ограничения: топология деформированного многообразия должна сохранятся. Это означает, что при выполнении преобразований исходного многообразия разрешены любые преобразования кроме склейки и разрывов.

Опишем процедуру построения упругой карты на примере стартового многообразия, представленного квадратом евклидовой двумерной плоскости. На первом шаге во множестве анализируемых данных определяются первая и вторая главные компоненты, и на них, как на осях, строится евклидова плоскость. Данные проецируются на плоскость; затем на плоскости берётся наименьший квадрат, содержащий все проекции. После этого необходимо выбрать тип карты. Тип

карты определяется числом более мелких квадратов, на которые она разбивается: жесткая карта — 12×12 квадратов, мягкая — 16×16 и детальная — 25×25. После того, как выбран тип карты, точка соединяется не с собственной проекцией, а с ближайшим к этой точке узлом.

На следующем этапе исходно жёсткая плоскость заменяется на упругую мембрану; при этом разрешены следующие деформации: изгиб по ребру, растяжение и сжатие. Заметим, что в ходе деформации каждый маленький квадрат будет деформирован в плоский четырёхугольник. При этом параметры упругости этой самой мембраны считаются заданными: растяжимость во всех направлениях одна и та же, и коэффициент упругости при изгибании двух соседних квадратов вдоль по соседнему ребру. Эту систему, состоящую из математических пружин и упругой мембраны, отпускают так, чтобы система достигла минимального уровня деформации энергии. Математические пружины являются математическими, имеют бесконечную растяжимость и не меняют своих свойств по мере растяжения. После этого на полученной поверхности определяются ортогональные проекции каждой точки. Далее пружины убираются, и карта распрямляется, образую плоскую карту, отчего положение определённых образов также меняет своё место, формируя кластеры.

Опишем идею выделения кластеров подробнее; для этого введём понятие локальной плотности. Снабдим каждый образ точки на упругой карте, представленной во внутренних координатах (то есть плоской) колоколообразной функцией, например, функцией Гаусса. Функция должна быть интегрируемой[3]:

$$f_j x = \exp\left\{-\frac{(x-r)^2}{\sigma^2}\right\} \ , \qquad (1)$$

здесь r_j — координата образа j-ой точки, а σ — свободный параметр, определяющий радиус корреляционного взаимодействия между образами точек. Локальная плотность $F(x)$ определяется как сумма функций (1) по всем точкам:

$$F(x) = \sum_{j=1}^{N} f_j(x) \qquad (2)$$

[3] Использование функции Гаусса не имеет ничего общего с нормальным распределением.

Уровни функции (2) и определяют уровень локальной плотности. Следует заметить, что возможны другие определения функции, кроме (1); единственное ограничение на эту функцию в том, что она должна быть интегрируемой. В нашей работе мы использовали функцию Гаусса.

Результаты

На рис. 1 показано распределение генов тРНК в пространстве частот триплетов по отношению к кодируемым ими аминокислотам во внутренних координатах эластичной карты, построенной в 63-мерном евклидовом пространстве частот триплетов; справа от карты представлена легенда, в которой разными маркерами обозначены соответствующие аминокислоты. Хорошо видно, что в состав кластеров входят гены изоакцепторов одной и той же аминокислоты; этот факт и составляет основное содержание настоящей работы.

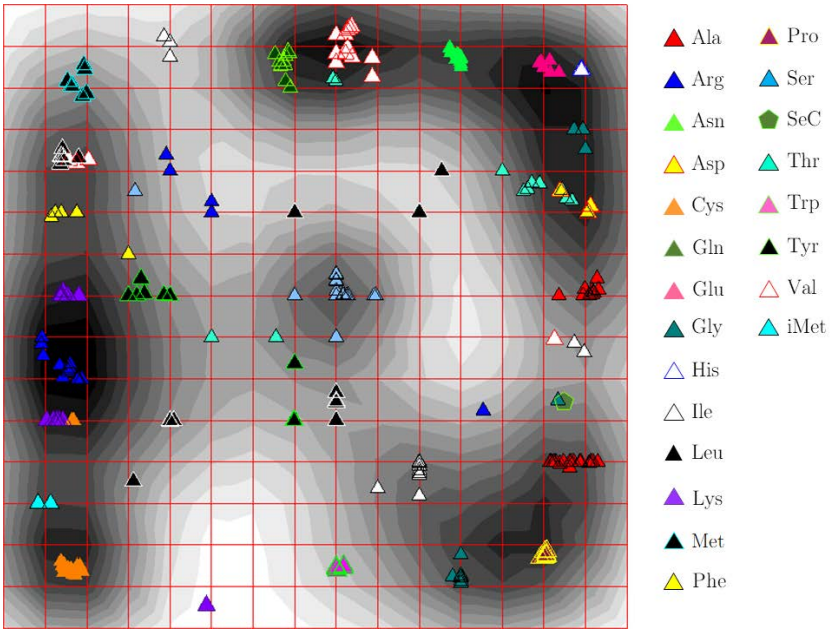


Рис. 1. Общая картина распределения генов тРНК человека по упругой мягкой (16×16) карте для 21 аминокислоты. Распределение генов индивидуальных синонимичных кодонов не показано.

Хорошо известно, что некоторые аминокислоты в матричной РНК кодируются разными кодонами; такие кодоны называются синонимичными. Здесь возникает вопрос: во-первых, как распределяются гены изоакцепторов, содер-

жащих разные антикодоны, однако соединяющихся с одинаковыми аминокислотами в процессе трансляции; во-вторых, что происходит с генами изодекодеров, несущих абсолютно одинаковые антикодоны, однако различающихся по первичной структуре (последовательности нуклеотидов), с точки зрения паттерна кластеризации.

Начнем с анализа распределения изодекодеров фенилаланина, гистидина, аспарагиновой кислоты, аспарагина и цистеина; как правило, данные аминокислоты кодируются двумя синонимами. Однако для каждой упомянутой аминокислоты в нашей базе данных имеется по одному антикодону. Гены тРНК, несущих гистидин, аспарагин и аспарагиновую кислоту, собраны в плотные кластеры, отдельные для каждой аминокислоты.

Гены тРНК, несущих цистеин и фенилаланин, демонстрируют самое

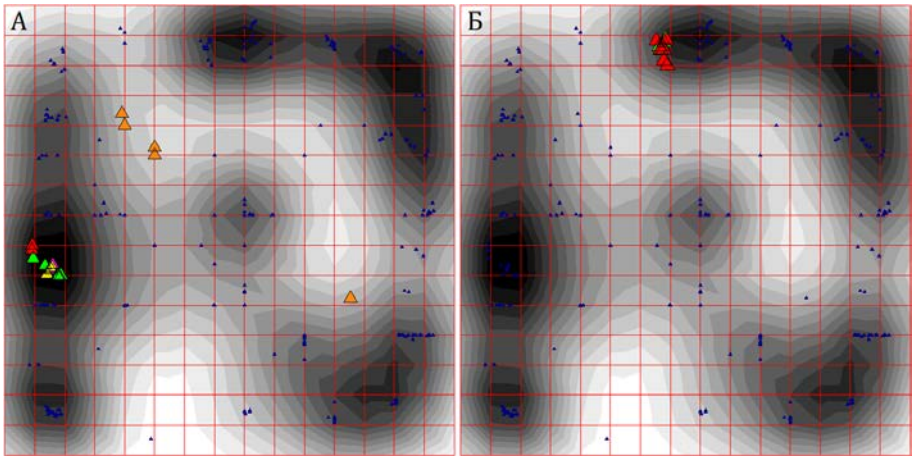


Рис. 2. Сравнение разных паттернов распределения генов разных изодекодеров. Гены изодекодеров Gln (2Б) демонстрируют монокластерный характер распределения, аналогичный таковому для глутаминовой кислоты и пролина, тогда как гены одного из изодекодеров Arg — AGA (2А) находятся на значительном отдалении от основного кластера. Это может быть объяснено фактом присутствия в генах данных изодекодеров интронов.

большое уклонение (самый дальний выход из соответствующего им кластера). Из данных наблюдений следует, что для генов тех изоакцепторов, которые представлены одной группой изодекодеров, характерен единый паттерн распределения. Такой же характер распределения присущ генам тРНК, несущих аминокислоты, не имеющие синонимичных кодонов вовсе. К таковым относятся триптофан, представленный одним кодоном UGG, и селеноцистеин, кодируемый стоп-кодоном UGA.

Изодекодеры всех изоакцепторов глутаминовой кислоты и пролина также собраны в один сплошной кластер, демонстрирующий модель распределения, подобную глутаминовой. При этом гены тРНК некоторых аминокислот, переносимых более чем одним типом изодекодеров, демонстрируют отличный от вышеописанного характер распределения в пространстве частот.

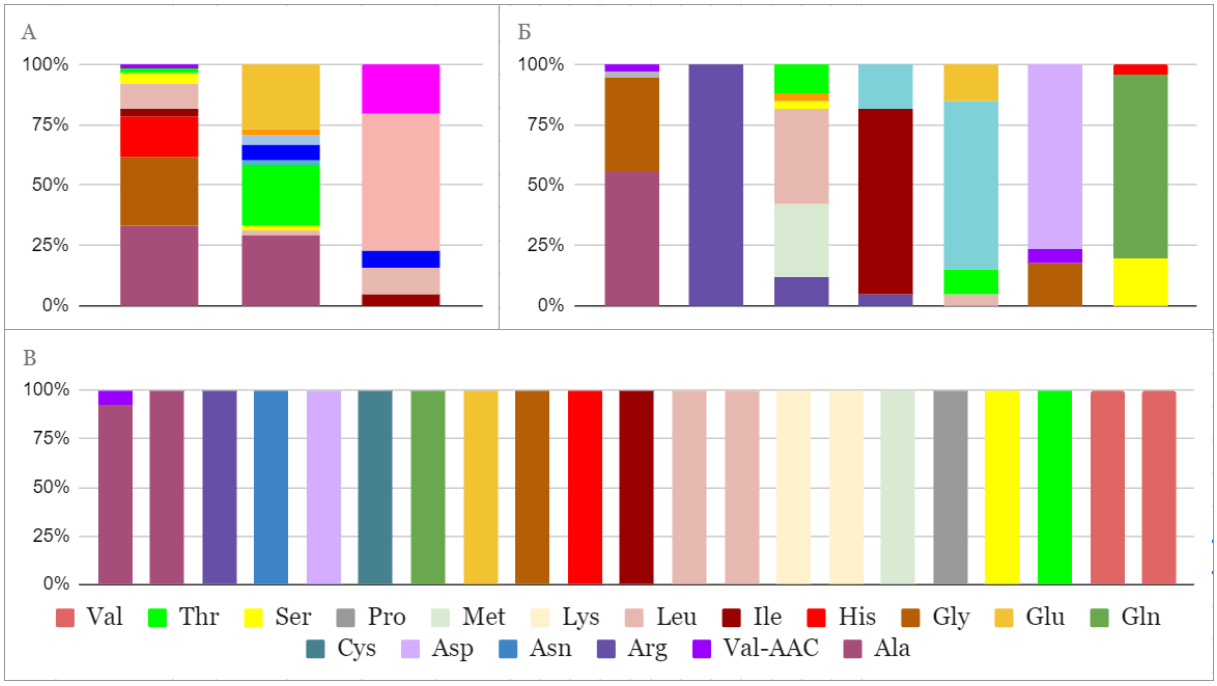


Рис. 3. Нормированные гистограммы, изображающие состав устойчивых групп при делении на 3, 7 и 21 класс (3А, 3Б, 3В соответственно). Заметно, что при увеличении количества классов группы становятся более однородными. Гены тРНК некоторых аминокислот не вошли ни в один кластер. Наиболее устойчивыми оказались группы, содержащие гены тРНК-Ala, Arg, Cys, Leu, Ile.

На рис. 2 показаны два различных паттерна кластеризации изоакцепторов и изодеко-деров для генов тРНК, кодирующих глутамин (2А) и аргинин (2Б). Как правило, гли-цин кодируется двумя синонимичными кодонами и переносится изоакцепторами двух типов. Гены для каждой из разновидностей представлены в нашей базе данных (6 последовательностей генов для изодекодеров CAA, обозначены как зеленые треугольники, и 13 — для CAG, красные треугольники). Набор изодекодеров глутамина демонстрирует монокластерный характер распределения: гены обоих изодекодеров формируют плотную группу. Напротив, распределение генов изодекодеров аргинина демонстрирует противоположную картину.

Аминокислота аргинин кодируется шестью синонимичными кодонами и переносится пятью различными изоакцепторами (тРНК с антикодоном CGC отсутствует). Изодекодеры для четырёх из пяти изоакцепторов сбиваются в чётко обозначенный кластер (CGU, на схеме обозначены как красные треугольники, CGG — желтые, AGG — зелёные треугольники, CGA — фиолетовые, AGA — оранжевые). Изодекодеры, составляющие набор для антикодона AGA, отличаются от других наличием интронов. Это может быть причиной столь большого уклонения точек, соответствующих генам всех тРНК с антикодоном AGA от основного кластера для аминокислоты аргинин. А гены тРНК-ARG демонстрируют типичный для интрон-содержащих генов тРНК. В организме человека к таковым относятся гены тРНК с антикододонами UAC (Tyr), AUA (Ile), TTG (Leu) и упомянутым выше AGA (Arg). Гены этих изодекодеров разбросаны по карте и не образуют плотного кластера. Стоит отметить, что интроны в генах тРНК весьма консервативны, и такого рода распределение можно считать неслучайным.

Для линейной кластеризации данных множество частотных словарей последовательно разбивалось на разное число классов по 10 раз. Необходимость многократного повторения для каждого разбиения на классы обусловлена тем, что при каждом новом разделении на классы результаты новой реализации будут несколько отличаться от предыдущих. Затем выявлялись устойчивые группы — такие наборы точек, которые при каждой реализации часто попадают в один и

тот же класс (при этом номер класса не имеет значения — главное, чтобы одни и те же точки оказывались в одном и том же классе). Частота попадания точек в один и тот же класс (устойчивостью распределения) в нашем случае составил не менее 75%.

Результаты линейной кластеризации представлены в виде нормированных гистограмм (рис. 3), отображающих процентный состав сформированных групп (количество вошедших в состав каждой группы точек в данном случае не столь важно). Кластеры становятся тем однороднее, чем на большее число классов разбиваются данные. Максимальная однородность классов наблюдается при разделении на 21 класс.

При этом большое число точек не вошло ни в одну из устойчивых групп. Число волатильных точек варьировало от 30 % до 63 % от количества всех анализируемых точек. Это объясняется особенностью данных и довольно высоким требуемым уровнем устойчивости (75 %).

Заключение

Независимо от метода анализа частотных словарей, составленных для исследуемых последовательностей генов тРНК, полученные результаты говорят о структурированности данных. Это означает, что распределение и разнообразие последовательностей генов тРНК человека неслучайно, и искать обоснование сформированной структуре данных действительно имеет смысл. В результате нелинейной кластеризации часть точек группируются в обособленные кластеры в соответствии с типом переносимой аминокислоты, часть — в соответствии с антикодоном будущей тРНК. Последовательности некоторых генов демонстрируют рассеянный характер распределения. Особенно это характерно для интронсодержащих генов тРНК, сильно уклоняющихся от кластера.

Поскольку основной целью настоящей работы являлся поиск взаимосвязи между структурой и функцией данных (а именно, ответа на вопрос, группируются ли в один кластер гены тРНК, несущих одну и ту же аминокислоту), постольку был изучен только функциональный состав кластеров, и поиск возможных свя-

зей между структурой и таксономией не осуществлялся, что определило и выбор генетического материала.

Результаты проделанной работы позволяют заключить, что искомая взаимосвязь действительно есть, но для спецификации определяющих её факторов необходимо провести дальнейшие исследования. Результаты линейной кластеризации (МДЯ) фактически показали, что никакой «наследственности» в кластеризации генов тРНК нет. Это может быть непрямым, но сильным свидетельством в пользу древнего расхождения эволюции генов тРНК.

Список литературы

1. Kozak, M. Initiation of translation in prokaryotes and eukaryotes. // Gene. — 1999. — V.234. — № 2. C.187 – 208.
2. Geslain, R., Pan, T. Functional analysis of human tRNA isodecoders. // Journal of molecular biology. — 2010. — V.396. — № 3. C.821 – 831.
3. Goldkamp, A.K., Li, Y., Rivera, R.M., Hagen, D. E. Characterization of tRNA expression profiles in large offspring syndrome. // BMC genomics. — 2022. — V.23. — No.1. C.1 – 16.
4. Kanai, A. *Molecular evolution of disrupted transfer RNA genes and their introns in archaea*. Evolutionary Biology: Exobiology and Evolutionary Mechanisms, 2013. C.181 – 193.
5. Goodenbour, J.M. Pan, T. Diversity of tRNA genes in eukaryotes. // Nucleic acids research. — 2006. — V.34. — No.21. C.6137 – 6146.
6. Chan, P.P., Lowe, T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. // Nucleic acids research. 2009. — V.37(suppl 1), C.D93 – D97.
7. Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., Chen, Y. The functional spectrum of low-frequency coding variation. // Genome biology. — V.12. — No.9. C.1 – 17.
8. Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., Eichler, E.E. 1000 Genomes Project. Diversity of human copy number variation and multicopy genes. // Science. — 2010. —

V.330. — No.6004. C.641 – 646.

9. Khurana E., Fu Y., Colonna V., Mu X.J., Kang H.M., Lappalainen T., Sboner A., Lochovsky L., Chen J., Harmanci A., Das J. Integrative annotation of variants from 1092 humans: application to cancer genomics. // Science. — 2013. — V.342. — No.6154. C.1235587.
10. Schmidt, C.A., Matera, A.G. tRNA introns: presence, processing, and purpose. // Wiley Interdisciplinary Reviews: RNA. — 2020. — V.11. — No.3. C.1583.
11. Arimbasseri, A.G., Maraia, R.J. RNA polymerase III advances: structural and tRNA functional views. // Trends in biochemical sciences. — 2016. — V.41. — No.6. C.546 – 559.
12. Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Yu., eds. Principal manifolds for data visualization and dimension reduction. Vol. 58. Berlin: Springer, 2008.
13. Фукунага К. Введение в статистическую теорию распознавания образов: Пер. с англ. — 1979. — Наука. 576 с.
14. Jain A., Murty M., Flynn P. Data clustering: A review // ACM Computing Surveys. — 1999. — Vol. 31, no. 3. — Pp. 264–323.

КЛАСТЕРИЗАЦИЯ БАКТЕРИЙ ПО ТРИПЛЕТНОМУ СОСТАВУ ГЕНОВ 5S РНК

Ю.И.Овчинникова[1], М.Г.Садовский[2,3,4]

[1]Сибирский федеральный университет, ИФБиБТ, ***july.l4o6@mail.ru***

[2]Институт вычислительного моделирования СО РАН, ***msad@icm.krasn.ru***

[3]Федеральный Сибирский научно-клинический центр ФМБА России,

[4]Красноярский государственный медицинский университет МЗ РФ

Связь между структурой нуклеотидных последовательностей, кодируемой ими функцией и таксономией носителя является важнейшим предметом исследования молекулярных биологов и биоинформатиков, а также специалистов по обработке больших массивов данных. В этом направлении существуют различные подходы; один их них связан с изучением связи структуры биологических макромолекул (конкретно — молекул рибосомальных РНК бактерий) и таксономии их носителей. Для указанных организмов классическим предметом исследования являются последовательности 16 S РНК. Однако не меньший интерес представляет и изучение связи между структурой и таксономией для других последовательностей. В рамках настоящей работы мы изучали связь между структурой и таксономией на примере последовательностей 5 S РНК бактерий.

Структура биологических макромолекул может быть определена многими различными способами; в рамках настоящей работы под структурой мы будем понимать частотный словарь триплетов (иногда также называемый частотным профилем) W_j; здесь индекс j перечисляет рассматриваемые последовательности. Под частотным словарём триплетов будем понимать список всех триплетов, встречающихся в изучаемой последовательности, с указанием частот этих триплетов. Построение частотного словаря преобразует нуклеотидную (символьную) последовательность в точки в 63-мерном метрическом пространстве, делая их математическими объектами, позволяющими для своего изучения применять весь арсенал соответствующих средств.

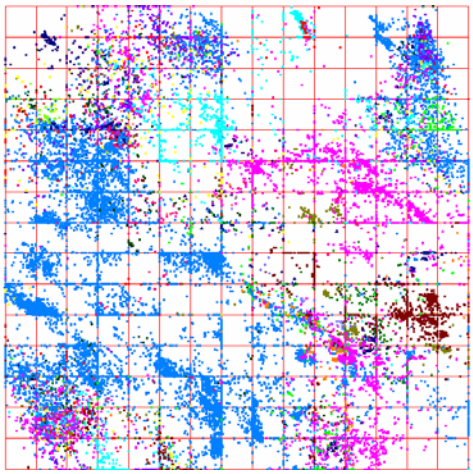
Настоящая работа имеет две цели:

1) изучить особенности кластеризации частотных словарей триплетов генов 5 S РНК бактерий в части связи между такой кластеризацией и таксономическим составом кластеров (если они будут обнаружены). Заметим, что 5 S РНК бактерий не являются самым широко распространённым объектом для такого рода исследований и сравнение полученных результатов с другими, например, построенными по кластеризации на основе генов 16 S РНК бактерий, определяет новизну данной работы;
2) изучить устойчивость наблюдаемой кластеризации (если она будет обнаружена) к случайному отбору исключаемых сверхпредставленных таксонов.

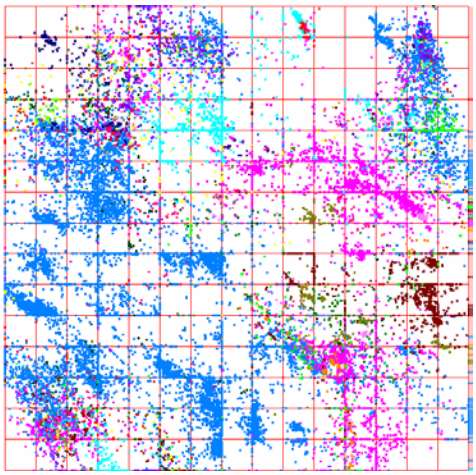
Поясним вторую цель подробнее. Ранее [1 – 3] было показано, что триплетный состав различных генетических систем очень хорошо коррелирует с таксономическим положением носителей соответствующих генов. Однако зачастую видовой состав баз генетических данных (в нашем случае — базы SILVA) весьма смещён: некоторые таксоны низкого уровня представлены непропорционально большим количеством записей (например, на уровне штаммов и т.п.), что приводит к искажению картины распределения частотных словарей в метрическом пространстве.

Стандартным выходом из такой ситуации является индексация базы данных: удаление части записей, находящихся в сверхпредставленных группах. При этом возникает закономерный вопрос: каково влияние выбора удаляемых объектов на кластеризацию? Ответ на вопрос заранее неизвестен, в общем случае, и одна из задач настоящей работы — проверить, насколько велико такое влияние.

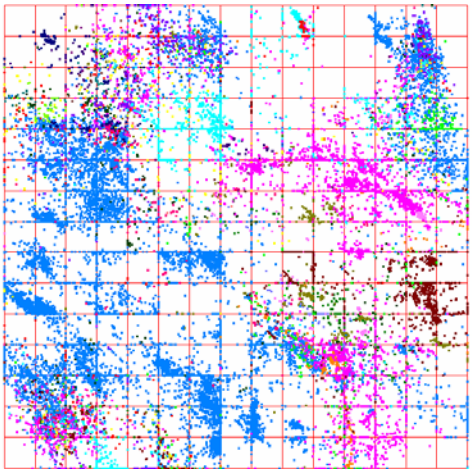
Перейдём к описанию самой работы и результатов. База генов SILVA содержит в общей сложности 182697 записей генов 5 S РНК бактерий; это число генов содержалось в базе данных до индексации. Число записей, использовавшихся в работе, представлено в Таблице 1. В этой таблице представлены значения числа генов в исходной базе до и после индексации. Для одной из реализаций индексированной базы данных число записей в ней составило 49535 генов.



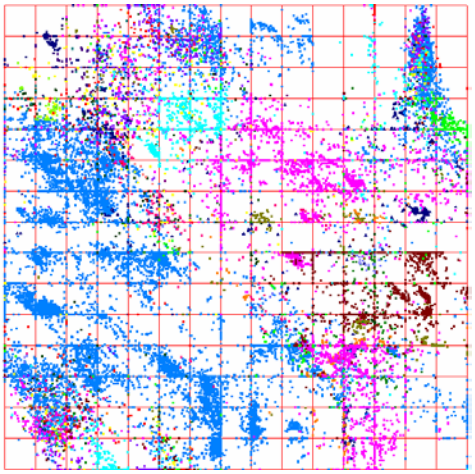
База 1



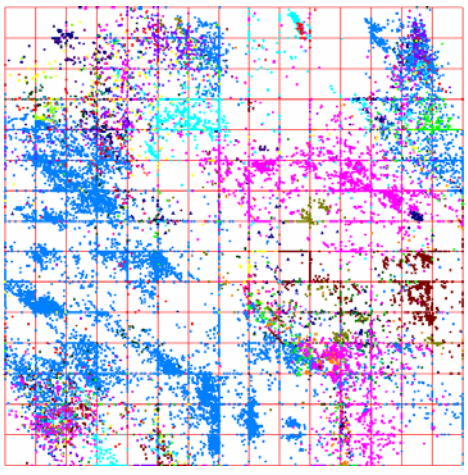
База 2



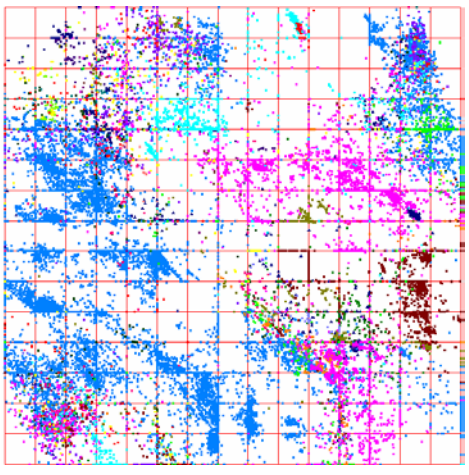
База 3



База 4



База 5



База 6

Рис.1. Распределение 49535 генов 5 S РНК всех двадцати трёх отделов по упругой карте.

Таблица 1

Состав базы генов 5 S РНК бактерий SILVA на уровне отделов. N — число генов в исходной базе, N_1 — число генов после индексации.

Отдел	N	N_1	Отдел	N	N_1
Actinobacteriota	562	291	*Fusobacteriota*	694	694
Bacteroidota	855	491	*Gemmatimonadota*	110	55
Campylobacterota	7744	2589	*Myxococcota*	732	555
Chloroflexi	3180	2429	*Nitrospirota*	131	33
Cyanobacteria	3090	2316	*Patescibacteria*	776	360
Deferribacterota	118	59	*Planktomycetota*	432	296
Deferrisomatota	101	44	*Proteobacteria*	158610	27612
Deinococcota	432	212	*Spirochaetota*	1567	1433
Desulfobacterota	708	451	*Synergistota*	78	78
Elusimicrobiota	126	68	*Thermotogota*	179	179
Fibrobacterota	74	65	*Verrucomicrobiota*	1318	1180
Firmicutes	12748	8336			

Изучение устойчивости кластеризации по отношению к случайному выборов элементов для индексации базы генов 5 S РНК бактерий проводилось с помощью метода упругих карт — см. подробнее [4 – 6]. Это нелинейный метод, позволяющий уменьшать размерность данных и выявлять кластеры. Следует также подчеркнуть, что для исходных данных линейно независимыми являются лишь 63 триплета — это обусловлено тем, что сумма частот всех триплетов, определяемых для каждого гена, равна единице. Тем самым, из анализа следует исключать какой-то один триплет; формально исключённым может быть любой триплет, однако на практике исключать следует такой, для которого стандартное отклонение, определяемое по всей базе исследуемых генов, является минимальным. Такой выбор обусловлен тем, что этот триплет вносит наименьший вклад в различимость генов.

Кластеризация генов 5 S РНК бактерий по их частотным словарям триплетов проводилась с помощью свободно распространяемого ПО *VidaExpert* [5 – 7].

На Рисунке 1 показано распределение всех двадцати трёх отделов изученных бактерий по упругой карте; для изучения распределения была выбрана мягкая (16 × 16) карта; все параметры карты были выбраны по умолчанию. Самый важный вывод из результатов, представленных на Рисунке 1, состоит в том, что 5 S РНК бактерий столь же чувствительны к таксономическому составу, как и их 16 S РНК; вопрос об относительной чувствительности и точности требует дополнительных исследований.

На Рисунке 1 показаны результаты кластеризации методом упругих карт, полученные для шести разных версий базы генов 5 S РНК бактерий. Напомним, что сравнительное исследование этих кластеризаций, полученных на разных базах данных, и есть основной результат нашей работы. На этом рисунке точки, соответствующие каждому из отделов, перечисленных в Таблице 1, показаны своим цветом (см. Таблицу 2). Может создаться впечатление, что на этом рисунке число цветов, использованных для выделения маркеров, меньше 23. Однако это не так. Кажущееся уменьшение числа цветов объясняется тем, что некоторые из точек проектируются одна на другую и на упругих картах (см. Рис. 1), представленных во внутренних координатах, налагаются одна на другую и становятся не видны.

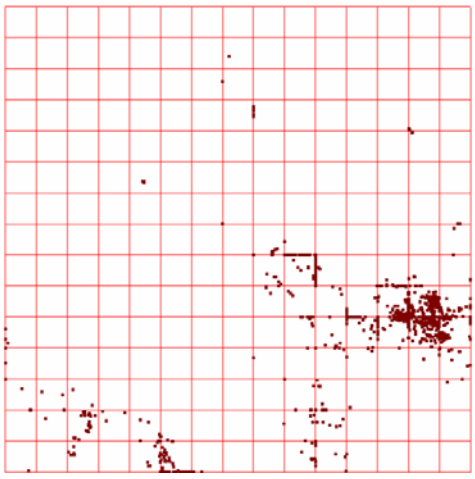
Таблица 2

Расшифровка цветовой кодировки отделов бактерий, рассмотренных в работе

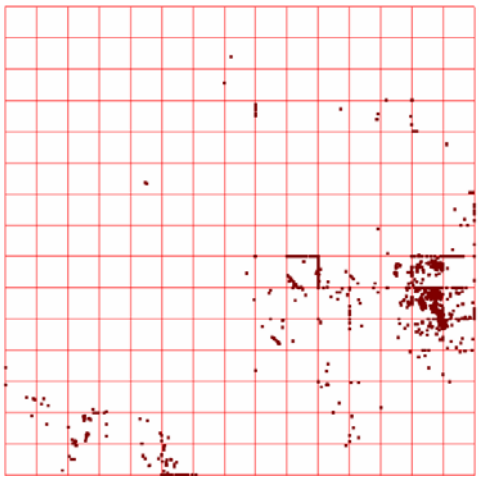
	R	G	B	цвет
Actinobacteriota	255	0	0	
Bacteroidota	0	255	0	
Campylobacterota	128	0	0	
Chloroflexi	255	255	0	
Cyanobacteria	0	255	255	
Deferribacterota	128	128	128	
Deferrisomatota	128	0	128	
Deinococcota	0	0	128	

Desulfobacterota	255	0	128	
Elusimicrobiota	128	128	255	
Fibrobacterota	128	255	0	
Firmicutes	255	0	255	
Fusobacteriota	255	128	0	
Gemmatimonadota	0	128	128	
Myxococcota	128	0	255	
Nitrospirota	255	128	128	
Patescibacteria	0	128	0	
Planctomycetota	0	64	0	
Proteobacteria	0	128	255	
Spirochaetota	128	128	0	
Synergistota	128	64	64	
Thermotogota	64	0	128	
Verrucomicrobiota	255	128	255	

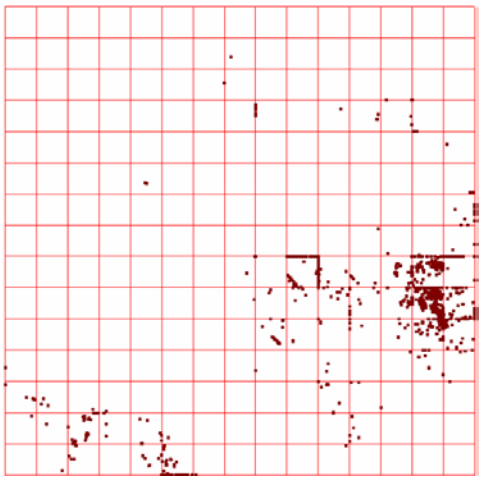
Прежде чем описывать результаты исследования устойчивости кластеризации к составу базы генов, укажем на один важный и нетривиальный результат. Ранее [2, 3] было показано, что кластеризация методом упругих карт и классификация методом динамических ядер генов 16 S РНК бактерий выявляет очень сильную связь между классами либо кластерами, и таксономией носителей этих генов. Подобного рода поведение естественно ожидать и от других типов генов; результаты, представленные в настоящей работе, однозначно доказывают, что для генов 5 S РНК бактерий наблюдается полностью аналогичная картина кластеризации. Очевидно, что кластеры, выделяемые по частотам триплетов генов 5 S РНК бактерий, столь же чувствительны к таксономическому положению носителей этих генов. Иными словами, так же, как и для генов 16 S РНК бактерий, кластеры, выделяемые на основе значений частот триплетов генов 5 S РНК бактерий, также формируются с явным таксономическим предпочтением. В частности, из этого совпадения следует, что отдельного внимания заслуживает одновременное сравнительное исследование результатов классификации



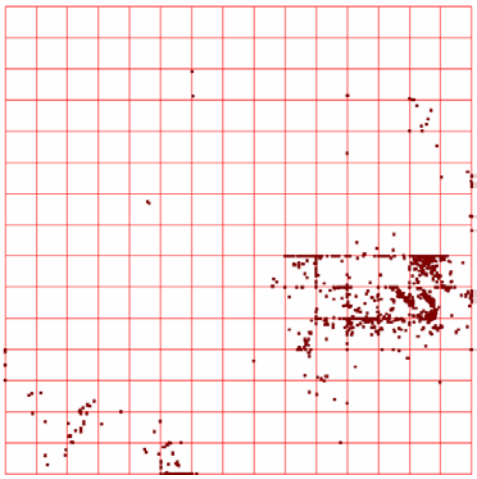
База 1



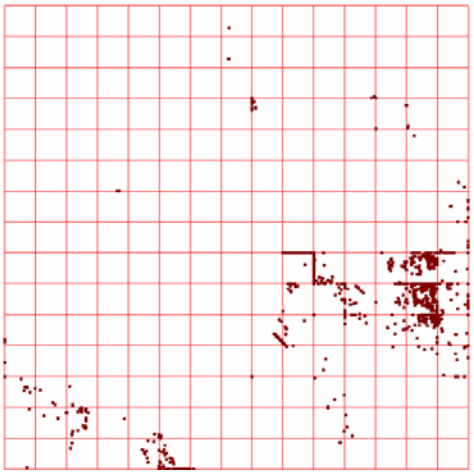
База 2



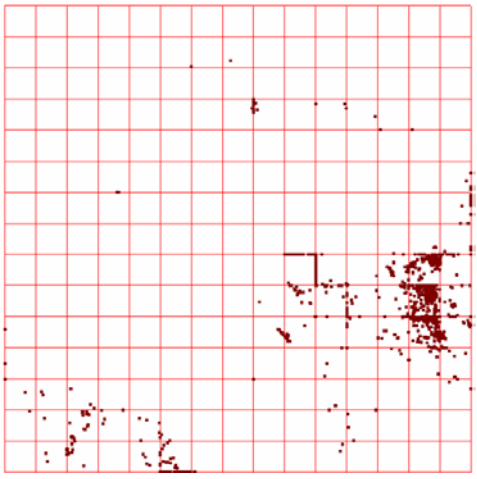
База 3



База 4



База 5



База 6

Рис.2. Индивидуальное распределение генов 5 S РНК бактерий отдела *Campylobacterota* на мягкой упругой карте 16 × 16, для шести разных реализаций баз данных.

(кластеризации) **одних и тех же** организмов — бактерий в нашем случае, — получаемых для разных генов их S РНК. Однако детальное обсуждение этих исследований выходит за рамки настоящей работы.

Перейдём теперь к описанию основного результата настоящей работы — исследованию устойчивости результатов кластеризации к случайному удалению записей в тех таксонах, которые являются сверхпредставленными в исходной базе данных. Как уже говорилось выше, для этого были созданы шесть баз данных, в которых записи генов в таксонах с малой представленностью (т.е. с небольшим числом видов в них, до 30 записей) были исключены из анализа, поскольку такие малопредставленные данные не способны породить сигнал должной силы, однако формируют шум, который существенно искажает картину кластеризации.

Записи генов в сверхпредставленных таксонах были частично удалены, и при этом выбор удаляемых генов осуществлялся случайным образом (индексирование). Индексирование сверхпредставленных таксонов необходимо для того, чтобы подавить слишком сильный сигнал от такой группы организмов, который также способен существенно исказить результаты кластеризации.

Каждая из шести баз данных (см. Рисунок 1) были обработаны упругой картой с одними и теми же значениями параметров. Сравнительный анализ распределений, показанных на Рисунке 1, позволяет утверждать, что влияние состава базы, по которой строится кластеризация, есть; однако это влияние не велико. Действительно, если взять любую пару карт из представленных на Рисунке 1, и сравнить их между собой, то можно видеть, что буквального, попиксельного соответствия не наблюдается. Иными словами, если две такие карты наложить одна на другую, то они не совпадут.

С другой стороны, видимые различия между этими картами весьма невелики. Такого рода расхождения в наблюдаемых распределениях могут быть строго описаны и формализованы, однако мы не будем заниматься этим в данной работе. Мы ограничимся тем, что все эти карты очевидным образом визуально очень близки между собой. Собственно, это наблюдение и даёт ответ на основ-

ной вопрос настоящей работы: верно ли, что разные базы, полученные случайным образом (по процедуре, описанной выше), дают близкий или даже совпадающий результат.

Это полностью подтверждает Рисунок 2, на котором представлены карты, показывающие один и тот же отдел (*Campylobacterota*), для разных реализаций баз данных. На этом рисунке показано распределение генов 5 S РНК бактерий, полностью повторяющее то, что показано на Рисунке 1. Однако на этом рисунке маркеры всех отделов, кроме *Campylobacterota*, имеют нулевой размер. Иными словами, картина плотности распределения и само распределение остаются прежними, но видимым являются только гены указанного отдела. Представители этого отдела достаточно многочисленны, чтобы увидеть характерные изменения в форме кластеров. Хорошо видно, что в целом все шесть карт весьма близки друг к другу: показанный на них кластер имеет почти совпадающую форму. Более того, видно, что пары баз (2 – 3) и (5 – 6) дают фактически идентичные картины распределения данного отдела. Такая близость структур говорит об очень малом влиянии конкретного набора генов в индексированной базе данных на результаты кластеризации.

Как видно из представленных результатов, ответ на вопрос об отсутствии сильного влияния случайной индексации на картину кластеризации с точки зрения таксономического состава выделяемых кластеров положителен: такое влияние следует считать очень слабым и несущественным для дальнейших исследований. Полученные в настоящей работе результаты доказывают, что состав базы, определяемый случайным исключением сверхпредставленных записей, не оказывает сколько-нибудь существенного влияния на результаты кластеризации. Это наблюдение позволяет впредь использовать любую из индексированных баз, получаемых случайным исключением сверхпредставленных таксонов, для исследования связи структуры, функции и таксономии.

Работа выполнена при частичной (ОЮИ) поддержке гранта Министерства науки и высшего образования РФ (075-15-2022-1121).

Список литературы

1. Sadovsky M., Putintseva Yu., Chernyshova A., Fedotova V. Genome structure of organelles strongly relates to taxonomy of bearers // International Conference on Bioinformatics and Biomedical Engineering / Springer. — 2015. — P. 481–490.
2. Gorban A., Popova T., Sadovsky M. Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy // Open Systems & Information Dynamics. — 2000. — Vol. 7, no. 1. — P. 1–17.
3. A. Teterleva, V. Abramov, A. Morgun, I. Larionova, M. Sadovsky Unsupervised Classification of Some Bacteria with 16S RNA Genes // LNBI, vol.13346, Part I, 2022, pp.205 – 215.
4. Gorban A., Zinovyev A. Principal Manifolds for Data Visualisation and Dimension Reduction // Lecture Notes in Computational Science and Engineering — Berlin – Heidelberg – New York: Springer, 2007. — Vol. 58. — P. 153–176.
5. Gorban A., Zinovyev A. Fast and user-friendly non-linear principal manifold learning by method of elastic maps // 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015. — 2015. — P. 1–9. — Access mode: https://doi.org/10.1109/DSAA.2015.7344818.
6. Gorban A., Zinovyev A. Elastic principal graphs and manifolds and their practical applications // Computing 75, no. 4 (2005): 359-379.
7. Gorban, Alexander N., Alexander Pitenko, and Andrei Zinovyev. "ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data." arXiv preprint arXiv:1406.5550 (2014).

МЕТАМОДЕЛЬ ДЕТАЛИЗАЦИИ ИНТЕГРАЛЬНЫХ ОЦЕНОК ДЛЯ ОПРЕДЕЛЕНИЯ ПРИЧИН СОСТОЯНИЯ ПРИРОДНО-ТЕХНОГЕННОЙ БЕЗОПАСНОСТИ ТЕРРИТОРИЙ

Т.Г. Пенькова, В.В. Ничепорчук

Институт вычислительного моделирования СО РАН, ***penkova_t@icm.krasn.ru, valera@icm.krasn.ru***

Введение

Важнейшими этапами планирования мероприятий долгосрочного развития территорий является сбор и анализ показателей их состояния, а также определение целей, которые планируется достичь. Развитие информационных технологий привело к естественному росту количества показателей, позволяющих детально описывать необходимые аспекты развития сложных систем. Особенности восприятия человеком информации обуславливают необходимость ограничения объёма информации и принятие решений, как правило, сопровождается постоянным обобщением или редуцированием данных. При этом использование разных методов «сжатия» информации может привести к существенным искажениям реальности, что сказывается на качестве решений – отклонении результатов мероприятий от ожидаемых. Управление большими территориями требует дифференцированного подхода даже в рамках одного региона. Принятие обоснованных решений должно основываться на анализе большого числа разнородных показателей с учётом особенностей конкретных территорий [1, 2].

Предупреждение чрезвычайных ситуаций является одной из важнейших задач территориального управления. Указом Президента РФ от 16.10.19 г. № 501 утверждены целевые показатели безопасности, которые необходимо достичь к 2030 году [3]. В частности, поставлена задача снижения на 20% количества чрезвычайных ситуаций (ЧС) и потерь от них. Для повышения безопасности жизнедеятельности населения и территорий активно внедряются системы оперативного управления. Созданы обширные сети мониторинга потенциальных источников ЧС, наблюдения за параметрами окружающей среды, внедряются чувстви-

тельные датчики контроля технологических процессов, системы видеомониторинга [4, 5].

Наряду с инструментальными средствами мониторинга развиваются методы обработки больших массивов данных, поддержки принятия решений, в том числе в сфере стратегического управления территориальной безопасностью [6]. В России и в мире ведётся большое количество исследований методов анализа рисков, оценивания текущего состояния социально-природно-техногенных систем и прогнозирования развития ситуаций [7]. Однако, в большинстве случаев, природные и техногенные процессы рассматриваются независимо, что не позволяет оценивать обстановку комплексно с учётом влияния многих факторов. Для оценивания состояния безопасности используется три основных подхода. *Вероятностный подход* позволяет рассчитывать оценку риска возникновения ЧС с помощью математических моделей, связывающих предпосылки с вероятностью их проявления. Методы данного типа используются для расчёта индивидуальных, коллективных и социальных рисков и ориентированы, как правило, на производственные объекты определённого вида [8]. Их применение для оценок территорий требует совершенствования нормативной базы и серьёзной адаптации расчетных моделей. *Статистические методы* позволяют формировать оценки на основе анализа данных за определенный период наблюдения. Их достоинствами является объективность, возможность исследовать динамику изменений наблюдаемых параметров и формировать сводные показатели. Однако такие методы не могут быть применены для редко наблюдаемых событий. *Эвристический подход* используется при формировании оценки, когда формальные методы слишком сложны, а исходная база данных недостаточна для получения однозначного аналитического решения. Однако применение методов данного типа зачастую ведёт к ошибкам субъективного характера. Таким образом, подтверждается востребованность *гибридного подхода*, позволяющего получать комплексные оценки безопасности территорий с учетом взаимного влияния различных факторов риска и их временн*о*го развития.

С учетом требований, предъявляемых к управлению безопасностью боль-

ших территорий, и преимуществ существующих подходов к анализу рисков, авторами был предложен *метод интегрального аналитического оценивания* состояния территорий, обеспечивающий формирование комплексного показателя природно-техногенной безопасности по данным мониторинга состояния окружающей среды и объектов техносферы [9, 10]. Интегральная оценка комплексного показателя определяется по иерархии оценок, рассчитываемых на основе территориально-ориентированных нормативов. Иерархия оценок позволяет получать обобщенные количественные характеристики состояния безопасности территорий, выполнять их сравнительный анализ и, в случае необходимости, детализировать оценки до конкретных показателей, что даёт возможность определять первопричины текущего состояния и формировать целевые управляющие рекомендации. В данной работе рассматривается метамодель и алгоритм детализации интегральных оценок для определения причин состояния природно-техногенной безопасности территорий и формирования управляющих рекомендаций, основанные на иерархических зависимостях причин возникновения проблемных ситуаций.

Принципы формирования иерархии интегральных оценок природно-техногенной безопасности территорий

Комплексный показатель природно-техногенной безопасности территорий формируется путем построения иерархического представления оценок показателей, характеризующих природные и техногенные факторы риска возникновения ЧС. Расчёт оценок выполняется на основе территориально-ориентированной нормативной модели. Нормативная модель определяет «желаемый» уровень безопасности с учетом индивидуальных особенностей территорий и реальных возможностей его достижения. Нормативная модель включает иерархическую систему показателей, коэффициенты значимости показателей, функции агрегирования оценок, нормативные значения показателей и коэффициенты чувствительности оценок [11].

Основой интегрального оценивания является иерархия показателей, кото-

рая содержит два типа показателей: *базовые показатели*, представляющие нижний уровень иерархии, и *комплексные показатели*, формирующие остальные уровни иерархии по принципу агрегирования. На рисунке 1 представлен фрагмент иерархии показателей, демонстрирующий комплексные показатели. Данная иерархия разработана для оценки природно-техногенной безопасности территорий Красноярского края и учитывает опасные события техногенного и природного характера, а также результаты мониторинга состояния окружающей среды. Коэффициенты значимости показателей применяются для расчёта интегральных оценок комплексных показателей и определяют относительные весовые коэффициенты, которые характеризуют вклад показателей нижнего уровня иерархии в показатели верхнего уровня. Функции агрегирования оценок применяются при расчете оценок базовых показателей и определяют тип функции, обеспечивающей переход от многомерных оценок, рассчитанных по нескольким пунктам наблюдения, к одномерным значениям, рассчитанным для территории в целом. Нормативные значения показателей являются целевыми значениями базовых показателей и определяются индивидуально для каждой территории по результатам анализа многолетних наблюдений. Коэффициенты чувствительности оценок применяются также для расчета оценок базовых показателей и задают скорость изменения оценок при отклонении фактических значений от заданных нормативных значений. Все характеристики нормативной модели разрабатываются на основе спецификаций территорий с учетом их физико-географических и социально-экономических особенностей с привлечением экспертов.

Формирование иерархии интегральных оценок состоит в последовательном расчете оценок базовых и комплексных показателей. Оценки базовых показателей определяют соответствие фактических значений нормативу, учитывая тенденцию показателя, и позволяют оценить степень изменения показателя по отношению к нормативным значениям с учетом заданной чувствительности оценок. Расчет оценок комплексных показателей выполняется на основе полученных оценок показателей и их значимости снизу вверх по дереву иерархии. Дополнительно, выполняется интерпретация оценок показателей путем преобразо-

вания количественных значений в эквивалентные качественные значения по оценочной шкале. Оценочная шкала определяется как нечеткая логико-лингвистическая переменная, соответствующая рассматриваемому показателю и принимающая множество значений: «Улучшенный», «Хороший», «Приемлемый», «Удовлетворительный», «Пониженный», «Низкий», «Критический».

Рис. 1. Фрагмент иерархии показателей
природно-техногенной безопасности территорий

Получаемая иерархия оценок, с их количественными и качественными значениями, дает возможность лицу принимающему решения, с одной стороны, оперировать обобщенной информацией, выполнять сравнительный анализ территорий на любом уровне агрегирования, отслеживать динамику изменения комплексных показателей для планирования стратегических мероприятий управления безопасностью и, с другой стороны, детализировать оценки до конкретных показателей, переходя на уровень управления рисками возникновения чрезвычайных ситуаций. Более подробное описание процессов формирования нормативной модели и интегрального оценивания природно-техногенной безопасности территорий представлено в работах [9 - 11].

Метамодель детализации интегральных оценок для определения причин состояния природно-техногенной безопасности территорий

Определение причин текущего состояния природно-техногенной безопасности территорий основано на анализе сложившейся ситуации по результатам интегрального оценивания. Возникновение проблемных ситуаций отражается в оценках комплексных показателей. Поэтому метамодель определения причин состояния природно-техногенной безопасности территорий строится на основе иерархии оценок и представляет собой *когнитивную карту*, позволяющую анализировать сложившуюся ситуацию с выделением конкретных проблем и причин (рисунок 2).

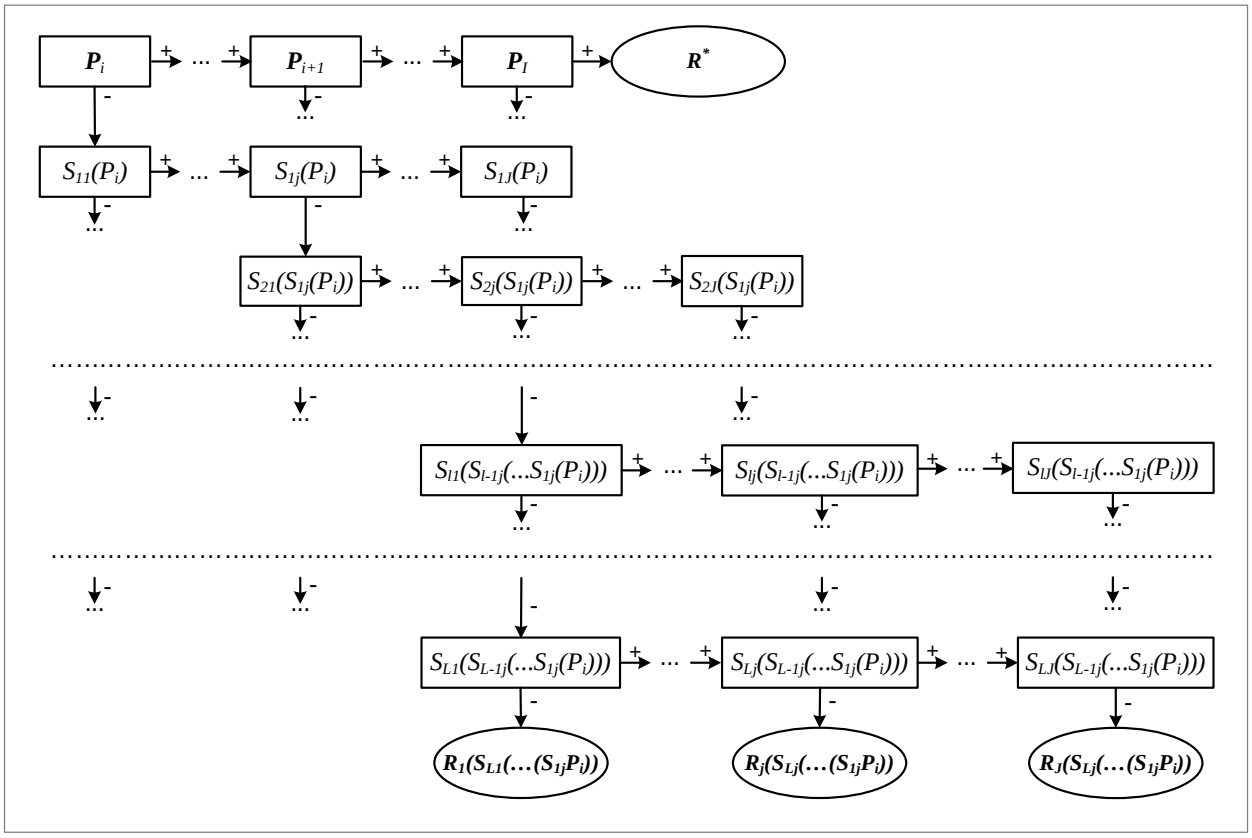


Рис. 2. Метамодель определения причин
состояния природно-техногенной безопасности территорий

Согласно теории когнитивного анализа и моделирования [12], под «ситуацией» понимается сочетание условий и обстоятельств, создающих определенную обстановку, в которой возникла «проблема». Под «проблемой» понимается несоответствие желаемого и фактического уровней достижения целей, возникшее в результате определенных «причин». Для определенного сочетания проблем и

причин формируются управляющие рекомендации, направленные на достижение желаемого уровня целей. Особенность предложенной метамодели состоит в применении иерархических зависимостей между причинами возникновения проблемных ситуаций.

В ходе моделирования ситуации определяется показатель «проблема» P_i, характеризующий текущее состояние, где $i = \overline{1,I}$ – количество возникших «проблем». Для каждого показателя «проблема» определяется совокупность иерархически подчиненных показателей «причина» $S_{lj}(S_{l-1j}(\dots S_{1j}(P_i)))$, которые характеризуют возникшую проблемную ситуацию, где $l = \overline{1,L}$ – количество уровней иерархии «причин», $j = \overline{1,J}$ – количество показателей «причина» на уровне иерархии. Так, например, определена «проблема» P_1: оценка комплексного показателя «Природно-техногенная безопасность» имеет значение «Пониженный». Согласно иерархии комплексных показателей (рисунок 1), данная «проблема» связана со следующими причинами: «Техногенная безопасность» $S_{11}(P_1)$ и «Природная безопасность» $S_{12}(P_1)$. В свою очередь, каждый из указанных показателей «причина», содержит дочерние показатели «причина». Так, для показателя $S_{11}(P_1)$ определяются «причины»: $S_{21}(S_{11}(P_1))$ «Обстановка на объектах техносферы», $S_{22}(S_{11}(P_1))$ «Транспортная обстановка», $S_{23}(S_{11}(P_1))$ «Обстановка на объектах ЖКХ», $S_{24}(S_{11}(P_1))$ «Пожарная обстановка» и $S_{25}(S_{11}(P_1))$ «Радиационная обстановка». Для показателя $S_{12}(P_1)$ определяются «причины»: $S_{21}(S_{12}(P_1))$ «Метеорологическая обстановка», $S_{22}(S_{12}(P_1))$ «Гидрологическая обстановка», $S_{23}(S_{12}(P_1))$ «Геофизическая обстановка» и $S_{24}(S_{12}(P_1))$ «Лесопожарная обстановка». И так далее, спускаясь вниз по иерархии до базовых показателей. При этом, показатель «проблема» может задаваться на любом уровне иерархии комплексных показателей.

В когнитивной модели используется два типа причинно-следственных связей: положительные (+) и отрицательные (-), которые соответствуют состоянию показателей: «норма» и «не норма». Состояние «норма» и «не норма», обеспечивающее направленный переход между концептами метамодели, определяется исходя из значений интегральных оценок. Например, в качестве состояния «не

норма» могут рассматриваться значения оценок «Пониженный», «Низкий» или «Критический», либо ухудшение оценки показателя по сравнению с предыдущим отчетным периодом. В случае, когда все показатели «проблема» P_i в «норме», формируется общая рекомендация R^+. Если «проблема» P_i в «не норме», то на каждом уровне иерархии существует показатель «причина» в состоянии «не норма», что обеспечивает движение вниз по иерархии. Рекомендация $R_j(S_{Lj}(\dots S_{1j}(P_i)))$ формируется для случая, когда показатель «причина» $S_{Lj}(S_{l-1j}(\dots S_{1j}(P_i)))$ уровня иерархии L в состоянии «не норма». Управляющие рекомендации могут меняться в зависимости от конкретного значения оценки.

Алгоритм детализации интегральных оценок состояния природно-техногенной безопасности территорий

В основе алгоритма детализации интегральных оценок лежит «дедуктивный» принцип реализации стратегического управления природно-техногенной безопасностью территорий. На начальном уровне определяются *территориальные приоритеты* на основе расчета оценок интегральных показателей и составления картограмм снижения рисков. Формирование рекомендаций и решений по планированию управляющих мероприятий выполняется с привлечением специалистов конкретных муниципальных образований. На следующем уровне осуществляется распределение *опасностей по видам ситуаций* с использованием классификатора МЧС России [13], имеющего иерархическую структуру, сопоставимую с иерархией показателей для интегрального оценивания. Такая детализация позволяет определить зону ответственности ведомств и отраслей, полномочия которых включают обеспечение безопасности по видам обстановок. Далее анализируются факторы, являющиеся *причинами* опасных событий или возникновения ЧС [14]. Для каждого вида обстановки причины определяются исходя из разных способов группировки факторов рисков (управляемые / не управляемые; природные / антропогенные / социальные; детерминированные / вероятностные и т.п.). На основе выявленных факторов риска, с привлечением экспертов в конкретной области, проводится обоснование объёмов планируемых мер и управля-

ющих мероприятий.

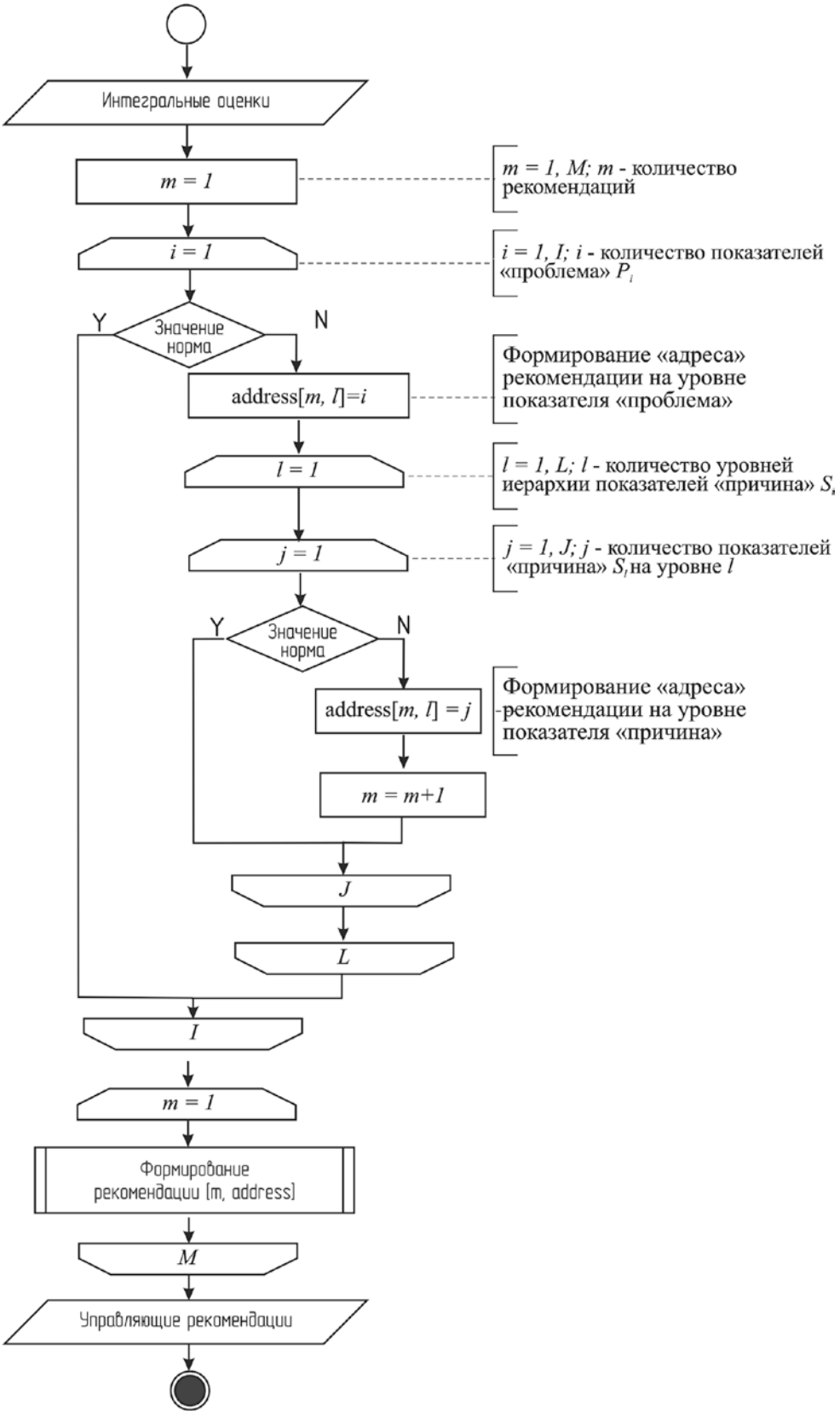


Рис. 3. Алгоритм детализации интегральных оценок природно-техногенной безопасности территорий.

На основе предложенной метамодели с учетом изложенных принципов стратегического управления разработан *алгоритм детализации интегральных оценок* с формированием управляющих рекомендаций (рисунок 3).

Работа алгоритма заключается в определении причин возникновения проблемы путем проверки значений интегральных оценок на соответствие «норме» и формировании рекомендаций по нивелированию негативных факторов. Движение по иерархии показателей обеспечивают три цикла. Сначала проверяется соответствие значений интегральных оценок для показателей «проблема». В случае соответствия всех показателей формируется рекомендация о том, что изменений мероприятий не требуется и алгоритм прекращает работу. Затем, при несоответствии одного или более показателей «проблема», запускается «маршрут» движения по уровням иерархии показателей «причина» с проверкой значений их оценок на каждом уровне. Как правило, показатель «проблема» имеет несколько показателей «причина», у которых оценка не соответствует норме. После проверки всех показателей «причина», выполняется формирование управляющих рекомендаций. Число рекомендаций равно числу выявленных «причин» последнего уровня.

Процесс формирования *управляющих рекомендаций* основывается на анализе факторов риска для базовых показателей. Например, в результате оценивания природно-техногенной безопасности территорий Красноярского края выявлена проблема: показатель «Техногенная обстановка» имеет значение «Низкий» для двух территорий. Одна из территорий относится к типу пригородный район, другая – удаленные территории. Детализация интегральных оценок позволила выявить несоответствующие норме показатели «причины»: «Пожарная обстановка» – «Бытовые пожары». После чего, на этапе формирования управляющих рекомендаций, выполняется анализ факторов риска для бытовых пожаров, что позволит выявить первопричины. Так, для территории пригородного района определены факторы риска – пожары в загородных домах (садовых участках), для удаленных территорий – пожары в жилых домах. На основе статистических данных о пожарах, включая результаты дознания, определены первопричины:

для первого случая – неконтролируемое сжигание сухой травы поздней весной, отсутствие противопожарных разрывов в застройке, дефицит средств тушения пожаров; для второго случая – износ электропроводки, позднее прибытие противопожарных формирований, недостаточная обеспеченность водой и системами противопожарной сигнализации. В результате, для выявленных причин сложившейся проблемной ситуации формируются управляющие рекомендации:

- усиления информирования населения и надзорной деятельности в пожароопасный период;
- создание и обновление минерализованных полос для недопущения перехода огня на жилые строения;
- субсидирование приобретения пожарной техники и оборудования;
- строительство и комплексное использование источников водоснабжения;
- поддержка добровольных пожарных;
- развитие систем оповещения и противопожарной сигнализации.

Накопленные данные мониторинга опасных событий и паспорта безопасности территорий позволяют обосновать применение рекомендаций для конкретного объекта, оценить необходимые ресурсы и ожидаемый эффект.

Таким образом, предложенные метамодель и алгоритм позволяют на основе иерархии интегральных оценок проанализировать текущее состояние природно-техногенной безопасности территорий, выявить причины возникновения проблемных ситуаций и, двигаясь от общих обстановок до конкретных факторов риска, сформировать обоснованные управляющие рекомендации.

Заключение

В работе предложены метамодель и алгоритм, позволяющие на основе анализа иерархии интегральных оценок природно-техногенной безопасности территорий определить первопричины проблемной ситуации и сформировать управляющие рекомендации. Представленная метамодель и алгоритм детализации интегральных оценок может иметь широкую сферу применения. Иерархическая организация территориального и отраслевого управления требует для при-

нятия решений достоверной агрегации больших объёмов первичной информации, а ее детализация позволит принимать обоснованные решения. В данный момент выполняется апробация предложенных решений для задач оценивания природно-техногенной безопасности Красноярского края в рамках мероприятий Стратегии цифровизации экономики России и повышения качества управления.

Благодарности. Работа выполнена в рамках проекта государственного задания ФИЦ КНЦ СО РАН программы фундаментальных исследований Российской Федерации (рег. № 0287-2021-003).

Список литературы

1. Олтян И. Ю., Арефьева Е. В., Коровин А. И. Совершенствование оценки состояния защиты населения субъектов Российской Федерации от чрезвычайных ситуаций природного и техногенного характера //Технологии гражданской безопасности. – 2021. – Т. 18. – №. S. – С. 35-41.

2. Penkova T., Nicheporchuk V., Metus A. Comprehensive operational control of the natural and anthropogenic territory safety based on analytical indicators // LNCS, International Joint Conference on Rough Sets. – Springer, Cham, 2017. – С. 263-270.

3. Указ Президента РФ от 16 октября 2019 г. № 501 «О Стратегии в области развития гражданской обороны, защиты населения и территорий от чрезвычайных ситуаций, обеспечения пожарной безопасности и безопасности людей на водных объектах на период до 2030 года».

4. Качанов С.А., Нехорошев С.Н., Попов А.П. Информационные технологии поддержки принятия решений в чрезвычайных ситуациях: Автоматизированная информационно-управляющая система Единой государственной системы предупреждения и ликвидации чрезвычайных ситуаций: вчера, сегодня, завтра: М: Деловой экспресс. – 2011. – 400 с.

5. Karki P., Lea P. Internet of things for architects. – Packt Publishing, 2018. – 454 с.

6. Измалков В. А. Развитие АИУС РСЧС как динамической автоматизированной системы //Технологии гражданской безопасности. – 2017. – Т. 14. – №. 2 (52). – С. 26-31.

7. Crozier M. J., Glade T. Landslide hazard and risk: issues, concepts and approach //Landslide hazard and risk. – 2005. – С. 1-40.

8. Гражданкин А. И., Печёркин А. С., Сидоров В. И. Допустимый риск-мера неприемлемой опасности промышленной аварии //Безопасность труда в промышленности. – 2015. – №. 3. – С. 66-70.

9. Пенькова Т.Г., Метус А.М., Ничепорчук В.В. Метод интегрального аналитического оценивания природно-техногенной безопасности территорий (на примере Красноярского края) // Проблемы анализа риска. – 2018. – Т.15. – №5. – С. 16-25. DOI: 10.32686/1812-5220-2018-15-5-16-25

10. Penkova T.G., Metus A.M., Nicheporchuk V.V. Method of integral analytical estimation of the natural and anthropogenic territory safety (in case of Krasnoyarsk region) // Issues of Risk Analysis. – 2018. – 15. – p. 16-25, DOI: 10.32686/1812-5220-2018-15-5-16-25.

11. Ничепорчук В.В., Пенькова Т.Г., Метус А.М. Формирование стандарта природно-техногенной безопасности территорий Красноярского края // Проблемы безопасности и чрезвычайных ситуаций. – 2018. – №.2 – С. 41-52.

12. Абрамова Н. А., Авдеева З. К. Когнитивный анализ и управление развитием ситуаций: проблемы методологии, теории и практики // Проблемы управления. – 2008. – №. 3. – С. 85-87.

13. Приказ МЧС России № 329 от 8.07.2004 «Критерии информации о чрезвычайных ситуациях»

14. Ничепорчук В.В., Пенькова Т.Г. Комплексный анализ факторов территориальных рисков // Проблемы анализа риска, 2019. – Т.16. №4. – С. 52-62, DOI: 10.32686/1812-5220-2019-16-4-0-0.

СРАВНИТЕЛЬНАЯ ОЦЕНКА ЗДОРОВЫХ ПАЦИЕНТОВ И ПАЦИЕНТОВ С ДМПП ПО ИХ ЭХОКАРДИОГРАФИЧЕСКИМ ПОКАЗАТЕЛЯМ

В.В.Сакович[1], Т.Е.Забродская[3], М.Г.Садовский[1,2,4]

[1]Красноярский государственный медицинский университет МЗ РФ

[2]Институт вычислительного моделирования СО РАН, ***msad@icm.krasn.ru***

[3]КГБУЗ «Красноярский краевой клинический онкологический диспансер им. А.И. Крыжановского», ***ng286329@mail.ru***

[4]Федеральный Сибирский научно-клинический центр ФМБА России

Дефект межпредсердной перегородки (ДМПП) — это врожденный порок сердца (ВПС), характеризующийся наличием сообщения между правым и левым предсердиями. Изолированные ДМПП различных типов встречаются с примерной частотой один случай на полторы тысячи живых новорожденных, составляя около 40% всех ВПС [2, 3]. Гемодинамические изменения, характерные для ДМПП, обусловлены аномальным сбросом крови из левого предсердия (ЛП) в правое, что приводит к появлению постоянно циркулирующего дополнительного объема крови в малом круге кровообращения (МКК), который ведёт к перегрузке правых отделов сердца и развитию легочной гипертензии (ЛГ) [4].

Основным инструментом в постановке диагноза ДМПП является двумерная трансторакальная эхокардиография (ЭХОКГ) с цветным доплеровским картированием, которая позволяет определить топографию и размер дефекта, объем и направление шунтирования крови, увеличение размеров камер сердца, наличие и степень недостаточностей трикуспидального клапана (ТК) и клапана легочной артерии (ЛА). Динамика указанных показателей может выступать в качестве показателей целесообразности и адекватности выбранной тактики ведения пациента [4, 5]. Тем самым, целью настоящей работы является выделение информационно значимых показателей, получаемых применением метода ЭХОКГ, для задач диагностики ДМПП с помощью линейных и нелинейных методов статистики.

Материалы и методы

Исследование проводилось на базе Федерального центра сердечно-сосудистой хирургии г. Красноярска с 2018 по 2022 гг. Всего обследовано 111 детей в возрасте до 18 лет среди них 52 мальчика (46,85 %) и 59 девочек (53,15 %). Группа условно здоровых пациентов состояла из 23 мужчин (62,16 %) и 14 женщин (37,84 %). В группу исследования вошли 74 пациента с установленным диагнозом «изолированный ДМПП», из них 29 мужчин (39,19 %) и 45 женщин (60,81 %). Средний возраст пациентов контрольной группы составляет 9,42 ± 3,48, что статистически значимо выше по сравнению с группой исследования, медиана, 1 и 3 квартили которой составили 5,00 [1,00; 8,00] (p = 0,005). У всех пациентов проводилась регистрация структурных и функциональных эхокардиографических показателей. Структурные показатели включали линейные размеры и объемы полостей сердца, массу миокарда левого желудочка (ММ ЛЖ), толщину межжелудочковой перегородки и задней стенки левого желудочка (ЗСЛЖ).

Функциональные показатели включали в себя фракцию выброса левого желудочка и фракцию изменения площади правого желудочка (ФВ ЛЖ и ФИ S ПЖ), время выброса крови в аорту (t выброса в аорту), амплитуды движения латеральных сегментов митрального и трикуспидального клапанов (MAPSE и TAPSE) и скорости движений базальных сегментов латеральных отделов ЛЖ и ПЖ (LVs и RVs), пиковые скорости и продолжительность фаз пассивного (E) и активного (E′) наполнения желудочков. Регистрация ЭХОКГ параметров проводилась синхронизировано с электрической активности сердца. Оценка продольной деформации миокарда камер сердца (Global Longitudinal Strain – GLS) проведена с помощью новейшей ультразвуковой методики оценки деформации и скручивания миокарда по двухмерному отслеживанию пятен серой шкалы ультразвукового изображения — Speckle-tracking echocardiography в режиме двухмерной эхокардиография с последующей постобработкой в программном пакете QLAB. Полученные результаты были занесены и оформлены в базу данных Microsoft Excel.

Статистическая обработка осуществлялась с помощью программы SPSS. Проводилась проверка на нормальность распределения с помощью критерия Шапиро-Уилка. Для данных, подчиняющихся закону нормального распределения, было найдено среднее значение и стандартное отклонение ($X \pm \sigma$). Для ненормально распределенных данных находилась медиана, 1 и 3 квартили (*Me [Q1; Q3*]). Различия между группами находились с помощью критерия Стьюдента для данных, подчиняющихся закону нормального распределения, а для ненормальных данных использовался критерий Манна-Уитни. Различия считались статистически значимыми при уровне значимости $p < 0{,}05$.

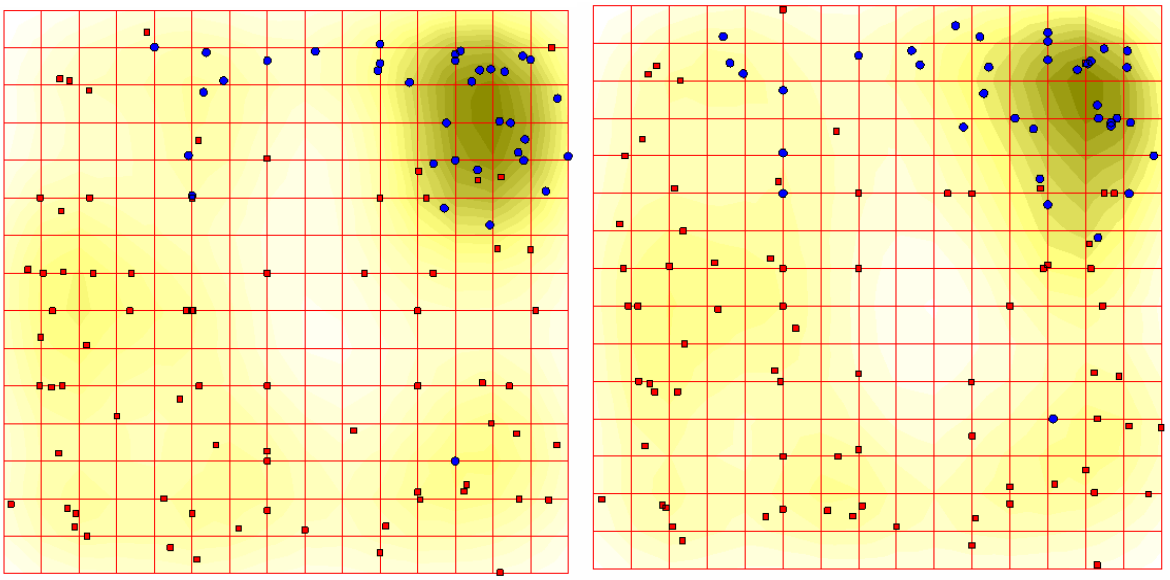
Выделение наиболее значимых диагностических признаков проводилось с помощью анализа площади под ROC-кривыми (AUC). Для сокращения размерности и визуализации данных был использован нелинейный метод упругих карт в программе VidaExpert. Суть данного метода состоит в том, что многомерные данные проецируются на двумерную поверхность и отображаются на ней, как на карте. В результате, хорошо визуализируется распределение объектов по карте и образование отдельных кластеров. Идентификация кластеров осуществлялась по локальной плотности точек [7, 9]. В ходе работе были построены мягкие карты размером 16 на 16 (soft map 16 × 16).

Результаты

Среди структурных и функциональных показателей, полученных методом ЭХОКГ, особый интерес представляют пиковые скорости кровотока. При ДМПП пиковые скорости на трикуспидальном клапане были выше по сравнению с контрольной группой: E ТК (77,00 [58,00; 94,00] против 62,76 ± 11,11 в контрольной группе, $p = 0{,}001$), E′ ТК (60,00 [48,00; 77,00] против 38,48 ± 7,96 в контрольной группе, $p < 0{,}001$). Увеличение показателей в правых камерах при ДМПП закономерно и связано с тем, что скорость кровотока определяется как площадью поперечного сечения, так и объёмом проходящей крови. Также у пациентов с ДМПП отмечается статистически значимое повышение систолического давления в ЛА (27,00 [23,00; 34,00] против 23,76 ± 2,27 в контрольной группе, $p = 0{,}014$),

что является результатом циркуляции «дополнительного» объёма крови в малом круге кровообращения, возникающего при данном ВПС и способствующего перегрузке правых отделов сердца и развитию легочной гипертензии. При анализе результатов тканевой доплерографии для правого и левого желудочков статистически значимых различий выявить не удалось ($p = 0{,}980$ и $p = 0{,}985$ соответственно).

Выявление наиболее значимых показателей проводилось с помощью ROC-анализа. О диагностической значимости исследуемого показателя свидетельствует статистически значимая ($p < 0{,}05$) площадь под ROC-кривой AUC > 0,5. Из 51 показателя диагностически значимыми при бинарной классификации оказались следующие параметры: ЧСС, скорость второго пика на митральном клапане (E′ МК), пиковый градиент и пиковые скорости на ТК, систолическое давление легочной артерии (СДЛА).



а *б*

Рис.1. Упругие карты, построенные по всем переменным (*а*) и с исключением антропометрических данных (*б*)

Классические методы статистики и ROC-анализ продемонстрировали диагностически значимые ЭКГ и ЭХОКГ показатели, но так как задача построения диагностического инструмента требует определения класса заболевания только по данным этих измерений, учитывая, что априорное знание о наличии или от-

сутствии заболевания в реальной клинической практике отсутствует, необходимо обратиться к методам кластеризации пациентов (обучение без учителя). Для выделения кластеров в полученной базе данных был использован нелинейный метод упругих карт. На Рис.1 представлены упругие карты распределения обследуемых с учётом всех показателей (на Рис.1б, исключая антропометрические). На данных картах здоровые пациенты обозначены синим кружком, пациенты с установленным диагнозом «изолированный ДМПП» — красным квадратом.

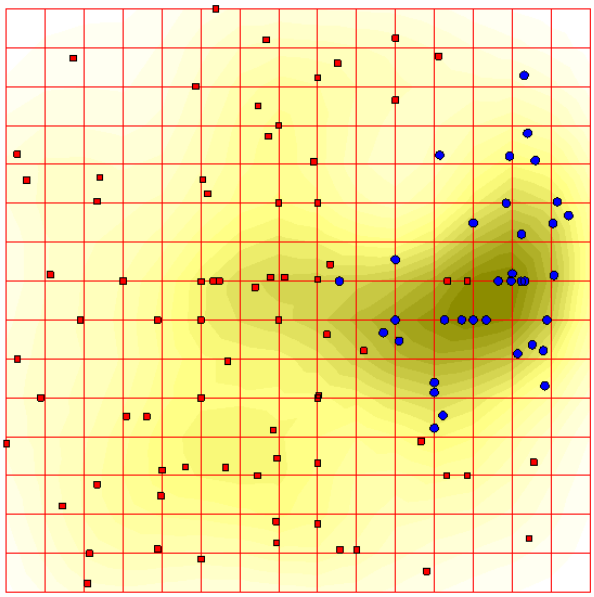


Рис.2. Упругая карта, построенная по диагностически значимым признакам, выделенным с помощью ROC-анализа

Данные рисунки показали, что здоровые пациенты формируют кластер с высокой плотностью, который объединил в себе около 3/4 условно здоровых пациентов. При этом часть здоровых пациентов формируют кластер с небольшой плотностью. Внутренняя структура кластеров больных пациентов также неоднородна — по данным рисункам определяются 3 кластера с низкой плотностью. Несмотря на то, что здоровые и больные пациенты достаточно хорошо отделились друг от друга, часть пациентов с ДМПП попали в центр кластера здоровых пациентов. После удаления антропометрических данных больные пациенты отдалились от центра кластера здоровых, что понижает степень ложноотрицательного срабатывания (Рис.1*б*). Удаление всех переменных за исключением диагно-

стически значимых, выявленных с помощью ROC-анализа, снизило качество кластеризации — несмотря на то, что здоровые пациенты объединились в один большой кластер, процент ложно-отрицательности увеличился с 5,4 % (Рис. 16) до 16,2 % (Рис.2).

В центре кластера оказались 2 пациента с ДМПП, выделенные диагностически значимые показатели которых находятся в пределах значений, характерных для здоровых пациентов. Пациенты с ДМПП, входящие в состав кластера с высокой плотностью, но расположенные относительно далеко от центра кластера, имеют значения СДЛА, характерные для легочной гипертензии (25,11 ± 3,29), остальные показатели находятся в пределах возрастной нормы.

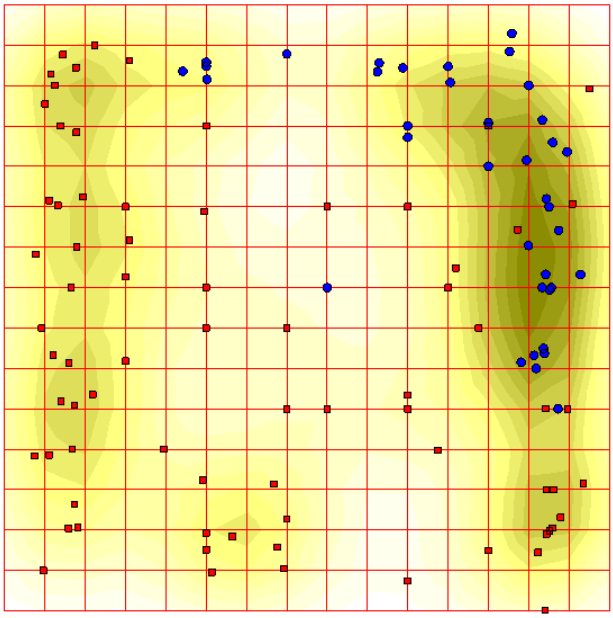


Рис.3.Упругая карта, построенная с учетом коррелирующих переменных

Корреляционная матрица, построенная по всем переменным, выявляет наличие высокой статистически значимой зависимости между переменными, в связи с чем для устранения последствий мультиколлинеарности среди признаков, сильно коррелирующих друг с другом, в исследование включался только один из них. Таким образом, в качестве исследуемых признаков были отобраны

- возраст,
- площадь правого предсердия,
- конечно-диастолический объем левого желудочка,

– продолжительность пиков трикуспидального клапана и
– скорость первого пика ТК.

Упругая карта, построенные с учётом коррелирующих переменных, представлены на Рисунке 3.

Согласно рисунку 3, на упругой карте можно выделить 3 кластера с высокой плотностью (кластер здоровых пациентов и 2 кластера пациентов с ДМПП). Кроме того, на упругой карте заметны ещё 2 кластера с низкой плотностью распределения. Дополнительное деление кластеров говорит о неоднородности внутренней структуры, как здоровых пациентов, так и пациентов с ДМПП. Процент ложноотрицательных результатов составил 6,76 %.

Обсуждение

Таким образом, несмотря на многочисленные рекомендации, довольно часто определение показаний к операции и выбор методики лечения зависит от опыта специалистов: рентгенхирургов, врачей функциональной диагностики и кардиологов, приоритетов клиники, доступности кардиохирургической помощи и массы других субъективных и объективных факторов, лежащих вне общепринятых показаний к оперативному лечению. [1, 7]. Многочисленные публикации посвящены описанию деформации левого предсердия и желудочка и правого желудочка у взрослых до и после РЭО ДМПП [6, 10]. В настоящее время использование методик изучения деформации в детской ЭХОКГ (включая strain и strain rate) находится на ранних этапах изучения. Для использования в клинической практике необходимо определение и принятие нормативных показателей.

В ходе исследования были обнаружены статистически значимые различия в функциональных показателях. У пациентов с ДМПП наблюдались высокие значения пиковых скоростей на ТК и статистически значимо высокое СДЛА. Кластеризация пациентов нелинейным методом упругих карт показало, что диагностически значимые признаки улучшают качество кластеризации, образовав единый кластер здоровых пациентов, однако процент пациентов с ДМПП, попавших в кластер здоровых пациентов тоже увеличивается, что заставляет обра-

тить на них особое внимание, уточнить достоверность поставленного им ранее диагноза. Наименьший процент ложноотрицательных результатов кластеризации продемонстрировали упругие карты, построенные с учетом коррелирующих признаков.

В дальнейшем планируется изучение Strain предсердий, потому что при ДМПП именно предсердия подвергаются патологической деформации. Полученные показатели продольной систолической деформации наряду с выделенными диагностическими признаками могут стать основой для нового протокола медицинского обследования больных с сердечно-сосудистыми заболеваниями.

Список литературы

1. Ассоциация сердечно-сосудистых хирургов России. Клинические рекомендации. Дефект межпредсердной перегородки. – 2018. – 34 с.
2. Бураковский, В. И. Сердечно-сосудистая хирургия: руководство / В. И. Бураковский, Л. А. Бокерия и др.; под ред. акад. АМН СССР В. И. Бураковского, проф. Л. А. Бокерия. – М.: Медицина, 1989. – 752 с.
3. Купряшов, А. А. Дефект межпредсердной перегородки. Частичный аномальный дренаж легочных вен / А. А. Купряшов // Детская кардиохирургия. – 2016. – С. 294 – 312.
4. Шабан И. К., Адамович Е. А. Патофизиологические аспекты дефекта межпредсердной перегородки и принципы его инструментальной диагностики у детей / И. К. Шабан, Е. А. Адамович // Молодежь – практическому здравоохранению. – 2019. – С. 194 – 197.
5. Burbano, N. Understanding Echo Through Embryology / N. Burbano. – DOI 10.1053/j.jvca.2018.06.029 // Journal of Cardiothoracic and Vascular Anesthesia. – 2019. – Vol. 33., №. 4. – P. 1048 – 1049.
6. Bussadori C. et al. Right and left ventricular strain and strain rate in young adults before and after percutaneous atrial septal defect closure //Echocardiography. – 2011. – Vol. 28., №. 7. – P. 730-737.
7. Gorban, A. N. Elastic maps and nets for approximating principal manifolds and their application to microarray data visualization / A. N. Gorban, A. Y. Zino-

vyev . – DOI 10.1007/978-3-540-73750-6_4 // Principal manifolds for data visualization and dimension reduction. – Berlin: Springer, 2008. – P. 96 – 130.

8. Jategaonkar S. R. et al. Two-dimensional strain and strain rate imaging of the right ventricle in adult patients before and after percutaneous closure of atrial septal defects //European Journal of Echocardiography. – 2009. – Vol. 10., №. 4. – P. 499-502.
9. Gorban, A. N., Pitenko A., Zinovyev A. ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data. / A. N. Gorban, A. Pitenko , A. Zinovyev. – DOI 10.48550/arXiv.1406.5550 // Mathematical Software. – 2014. – Vol. 3, № 8. – P. 1 – 9.
10. Ozturk O., Ozturk U., Karahan M. Z. Assesment of right ventricle function with speckle tracking echocardiography after the percutaneous closure of atrial septal defect //Acta Cardiologica Sinica. – 2017. – Vol. 33., №. 5. – P. 523.

ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ КЛАССИФИКАЦИИ БАКТЕРИЙ ПО ГЕНАМ 16S РНК ПО ЧАСТОТНОМУ СОСТАВУ ТРИПЛЕТОВ

А.А.Тетерлева[1], М.Г.Садовский[2,3,4]

[1]Сибирский федеральный университет, ИФБиБТ, ***tenth_smith@mail.ru***

[2]Институт вычислительного моделирования СО РАН, ***msad@icm.krasn.ru***

[3]Федеральный Сибирский научно-клинический центр ФМБА России,

[4]Красноярский государственный медицинский университет МЗ РФ

Микроорганизмы играют жизненно важную роль в живых системах во многих отношениях. В почве или океане микробы участвуют в различных процессах, таких как круговорот углерода и азота, циркуляция питательных веществ и получение энергии. Связь между микробным дисбактериозом и развитием болезней широко изучалась ранее. Окончательная идентификация бактериальных патогенов человека с использованием целевого частичного секвенирования 16S РНК использовалась в клинических микробиологических лабораториях в течении последних 30 лет [1 – 3]. В частности, микробные сообщества в кишечнике человека связаны с патофизиологией ряда хронических заболеваний, таких как болезнь Паркинсона и болезнь Альцгеймера, поэтому анализ связи состава микробиоты с состоянием здоровья человека является ключевым шагом в поиске предиктора ряда заболеваний.

С появлением технологии секвенирования следующего поколения было получено огромное количество метагеномных данных о некультивируемых микробах в дополнение к культивируемым микробам. Новейшие методы секвенирования позволили снизить затраты, что привело к значительному росту числа секвенированных последовательностей ДНК/РНК. Важная задача здесь состоит в том, чтобы сгруппировать данные последовательности, используя стабильные, быстрые и точные методы. Для данных секвенирования микробиома часто используются гены 16 S рибосомной РНК. Кластеризация секвенированных последовательностей является основной задачей биоинформатики, которая привлекает к себе внимание с развитием метагеномики и микробиомики. Так, например, су-

ществует огромное количество доступных (косвенных) доказательств связи состава микробиоты кишечника человека и таких заболеваний, как аллергии, воспалительные заболевания кишечника, метаболические заболевания и даже психические заболевания.

Секвенирование гена 16 S рРНК позволяет оценивать филогенетические отношения микробов, поскольку ген 16 S рРНК кодирует РНК-компонент малой субъединицы (SSU) прокариотических рибосом. 16 S рРНК выполняет несколько функций, включая структурную роль, а также имеет решающее значение для синтеза белка. Наряду с 23 S РНК, он формирует каркас рибосомы, помогающий связывать 50 S и 30 S рибосомные субъединицы, а также определять положения рибосомных белков. 16 S РНК обладает низким темпом эволюции. Сам ген 16 S рРНК имеет размер примерно 1500 пар оснований, не очень значительно варьируясь по длине, а его генетическая структура включает 9 высококонсервативных и 9 гипервариабельных областей (V1–V9). Консервативные области могут служить универсальными сайтами связывания праймеров для ПЦР-амплификации фрагментов генов, тогда как гипервариабельные области содержат значительное разнообразие последовательностей, полезное для целей идентификации прокариот [4]. Все микроорганизмы имеют по крайней мере одну копию гена 16 S, что делает его повсеместным.

Настоящая работа посвящена работы выявлению особенностей связи таксономии бактерий с триплетным составом их генов 16 S РНК. Для этого была сформирована база генетических данных на основе открытой базы генов 16 S РНК SILVA, проведены вычисления по выявлению особенностей классификации и кластеризации генов, преобразованных в частотные словари триплетов и описана таксономическая структура образовавшихся классов и выявляемых кластеров.

Материалы и методы

Для целей нашего исследования использовались гены 16 S РНК бактерий. Генетический материал брался из открытого источника данных SILVA

(https://www.arb-silva.de/). Таксономическая принадлежность является основным компонентом анализа микробного сообщества. Таким образом, выбор базы данных 16 S также важен, поскольку он может повлиять на постанализ и интерпретацию состава сообщества.

Таблица 1

Количественный и качественный состав порядков после индексирования исходной базы генетических данных; N — число видов.

Тип	Класс	Порядок	Семейство	N
Acidobacteriales				
Acidobacteriota	*Acidobacteriae*	*Acidobacteriales*	*Acidobacteriaceae*	31
Acidobacteriota	*Acidobacteriae*	*Acidobacteriales*	*Koribacteraceae*	1
Acidobacteriota	*Acidobacteriae*	*Solibacterales*	*Solibacteraceae*	2
Acidimicrobiales				
Actinobacteriota	*Acidimicrobiia*	*Acidimicrobiales*	*Acidimicrobiaceae*	24
Chlamydiales				
Verrucomicrobiota	*Chlamydiae*	*Chlamydiales*	*Chlamydiaceae*	49
Verrucomicrobiota	*Chlamydiae*	*Chlamydiales*	*Parachlamydiaceae*	39
Firmicutes	*Bacilli*	*Mycoplasmatales*	*Mycoplasmataceae*	13
Bacteroidia				
Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Bacteroidaceae*	163
Verrucomicrobiota	*Chlamydiae*	*Chlamydiales*	*Chlamydiaceae*	101
Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Prevotellaceae*	147

Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Tannerellaceae*	94
Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Porphyromonadaceae*	106
Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Rikenellaceae*	69
Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Dysgonomonadaceae*	58
Bacteroidota	*Bacteroidia*	*Bacteroidales*	*Marinifilaceae*	58
Bacillales				
Firmicutes	*Bacilli*	*Staphylococcales*	*Staphylococcaceae*	150
Firmicutes	*Bacilli*	*Bacillales*	*Bacillaceae*	151
Firmicutes	*Bacilli*	*Lactobacillales*	*Listeriaceae*	148
Firmicutes	*Bacilli*	*Paenibacillales*	*Paenibacillaceae*	124
Firmicutes	*Bacilli*	*Bacillales*	*Planococcaceae*	101
Firmicutes	*Bacilli*	*Brevibacillales*	*Brevibacillaceae*	147
Firmicutes	*Bacilli*	*Exiguobacterales*	*Exiguobacteraceae*	191
Firmicutes	*Bacilli*	*Alicyclobacillales*	*Alicyclobacillaceae*	176

В базе данных SILVA содержится более 190000 последовательностей, разбитых по таксономическому признаку на следующие разделы: домен, отдел, класс, порядок, семейство, род, вид и штамм (не обязательно). В этой базе данных содержатся записи некоторых микроорганизмов, которые встречаются неоднократно. Это связано с тем, что к настоящему времени классификация бактерий не является устоявшейся, и разные исследователи относят один и тот же организм к разным таксономическим группам. Другая причина этого эффекта — собственный недостаток базы SILVA: дело в том, что таксономическое положение большинства содержащихся в ней генов определяется средствами биоинформатики, и это определение имеет высокий (до 25 %) уровень ошибок [5]. Это приводит к тому, что данный организм (или даже таксономическая единица) учитывается в данной базе более одного раза.

Мы изучали последовательности следующих порядков бактерий: *Acidobacteriales* (34 гена), *Acidimicrobiales* (24 гена), *Bacteroidia* (3017 генов), *Chlamydiales* (820 генов) и *Bacillales* (48579 генов). Такой большой разброс может приводить к существенному искажению результатов кластеризации: сверхпредставленные таксоны формируют очень сильный «сигнал», искажающий картину кластеризации генов тех порядков, для которых в базе содержится малое число записей (видов).

Для того чтобы избежать такого искажения, мы индексировали нашу базу данных, исключая все «малочисленные» таксоны. Аналогично, в сверхпредставленных таксонах (в нашем случае это порядки *Bacteroidia, Chlamydiales* и *Bacillales*) были случайным образом исключены некоторые записи, так чтобы число оставшихся записей не очень превышало число записей в других порядках. В результате общее число последовательностей в индексированной базе (по которой, собственно, проводилось данное исследование) составило 2143 записи. Окончательный состав изученных представлен в Таблице 1. Последовательности рРНК, соответствующие бактериям в этих порядках, скачивались из базы данных. Затем с помощью *ad hoc* программы преобразовывались в частотные словари триплетов; данная конструкция хорошо известна, и мы не будем здесь останавливаться на её детальном описании [6 – 9]. Внутренняя структурированность набора генов 16 S РНК бактерий изучалась с помощью упругих карт [11, 12].

Результаты

При анализе исключался один триплет; это связано с тем, что все 64 триплета являются линейно зависимыми. Формально можно исключить любой триплет, однако мы исключали тот, для которого значение стандартного отклонения, определяемого по всей базе генов, было минимальным. Такой выбор объясняется тем, что в пределе (стандартное отклонение равно нулю и, следовательно, частота этого триплета у всех генов одна и та же) такой триплет не даёт никакого вклада в различимость объектов. Соответственно, триплет с минимальным значением стандартного отклонения даёт наименьший вклад в различимость генов;

в нашем случае таким триплетом был CAC. Для построения упругой карты использовалось свободно распространяемое ПО *VidaExpert* [12, 13]. Мы использовали упругую карту размером 16 × 16 («мягкую» в терминологии разработчиков ПО), на которой и изучалась кластерная структура данных.

Следующий шаг заключался в выделении кластеров на полученной упругой карте. Кластеры выделялись по локальной плотности. Радиус корреляции при построении локальной плотности был равен 0,15, число градаций уровня локальной плотности — 15; значения всех остальных параметров были выбраны по умолчанию. Участки, имеющие наибольшую плотность (визуально более интенсивную темную окраску на используемой карте по сравнению с другими), идентифицировались нами как кластеры. На исследуемой карте мы выделили 9 кластеров: рисунок 1 содержит прорисовку этих кластеров.

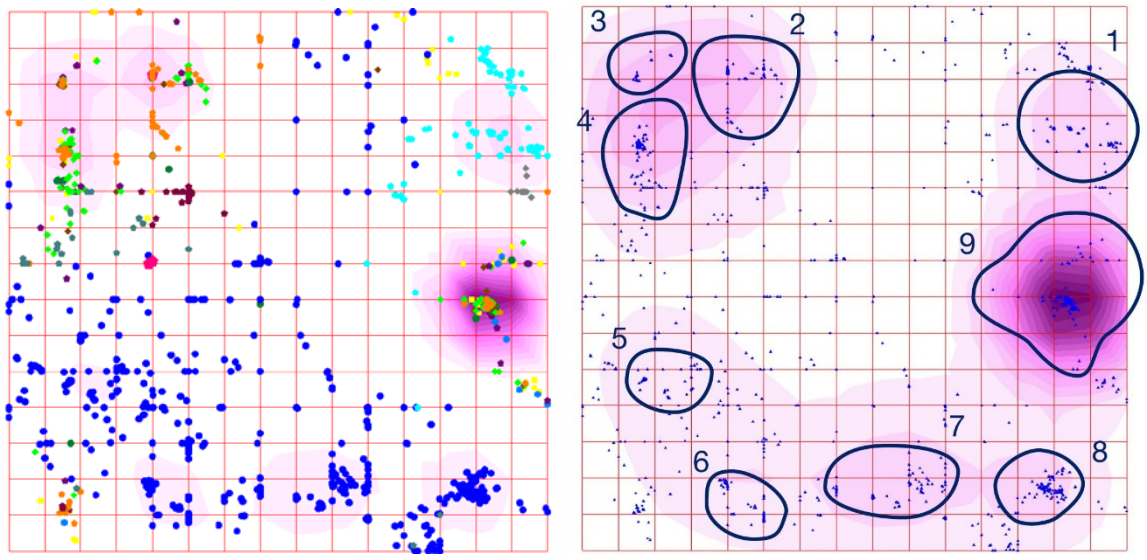


Рис.1. Распределение генов разных таксонов по кластерам.

На следующем этапе мы исследовали таксономический состав данных кластеров. Так, например, всего было 1200 последовательностей генов 16 S РНК типа *Fermicutes* в индексированной базе, из них по кластерам разошлись 1093 последовательности. Разбиение на кластеры носит неслучайный характер, т.е. в один кластер попадают последовательности одного таксона или близких к нему; так, например, патогенные бактерии (*Chlamydiales*) образуют на карте характерный кластер (кластер 1, см. Рис. 1). На карте он выглядит очень обособленно от

других кластеров, являясь практически моносоставным (бактерии *Chlamydiales* составляют в нём 98,62 %, а бактерии *Bacilli* — 1,38 %). Общая картина таксономического состава каждого кластера отражена в таблице 2.

Исходя из вышеприведенных данных, мы предполагаем, что при дальнейшем развитии методики с помощью метода упругих карт можно будет проводить диагностику заболеваний различного спектра по составу микробиоты кишечника, фиксируя на индивидуальной карте человека кластеры тех или иных бактерий кишечной флоры и определяя по кластерной принадлежности патогенность.

Таблица 2

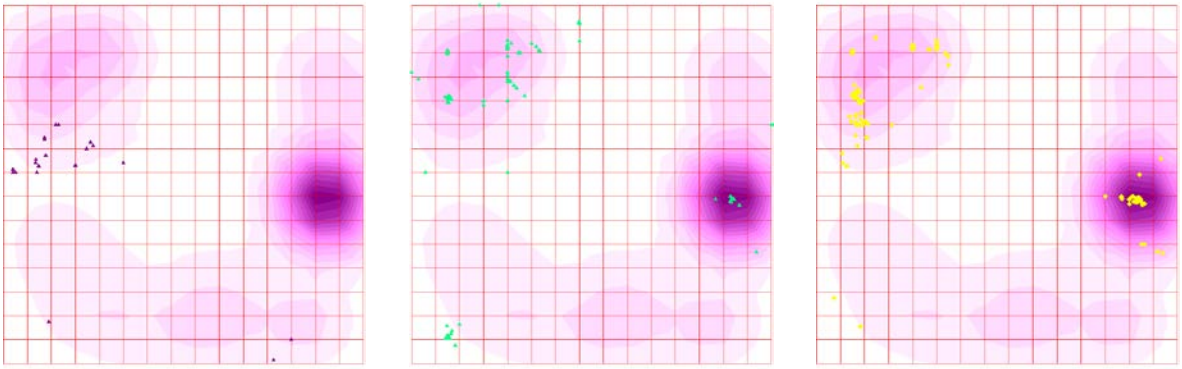
Таксономический состав кластеров (см. Рис. 1); здесь *N* —число последовательностей из индексированной базы, устойчиво распределившихся по кластерам, *M* — число последовательностей таксона, попавших в данный кластер

Таксон	*N*	Таксон	*M*	Таксон	*M*	Таксон	*M*	
Кластер 1								
Firmicutes	1093	*Bacilli*	1093	*Mycoplasmatales*	132	*Mycoplasmataceae*	2	2
Verrucomicrobiota	143	*Chlamydiae*	143	*Chlamydiales*	143	*Chlamydiaceae*	129	129
						Parachlamydiaceae	14	13
Кластер 2								
Firmicutes	1093	*Bacilli*	1093	*Alicyclobacillales*	145	*Alicyclobacillaceae*	145	92
				Bacillales	242	*Bacillaceae*	144	39
						Planococcaceae	99	15
				Brevibac	143	*Brevibacill*	143	1

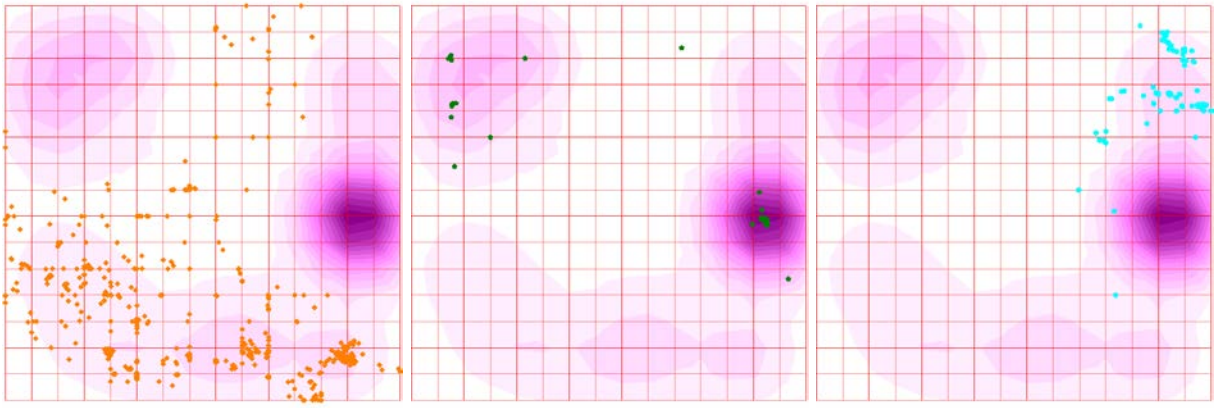
				illales		*aceae*		
				Exiguobacterales	187	*Exiguobacteraceae*	187	2
				Lactobacillales	141	*Listeriaceae*	141	3
				Staphylococcales	112	*Staphylococcaceae*	112	11
-		-		-		*Bacillaceae*	144	1
Кластер 3								
Firmicutes	1093	*Bacilli*	1093	*Alicyclobacillales*	145	*Alicyclobacillaceae*	145	12
				Bacillales	242	*Bacillaceae*	144	18
						Planococcaceae	99	8
				Brevibacillales	143	*Brevibacillaceae*	143	10
				Exiguobacterales	187	*Exiguobacteraceae*	187	30
				Lactobacillales	141	*Listeriaceae*	141	9
				Staphylococcales	112	*Staphylococcaceae*	112	18
Кластер 4								
Actinobacteriota	3	*Acidimicrobiia*	3	*Acidimicrobiales*	3	*Acidimicrobiaceae*	3	2
Firmicutes	1093	*Bacilli*	1093	*Alicyclobacillale*	145	*Alicyclobacillaceae*	145	27

				s				
				Bacillales	242	*Bacillaceae*	144	42
						Planococcaceae	99	27
				Brevibacillales	143	*Brevibacillaceae*	143	7
				Exiguobacterales	187	*Exiguobacteraceae*	187	3
				Lactobacillales	141	*Listeriaceae*	141	12
				Staphylococcales	112	*Staphylococcaceae*	112	10
Кластер 5								
Bacteroidota	343	*Bacteroidia*	343	*Bacteroidales*	343	*Prevotellaceae*	61	11
						Tannerellaceae	50	13
Кластер 6								
Bacteroidota	343	*Bacteroidia*	343	*Bacteroidales*	343	*Bacteroidaceae*	131	28
						Prevotellaceae	61	49
Кластер 7								
Bacteroidota	343	*Bacteroidia*	343	*Bacteroidales*	343	*Dysgonomonadaceae*	4	4
						Porphyromonadaceae	94	94
						Rikenellaceae	3	3

						Tannerellaceae	50	37
Кластер 8								
Actinobacteriota	3	*Acidimicrobiia*	3	*Acidimicrobiales*	3	*Acidimicrobiaceae*	3	1
Bacteroidota	343	*Bacteroidia*	343	*Bacteroidales*	343	*Bacteroidaceae*	131	103
						Prevotellaceae	61	1
Кластер 9								
Firmicutes	1093	*Bacilli*	1093	*Alicyclobacillales*	145	*Alicyclobacillaceae*	145	14
				Bacillales	242	*Bacillaceae*	144	44
						Planococcaceae	99	49
				Brevibacillales	143	*Brevibacillaceae*	143	125
				Exiguobacterales	187	*Exiguobacteraceae*	187	152
				Lactobacillales	141	*Listeriaceae*	141	117
				Paenibacillales	120	*Paenibacillaceae*	120	120
				Staphylococcales	112	*Staphylococcaceae*	112	73
Verrucomicrobiota	143	*Chlamydiae*	143	*Chlamydiales*	143	*Parachlamydiaceae*	14	1



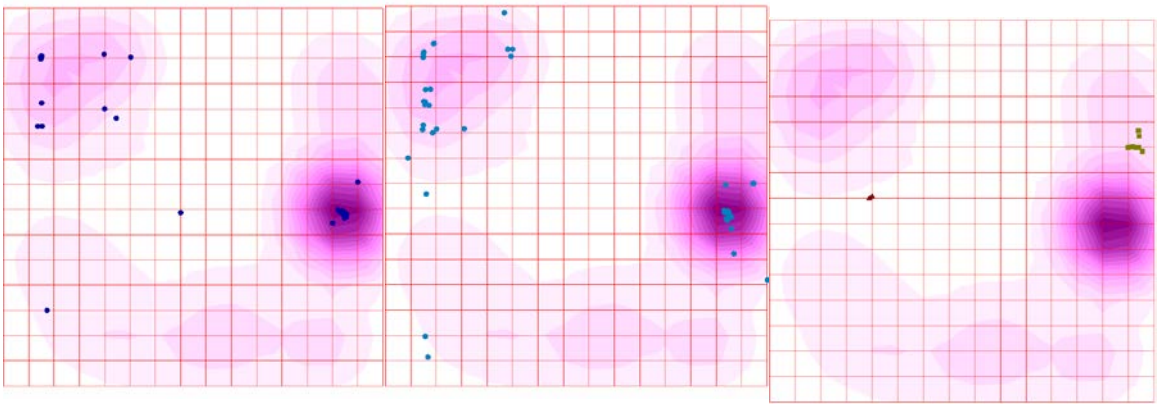
(а) *Acidimicrobiales* (б) *Alicyclobacillales* (в) *Bacillales*



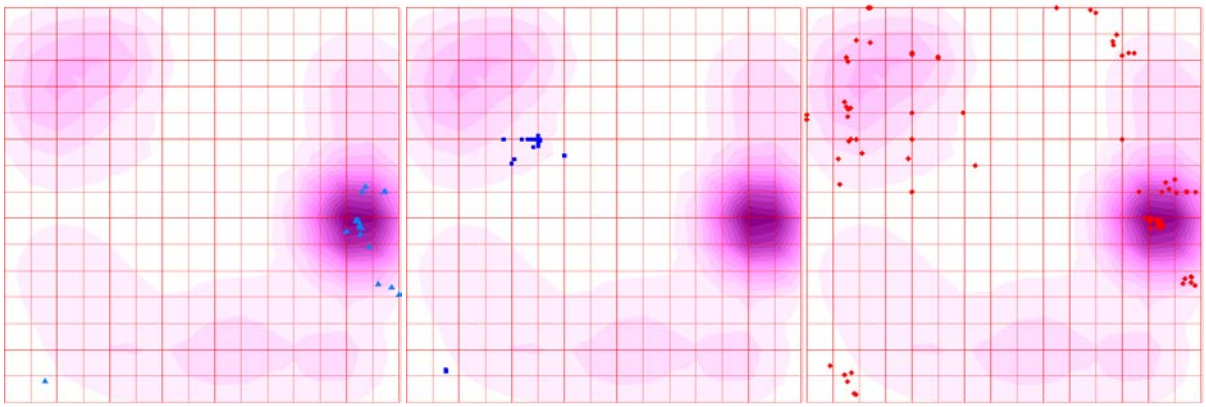
(г) *Bacteroidales* (д) *Brevibacillales* (е) *Chlamydiales*

Рис. 2. Распределения отдельных порядков по упругой карте

Основной задачей данной работы является проверка гипотезы о наличие связи структуры нуклеотидной последовательности с таксономией ее носителя и существование иерархической классификации, отражающей связь между классами различных уровней и таксономией более низких таксономических единиц. Для решения задачи мы выявляли классы методами классификации без учителя. Ожидаемый результат при исследовании методом упругих карт состоял в том, что будут выделяться кластеры и эти кластеры будут включать в основном таксономически близкие организмы. Прямым подтверждением справедливости этого предположения является картина распределения отдельных порядков по (мягкой) упругой карте 16 × 16, показанная на Рисунках 2 и 3. На этих рисунках хорошо видно, что гены 16S РНК бактерий разных отрядов располагаются на упругой карте совершенно неслучайно. При этом неслучайность распределения следует понимать как сильное предпочтение в распределении генов по кластерам.



(а) *Exiguobacterales* (б) *Lactobacillales* (в) *Mycopl+Solibact.**



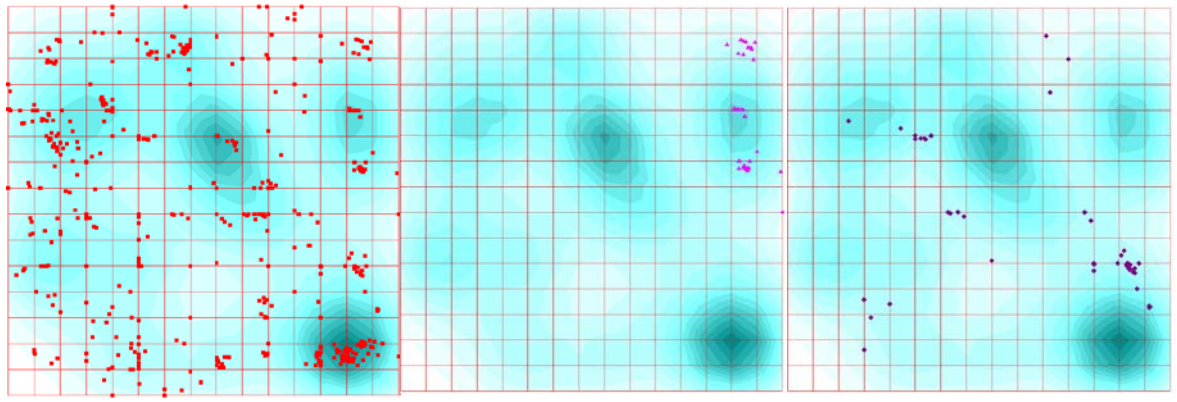
(г) *Paenibacillales* (д) *Acidobacteriales* (е) *Staphylococcales*

Рис. 3. Распределения отдельных порядков по упругой карте

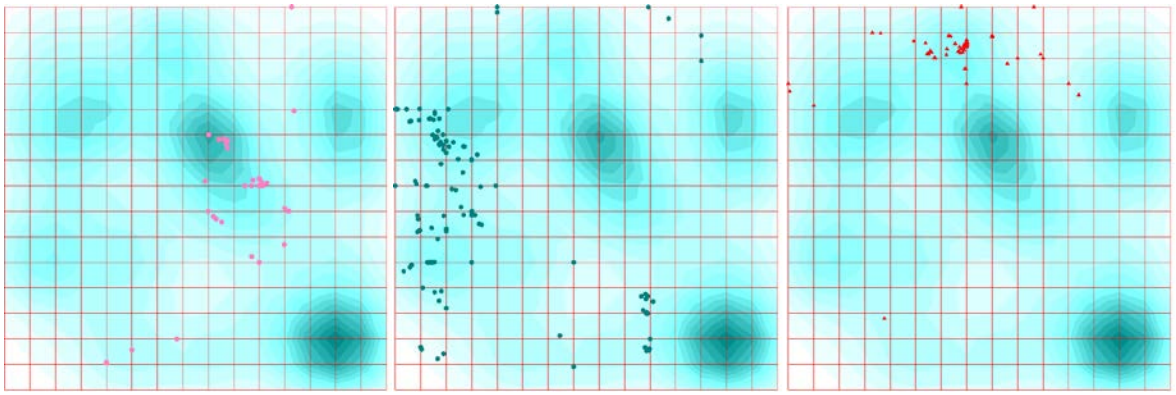
Иными словами, на рисунке 2 видны два типа неслучайности в распределении генов: сам характер распределения генов в пространстве частот триплетов — видно, что гены образуют явно видимые кластеры (группы точек повышенной плотности), и эти кластеры чётко отделены один от другого, и второй тип неслучайности — характер распределения точек (генов), принадлежащих одной и той же таксономической единице, собственно по кластерам. В задачи настоящей работы не входила формальная проверка различимости кластеров, однако она видна, что называется, невооружённым взглядом.

Обратим внимание на один важный эффект, проявляющийся в рисунках 2 и 3. На них хорошо видно, что гены бактерий некоторых порядков — а именно, *Alicyclobacillales, Bacillales, Brevibacillales* и *Staphylococcales* — распределены по двум разным кластерам. Причины такого поведения генов могут быть разны-

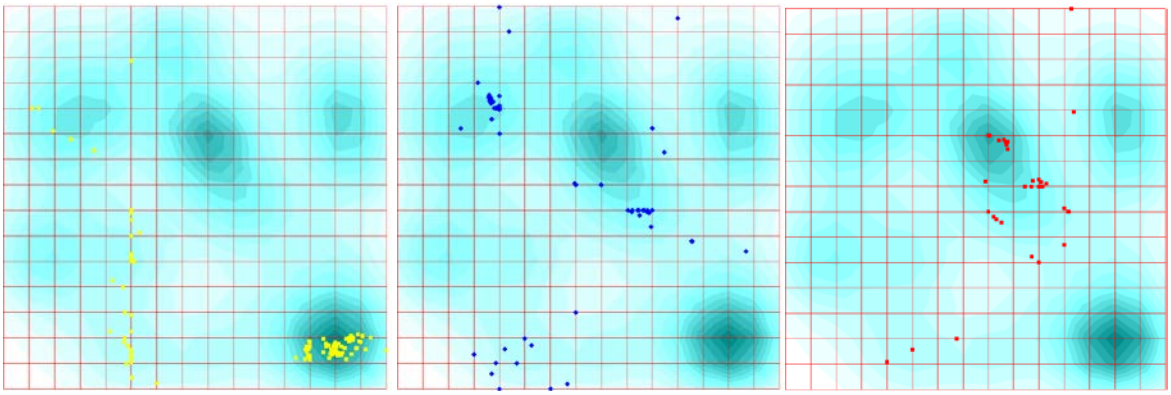
ми.



(а) *Bacterioidia all* (б) *Chlamydiaceae* (в) *Dysgonomonadaceae*



(г) *Marinifilaceae* (д) *Prevotellaceae* (е) *Rikenellaceae*



(ж) *Bacteroidaceae* (з) *Tannerellaceae* (и) *Porphyromonadaceae*

Рисунок 3 – Распределения отдельных семейств на упругой карте в пределах порядка *Bacteoidales*

Первой причиной естественно считать биологические свойства микроорганизмов этих порядков; проблема, однако, в том, что сформулировать гипотезу о биологических причинах такого различия весьма затруднительно. Так, напри-

мер, совершенно непонятно, почему такие биологические свойства наблюдаются только у избранных порядков. Гораздо более естественным является то, что наблюдаемое «расщепление» генов по двум кластерам является артефактом. А именно, наблюдаемый эффект обусловлен типом упругой карты, выбранной для анализа.

Метод упругих карт приближает многомерные данные многообразиями малой размерности; в нашем случае размерность равна двум. Но фиксация размерности ещё не определяет однозначно само многообразие, которое будет использоваться в процедуре приближения; разнообразие возможных типов двумерных многообразий очень велико. Использовавшееся нами программное обеспечение позволяет работать с двумя типами двумерных многообразий: квадрат и сфера. Собственно, эффект, о котором идёт речь, мог быть порождён тем, что в качестве начального многообразия использовался именно квадрат. Вполне возможна ситуация, при которой использование в качестве начального многообразия сферы приведёт к слиянию двух кластеров, находящихся в разных частях квадрата (во внутренних координатах) в один. Более детальное изучение этого феномена выходит за рамки настоящей работы.

Обсуждение

Основной задачей данной работы является проверка двух гипотез: наличие связи структуры нуклеотидной последовательности с таксономией ее носителя и существование иерархической классификации, отражающей связь между классами различных уровней и таксономией более низких таксономических единиц. Для решения обеих задач мы выявляли классы методами классификации без учителя. Ожидаемый результат при исследовании методом упругих карт состоял в том, что будут выделяться кластеры и эти кластеры будут включать в основном таксономически близкие организмы. Во второй задаче ожидаемый результат состоял в том, что анализ распределения генов 16S РНК бактерий по частотам триплетов покажет кластеризацию на каждом таксономическом уровне. Поясним сказанное. Как было показано выше, изучение совместного распределения ука-

занных генов для бактерий нескольких порядков приводит к тому, что на упругой карте выявляются кластеры, причём каждый кластер содержит не случайный набор организмов, а тех, что принадлежать одному порядку. Установлено, что такое поведение характерно для любого таксономического уровня: если проводить кластеризацию только в пределах одного таксона, то образующиеся кластеры будут содержать гены бактерий, принадлежащих таксономическим группам следующего «вниз» уровня.

Выше представлены результаты анализа кластеризации в виде таблиц с таксономическим составом выделенных кластеров. Мы проверяли состав кластеров сначала «прямым» способом: визуализировали на мягкой карте (16 × 16) с заданной локальной плотностью отдельные таксоны, чтобы определить их распределение. Выяснилось, что в основном таксоны (низкого уровня типа порядка или семейства) образуют чётко отграничиваемые скопления в плоскости карты, но есть и таксоны-исключения. Первые в таком случае назовем плотными кластерами, а вторые — рыхлыми. Примеры этих кластеров показаны на рисунках 3(е), 4(в), 4(г) (плотные) и на рисунках 3(г), 4(д) (рыхлые). Однако есть и промежуточный вариант распределения, в котором последовательности не сосредоточены в одном кластере, но и не разбросаны по всей карте. То есть порядок в распределении точек, соответствующих тем или иным таксонам, всё же прослеживается в характере структуры двух или более кластеров. Примерами таких таксонов являются *Bacillales* (рисунок 3(в)), *Brevibacillales* (рисунок 3(д)) и *Exiguobacterales* (рисунок 4(а)).

Наряду с этим использовался и «обратный» подход: на карте выделялся отдельный кластер и изучалось его таксономическое наполнение; результаты такого анализа показывают, что выделяются плотные кластеры, состоящие в основном из одного или двух семейств, а также рыхлые, состоящие из бóльшего количества семейств. Тем самым, в рамках нашей работы можно ограничиться любым из этих подходов по желанию.

Особого внимания заслуживают таксоны-исключения, упомянутые выше. Их поведение на карте может быть обосновано рядом причин:

- ошибки в заполнении базы;
- ошибки исследователей в идентификации вида;
- отражение действительных биологических особенностей.

Первая причина подразумевает, что исследователи, наполнявшие базу данных, допустили ряд опечаток. Вторая причина может свидетельствовать об ошибках в идентификации и классификации некоторых микроорганизмов. Заметим, что в этом случае наши результаты являются основанием для независимой повторной перепроверки качества определения этого микроорганизма. Третья причина выглядит биологически наиболее интересной. Возникает вопрос, чем отличаются бактерии, не подчиняющиеся упорядоченному распределению, от тех, которые образуют на карте чёткие кластеры и почему доля таких сильно отличающихся видов крайне невелика. Ответы на эти вопросы требуют дальнейших исследований.

Список литературы

1. Tang, Yi-Wei, Nicole M. Ellis, Marlene K. Hopkins, Douglas H. Smith, Deborah E. Dodge, and David H. Persing. Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. // Journal of clinical microbiology 36, no. 12 (1998): 3674-3679.
2. Clarridge III, Jill E., Silvia M. Attorri, Qing Zhang, and John Bartell. 16S ribosomal DNA sequence analysis distinguishes biotypes of *Streptococcus bovis*: *Streptococcus bovis* biotype II/2 is a separate genospecies and the predominant clinical isolate in adult males // Journal of clinical microbiology 39, no. 4 (2001): 1549-1552.
3. Clarridge III, Jill E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases // Clinical microbiology reviews 17, no. 4 (2004): 840-862.
4. Van de Peer, Yves, Sabine Chapelle, and Rupert De Wachter. A quantitative map of nucleotide substitution rates in bacterial rRNA // Nucleic acids research 24, no. 17 (1996): 3381-3391.
5. Edgar R. 2018. Taxonomy annotation and guide tree errors in 16S rRNA databases // PeerJ 6:e5030 https://doi.org/10.7717/peerj.5030

6. Gorban A.N., Popova T.G., Sadovsky M.G., and Wunsch D.C. Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts // Intelligent Engineering Systems Through Artificial Neural Networks. — American Society of Mechanical Engineers (ASME), 2001. — P. 657–663.
7. Sadovsky M., Putintseva Yu., Chernyshova A., Fedotova V. Genome structure of organelles strongly relates to taxonomy of bearers // International Conference on Bioinformatics and Biomedical Engineering / Springer. — 2015. — P. 481–490.
8. Gorban A., Popova T., Sadovsky M. Classification of symbol sequences over their frequency dictionaries: towards the connection between structure and natural taxonomy // Open Systems & Information Dynamics. — 2000. — Vol. 7, no. 1. — P. 1–17.
9. Sadovsky M., Senashova M., Putintseva Yu. Eight Clusters, Synchrony of Evolution and Unique Symmetry in Chloroplast Genomes: The Offering from Triplets // Chloroplasts and Cytoplasm: Structure and Functions. — Nova Science Publishers, Inc., 2018. — P. 25–95. — ISBN: 978-1-53614-127-????
10. Fukunaga K. Introduction to statistical pattern recognition. — London: Academic Press, 1990.
11. Gorban A., Zinovyev A. Principal Manifolds for Data Visualisation and Dimension Reduction // Lecture Notes in Computational Science and Engineering — Berlin – Heidelberg – New York: Springer, 2007. — Vol. 58. — P. 153–176.
12. Gorban A., Zinovyev A. Fast and user-friendly non-linear principal manifold learning by method of elastic maps // 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015. — 2015. — P. 1–9. — Access mode: https://doi.org/10.1109/DSAA.2015.7344818.
13. Gorban A., Zinovyev A. Elastic principal graphs and manifolds and their practical applications // Computing 75, no. 4 (2005): 359-379.

СВЁРТКА ДАННЫХ ОБ АКТИВНОСТИ СЛОЖНЫХ СИСТЕМ С ПОМОЩЬЮ ПИКТОГРАФИКИ В НОТАЦИИ UGVA

В.А. Углев

Сибирский федеральный университет, ***uglev-v@yandex.ru***

Сложные системы, анализ функционирования которых необходимо наблюдать (осуществлять мониторинг) и принимать своевременные управляющие решения, описываются определёнными классами моделей. Для математического описания их активности применяются цепи Маркова, конечные автоматы, и даже наборы эвристических правил (деревья принятия решений). Все их можно свести к графовому описанию, оперирующему однородной структурой. Даже описание свёрточных нейронных сетей в определенном смысле можно задать в виде графа. Эта однородность подкупает своей простотой, но не согласуется с формальной постановкой задачи принятия решений, осуществляемой человеком. А имея упрощенную структуру её сложно объяснять. Поэтому при создании интеллектуальных систем актуальной является концепция объяснимого искусственного интеллекта (XAI [1]). А так как объяснение человеку сложных решений или поддержка их принятия лучше осуществлять в графическом виде, то рассмотрим подход к визуализации неоднородных структур данных об активности систем, опираясь на пиктографику.

Осуществление акта принятия решений, согласно модели афферентного синтеза П.К. Анохина [2], должны иметься следующие исходные данные: целевые установки (доминирующая мотивация), параметры текущей ситуации (обстановочная афферентация) и имеющийся опыт (память). Упрощенная схема этого процесса показана на рис. 1. По сути, в едином пространстве объединяются данные о настоящем, прошедшем и будущем (ожидаемом целевом состоянии). Очевидно, что все эти три компоненты могут и должны различаться во входном потоке данных при реализации механизма принятия решений.

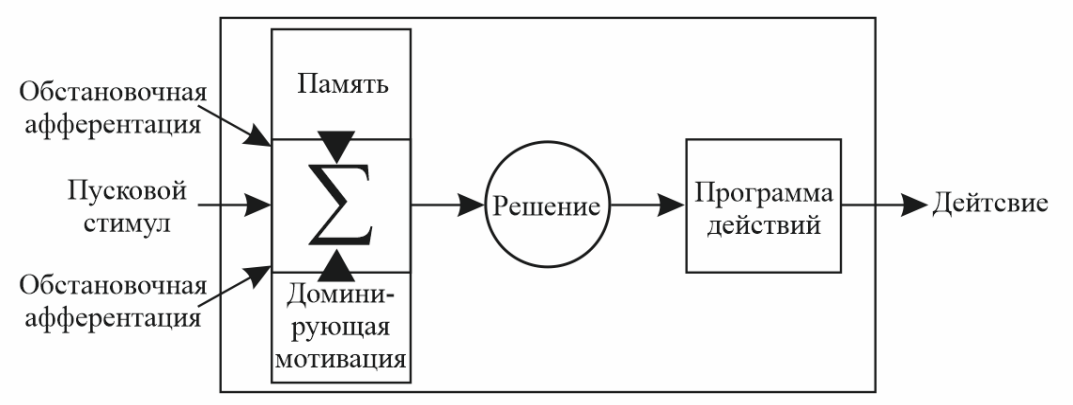


Рис. 1. Схема афферентного синтеза (процесса принятия решений) по [2]

Активность систем, как правило, проявляется в виде динамики её деятельности (филогенез) и в истории её становления (онтогенез). Так как любая эволюционирующая система вынуждена приспосабливаться (адаптироваться) к меняющимся условиям окружающей среды [3], то для понимания динамики системы нужно разделять вклад текущей активности, активности в прошлом и целевых/нормативных показателей. В результате имеется три графовых структуры, которые должны влиять на результирующее решение. Рассмотрим подход к их визуализации.

Единовременное визуальное представление нескольких графов человеку для сопровождения принятия единичных решений или их объяснения создаёт сложности в их восприятии. Как и с таблицами, такая форма представления приводит к «перегрузке» внимания лица, принимающего решение: с виду структуры сложны и однородны, но семантическую нагрузку они несут различную. Попытки их визуализации с помощью кластеров, карт в пространстве малой размерности, глифов, графиков в параллельных координатах и пр. методах визуализации [4] значительно упрощают ситуацию, но не решают проблему. Происходит либо предельное обобщение (все данные об экземпляре целевой системы становятся точкой в гиперпространстве факторов); либо имеющееся разнообразие факторов группируется на графическом образе (число адекватно воспринимаемых параметров многократно превышает возможности восприятия [5]). В обоих случаях объяснительная способность значительно снижается.

Альтернативным подходом является такой вид пиктографики, как лица Чернова [6]. Но и он имеет ограничения: необходимость в зеркальной симметрии данных, наличие малых искажений образа, малый объём кодирующих признаков [7]. Развитием этого подхода сначала стал метод бодикодера (антропоморфной фигуры по [8]), а затем метода унифицированного графического воплощения активности (UGVA, подробнее см. [9]). Раскроем некоторые особенности формирования таких образов и покажем и преимущества для визуализации активности сложных систем с разнородными наборами исходных данных при сопровождении процесса принятия решений.

Модель визуализации в нотации UGVA базируется на том, что для конкретной задачи принятия решений формируются последовательно архитектурный образ, базовый образ, частный образ, индивидуализированный образ экземпляра системы. Описание ключевых характеристик антропоморфного образа включает определение числа и типов осей данных, типов симметрии образа, наличие интегральной (оценочной) зоны и специфику отображения времени. Приведем, для примера, описание типов осей и видов симметрии. Для пояснения возьмем в качестве иллюстрирующего объекта космический аппарат (искусственный спутник Земли для обеспечения функций навигации и связи), чьи параметры необходимо визуализировать на антропоморфном образе.

Пусть имеется множество ключевых параметров объекта $\{S_j\}$ с порядковыми номерами $n, m, \ldots, k \in j$ которые зафиксированы в момент времени p_k. Тогда параметры S_j будут ассоциированы с отдельной осью или её частью, входящих в состав антропоморфного образа (в первую очередь это конечность). Выделим пять типов организации осей (см. рис. 2):

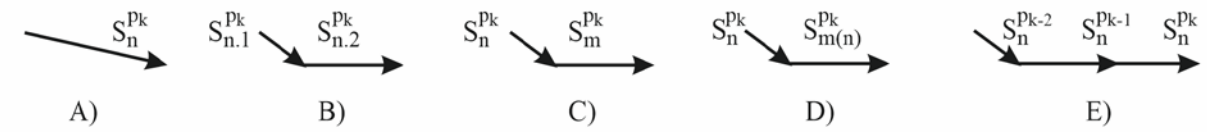


Рис. 2. Графические шаблоны различных типов осей в составе образов UGVA

- простая ось (тип *A*), когда количественное значение ключевого признака

S_j накладывается на один графический примитив; например, в виде параллелограмма-конечности, где массу аппарата ставим в соответствие длине фигуры, а объём – её высоте;

- составная структурная (тип *B*), когда структурные составляющие количественного показателя одного обобщенного признака S_j накладываются на последовательность графических примитивов так, чтобы продемонстрировать специфику их совокупности; например, «надежность» аппарата может задавать длину первого сегмента конечности для коэффициента готовности, дину второго для вероятности безотказной работы;

- составная разнородная ось (тип *C*), когда количественные показатели разных S_j накладываются на последовательность графических примитивов в рамках одной оси, но относящиеся к различным аспектам оценки объекта анализа; например, число принимающих антенн модуля полезной нагрузки определяет длину одного сегмента конечности, а число «ретрансляторов» – длину второго;

- составная причинно-следственная ось (тип *D*), когда выбор количественных показателей разных S_j, накладывающийся на последовательность графических примитивов в рамках одной оси, выбирается таким образом, что они демонстрируют причинно-следственную зависимость (предпосылка-следствие) между компонуемыми признаками; например, длина первого сегмента демонстрирует вероятность безотказной работы бортовой командно-измерительной системы аппарата, а форма и длина второго – тип резервирования и число резервных комплектов.

- составная временная ось (тип *E*), когда каждый сегмент оси последовательно соответствует состояниям признака S_j в заданные моменты времени *p*; например, значение коэффициента полезного действия солнечной батареи определяет длину четырех сегментов оси на момент вывода аппарата на орбиту, через 5, 10 и 15 лет активного существования соответственно.

Теперь поясним другой параметр образа – тип симметрии. Выделим следующие варианты (см. рис. 3), опираясь на тот же пример со спутником:

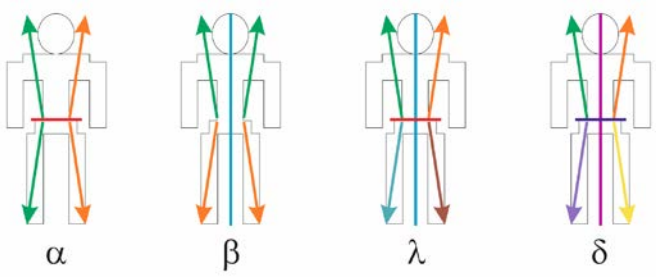


Рис. 3. Графические шаблоны различных типов симметрирования для образов UGVA

- билатеральная осевая симметрия по вертикали образа (тип α), позволяет противопоставить верхний набор осей нижнему, подчеркнув динамику изменения (как правило во времени); например, соотношение показателей исходных характеристик к аппарату на этапе утверждения исходного технического задания (например, с электро-радио изделиями класса space&military импортного производства) и их фактических значений на этапе приёмки (после импортозамещения);

- билатеральная осевая симметрия по горизонтали образа (тип β), позволяет противопоставить правый набор осей левому, подчеркнув соотношение между наборами различных признаков, имеющих схожий вклад в параметрическую модель объекта; например, сопоставление потребляемых ресурсов электроэнергии аппаратом для модуля служебных систем (исключая саму подсистему энергообеспечения) и модуля полезной нагрузки;

- комбинированная осевая симметрия (тип λ), включающая независимо интерпретируемые вертикальную и горизонтальную (обе) оси симметрии в составе образа; для космического аппарата примером может служить комбинация вариантов из α и β;

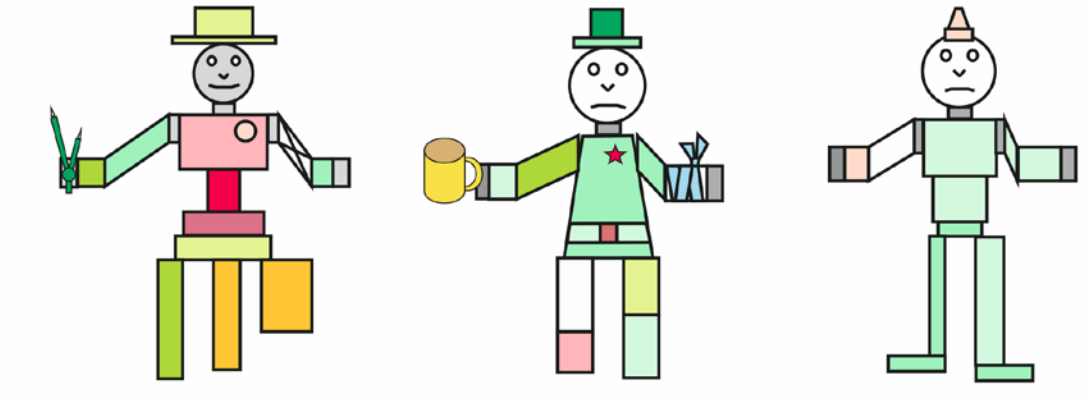
- аксиальная (лучевая) симметрия (тип δ), позволяющая интерпретировать композицию осей относительно центра вращения (все оси имеют «равноправный» вклад в составе образа); например, отображение для каждой подсистемы модуля служебных систем аппарата осей типа E, сформированных относительно вклада ресурсов на этапах проектирования, реализации и тестирования.

Выбор типа каждой оси и типа симметрирования напрямую зависит от сложности системы, специфики решаемой задачи и выбора точки зрения лицом, принимающем решение [10]. Приведенные выше примеры для космического аппарата носят иллюстративный характер, но позволяют сделать общее представление о различии соответствующих типов. С целью кратко обозначения архитектуру образа в нотации UGVA, закодируем его следующим образом: греческой буквой опишем тип плоскостной симметрии; далее цифрой число основных осей; потом тип осей буквой латинского алфавита; затем индикаторы наличия оценочной зоны (*i*) и учета закодированной активности (*t*). Налагаемые артефакты и оперативные данные на этапе формирования архитектуры образа не кодируются (примеры будут приведены далее).

Очевидно, что графовые структуры, являющиеся исходными данными для задачи принятия решений по схеме с рис. 1 будут нуждаться в группировке, обобщении и выборе оценочных (интегральных) показателей. Проиллюстрируем характеристику различных образов и их графическое представление в нотации UGVA для трех сложных объектов социальной природы, учет активности которых важен при выработке управляющих решений.

Первый пример – учащийся (студент), дистанционно взаимодействующий с интеллектуальной средой электронного обучения (ITS) по всем предметам учебного плана, которая должна не только вырабатывать управляющие воздействия. *Второй пример* – работник промышленного предприятия (подразделение монтажников), непосредственный начальник которого использует интеллектуальную среду поддержки принятия решений, которая должна давать рекомендации по мотивации сотрудников по результатам учёта их деятельности. *Третий пример* – проектная команда, выносящая свои научно-исследовательские и опытно-конструкторские работы в виде проектов (заявки) на конкурсы грантового фонда для финансирования. Здесь также применяется интеллектуальная система поддержки принятия решений. Во всех трех случаях необходимо не только предложить определенные решения, но и сопроводить их образами в нотации UGVA для того, чтобы лицо, принимающее решение, могло убедиться в их адек-

ватности и сделать наиболее предпочтительный выбор. Характеристики рассмотренных объектов приведены в таблице, а примеры индивидуализированных образов показаны на рис. 4 а, б и в для второго, третьего и четвертого столбцов таблицы соответственно.



а) б) в)

Рис. 4. Примеры экземпляров индивидуализированных образов в нотации UGVA

Таблица.

Характеристика условий и параметры исходных данных для построения образов

Объект	Учащийся	Работник	Проектная команда
Базовые характеристики образа	Ключевые группы умений	Ключевые группы деятельности / обязанностей	Группы ключевых работ / задач в проекте (заявке)
Ключевой интересант	Учащийся, руководитель программы	Работодатель (руководитель)	Эксперт фонда
Ведущая задача	Управление учебным процессом	Мотивация к работе (включая обоснование премирования)	Оценка перспектив проекта (включая отбор на финансирование)
Целевой	Нормативные по-	Плановый объём	Требования

компонент	казатели обучения из учебного плана	работ на отчетный период	конкурсной документации
Обстановочный компонент	Оценки знаний и компетентностный профиль	Данные журнала о выполнении работ	Заявленные ресурсы на выполнение проекта
Опыт	Структура текущего и предыдущего учебных планов	Данные о квалификации и её повышении	Данные о ранее выполненных проектах членами команды
Тип симметрирования	Комбинированная осевая симметрия	Аксиальная симметрия	Аксиальная симметрия
Интегральной (оценочный) показатель	Баланс учебной нагрузки по ключевым группам умений	Отклонение от значения максимального премирования	Оценка соответствия опыта и компетенций команды условиям конкурса
Число осей	По 4 в нижней и средней страте	Без выделения страт (4, по числу недель)	Без выделения страт (5 по группам S_j)
Типы осей	Составная структурная (*B*)	Составная разнородная (C)	Составная причинно-следственная (D)
Исходные данные для осей	Распределение нагрузки в плане	Структура работ по заданиям на смену	Распределение работ и задач
Исходные данные для тепловой карты	Текущие оценка знаний / компетентностный профиль	Отметки о выполнении работ	Компетентностный профиль исполнителей
Примеры признаков для	Доклад на конференции (значок)	Нарушение распорядка (кружка)	Наличие курирующей

артефактов	Спецкурс работодателя (циркуль)	Больничный (бинт)	организации (шляпа)
Охват временного интервала на визуальном образе	От данных о предыдущей ступени подготовки до результата проведения текущего контрольно-измерительного мероприятия	За отчетный месяц	От первой заявки исполнителей в фонд до момент проведения конкурсного отбора
Обозначение архитектуры образа	λ5Bit	𝛿6Bit	𝛿5Bit

В 2022 году нами были использованы образы в нотации UGVA для решения ряда задач, связанных с сопровождением учебного процесса в Сибирском федеральном университете для магистратуры специальности 09.04.01.03: отбора (сравнения) абитуриентов при проведении приемной компании в магистратуру, оценки и мониторинга успеваемости группы учащихся, актуализации учебных планов (например, см. [11]). Ведется работа по применению метода для оптимизации бизнес-процессов, связанных с деятельностью работников отдельных подразделений на градообразующих предприятиях.

В заключении следует отметить, что метод унифицированного графического воплощения активности (UGVA) зарекомендовал себя в задачах поддержки принятия решений человеком для моделей, ориентированных на функциональный подход (визуализация активности). Его применение для отображения сложных систем относительно их структуры в большинстве случаев будет иметь меньшую результативность относительно иных методов когнитивной визуализации из Data Mining.

Список литературы

1. Arrieta A.B., D´ıaz-Rodr´ıguez N., Del Ser J. et al. Explainable Artificial Intelli-

gence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 2020, pp. 82–115.

2. Анохин П.К. Узловые вопросы теории функциональных систем – М.: Наука, 1980. – 200 с.
3. Балашов Е.П. Эволюционный синтез систем. – М.: Радио и связь, 1985. – 328 с.
4. Han J., Kamber M. (2006). Data mining: concepts and techniques, 2nd. University of Illinois at Urbana Champaign: Morgan Kaufmann. 2006. – 560 p.
5. Miller G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 1956. vol. 63. №. 2. – pp. 81-97.
6. Chernoff H. The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, vol. 68. 1973. pp. 361-368.
7. Углев В.А. Антропоморфные образы в нотации UGVA как средство поддержки принятия решений // Бионика – 60 лет. Итоги и перспективы: Материалы I Международной конференции. – М.: Ассоциация технических университетов, 2022. – С. 231-242.
8. Филимонов В.А. Способ когнитивной визуализации многопараметрических компонентов системы // Робототехника и искусственный интеллект: Материалы XIII Всероссийской конференции с международным участием. – Красноярск: Литера-принт, 2021. – 136-140. https://aesu.ru/local/conference/_docs/2021/RAI-21_print.pdf
9. Углев В.А. Метод унифицированного графического воплощения активности (UGVA) / В. А. Углев // Робототехника и искусственный интеллект: Материалы XI Всероссийской конференции с международным участием. – Красноярск: ЛИТЕРА-принт, 2019. – С. 161-172. doi: 10.13140/RG.2.2.15680.46088.
10. Тарасенко Ф.П. Прикладной системный анализ: учеб. Пособие. – М.: Кнорус, 2010. – 224 с.
11. Углев В.А. Визуальная поддержка принятия решений при синтезе учебных планов с помощью метода UGVA // Искусственный интеллект и принятие решений. – 2022. – №2. – С. 51-61.

СИСТЕМА AMI – ИНСТРУМЕНТ АНАЛИЗА ЭФФЕКТИВНОСТИ ЦИФРОВОЙ НАРУЖНОЙ РЕКЛАМЫ СРЕДСТВАМИ КОМПЬЮТЕРНОГО ЗРЕНИЯ

О.В. Усманов[1], М.Г. Доррер[2]

[1] ООО Смарт Диджитал, ***inbox@hi-tech24.ru***

[2] Сибирский государственный университет науки и технологий им. М.Ф. Решетнева, ***dorrer_mg@sibsau.ru***

Введение

Оценка эффективности рекламы является больным местом большинства предприятий. И это вполне понятно, на рекламу тратятся значительные суммы, а результативность от нее часто непонятна не только руководителям, но и маркетологам. Можно выделить несколько основных причин низкой эффективности рекламы:

- Отсутствие информации о целевом потребителе и источниках получения им информации (каналы доступа к потребителю).
- Отсутствие обратной связи с потребителем.
- Ошибки сегментации.

Использующиеся в настоящий момент каналы обратной связи:

Традиционные каналы обратной связи с потребителями:

- Платформы социальных сетей
- Интернет-мониторинг
- Сторонние сайты
- Отделы продаж и техподдержки
- Социология
- WIFI аналитика

Однако при работе с DOOH (Digital Out-of-Home) рекламой применение таких инструментов невозможно, либо неприемлемо дорого, или же необъективно. Вместе с тем, с развитием систем компьютерного зрения формируется возможность отслеживания для визуальной рекламы формата DOOH метрик, харак-

терных до настоящего момента для интернет-маркетинга:

- Reach (охват аудитории).
- Просмотры.
- Клики.
- Демографические характеристики аудитории (пол, возраст).

Связанные работы

Следует отметить, что задача оценки обратной связи от потребителей – реакции на товар, услугу, рекламные мероприятия, не только является актуальной практической задачей, но и популярным объектом научных исследований.

Так, авторы работы [1] изучают реакции потребителей на цифровую рекламу с использованием гипотез о переносе аффекта и включения поискового поведения.

В работе [2] авторы провели метаанализ с целью изучения взаимосвязи между медиа-контекстом и рекламной памятью в количественных исследованиях.

Работа [3] исследует феномен «отложенной конверсии» рекламного материала и реализует модель анализа отложенных конверсий с учетом времени задержки на основе моделей машинного обучения

В исследовании [4] изучалось влияние таких характеристик рекламы как длина, юмор и информативность на воспринимаемую навязчивость рекламы и, следовательно, на маркетинговые результаты как для онлайн-рекламодателей, так и для владельцев веб-сайтов.

Автор работы [5] выдвинул идею использования интеллектуального анализа данных, в частности инструмента ассоциативных правил при анализе данных реляционной базы данных компании по недвижимости с целью выбора рекламных носителей на основе оценки эффективности рекламы. Подобный подход применялся и в работе [6]. Кроме того анализ данных о потребительском поведении может проводиться для выделения групп потребителей с помощью инструментов кластеризации [7].

Следует отметить обширный корпус статей, использующих методы обработки естественного языка (NLP) для оценки реакций потребителей на рекламные материалы, высказанных в социальных сетях и системах обратной связи по качеству товаров и услуг.

Так, работа [8] посвящена сравнительному исследованию подходов к оценке эффективности рекламных кампаний на основании анализа текстов отзывов в социальных сетях.

В статье [9] была предпринята попытка представить эффективный способ, оценки потребительского рынка, например, что люди ищут при покупке продуктов и каковы типичные причины разочарования в продукте или процессе его покупки.

Работа [10] описывает методологию, основанную на машинном обучении и анализе текстов на естественном языке, для измерения настроений пользователей, выраженных в текстовой обратной связи.

Подобную задачу в масштабе корпоративной системы управления качеством решали авторы работы [11]. В этом исследовании описывается прототип механизма анализа отзывов клиентов на основе алгоритмов машинного обучения без учителя, обеспечивающего оперативное выявление проблем с качеством продукта.

Отдельно следует выделить публикации, посвященные прогнозу эффективности рекламных акций.

Так, исследование [12] направлено на анализ эффективности рекламы в контексте внедрения возобновляемых источников энергии. Инструментом анализа и прогноза выступили искусственные нейронные сети.

Подобная задача в работе [13] решается при помощи рекуррентных нейронных сетей архитектуры LSTM.

Методы решения задач

Модуль компьютерного зрения системы AMI базируется на комплексном решении, включающем несколько архитектур глубоких сверточных нейронных

сетей, решающих частные задачи анализа аудитории рекламных поверхностей.

Подсчет появления зрителей перед рекламной поверхностью (оценка охвата аудитории) реализуется за счет применения архитектуры MediaPipe от Google [14] в сочетании с разработанным в рамках работы над AMI трекером объектов.

Метрикой оценки просмотра рекламного материала могут быть эмоции, испытываемые зрителем. Измерение этого показателя достигается применением предобученных моделей распознавания эмоций, например EmoPy [16] и подобных им. В представленном решении использовалась полностью оригинальная сверточная нейронная сеть распознавания эмоций.

Задача анализа положения тела и жестов зрителей решается за счет применения библиотек типа OpenPose [18], основанных также на инструментах CNN. В предложенном решении использовалась архитектура MediaPipe Pose [19].

Анализ демографических характеристик аудитории позволит выполнить применение CNN архитектуры Gender_net VGG16 [20] или подобной. В представленном решении для решения этой задачи использована архитектура из работы [17], переконфигурированная по числу выходов в соответствии с количеством категорий возраста, выявляемых системой, и дообученная на датасете детекции возраста по лицам.

Для анализа пола зрителей также используется вариант архитектуры из работы [17] в котором изменено количество нейронов в выходном слое, а также использована отличная от базового варианта архитектуры функция активации на выходном слое и функция потерь (применена бинарная кросс-энтропия вместо категориальной).

Оценка числа просмотров требует оценки фокусировки внимания зрителей на рекламной поверхности. В предложенном решении данная задача решается при помощи детектора лиц Mediapipe FaceDetection [21].

Обзор функциональных возможностей системы

Описанные в предыдущем разделе методы и архитектуры были использованы при разработке системы AMI. Данная система представляет собой интеллектуальный сервис видеоаналитики, который позволяет подсчитывать количество людей, просмотревших рекламу, определять их пол, эмоции, примерный возраст, время просмотра, а также показывать таргетированную рекламу в режиме реального времени.

AMI устанавливается на цифровой экран или медиаплеер с HD вебкамерой или IP камерой.

В процессе эксплуатации AMI отслеживает каждый направленный на экран взгляд, вычисляет продолжительность просмотра, определяет возрастную группу, пол смотрящего, его эмоции при этом сохраняя его полную анонимность.

Собранные данные отправляются на сервер в режиме реального времени. Система обеспечивает просмотр статистики онлайн.

Аналитические возможности системы

Аналитика просмотров ведется в разрезе списка подключенных устройств и временных интервалов.

Общая статистика отражается в разрезе следующих показателей:

- Количество людей, зафиксированных перед камерой.
- Количество людей, посмотревших на объект интереса (например, экран с медиаконтентом).
- Отношение просмотревших к общему числу определенных людей (конверсия потока потенциальных зрителей в просмотры).
- Общее время, в течение которого люди находились перед камерой.
- Общее время, в течение которого люди смотрели на объект интереса.
- Коэффициент вовлеченности в контент – отношение времени просмотра к общему времени нахождения зрителей перед объектом интереса.

Анализ демографических показателей аудитории позволяет собрать данные по просмотрам относительно пола, возраста и эмоций за указанный временной интервал на подключенных к системе устройствах.

Демографическая статистика собирается в разрезе следующих показателей:

- Доля просмотров, совершенных мужчинами;
- Доля процент просмотров, совершенных женщинами;
- Категория зрителей, на долю которой пришлось максимальное число просмотров с разбивкой по демографическим показателям;
- Средняя продолжительность одного просмотра у мужчин;
- Средняя продолжительность одного просмотра у женщин;
- Категория зрителей, на долю пришлись самые длительные просмотры с разбивкой по демографическим показателям;

Анализ привлекательности медиа материалов анализируется путем сбора и представления статистики просмотров медиафайлов, демонстрировавшихся на подключенных к системе устройствах за исследуемый период.

Анализ по каждому медиафайлу ведется в разрезе следующих ключевых показателей:

- Количество потенциальных зрителей, зафиксированных перед камерой во время проигрывания медиафайлов (в том числе те, кто не смотрел на объект внимания).
- Количество просмотров, зафиксированных перед камерой во время воспроизведения каждого медиафайла.
- Медиафайл с самой продолжительной общей длительностью просмотров.
- Медиафайл, чаще всего вызывавший положительные эмоции у зрителей.
- Среднее время нахождения потенциальных зрителей перед объектом внимания во время воспроизведения медиафайла.
- Средняя длительность всех просмотров, зафиксированных во время

воспроизведения медиафайла.

В число статистических параметров по каждому медиаматериалу входят:

- Количество людей, зафиксированных перед камерой во время воспроизведения медиафайла.
- Количество людей, посмотревших на объект интереса (например, экран с медиаконтентом) во время воспроизведения медиафайла.
- Отношение просмотревших к общему числу определенных людей (конверсия потока потенциальных зрителей в просмотры).
- Общее время, в течение которого люди находились перед камерой во время воспроизведения медиафайла.
- Общее время, в течение которого люди смотрели на объект интереса во время воспроизведения медиафайла.
- Дата последнего зафиксированного просмотра данного медиафайла.

Выводы

Таким образом, описанная в данной работе система AMI за счет применения комплекса архитектур сверточных нейронных сетей автоматизировать задачу анализа реакции потребителей на медиаматериалы различного назначения. Система графиков позволяет проанализировать эффективность рекламы или измерить поток покупателей в торговом центре с точностью и подробностью, которые ранее были доступны лишь при проведении полноценного маркетингового исследования целой группой полевых исследователей – маркетологов, социологов и психологов.

Следует отметить, что предложенное в системе AMI решение обладает существенными признаками оригинальности. Так, ближайшим аналогом, существующим на рынке, является решение Quividi [24]. В рамках данного решения предлагается измерять поток зрителей перед рекламными поверхностями, а также оценивать внимание зрителей к рекламным материалам за счет анализа положения головы зрителя.

AMI обладает существенными отличиями и преимуществами по сравне-

нию с решением Quividi:

- Оценка меры положительности отношения зрителей к демонстрируемым материалам за счет применения инструментов распознавания эмоций;
- Реализация немедленной обратной связи рекламной поверхности с реакцией зрителя, выражающаяся в подстройке сценария демонстрации рекламных материалов к поведению и эмоциям зрителя.

Следует отметить, что хотя результаты статистики аудитории сами по себе ценны для задач управления рекламными компаниями, однако работа обладает существенным потенциалом дальнейшего развития.

В настоящий момент ведется разработка алгоритмов предиктивной аналитики аудитории. Данный подход является существенным шагом в развитии функциональности системы, позволяющим повысить адресность и эффективность рекламных сообщений. Рекламодатель сможет редактировать свой контент под целевую аудиторию, повышая охват и вовлеченность, а следовательно – и конечный эффект от рекламы. Также, анализируя статистику в динамике, выполняя прогноз оцениваемых параметров аудитории, ее реакцию в привязке к временным диапазонам, к различным демографическим группам аудитории, и учитывая влияние других факторов, система позволит повысить точность предсказания расходов на рекламный бюджет. Таким образом, система обладает значительным потенциалом для дальнейшего совершенствования.

Список литературы

1. Stewart K. et al. Examining digital advertising using an affect transfer hypothesis // J. Res. Interact. Mark. 2018. Vol. 12, № 2. P. 231–254.
2. Kwon E.S. et al. Impact of Media Context On Advertising Memory // J. Advert. Res. 2019. Vol. 59, № 1. P. 99–128.
3. Chapelle O. Modeling delayed feedback in display advertising // Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2014. P. 1097–1105.

4. Goodrich K., Schiller S.Z., Galletta D. Consumer Reactions to Intrusiveness Of Online-Video Advertisements // J. Advert. Res. 2015. Vol. 55, № 1. P. 37–50.
5. Wang Z.-Y., Liu Q. Feedback analysis of real estate advertising effectiveness based on association rules. 2012. Vol. 34. P. 98–102.
6. Ogurtsov D.A., Dorrer M.G. Application of association rules learning for studying the store history of a large retail chain // Journal of Physics: Conference Series. 2019. Vol. 1399, № 3.
7. Dorrer M.G., Fomin A. V, Loginov D.A. Clustering of participants in the {MaxBonus} loyalty system using Kohonen's self-organizing maps // J. Phys. Conf. Ser. {IOP} Publishing, 2020. Vol. 1679. P. 42010.
8. Raudeliūnienė J. et al. Evaluation of Advertising Campaigns on Social Media Networks // Sustainability. 2018. Vol. 10, № 4. P. 973.
9. Pal S. Customer Feedback Analysis using NLP // Indian J. Comput. Sci. 2021. Vol. 6, № 1. P. 17.
10. Kumar A., Jain R. Uniform Textual Feedback Analysis for Effective Sentiment Analysis. 2021. P. 273–289.
11. Lin M.S., Wen R. Customer Feedback Analysis Engine for Manufacturing Quality Management // SSRN Electron. J. 2022.
12. Sharifi M. et al. Forecasting of advertising effectiveness for renewable energy technologies: A neural network analysis // Technol. Forecast. Soc. Change. 2019. Vol. 143. P. 154–161.
13. Доррер М.Г. et al. Применение рекуррентных нейронных сетей для прогнозирования эффекта от применения стимулирующих акций в торговых сетях // Моделирование неравновесных, адаптивных и управляемых систем : Материалы XXIV Всероссийского семинара, Красноярск, 01–03 октября 2021 года / ed. Сенашова М.Ю. Красноярск: Институт вычислительного моделирования Сибирского отделения Российской академии наук, 2021. P. 44–52.
14. Google. MediaPipe [Electronic resource]. URL: https://mediapipe.dev/.
15. Redmon J. et al. You Only Look Once: Unified, Real-Time Object Detection.

2015.

16. ThoughtWorks Arts. EmoPy 0.0.5 [Electronic resource]. URL: https://pypi.org/project/EmoPy/.
17. Sethi K. Emotion Detection Using OpenCV and Keras [Electronic resource]. URL: https://medium.com/swlh/emotion-detection-using-opencv-and-keras-771260bbd7f7.
18. CMU-Perceptual-Computing-Lab. OpenPose [Electronic resource]. URL: https://github.com/CMU-Perceptual-Computing-Lab/openpose.
19. Google. MediaPipe Pose [Electronic resource]. URL: https://google.github.io/mediapipe/solutions/pose.
20. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.
21. Google. MediaPipe Face Detection [Electronic resource]. URL: https://google.github.io/mediapipe/solutions/face_detection.
22. Krafka K. et al. Eye Tracking for Everyone. 2016.
23. Canu S. Object Tracking with Opencv and Python [Electronic resource]. URL: https://pysource.com/2021/01/28/object-tracking-with-opencv-and-python/.
24. Quividi. Quividi. Digital Signage Analytics and Interactivity [Electronic resource]. URL: https://quividi.com/.

ПРИМЕНЕНИЕ ИНСТРУМЕНТОВ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ ОЦЕНКИ ГАБАРИТОВ ДВИЖУЩЕГОСЯ КОНТЕЙНЕРА

А.Е. Алехина, М.Г. Доррер

Сибирский государственный университет науки и технологий им М.Ф. Решетнева, ***dorrer_mg@sibsau.ru***

Введение

Сельское хозяйство является широчайшим сектором экономики и играет важную роль в общем экономическом развитии страны. Технологические достижения в области сельского хозяйства позволят повысить конкурентоспособность многих видов сельскохозяйственной деятельности [1]. Вместе с тем, современное сельское хозяйство сталкивается с многочисленными проблемами, в том числе: постоянно растущий спрос на качественные продукты питания, нехватка рабочей силы и пахотных земель, сокращение поливной воды, повышенное загрязнение почвы, потеря урожая из-за болезней растений и вредителей. В таких условиях для поддержания эффективности сельскохозяйственной отрасли сектору необходимо прибегать к новейшим сетевым технологиям и методам искусственного интеллекта (ИИ) для оптимизации ресурсов и устойчивого производства качественных и экологически чистых продуктов питания.

В рамках данного круга проблем рассмотрим задачу автоматического управления силосопроводом комбайна. Силосопровод – это желоб, по которому перемещается измельченная зеленая (силос) масса из бункера комбайна в кузов грузовика, движущегося рядом с той же скоростью. Проблема заключается в том, что в процессе уборки урожая в поле комбайнер управляет одновременно и комбайном, и силосопроводом, что снижает эффективность выполнения им основной задачи.

Снизить нагрузку на комбайнера и улучшить показатели уборки кормовых культур призвана помочь система автоматической выгрузки кормоуборочной массы через силосопровод в кузов движущегося рядом грузовика. Для этого на силосопровод устанавливается стереокамера, а на борт комбайна — модуль ви-

деообработки. Модуль анализирует полученные с камеры данные и отправляет сигнал в бортовую сеть комбайна для управления силосопроводом.

Следует отметить, что вопрос применения инструментов компьютерного зрения на производственных площадках различного назначения является в настоящее время достаточно актуальным. Так, в работе [2] предложен метод распознавания фруктов для роботизированных систем для идентификации груш в сложной среде сада с использованием 3D-стереокамеры в сочетании с технологией глубокого обучения Mask Region-Convolutional Neural Networks (Mask R-CNN). Авторы статьи [3] для управления колесными роботами предложили алгоритм обнаружения объектов, основанный на сочетании улучшенного YOLOv4 и улучшенного GhostNet. В работе [4] представлен и обоснован новый подход к созданию аппаратно-эффективной автоматизированной системы распознавания номерных знаков для ограниченной среды с ограниченными ресурсами на базе сверточных нейронных сетей. Работа [5] предлагает основанную на глубоком обучении сверточную нейронную сеть Vision Transformer Particle Region (VitP-RCNN), обеспечивающую ускорение извлечения признаков для задачи визуального управления посадкой беспилотных летательных аппаратов. Автор работы [6] предлагает решение, позволяющее применить технологии интеллектуальной обработки изображений для извлечения признаков образцов художественного стиля стекла.

Авторы данной работы также обращались к задачам применения компьютерного зрения для автоматизации мониторинга производственных площадок [7], для управления умными торговыми автоматами [8] и сельскохозяйственными роботами [9].

Таким образом, представляет интерес решение задачи автоматизации операций управления силосопроводом комбайна при помощи инструментов компьютерного зрения

Материалы и методы

Растровые данные были получены на соревновании Computer vision

competition AgroHack Code 2022 [10]. Были представлены как изображения в формате png, так и облако точек, привязанных к бункеру комбайна. В нашей работе были использованы только образы, так как облако точек было очень сильно зашумлено и не представлялась привлекательным источником данных для качественного выполнения дополнительного поиска центроида движущегося объекта.

Для определения контуров контейнера были выбраны классические методы фильтрации изображения, такие как адаптивная фильтрация по среднему и адаптивная гауссовская фильтрация (adaptive mean threshold, adaptive gaussian threshold) [11].

Одним из примененных при решении задачи инструментов компьютерного зрения стала U-Net – свёрточная нейронная сеть, созданная в 2015 году для сегментации биомедицинских изображений в отделении Computer Science Фрайбургского университета [12].

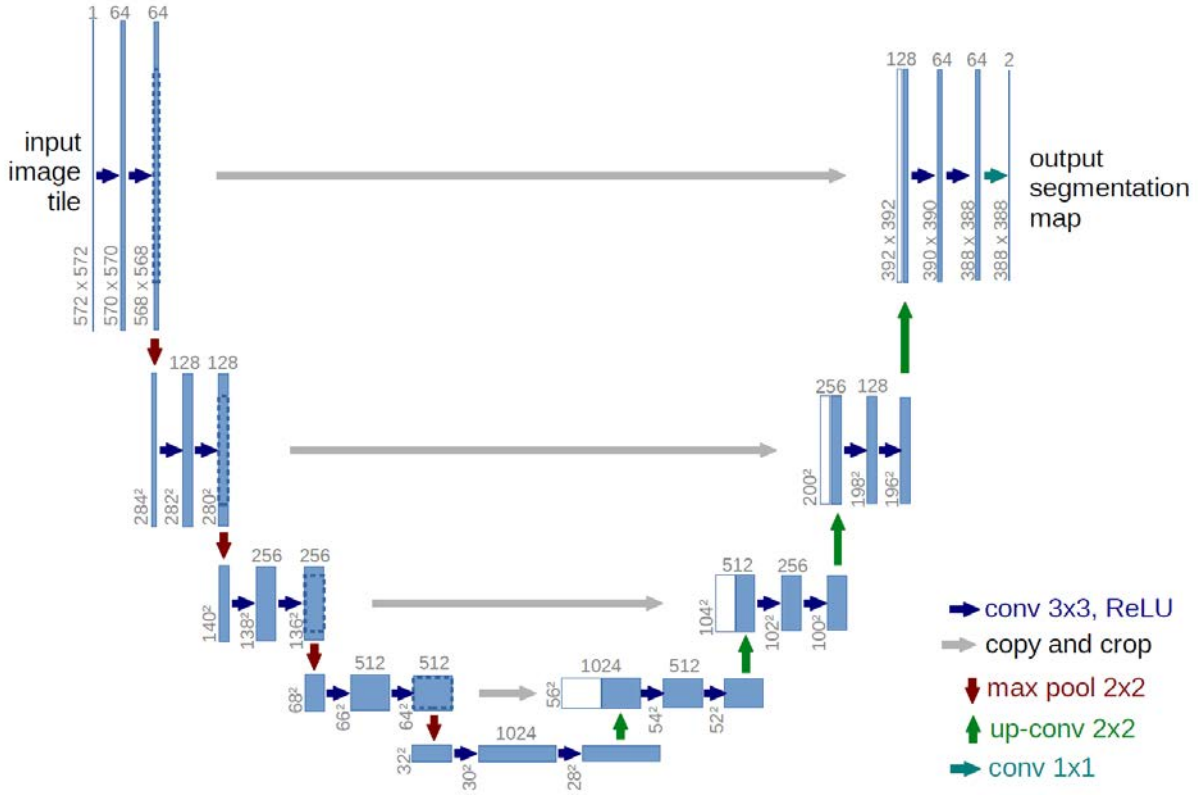


Рис. 1 Архитектура U-Net [12]

В качестве альтернативы для сравнения рассматривалось применение модели DeepLab v3 [13]. Семейство моделей DeepLab, разработанное Google, предназначено для решения задачи семантической сегментации. Качество прогноза достигается простым повышением частоты дискретизации последнего слоя свертки и вычислением пиксельных потерь.

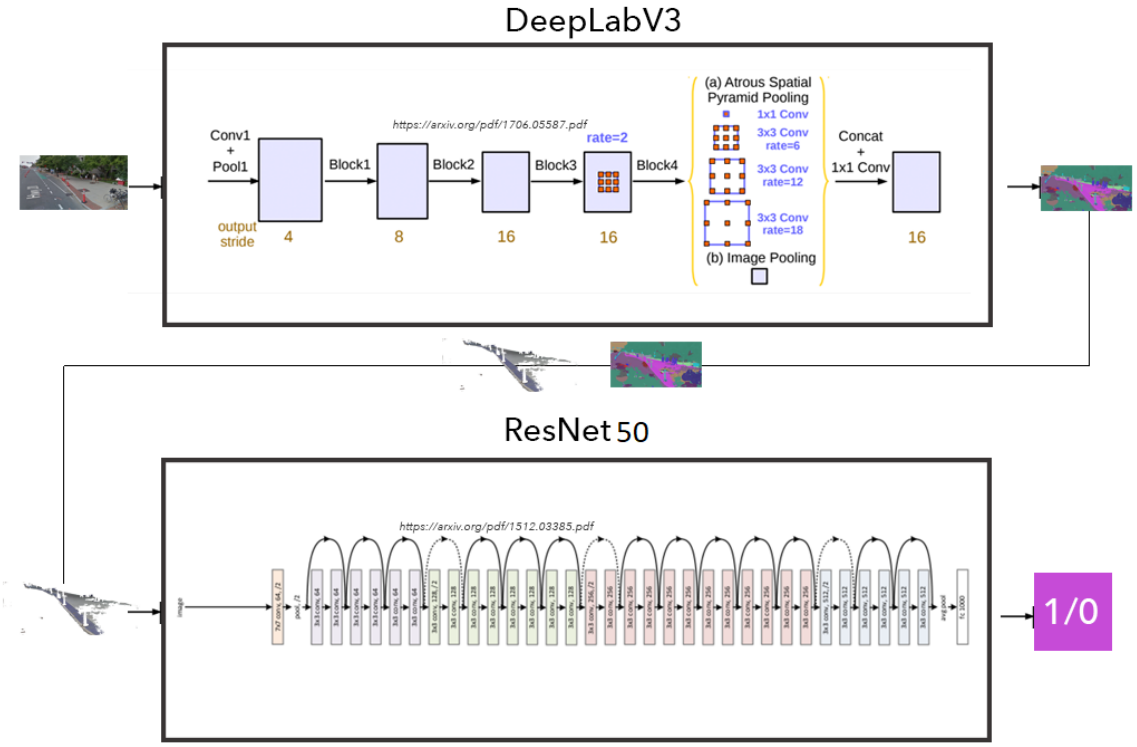


Рис 2. Архитектура Deeplab v3 [13]

Для определения качества обучения сети была выбрана метрика среднеквадратичной ошибки (MSE loss) и коэффициент Сорренсена-Дайса [14].

После бинаризации выбранного изображения и поиска замкнутого контура с 90помощью выбранных фильтров мы находим самый большой замкнутый контур, доводим его до граней относительного прямоугольника и рассчитываем относительные координаты центроида найденной области по формуле:

$$p_t = \left(\left[\frac{m_2}{m_0}\right], \left[\frac{m_1}{m_0}\right]\right), \text{ где} \quad (1)$$

$x = p_t(0)$,

m_0 – первый момент переменной, математическое ожидание, которое

представляет центр тяжести распределения,

m_1 – второй момент переменной – дисперсия,

m_2 – третий момент переменной – асимметрия.

Выбранные модели нейронных сетей обучались на платформе AMD Ryzen 7 4800H Geforce GTX 1660 Ti 6GB 16GB ОЗУ. Общее время работы занимало 5 часов.

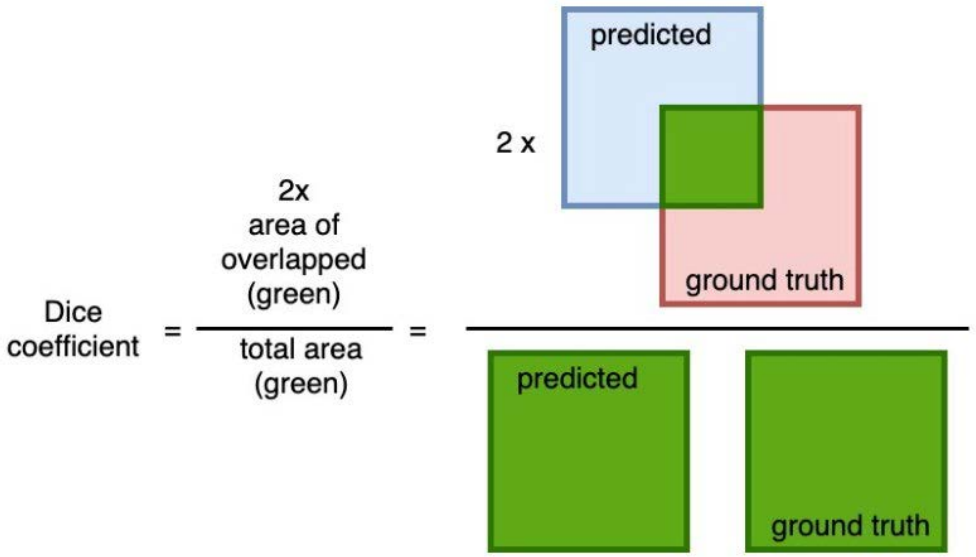


Рис. 3. Пример работы вычисления коэффициента Соренсена-Дайса

Результаты

Для итогового поиска нами был выбран фильтр adaptive mean treshold, так как при использовании данного фильтра на обработанных изображениях замкнутые контуры получались более качественными в отличии от преобразования Гаусса. Это было необходимо, так как в соответствии с одним из условий решения задачи необходимо было найти грани не самого бункера, а его оградительной решетки, которая находится выше основного бункера.

На этапе предобработки изображений к изображению, приведенному к оттенкам серого цвета, были применены алгоритмы фильтрации по среднему и гауссовский. Результат можно видеть на рис. 4.

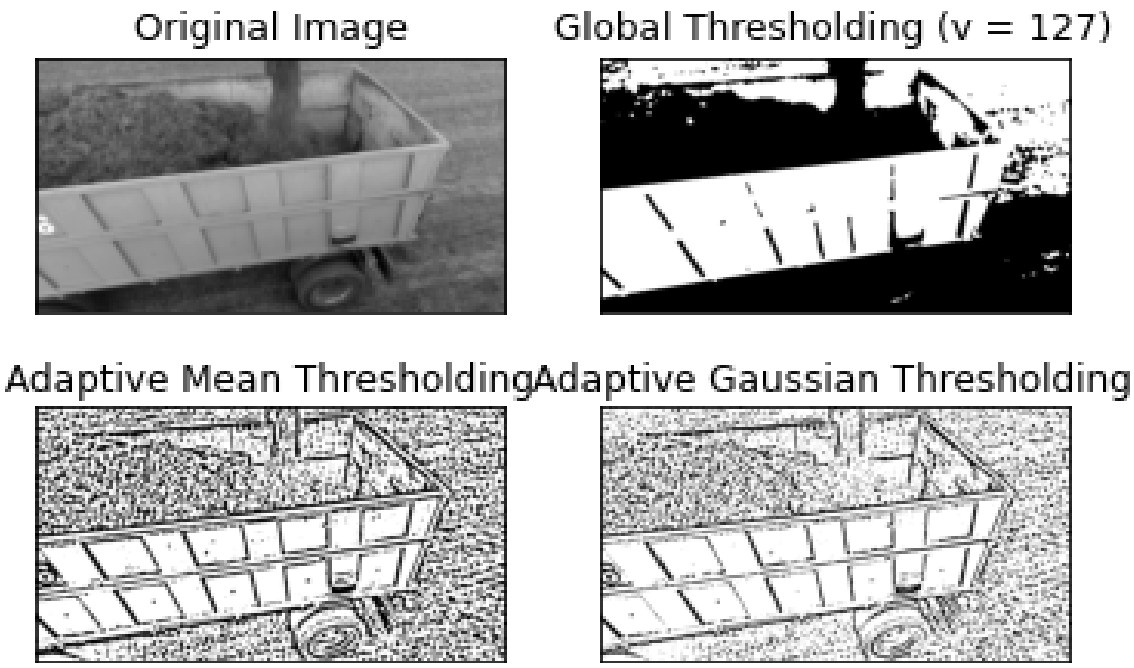


Рис. 4. Результаты работы выделения контуров на изображении с помощью фильтров

После использования выбранного фильтра мы получили замкнутый контур максимально похожий на прямоугольник. Всего было обработано 4600 изображений размером 680x480 пикселей. Вручную было отобрано 1200 «хороших» разметок. Датасет строился по принципу пропорции 8:1:1.

На рис. 5 показаны результаты работы алгоритма полуавтоматической разметки изображения, выделяющей контур силосного бункера автоприцепа.

а)	б)

Рис. 5 Результаты работы алгоритма полуавтоматической разметки данных, где а) – разметка не вошла в обучающую выборку, б) – разметка вошла в обучающую выборку

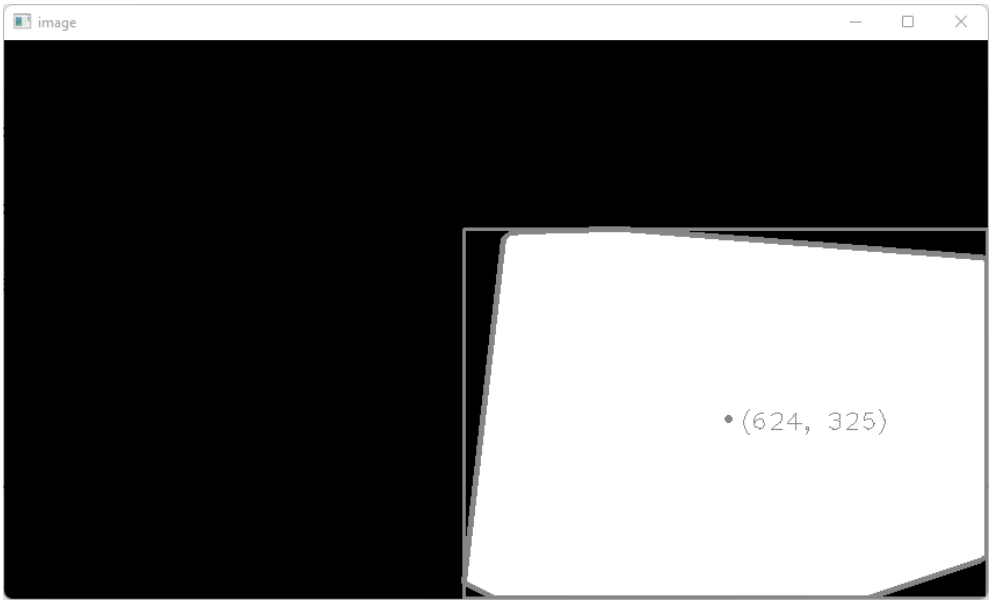


Рис. 6. Пример расчет центра сегментированной области

Для получения лучшей модели сегментации грани контейнера выбранные архитектуры обучались на отобранных изображениях последовательно. Точность и ошибки U-Net и Deeplabv3 составили соответственно 57 и 62 процента точности, величина функции ошибки составила 0.0537 и 0.0287 соответственно. На рисунках 7 и 8 показаны результаты работы моделей на валидационной выборке.

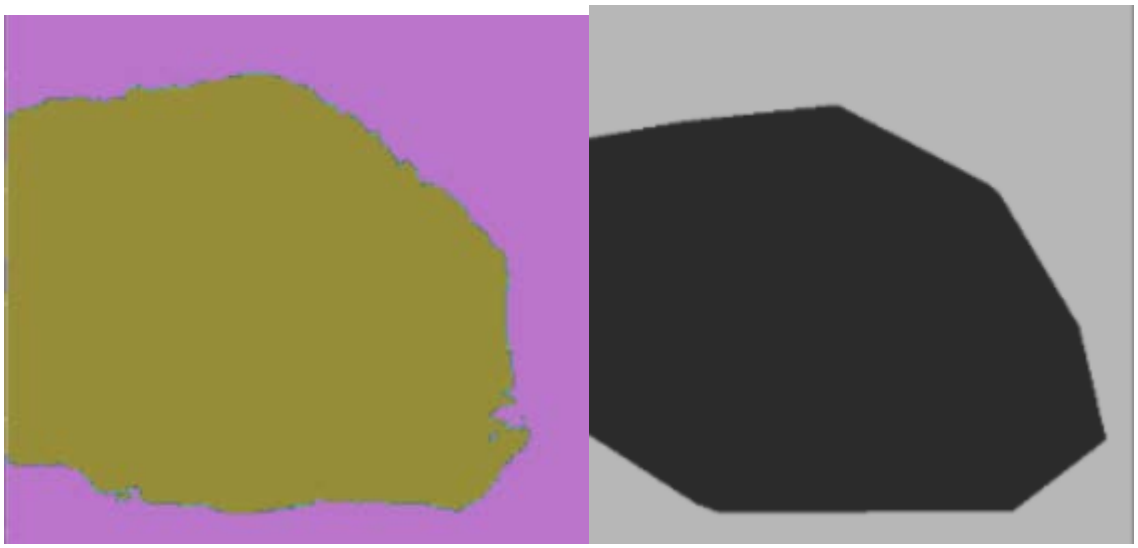


Рис. 7. Результаты U-Net на валидационной выборке

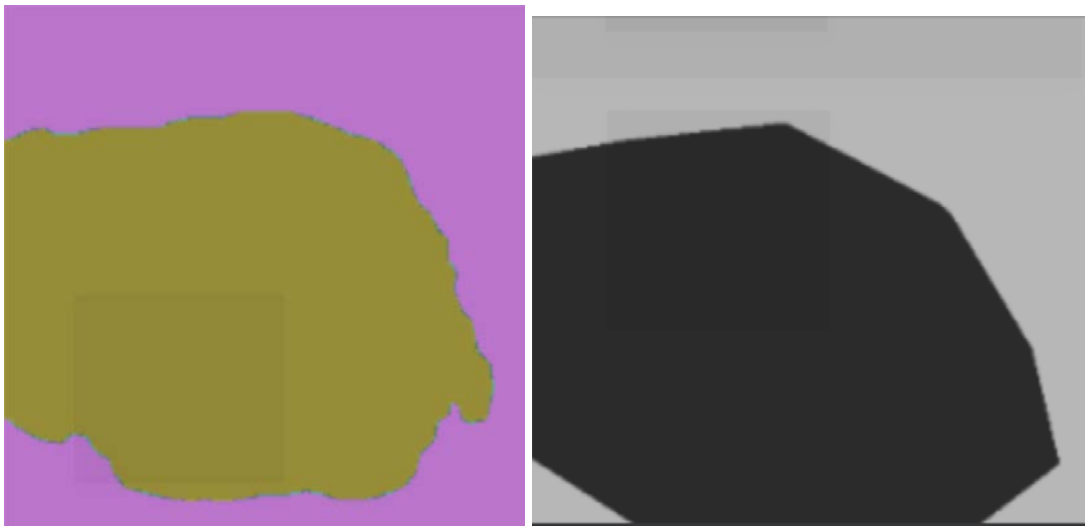


Рис. 8. Результаты DeepLab v3 на валидационной выборке

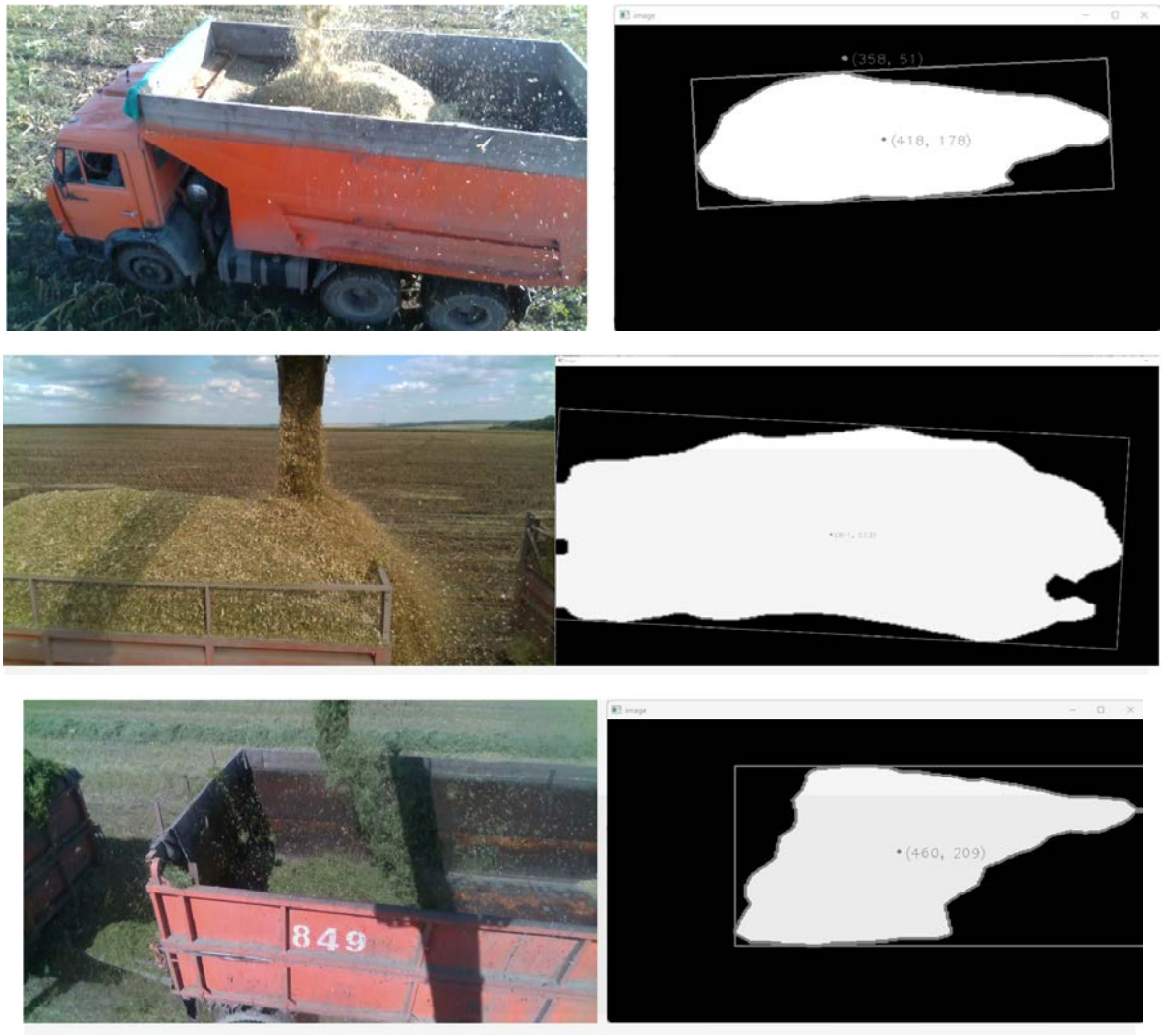


Рис. 9. Результаты работы модели Deeplabv3 при расчете габаритов и заполненного объема бункера прицепа

На рисунке 9 показаны результаты работы модели Deeplabv3 и дальнейшего расчета габаритов кузова в разных условиях и этапах работы силоса. Как видно по результатам минимальные условия необходимые для решения задачи выполняется на разных ракурсах, на разном сырье и при различных уровнях заполнения бункера.

Выводы

Таким образом, нами была решена задача расчета габаритов бункера для сбора сырья в процессе уборке при использовании в качестве входных данных фотографии бункера и силосопровода комбайнв. Минимальные требования решения задачи были выполнены, условие аппаратной обработки на jetson xavier tx2 в 2-3 кадра удовлетворено. Итоговый алгоритм обрабатывал поток изображений 9 кадров в секунду.

Проблема фиксации центроида бункера во время движения остается, так как анализируемый объект не всегда попадает в кадр. Для того, чтобы решить эту задачу, необходимо добавить дополнительный фильтр из области точек и карт глубин. На данный момент данный фильтр нельзя было использовать, так как исходные изображения засорены разного типа шумами и искажены неправильным углом зрения, так как сама стереокамера находится над контейнером.

Рис. 10. Облако точек бункера с шумом и под неправильным углом зрения

Таким образом, нами было получено приемлемое решение задачи распознавания заполненного бункера силосного прицепа для автоматического управления силосопроводом.

Наилучшее качество решения задачи предварительной фильтрации показал алгоритм трешхолдинга adaptive mean treshold. Лучшее качество решения задачи семантической сегментации деталей бункера на изображении с целью расчета координат стенок бункера и координат центроида, аппроксимирующего усредненное положение бункера, показала архитектура сверточной нейронной сети Deeplabv3.

Список литературы

1. Jagtap O.J. et al. Smart Farming // Int. J. Res. Appl. Sci. Eng. Technol. 2022. Vol. 10, № 6. P. 331–346.
2. Pan S., Ahamed T. Pear Recognition in an Orchard from 3D Stereo Camera Datasets to Develop a Fruit Picking Mechanism Using Mask R-CNN // Sensors. 2022. Vol. 22, № 11. P. 4187.
3. Hu Y. et al. Object Detection Algorithm for Wheeled Mobile Robot Based on an Improved YOLOv4 // Appl. Sci. 2022. Vol. 12, № 9. P. 4769.
4. Padmasiri H. et al. Automated License Plate Recognition for Resource-Constrained Environments // Sensors. 2022. Vol. 22, № 4. P. 1434.
5. Yuan B., Ma W., Wang F. High Speed Safe Autonomous Landing Marker Tracking of Fixed Wing Drone Based on Deep Learning // IEEE Access. 2022. Vol. 10. P. 80415–80436.
6. Yu Y. Application of Smart Image Processing Technology in Feature Extraction of Glass Artistic Style Patterns // 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, 2022. P. 1715–1718.
7. Alekhina A., Gurenko A., Dorrer M. Application of Computer Vision Tools to Create a System for Monitoring the Work of Ground Equipment in Open Pits of Gold Mining Enterprises. 2022. P. 203–218.
8. Dorrer M.G., Tolmacheva A.E. Comparison of the YOLOv3 and Mask R-CNN architectures' efficiency in the smart refrigerator's computer vision // J. Phys. Conf. Ser. 2020. Vol. 1679. P. 042022.
9. Dorrer M.G., Popov A.A., Tolmacheva A.E. Building an artificial vision system

of an agricultural robot based on the DarkNet system // IOP Conf. Ser. Earth Environ. Sci. 2020. Vol. 548. P. 032032.

10. AgroCodeHub. Computer vision competition AgroHack Code 2022 [Electronic resource]. URL: https://hack.rshbdigital.ru/crop-unloading.
11. Roy P. et al. Adaptive thresholding: A comparative study // 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE, 2014. P. 1182–1186.
12. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.
13. Wang Y. et al. An improved Deeplabv3+ semantic segmentation algorithm with multiple loss constraints // PLoS One / ed. Zhang Q. 2022. Vol. 17, № 1. P. e0261582.
14. Dice L.R. Measures of the Amount of Ecologic Association Between Species // Ecology. 1945. Vol. 26, № 3. P. 297–302.

РАСПОЗНАВАНИЕ ПАТТЕРНОВ ПОВЕДЕНИЯ ПОСТУПАЮЩИХ НА ОНЛАЙН-ЭКЗАМЕНЕ В УНИВЕРСИТЕТ ПО ДАННЫМ ВИДЕОСЪЕМКИ НА ОСНОВЕ ПРИМЕНЕНИЯ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

М.В. Гунер

Сибирский федеральный университет, институт космических и информационных технологий, кафедра систем искусственного интеллекта,
gunermv@gmail.com

В последнее время стремительно набирают популярность и повсеместно внедряются онлайн-технологии, особо способствовала этому мировая пандемия COVID-19, начавшаяся в 2020 г. В образовательных учреждениях цифровые технологии широко используются в организации учебного процесса, в том числе при проведении процедуры оценки знаний учащихся и студентов [1-7].

Особенно важна процедура проведения вступительных экзаменов в университет, поскольку результаты таких экзаменов являются основой для конкурсного отбора и зачисления наиболее подготовленных. Здесь на университет одновременно возлагаются несколько задач: предоставление всем поступающим равных возможностей на поступление, пресечение грубых нарушений и мошеннических действий со стороны сдающих, и конечно, выполнение плана приема. Стоит отметить, что от качества контингента студентов напрямую зависят показатели их успеваемости и сохранности, а также показатели трудоустройства выпускников, научные показатели университета.

Цель работы – построить модели распознавания паттернов поведения поступающих на онлайн-экзамене в университет по данным видеосъемки на основе применения сверточных нейронных сетей.

Исходные данные – коллекция фотоизображений (кадров с видео) поступающих в Алтайский государственный технический университет им. И.И. Ползунова (АлтГТУ) во время сдачи ими вступительного онлайн-экзамена по математике в 2020 г.

Онлайн-экзамен проходил в авторской системе тестирования, разработанной по заказу АлтГТУ [8]. Во время экзамена все поступающие транслировали себя через веб-камеру (камеру). За прохождением онлайн-тестирования в реальном режиме времени следили модераторы от университета (рисунок 1).

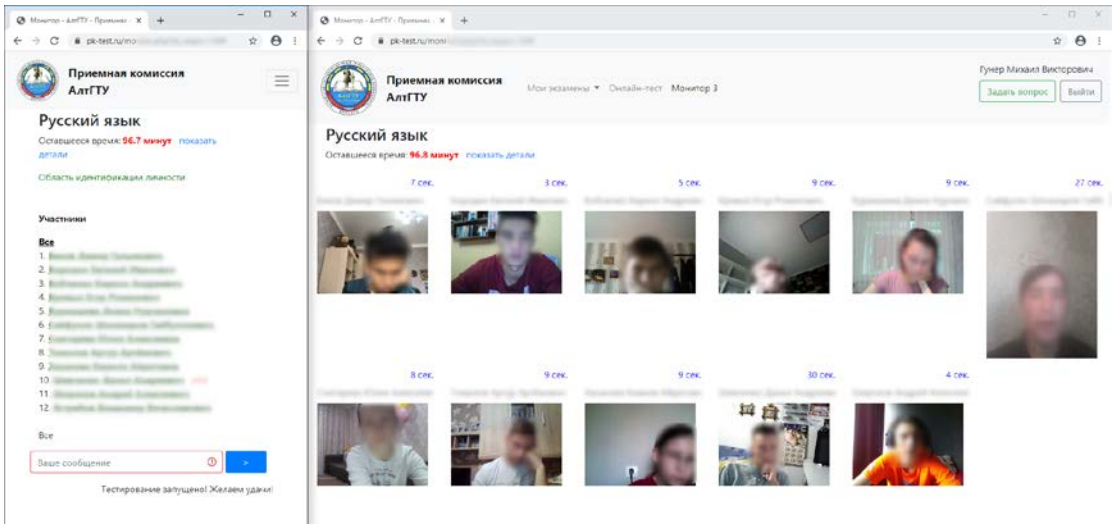


Рис. 1. Система онлайн-тестирования АлтГТУ. Вид со стороны модератора

Всего коллекция изображений насчитывала более 5000 штук (по 10 и более изображений на каждого из 553 абитуриентов, успешно прошедших вступительные испытания, в т.ч. онлайн-тестирование по математике, и зачисленных в АлтГТУ в 2020 г.). Каждое изображение содержало три цветовых канала (RGB) и было размером 640 на 480 пикселей, в абсолютном большинстве случаев ориентация изображения горизонтальная.

Обработка фотоизображений выполнялась в обезличенном виде, но с пометкой о величине расхождения входного рейтинга поступающего по математике и рейтинга по математическим дисциплинам по итогам экзаменационных сессий на первом курсе в ходе самого обучения в университете.

Выбор сверточных нейронных сетей для распознавания паттернов поведения поступающих обусловлен высокой эффективностью этих сетей в задачах классификации и сегментации изображений. В настоящей работе применялись остаточные сверточные нейронные сети ResNet от корпорации Microsoft, кото-

рые в 2015 г. с глубиной слоев 152 уровня стали чемпионом конкурса ILSVRC [9-14].

Проблема ослабевания градиента при обратном распространении ошибки в глубоких сетях ResNet решается при помощи так называемых остаточных связей или соединений быстрого доступа. Архитектура сети ResNet (на примере ResNet-50) показана на рисунке 2.

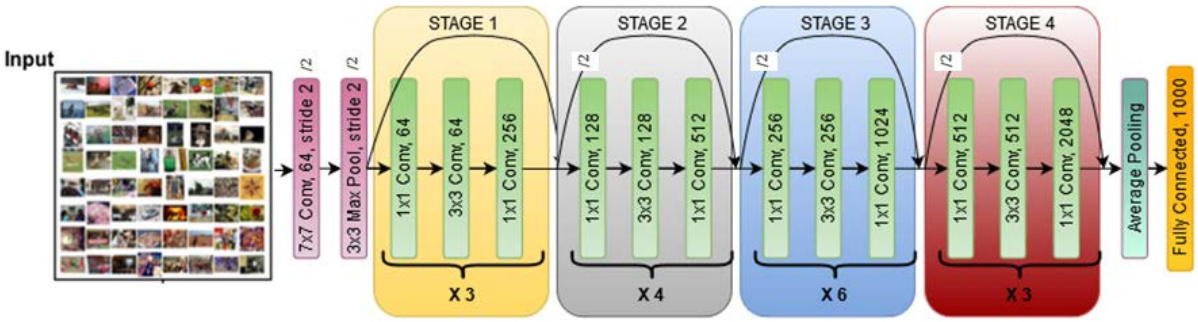


Рис. 2. Архитектура остаточной сверточной нейронной сети ResNet-50

Путем анализа изображений поступающих во время сдачи ими вступительного онлайн-экзамена по математике, а также рейтингов студентов до и после зачисления был выдвинут ряд гипотез по определению самостоятельности прохождения теста по данным видеосъемки. Наличие ручки (карандаша) и черновика в кадре, задумчивый взгляд рассматривались как «хорошее» поведение поступающего на экзамене, в то время как разговоры, взгляды по сторонам, отсутствие в кадре абитуриента или присутствие посторонних расценивалось как «плохое» (подозрительное) поведение, нарушение.

Первый шаг при решении любой задачи классификации изображений – подготовка датасета, или разметка данных. В настоящей работе разметка изображений осуществлялась в VGG Annotator [15].

В ходе разметки на каждом изображении выделялась область, охватывающая голову, туловище и руки абитуриента, а также предметы, находящиеся в его руках (при условии попадания в кадр) (рисунок 3). Затем для каждой области заполнялись значения атрибутов: качество изображения, поза поступающего,

паттерн поведения поступающего и т.д.

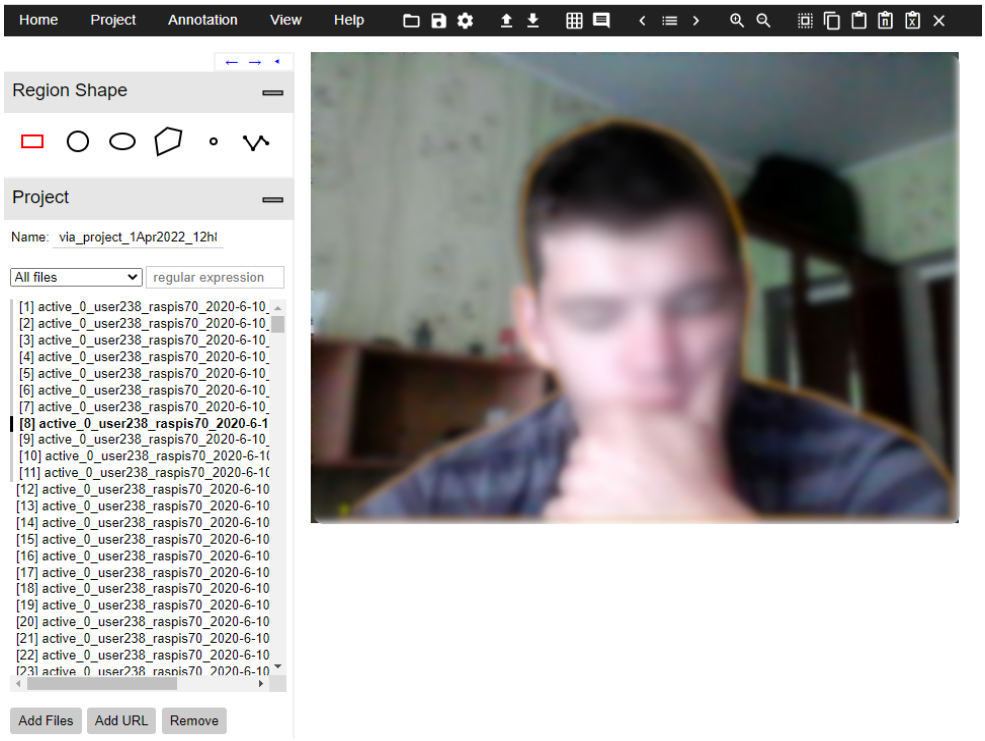


Рис. 3. Аннотирование изображений

Фрагмент коллекции изображений, где поступающий держит в руке ручку (карандаш), показан на рисунке 4.

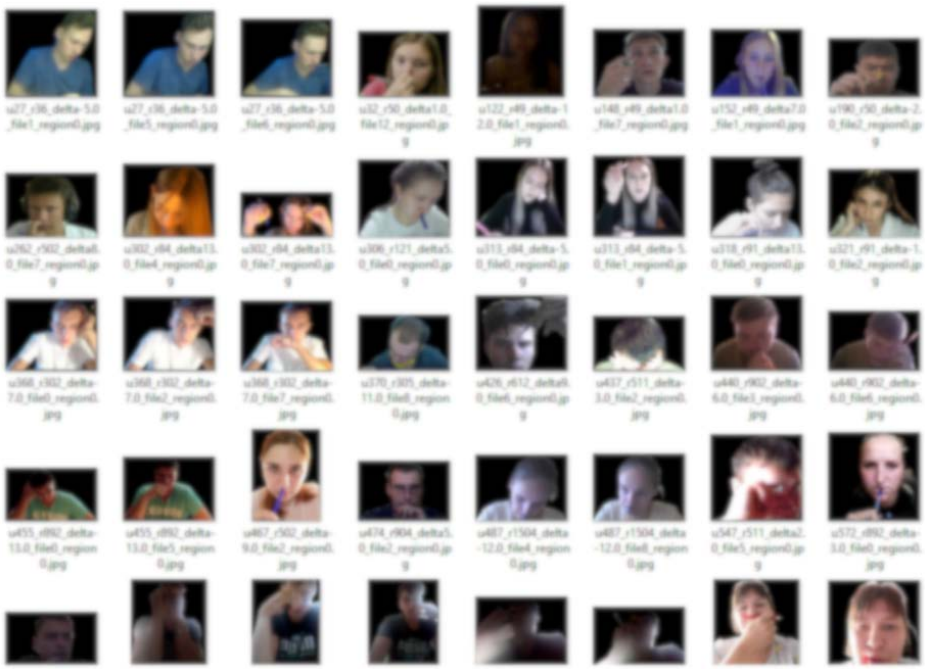


Рис. 4. Фрагмент коллекции изображений. Паттерн поведения поступающего «Держит ручку (карандаш) в руке»

Вычислительные эксперименты по обучению нейросетевых моделей распознаванию различных паттернов поведения поступающих и детектированию нарушений выполнялось в среде Google Colab на графическом процессоре NVidia Tesla K80.

Язык программирования Python 3.7. При предобработке изображений и проведении экспериментов использовались следующие библиотеки: torch 1.11.0+cu113, opencv-python, skimage, imutils, pillow, pandas, matplotlib, numpy, os, shutil, json, zipfile, tqdm.

Для целей распознавания паттернов поведения поступающих использовались различные архитектуры нейронных сетей ResNet, предобученные на более чем 1 млн. изображений из коллекции ImageNet. В качестве метода обучения использовался метод стохастической оптимизации Adam (метод адаптивной оценки момента).

Обучение проводилось по батчам, размер батча составлял 50 изображений. С целью повышения точности прогноза при обучении нейронных сетей применялась батч-нормализация по каналам изображения. Функция потерь (loss function) – кросс-энтропия (cross entropy).

Перед подачей изображений на вход остаточных сверточных нейронных сетей ResNet выполнялись процедуры аугментации и трансформации данных:

- зеркалирование по горизонтали;

- поворот изображения (на угол до 10 градусов);

- изменение размера изображения (на вход нейронных сетей всегда подавались RGB-картинки размером 224 на 224 пикселей);

- нормализация данных (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225].

Результаты экспериментов по обучению и тестированию остаточных сверточных нейронных сетей ResNet распознаванию различных паттернов поведения поступающих показаны в таблицах 1-2.

Таблица 1

Результаты экспериментов по распознаванию паттерна поведения поступающего «Держит ручку (карандаш) в руке». Точность распознавания

Архитектура нейронной сети	20 эпох		50 эпох	
	Обучение	Тест	Обучение	Тест
ResNet-18	0.84	0.69	0.88	0.68
ResNet-34	0.90	0.74	0.90	0.68
ResNet-50	0.87	0.75	0.93	0.72
ResNet-101	0.89	0.76	**0.93**	**0.76**
ResNet-152	0.89	0.72	0.94	0.73

Таблица 2

Результаты экспериментов по распознаванию паттерна поведения поступающего «Рукой держит голову, задумался, грызет пальцы». Точность распознавания

Архитектура нейронной сети	20 эпох		50 эпох	
	Обучение	Тест	Обучение	Тест
ResNet-18	0.87	0.79	0.90	0.78
ResNet-34	0.85	0.74	0.91	0.73
ResNet-50	0.87	0.78	0.91	0.79
ResNet-101	**0.91**	**0.86**	0.94	0.76
ResNet-152	0.88	0.83	0.92	0.75

Наилучшую точность распознавания паттерна поведения поступающих на онлайн-экзамене «Держит ручку (карандаш) в руке» показала нейронная сеть ResNet-101 при 50 эпохах обучения, паттерна «Рукой держит голову, задумался, грызет пальцы» - та же сеть ещё при 20 эпохах обучения. Более простые сверточные нейронные сети оказались не способы достичь такой точности, увеличение же количества эпох обучения приводило лишь к переобучению.

Попробуем ответить на вопрос: возможно ли вообще по изображениям с

камеры оценить самостоятельность прохождения онлайн-теста и предсказать, насколько успешно будет учиться студент в университете в случае зачисления? В качестве критерия успешности предсказания будем считать расхождение в рейтинге студента по математике до и после зачисления (по итогам вступительного вузовского онлайн-тестирования и по итогам экзаменационных сессий в ходе обучения в самом университете).

Для извлечения признаков из изображений вновь были обучены остаточные сверточные нейронные сети ResNet, однако функция потерь здесь находилась как среднее отклонение между ответом нейронной сети, одного выходного нейрона, и тем, что мы ожидали получить (вход: изображение, или матрица чисел; выход: то самое расхождение в рейтинге студента по математике до и после зачисления).

Как показано на рисунке 5, график изменения ошибок затухает после 20-й эпохи, предельная ошибка составила 23.2.

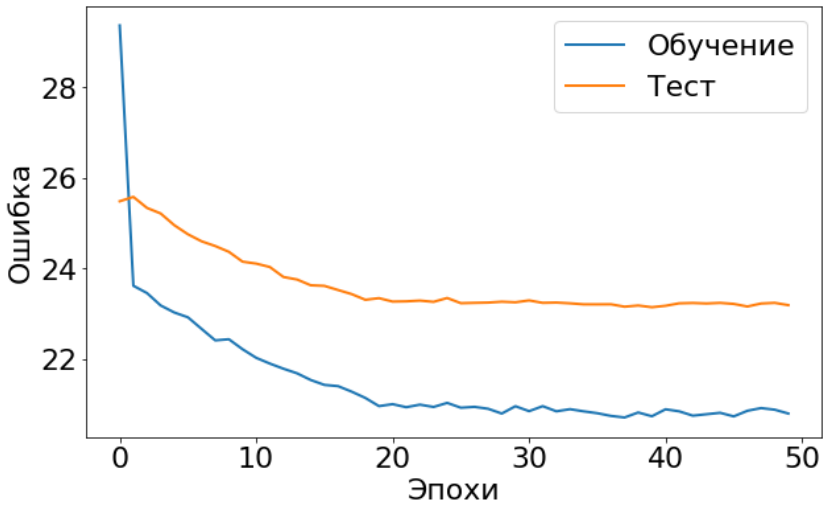


Рис. 5. График изменения ошибок обучения и обобщения на протяжении 50 эпох (на примере предобученной сети ResNet-50) в задаче предсказания по изображениям с камеры поступающих (во время онлайн-экзамена) расхождения в рейтинге по математике до и после зачисления

С учетом средней величины расхождения в рейтинге студентов по математике до и после зачисления (зачисленных в 2020 г. по итогам онлайн-тестирования) 45 баллов, лишь 48.4% такого расхождения (21.8 из 45) можно

объяснить (предсказать) по изображениям с камеры поступающего во время онлайн-экзамена, 51.6% (23.2 из 45) объяснить и предсказать по изображениям невозможно.

Исследование планируется продолжать с целью повышения точности распознавания паттернов поведения поступающих и выявления скрытых зависимостей в данных.

Список литературы

1. Jaap, A.; Dewar, A.; Duncan, C.; Fairhurst, K.; Hope, D.; Kluth, D. Effect of Remote Online Exam Delivery on Student Experience and Performance in Applied Knowledge Tests. BMC Medical Education 2021, 21, 86, doi:10.1186/s12909-021-02521-1.

2. Abdelrahim, D.Y.; Abdelrahim, D.Y. The Effects of COVID-19 Quarantine on Online Exam Cheating: A Test of COVID-19 Theoretical Framework. Journal of Southwest Jiaotong University 2021, 56.

3. Noorbehbahani, F.; Mohammadi, A.; Aminazadeh, M. A Systematic Review of Research on Cheating in Online Exams from 2010 to 2021. Educ Inf Technol (Dordr) 2022, 1–48, doi:10.1007/s10639-022-10927-7.

4. Online Exams and the COVID-19 Pandemic: A Hybrid Modified FMEA, QFD, and k-Means Approach to Enhance Fairness | SpringerLink Available online: https://link.springer.com/article/10.1007/s42452-021-04805-z

5. Lee, K.; Fanguy, M. Online Exam Proctoring Technologies: Educational Innovation or Deterioration? British Journal of Educational Technology 2022, 53, 475–490, doi:10.1111/bjet.13182.

6. Do Online Exams Facilitate Cheating? An Experiment Designed to Separate Possible Cheating from the Effect of the Online Test Taking Environment | SpringerLink Available online: https://link.springer.com/article/10.1007/s10805-014-9207-1

7. Teaching in the Post COVID-19 Era | SpringerLink Available online: https://link.springer.com/book/10.1007/978-3-030-74088-7

8. Приемная Комиссия АлтГТУ Available online: https://pk-test.ru/login.php

9. PyTorch - ResNet Available online: https://pytorch.org/hub/pytorch_vision_resnet/

10. ResNet: остаточная CNN для классификации изображений Available online: https://neurohive.io/ru/vidy-nejrosetej/resnet-34-50-101/

11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2015, doi:10.48550/arXiv.1512.03385.

12. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 2016, doi:10.48550/arXiv.1602.07261.

13. He, F.; Liu, T.; Tao, D. Why ResNet Works? Residuals Generalize. 2019, doi:10.48550/arXiv.1904.01367.

14. Нейронные сети и компьютерное зрение. Сверточные нейронные сети.

Режим доступа: https://stepik.org/course/50352/syllabus

15. Dutta, A.; Zisserman, A. The VIA Annotation Software for Images, Audio and Video. In Proceedings of the Proceedings of the 27th ACM International Conference on Multimedia; ACM: Nice France, October 15 2019; pp. 2276–2279.

СПИСОК ДОКЛАДОВ

СПИСОК АВТОРОВ

Ф.И.О.	Место работы	Стр.
Абрамов Владислав Геннадьевич	660037, Красноярск, Коломенская ул., 26 к.2, ФГБУ «Федеральный Сибирский научно-клинический центр Федерального медико-биологического агентства России» E-mail: excalibr@mail.ru	3
Алехина Анна Евгеньевна	660000, Красноярск, просп. Имени газеты Красноярский Рабочий, 31, Сибирский государственный университет науки и т ехнологий им. М.Ф. Решетнева E-mail: a.tolmacheva@solutionfactory.ru	161
Барцев Сергей Игоревич	660036, Красноярск, ул. Академгородок, строение 50, Институт биофизики СО РАН E-mail: bartsev@yandex.ru	16
Бахшиев Александр Валерьевич	195251, Санкт-Петербург, Политехническая ул., 29, Санкт-Петербургский политехнический университет Петра Великого E-mail: palexab@gmail.com	44
Болсуновский Николай Александрович	662971, Железногорск, ул. Кирова, 12-а, Сибирский федеральный университет, филиал г. Железногорска E-mail: bz990@ya.ru	24
Буряк Никита Андреевич	660041, Красноярск, Свободный пр., 79, корпус 3, Сибирский федеральный университет, Институт математики и фундаментальной информатики E-mail: osv@krasn.ru	34
Володько Ольга Станиславовна	660036, г. Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: osv@krasn.ru	34
Галочкин Иван	630087, Новосибирск, пр-т. К. Маркса, д.30/1, офис	39

Андреевич	238, 241, Группа Компаний «Карбис» E-mail: galochkinia@mer.ci.nsu.ru	
Гунер Михаил Викторович	Сибирский федеральный университет, институт космических и информационных технологий, кафедра систем искусственного интеллекта E-mail: gunermv@gmail.com	172
Дергунов Александр Владимирович	660036, г. Красноярск, ул. Академгородок, 50, ФИЦ КНЦ СО РАН E-mail: alexdergunov@icm.krasn.ru	34
Доррер Михаил Георгиевич	660000, Красноярск, просп. Имени газеты Красноярский Рабочий, 31, Сибирский государственный университет науки и технологий им. М.Ф. Решетнева E-mail: dorrer_mg@sibsau.ru	151, 161
Дубровская Ольга Анатольевна	630090, Новосибирск, пр. Академика Лаврентьева, 6, Федеральный исследовательский центр информационных и вычислительных технологий E-mail: dubrovskaya_oa@list.ru	39
Ерёменко Елизавета Андреевна	195251, Санкт-Петербург, Политехническая ул., 29, Санкт-Петербургский политехнический университет Петра Великого E-mail: elizaveta.yeremenko@gmail.com	44
Забродская Татьяна Евгеньевна	660022, Красноярск, ул. Партизана Железняка, 1, Красноярский государственный медицинский университет. Имени профессора В.Ф.Войно-Ясенецкого E-mail: ng286329@mail.ru	115
Зотин Александр Геннадьевич	660000, Красноярск, просп. Имени газеты Красноярский Рабочий, 31, Сибирский государственный университет науки и технологий им. М.Ф. Решетнева E-mail: zotin@sibsau.ru	54
Корсаков Антон	194064, Санкт-Петербург, Тихорецкий пр., д. 21, Цен-	44

Михайлович	тральный научно-исследовательский и опытно-конструкторский институт робототехники и технической кибернетики (ЦНИИ РТК) E-mail: a.korsakov@rtc.ru	
Кругляков Алексей Сергеевич	660036, г. Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: piggsyy@gmail.com	54
Кудинов Виталий Юрьевич	630090, Новосибирск, Пирогова, 1, Новосибирский государственный университет E-mail: v.kudinov@g.nsu.ru	39
Ларионова Ирина Андреевна	660022, Красноярск, ул. Партизана Железняка, 1, Красноярский государственный медицинский университет имени профессора В.Ф.Войно-Ясенецкого E-mail: vova_lar@mail.ru	3
Маркова Галия Муратовна	660041, Красноярск, Свободный пр., 82А, Сибирский федеральный университет E-mail: irvingstone@bk.ru	16
Медиевский Алексей Владимирович	660022, Красноярск, ул. Партизана Железняка, 1, Красноярский государственный медицинский университет имени профессора В.Ф.Войно-Ясенецкого E-mail: amedievsky@yandex.ru	54
Можаров Даниил Андреевич	630099, г. Новосибирск, ул. Советская 30, ФГБУ «Научно-исследовательский центр космической гидрометеорологии «Планета» Сибирский центр E-mail: mozharov.daniil.a@gmail.com	39
Молявко Анна Андреевна	660041, Красноярск, Свободный пр., 79, Сибирский феде- ральный университет, ИМФИ E-mail: okvaylom@gmail.com	3
Моргун Андрей Васильевич	660022, Красноярск, ул. Партизана Железняка, 1, Красноярский государственный медицинский уни-	3

	верситет имени профессора В.Ф.Войно-Ясенецкого E-mail: 441682@mail.ru	
Мутовина Ольга Александровна	660041, Красноярск, Свободный пр., 82А, Сибирский федеральный университет, Институт фундаментальной биологии и биотехнологии E-mail: mutovina.ole4ka@mail.ru	61
Недорез Яна Владимировна	630090, Новосибирск, ул. Пирогова, 1, Новосибирский государственный университет, E-mail: y.nedorez@g.nsu.ru	76
Ничепорчук Валерий Васильевич	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: valera@icm.krasn.ru	102
Овчинникова Юлия Игоревна	660041, Красноярск, Свободный пр., 82А, Сибирский федеральный университет, ИФБиБТ E-mail: july.l4o6@mail.ru	92
Окунев Алексей Григорьевич	630058, Новосибирск, Русская ул., 35, Высший Колледж Информатики Новосибирского Государственного Университета E-mail: okunev73@mail.ru	39
Пенькова Татьяна Геннадьевна	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: penkova_t@icm.krasn.ru	102
Пермяков Егор Вчеславович	630099, г.Новосибирск, ул.Советская 30, ФГБУ «Научно-исследовательский центр космической гидрометеорологии «Планета» Сибирский центр E-mail: permyakovyegor1204@gmail.com	39
Похабов Дмитрий Владимирович	Федеральное государственное бюджетное учреждение «Федеральный Сибирский научно-клинический центр Федерального медико-биологического агентства» E-mail: neurodmit@mail.ru	3

Пронин Артем Дмитриевич	662971, Железногорск, ул. Кирова, 12-а, Сибирский федеральный университет, филиал г. Железногорска E-mail: artempronin96@list.ru	24
Садовский Михаил Георгиевич	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: msad@icm.krasn.ru	3, 61, 76, 92, 115, 124
Сакович Виталий Валерьевич	660022, Красноярск, ул. Партизана Железняка, 1, Красноярский государственный медицинский университет им. Проф. В.Ф. Войно-Ясенецкого Минздрава России E-mail: sakovichvitaly@gmail.com	115
Симонов Константин Васильевич	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: simonovkv@icm.krasn.ru	54
Тетерлева Агния Алексеевна	660041, Красноярск, Свободный пр., 82А, Сибирский федеральный университет, ИФБиБТ E-mail: tenth_smith@mail.ru	124
Туник Мария Евгеньевна	Красноярский государственный медицинский университет им. Проф. В.Ф. Войно-Ясенецкого Минздрава России E-mail: tsuprikova.mary.maria@yandex.ru	3
Туценко Ксения Олеговна	660022, Красноярск, ул. Партизана Железняка, 1, Красноярский государственный медицинский университет им. проф. В.Ф. Войно-Ясенецкого Минздрава России E-mail: kseniamkib@gmail.com	3
Углев Виктор Александрович	662971, Железногорск, ул. Кирова, 12-а, Сибирский федеральный университет, филиал г. Железногорска E-mail: uglev-v@yandex.ru	24

Усманов Олег Витальевич	ООО Смарт Диджитал E-mail: inbox@hi-tech24.ru	151

Научное издание

Нейроинформатика, её приложения и анализ данных

Материалы XXX Всероссийского семинара 30 сентября – 2 октября 2022 года

Редактор М.Ю. Сенашова
Компьютерная верстка: М.Ю. Сенашова

Подписано в печать «15» августа 2022 г. Формат 60 × 90/8. (А4)
Бумага офсетная. Печать плоская.
Усл. печ. л. 11,0. Уч.-изд. л. 11,3.
Тираж 500 экз. Заказ 12

Отпечатано в типографии ИВМ СО РАН
660036, Красноярск, Академгородок