САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ЭНЕРГЕТИЧЕСКИХ СИСТЕМ

Бузмаков Григорий Александрович

Выпускная квалификационная работа бакалавра

Теоретико-игровое моделирование в генетике

Направление 01.03.02

Прикладная математика и информатика

Научный руководитель,

кандидат физ.-мат. наук,

доцент

Лежнина Е. А.

Санкт-Петербург

2019

Содержание

Введение

Функции белков в клетках живых организмов разнообразны. Белки-ферменты катализируют протекание биохимических реакций и играют важную роль в обмене веществ. Некоторые белки выполняют структурную или механическую функцию, образуя цитоскелет, поддерживающий форму клеток. Также белки играют ключевую роль в сигнальных системах клеток, при иммунном ответе и в клеточном цикле. Начиная с клеточной оболочки и заканчивая всеми составляющими клетки - все построено с участием молекул белка. Чтобы синтезировать белок в клетке необходимы дезоксирибонуклеиновая (ДНК) и рибонуклеиновая кислоты (РНК). ДНК несет в себе всю генетическую информацию, следовательно, в зависимости от заложенных в ней данных и будет строиться тот или иной белок. РНК ведет роль посредника между ДНК и синтезируемым белком. Процесс преобразования наследственной информации от гена в РНК или белок называется экспрессией генов.

Хотя все клетки нашего организма имеют одну и ту же переданную по наследству геномную ДНК, каждая клетка транскрибирует различные гены в виде мРНК в соответствии с типом клетки, биологическими процессами, нормальным или патологическим состоянием и т.д. Данное разнообразие в профилях генной экспрессии интенсивно изучается ввиду его биологического и клинического значения. Еще в конце 1980-х годов была использована технология микрочипов, которая на сегодняшний день позволяет одновременно отслеживать экспрессию десятков тысяч генов, создавая молекулярный портрет клетки. Эти данные могут быть использованы для идентификации генов, отвечающих за конкретное заболевание.

В данной работе были рассмотрены некоторые приложения теории кооперативных игр для анализа данных, полученных с помощью технологии

микрочипов. Однако, такие подходы не дают ответы на нормативные вопросы, как и не дают советы группам генов о том, как они должны вести себя внутри биологической клетки, а используются для описания поведения генов и предсказания исхода их взаимодействия.

Цель работы: изучить применение теории кооперативных игр к вычислению силы генов; применить теорию на практическом примере с реализацией алгоритма.

Постановка задачи

Используя массив, содержащий множество образцов ДНК, имеется возможность за один эксперимент, получить уровни экспрессии десятков тысяч генов в клетке путем измерения количества мРНК, связанной с каждым сайтом в массиве. На данный момент существуют различные экспериментальные платформы, основанные на технологии микрочипов, однако, общая задача данной технологии – последовательная генерация матрицы данных экспрессии генов, в которой строки обозначают гены, а столбцы – образцы исследований, элементами матриц являются числа, представляющие значения экспрессии генов, которые, в свою очередь, количественно определяют уровень экспрессии генов в образцах.

Интерпретация взаимодействия генов в биологических сетях вынуждает значительно ранжировать сетевые элементы. Было предложено несколько подходов для идентификации «центральных» генов среди огромного количества данных, предоставляемых технологией микрочипов. Важным инструментом интерпретации взаимодействия генов являются аналитические методы ранжирования сетевых элементов в сетях ко-экспрессии. Эти методы ранжируют отдельные элементы в соответствии с их важностью в структуре сети.

Также для анализа данных по экспрессии генов может использоваться теория кооперативных игр. В 2007 году Стефано Моретти был введен класс игр с микрочипами для количественной оценки значимости каждого гена в создании или регулировании состояния, представляющего интерес (например, заболевания), с учетом наблюдаемых взаимосвязей во всех подгруппах генов [1].

Целью этой работы является решение проблемы определения относительной значимости генов в процессе зарождения и развития интересующего нас генетического заболевания, на основе данных,

полученных с помощью технологии микрочипов, с учетом уровня взаимодействия между генами.

Сложные экспериментальные процедуры по сбору данных из микрочипов, требуют предварительный анализ данных, такой как оценка качества грубых данных, процедуры нормализации и так далее, для сокращения систематических ошибок, которые возникают в результате нескольких экспериментов.

В последние несколько лет задача анализа микрочипов была направлена на уменьшение систематической погрешности, однако, проблема полного устранения экспериментальной изменчивости практически не решается, а значит, требуется статистическая обработка данных, полученных с помощью микрочипов. Вследствие этого в применяемом в этой работе подходе я ссылаюсь на наблюдаемый средний уровень взаимодействия группы генов, то есть среднее число образцов с болезнью Альцгеймера, такое что группа генов может считаться ответственной за появление болезни, согласно принципу причинности: чем выше число наблюдаемых образцов, тем ниже вероятность влияния случайности на выводы, полученные с помощью данной модели.

Основная идея этой модели основана на теории кооперативных игр. Используется тот же формальный язык для моделирования взаимодействия между игроками, в роли которых выступают гены, при определенном интересующем нас сценарии, таком как патогенез генетического заболевания, как в теории кооперативных игр.

Кооперативная игра, которую мы рассматриваем, основана на сравнении двух матриц данных экспрессии генов: одна матрица содержит образцы ДНК с заболеванием, представляющим интерес, другая - образцы ДНК без данного заболевания. В начале работы используется дискриминантный метод в каждом образце, разделяющий весь набор генов

на два: на аномально выраженные по отношению к нормальным образцам ДНК гены, и гены уровни экспрессии, которых соответствуют нормальным образцам ДНК. Затем вводится отношение причинности (также называемое принципом достаточности), непосредственно определяющее характеристическую функцию игры. В качестве биологического значения индекса значимости, используемого для определения «силы» генов, будем рассматривать такие решения кооперативной теории игр, как вектор Шепли и индекс Банзафа.

Обзор литературы

Биологические микрочипы (биочипы, ДНК-микрочипы) являются одним из новейших инструментов современной биологии и медицины. Прообразом современных микрочипов стал саузерн-блоттинг, изготовленный в 1975 г. Эдом Саузерном. Он использовал меченую нуклеиновую кислоту для определения специфической последовательности среди фрагментов ДНК, зафиксированных на твердой подложке.

История развития технологии микрочипов относительно коротка. В конце 1980-х годов началось обсуждение глобального проекта «Геном человека», и почти одновременно в нескольких лабораториях появилась идея секвенирования, то есть расшифровки последовательного расположения нуклеотидных оснований в геномах живых систем) с помощью гибридизации (СПГ). Технологических вариантов реализации этой идеи было несколько, В 1990 г. был опубликован патент югославских исследователей Р. Дрманач и Р. Чрквеняков [2], в котором они предложили иммобилизовать на двумерной поверхности анализируемую ДНК и проводить ряд последовательных гибридизаций с комплементарными олигонуклеотидными зондами (короткими отрезками ДНК или РНК, меченными радиоактивным или флуоресцентным соединением). Независимо от них российский ученый А. Д. Мирзабеков в 1988 году в работе [3] предложил иную модель гибридизации. Его идеей было определять последовательность ДНК с помощью обратного подхода, а именно иммобилизации набора коротких олигонуклеотидов, представляющих собой все возможные последовательности определенной длины, с которыми предполагалось гибридизовать анализируемый фрагмент меченой ДНК с неизвестной последовательностью. Реконструкция последовательности анализируемой ДНК должна была производиться с помощью математических методов анализа. В последующем данная технологическая модель микрочипа стала преобладающей в

конструировании биочипа. Более подробную информацию о развитии моделей микрочипа можно найти в [4].

Д. Амаратунга и Х. Кабрера в книге [5] описали процесс проведения различных видов экспериментов с биочипами, способы обработки полученных изображений и методы подсчета уровня экспрессии генов.

Калифорнийские ученые П. Балди и В. Хэтвилд в своей работе [6] предложили уже более прикладную информацию по статистической обработке получаемых в процессе экспериментов результатов. Авторы предложили использовать различные методы кластеризации, факторный анализ, деревья решений, байесовские и нейронные сети для отделения влияния генов друг на друга в показаниях микрочипов ДНК.

В литературе представлено большинство моделей для анализа данных: вывода из матрицы данных экспрессии генов, роли генов, их взаимодействий и их поведения при изменении состояния биологической системы. Одной из таких работ является книга [7]. Методы, представленные в книге, затрагивают все аспекты статистики анализа микрочипов, от аннотаций и фильтрации до кластеризации и классификации. Альтернативный метод анализа экспрессии генов на основе кооперативных игр был предложен группой итальянских ученых С. Моретти, Ф. Патроне, С. Бонасси в статье [1]. Основным преимуществом этого подхода является возможность вычислить числовой индекс, называемый индексом значимости, который показывает значимость каждого гена при интересующем условии (например, генетическом заболевании) с учетом поведения других генов. В своей работе авторы используют вектор Шепли для ранжирования генов по значимости и дают аксиоматическую характеристику вектора Шепли на классе так называемых игр микрочипов. Стефано Моретти опубликовал множество работ, описывающих подходы, использующие теорию игр для анализа экспрессии генов. Так в статье [8] для представления взаимодействий между всеми возможными парами генов им был внедрен метод, основанный на

задаче минимальных стоимостных остовных деревьев (MCST) и расширен для реализации понятия ассоциативных коалиций генов. В работе [9] он рассмотрел игры с генами в роли игроков для анализа способности групп генов классифицировать образцы в правильные классы (например, класс нормальных тканей или класс опухолевых тканей). Также С. Моретти с группой ученых представил биологическую оправданность использования вектора Шепли игр микрочипов в качестве индекса значимости генов в [10], где набор генов, участвующих в патогенезе нейробластических опухолей, был выбран в соответствии с вектором Шепли игры микрочипов. В работе [11] авторы рассмотрели другое важное одноточечное решение в теории игр, индекс Банзафа, а именно предложил множество свойств, характеризующих его, и доказал правомерность его использования для вычисления индекса значимости.

Глава 1. Математическая модель

1.1. Определения

Игра – математическая модель взаимодействия нескольких участников.

Пусть N – конечное множество игроков.

Коалицией $S \subset N$ называется непустое подмножество множества игроков N.

Кооперативной игрой называется пара (N, v), где v – характеристическая функция, ставящая в соответствие каждой коалиции S некоторое вещественное число $v(S)$, называемое выигрышем коалиции и выражаемое силу коалиции. Предполагается, что $v(\emptyset) = 0$.

Решением игры является отображение игры в соответствующее множество допустимых исходов, понимаемого как множество возможных распределений суммарных выигрышей игроков, т.е. множеств

$$\{x \in \mathbb{R}^N : \sum_{i \in N} x_i \leq v(N)\}.$$

Исходом кооперативной игры является вектор из множества

$$X(N, v) = \{ x \in \mathbb{R}^N \mid \sum_{i \in N} x_i \leq v(N)\}$$

Решением кооперативной игры называется подмножество

$\sigma(N, v) \subset X(N, v)$.

Вектор Шепли – это математическое ожидание вклада каждого игрока, при предположении, что коалиции S мощности s (при $0 \leq s \leq n - 1$) возникают с одинаковыми вероятностями, и что все коалиции одной и той же мощности s также равновероятны.

$$\varphi_i(v) = \sum_{S \subseteq N:\, i \in S} \frac{(s-1)!\,(n-s)!}{n!}\big(v(S) - v(S \backslash i)\big),$$

где n – мощность множества $|N|$.

Индекс Банзафа – это математическое ожидание вклада каждого игрока, при предположении, что все коалиции S возникают с одинаковыми вероятностями.

$$\beta_i(v) = \sum_{S \subseteq N:\, i \in S} \frac{1}{2^{n-1}} \big(v(S) - v(S \backslash i)\big),$$

где n – мощность множества $|N|$.

1.2. Формализация кооперативной игры в контексте микрочиповой игры

Данная модель была предложена Стефано Моретти. Ее идея основана на теории кооперативных игр.

Пусть $N = \{1, \dots, n\}$ множество из n генов, $S_R = \{1, \dots, r\}$ множество образцов клеток из здоровых тканей, и $S_D = \{1, \dots, d\}$ множество образцов из тканей с генетическим заболеванием.

Задача экспериментов с микрочипами состоит в сопоставлении каждому образцу $j \in S_D \cup S_R$ профиль экспрессии $A(j) = (A_{ij})_{i \in N}$, где $A_{ij} \in \mathbb{R}$ величина экспрессии гена i в образце j. Глобально эти значения экспрессии будут называться набором данных эксперимента микрочипов. Далее будет использоваться набор данных, который предварительно обработали методом, обычно называемым нормализацией, позволяющим сравнивать интенсивность экспрессии генов из разных образцов. Нормализация необходима для корректировки любого смещения, возникающего в результате изменения технологий микрочипов.

Набор данных может быть представлен в форме двух матриц экспрессии генов $A^{S_R} = (A_{ij}^{S_R})_{i \in N, j \in S_R}$ и $A^{S_D} = (A_{ij}^{S_D})_{i \in N, j \in S_D}$, где j - это столбец, являющийся профилем экспрессии образца j. Обозначим через $E = \langle N, S_R, S_D, A^{S_R}, A^{S_D} \rangle$ микрочиповую экспериментальную ситуацию (МЭС).

На первом этапе анализа необходимо определить, экспрессия каких генов больных клеток выражается аномально по отношению к матрице

здоровых, т.е. сильнее или слабее нормы, в соответствии с определенным методом дискриминации. Гену i, экспрессия которого в образце проявляется вне нормы, соответствует значение булевой переменной $B_{ij} \in \{0,1\}$ равное 1, гену с нормальной экспрессией $-$ значение 0. Булевым профилем экспрессии будем называть вектор $B^j = (B_{ij})_{i \in N}$.

Дискриминантный метод может быть выражен как отображение m, присваивающее каждому профилю экспрессии из больных образцов соответствующий аномальный профиль экспрессии. Значит, вся информация об отличиях в экспрессии генов в образцах из S_D от образцов из S_R может быть представлена через булевую матрицу экспрессии $B^{E,m} \in \{0,1\}^{N \times S_D}$ с применением определенного дискриминантного метода к МЭС. Далее, будем отождествлять МЭС E и дискриминантный метод m с матрицей $B^{E,m}$.

Существует множество различных дискриминантных методов. Один из них – наивный метод для двух классов: 0 и 1, где 1 обозначает аномально, а 0 - нормально выраженный ген:

$$(\widehat{m}(A^{S_D}, A^{S_R}))_{ij} = \begin{cases} 1, \text{если } A_{ij}^{S_D} \geq max_{h \in S_R} A_{ih}^{S_R} \\ \quad \text{или } A_{ij}^{S_D} \leq max_{h \in S_R} A_{ih}^{S_R} \\ 0, \text{иначе} \end{cases}$$

Другой метод более консервативный. Для каждого $i \in N$ и каждого $j \in S_R$ выполняется

$$(\widehat{m}(A^{S_D}, A^{S_R}))_{ij} = \begin{cases} 1, \text{если } A_{ij}^{S_D} \leq p_i^{25\%} \text{ или } A_{ij}^{S_D} \geq p_i^{75\%} \\ 0, \text{иначе} \end{cases}$$

где $p_i^{25\%}$ и $p_i^{75\%}$ являются соответственно 25-ым и 75-ым процентилями распределения экспрессии гена i в соответствующей матрице экспрессии A^{S_R} для каждого $i \in N$.

Введем понятие опоры вектора. Опора $W \in \{0,1\}^N$ – это множество $sp(W) = \{i \in \{1, \dots, n\} \mid W_i = 1\}$.

Рассмотрим МЭС $E = \langle N, S_R, S_D, A^{S_R}, A^{S_D} \rangle$ и дискриминантный метод m. Определим микрочиповую игру как кооперативную игру (N, v), где

- Конечное множество генов N будет конечным множеством игроков,
- v – характеристическая функция, такая что $v(\emptyset) = 0$, присваивающая каждой коалиции $T \in 2^N \setminus \{\emptyset\}$ среднее значение количества образцов интересующей нас болезнью, определяемое по T в соответствии с принципом достаточности для групп генов.

Более точно характеристическая функция будет вычисляться по формуле

$$v(T) = \frac{|\Theta(T)|}{|S_D|},$$

где $|S_D|$ – мощность множества образцов больных тканей, а $|\theta(T)|$ – мощность множества $\Theta(T) = \left\{k \in S_D \middle| sp\left(B^{E,m}(k)\right) \subseteq T, sp\left(B^{E,m}(k)\right) \neq \emptyset \right\}$.

Условие $sp\left(B^{E,m}(k)\right) \neq \emptyset$ обусловлено практическими соображениями относительно интерпретации принципа достаточности для групп генов на образцах, где гены не проявляют каких-либо аномальных свойств экспрессии. Предполагается, что такие образцы способствуют уменьшению уровня ассоциации больных образцов с нужной болезнью.

Класс микрочиповых игр обозначим символом $\mathcal{M}$.

Пусть $\mathcal{G}$ – класс всех кооперативных игр и $C \subseteq \mathcal{G}$ подкласс всех кооперативных игр. Тогда будем говорить, что рассматривая множество игроков N мы определяем класс $C^N \subseteq \mathcal{G}$ как класс кооперативных игр на C с множеством игроков N.

Характеристическая функция кооперативной игры (N, v) супераддитивна, если для всех коалиций $S, T \subseteq N,\ S \cap T = \emptyset$ выполняется

$$v(S) + v(T) \leq v(S \cup T).$$

Характеристическая функция кооперативной игры (N, v) монотонна, если для всех $S \subseteq T \subseteq N$ выполняется

$$v(S) \leq v(T).$$

Введем понятие личного вклада каждого гена в образование генотипа, определяемого формулой

$$m_i(v,S) = v(S) - v(S \setminus \{i\}),$$

где $S \subseteq N,\ S \neq \emptyset$ и $i \in S$.

Характеристическая функция кооперативной игры (N,v) является выпуклой, если личный вклад любого игрока в какую-либо коалицию не превышает его личный вклад в большую коалицию, т. е.

$$m_i(v,S) \leq m_i(v,T)$$

для всех $S \subseteq T \subseteq N$ и всех $i \in S$.

Нетрудно проверить, что выпуклость подразумевает супераддитивность, но не наоборот.

Предложение 1. Пусть даны МЭС $E = \langle N, S_R, S_D, A^{S_R}, A^{S_D} \rangle$ и дискриминантный метод m и пусть (N,v) – соответствующая микрочиповая игра в $\mathcal{M}^N$. Тогда характеристическая функция v является монотонной и выпуклой.

Доказательство приведено в [1].

Пусть $|N|$ мощность конечного множества N. Вектором выигрыша (распределения) $(x_1, \dots, x_n)$ кооперативной игры (N,v) будем называть $|N|$-размерный вектор, описывающий выигрыш игроков, такой что каждый игрок $i \in N$ получает x_i. Решением для класса кооперативных игр C будем называть функцию ψ определяющую вектор выигрыша $\psi: C^N \to \mathbb{R}^N$. В контексте микрочиповых игр, решение рассматривается как вектор ранжирования силы генов, то есть ген получивший наибольший «выигрыш» является наисильнейшим в данной выборке и так далее.

Одним из популярных решений кооперативной игры является вектор Шепли. Введем вектор Шепли на микрочиповую игру (N,v):

$$\varphi_i(v) = \sum_{S \subseteq N:\, i \in S} \frac{(s-1)!\,(n-s)!}{n!} m_i(v,S)\,,$$

где $i \in N, s = |S|$ и $n = |N|$ – мощности коалиций.

Введем понятие простой игры. Простой игрой будем называть игру (N,u_R) на $R \subseteq N$, где

$$u_R(T) = \begin{cases} 1, \text{если } R \subseteq T, \\ 0, \text{иначе.} \end{cases}$$

Любая кооперативная игра (N, v) может быть записана в виде линейной комбинации единогласных игр:

$$v = \sum_{S \subseteq N, S \neq \emptyset} \lambda_S(v)\, u_S,$$

где $\lambda_S = \frac{\bar{\lambda}_S}{|S_D|}$ – коэффициент единогласия, а $\bar{\lambda}_S = \left|\left\{k \in S_D \,\middle|\, sp\left(B^{E,m}(k)\right) = S\right\}\right|$ – количество возникновений коалиций S в качестве опоры булевой матрицы экспрессии $B^{E,m}$.

Через коэффициенты единогласия $\left(\lambda_S(v)\right)_{S \in 2^N \setminus \{\emptyset\}}$ игры (N, v) может быть дано модифицированное представление вектора Шепли:

$$\varphi_i(v) = \sum_{S \subseteq N:\, i \in S} \frac{\lambda_S(v)}{|S|} \qquad (1)$$

для каждого $i \in N$.

Другим одноточечным решением кооперативной игры является индекс Банзафа:

$$\beta_i(v) = \sum_{S \subseteq N:\, i \in S} \frac{1}{2^{n-1}} m_i(v, S)$$

для каждого $i \in N$.

Альтернативная формула индекса Банзафа:

$$\beta_i(v) = \sum_{S \subseteq N:\, i \in S} \frac{\lambda_S(v)}{2^{|S|-1}} \qquad (2)$$

для каждого $i \in N$.

Глава 2. Аксиоматическая характеристика решений кооперативной теории игр в применении к микрочиповым играм

2.1. Генные регуляторные сети как партнерство генов

Генные регуляторные сети (ГРС) представляют собой набор молекулярных регуляторов, взаимодействующих друг с другом и с другими веществами в клетке, чтобы управлять экспрессией генов. В качестве узлов сети можно рассматривать гены, причем входными данными являются белки, такие как факторы транскрипции, а выходными данными является уровень экспрессии генов. Значение узла зависит от функции, которая, в свою очередь, зависит от значения его регуляторов на предыдущих временных шагах. В эти функции входит выполнение некоторой обработки информации внутри клетки, которая определяет клеточное поведение. Исследование механизмов работы генных регуляторных сетей является важным шагом для понимания генетического контроля развития отдельных органов или их систем. Также, понимание этих механизмов предоставляет широкие возможности для направленного изменения транскрипционной активности конкретных генов, что помогает в лечении заболеваний различной этиологии. Более подробно ознакомиться с механизмами регуляции генов и различными биохимическими процессами можно в [12].

В понимании механизма работы ГРС большую сложность представляет большое число генов, вовлеченных в изучение микрочипов, которое может достигать десятков и даже сотен тысяч генов. Уменьшение этого количества генов достигается за счёт фильтрации «шумных», незначимых и излишних генов. Экспрессия «шумных» генов зависит от помех в измерении, появляющихся вследствие вариации экспериментов. Незначимые гены - это гены, одинаково выраженные как в образцах тканей с болезнью, так и в

здоровых. Излишние гены коррелируются с другими, следовательно, регулируются ими. В данной работе предполагается, что такие решения кооперативных игр как вектор Шепли и индекс Банзафа, могут быть использованы для определения биологически значимых генов. Докажем данное утверждение для вектора Шепли, используя аксиоматический подход, такой что решение микрочиповой игры характеризуется с использованием базовых свойств, определяемых тем, как индекс значимости генов должен вести себя в различных простых ситуациях взаимодействия генов.

На первом шаге требуется ввести значение ГРС в контексте микрочиповых игр, используя терминологию теории игр. В данном случае ключевую роль играет определение партнерства генов.

Пусть $(N, v) \in \mathcal{M}^N$. Коалиция $S \in 2^N \backslash \{\emptyset\}$ такая что для каждой $T \subsetneq S$ и каждой $R \subseteq N \backslash S$

$$v(R \cup T) = v(R)$$

называется партнерством генов в микрочиповой игре (N, v).

Существует, как минимум, две причины, по которым имеется возможность представить партнерство генов в качестве ГРС в контексте микрочиповых игр. Во-первых, оно не требует никакой априорной информации о соответствующих регуляторных механизмах среди генов внутри сети. Вследствие высокой сложности ГРС, этот тип информации все еще не доступен для многих генов. Во-вторых, определение партнерства требует невозможности выявления определенной подгруппы генов, непосредственно взаимодействующих с внешним геном или группой генов в провоцировании генетической болезни. Это необходимое условие для группы генов при составлении ГРС, рассматриваемой как уникальная сеть генов с определенным уровнем экспрессии. Однако, существует вероятность, что совместная сила коалиции, созданной двумя непересекающимися партнерствами, будет больше, чем просто сумма их единичных значений, с учетом возможности взаимодействия отдельных сигнальных путей внутри клетки.

2.2. Индекс значимости генов как решение кооперативной игры

Индексом значимости будем называть решение $F: \mathcal{M}^N \to \mathbb{R}^N$ в классе микрочиповых игр с набором генов N в качестве множества игроков. Далее приведены некоторые свойства индекса значимости, связанные с концепцией партнерства генов.

Пусть дано конечное множество генов $N, v \in \mathcal{M}^N$, коалиции $S, T \in 2^N \backslash \{\emptyset\}$ являются партнерствами генов в игре (N, v).

Свойство 1. Решение F удовлетворяет свойству партнерской рациональности (ПР), если

$$\sum_{i \in S} F_i(v) \geq v(S)$$

для каждой коалиции $S \in 2^N \backslash \{\emptyset\}$.

Свойство 2. Решение F удовлетворяет свойству осуществимости партнерства (ОП), если

$$\sum_{i \in S} F_i(v) \leq v(N)$$

для каждой коалиции $S \in 2^N \backslash \{\emptyset\}$.

Данные два свойства определяют соответственно нижнюю и верхнюю границы силы партнерства, т.е. общая значимость партнерства генов в обнаружении патогенеза болезни отдельно не может быть меньше, чем среднее количество случаев болезни вызванных партнерством, и больше, чем среднее количество случаев болезни вызванных самой большой из возможных коалиций. Эти свойства определяют неотрицательную меру для вычисления значимости генов, провоцирующих заболевания, присваивая значение 1 партнерству генов, которое в соответствии с принципом достаточности, ответственно за появление болезни во всех больных образцах.

Критерием, с помощью которого возможно сравнение значимости различных партнерств генов, является их значение в микрочиповой игре,

однако, к рассматриваемым генам могут быть также присвоены некоторые другие значения в соответствии с их ролью во всех возможных коалициях.

Свойство 3. Решение F удовлетворяет свойству монотонности партнерства (МП), если

$$F_i(v) \geq F_j(v)$$

для каждого $i \in S$ и каждого $j \in T$, где коалиции $S, T \in 2^N \backslash \{\emptyset\}$ такие что: $S \cap T = \emptyset, v(S) = v(T), v(S \cup T) = v(N), |S| \leq |T|$.

Пусть даны два непересекающихся партнерства генов, которые вызывают одинаковое среднее количество случаев заболевания на заданном множестве образцов. И пусть гены вне объединения этих двух партнерств незначимы, тогда гены в партнерстве, имеющим меньшее количество генов, должны иметь больший индекс значимости, чем гены из другого партнерства, в котором вероятность существования излишних генов больше.

Свойство 4. Пусть $v_1, \dots, v_k \in \mathcal{M}^N$, $k > 1$. Решение F удовлетворяет свойству равного разделения (РР), если

$$F\left(\frac{\sum_{i=1}^{k} v_i}{k}\right) = \frac{\sum_{i=1}^{k} F(v_i)}{k}.$$

Заметим, что $\frac{\sum_{i=1}^{k} v_i}{k} \in \mathcal{M}^N$. Доказательство этого приведено в статье [1]. Данное свойство утверждает, что средний индекс значимости генов в двух или более различных микрочиповых играх $v_1, \dots, v_k \in \mathcal{M}^N$ с одинаковым набором генов, возникающих из различных МЭС, проводимых в разных лабораториях, должен быть равен индексу значимости генов в усредненной игре $\frac{\sum_{i=1}^{k} v_i}{k}$.

Нулевым геном игры (N, v) будем называть ген $i \in N$ такой что

$$v(S \cup i) = v(S)$$

для каждой коалиции $S \subseteq N \backslash \{i\}$.

Свойство 5. Решение F удовлетворяет свойству нулевого гена (НГ), если для каждого нулевого гена $i \in N$

$$F_i(v) = 0.$$

То есть, если вклад гена в каждую из коалиций $S \in 2^N$ нулевой, то этот ген имеет нулевое значение.

Теорема 1. Вектор Шепли на классе простых микрочиповых игр $\mathcal{M}^N$ является единственным индексом значимости, удовлетворяющим свойствам ПР, ОП, МП, РР и НГ.

Доказательство данной теоремы приведено в статье [1].

Введем еще несколько свойств.

Свойство 6. Решение F удовлетворяет свойству симметрии (С) на $\mathcal{M}^N$, если для каждого $i, k \in S$,

$$F_i(v) = F_k(v).$$

Свойство 7. Решение F удовлетворяет свойству индивидуальной согласованности (ИС), если

$$F_i\left(u_{\{i\}}\right) = 1$$

для каждого $i \in N$.

Свойство 8. Пусть $v = (v^1, \dots, v^m) \in \mathcal{M}^N$, пусть $l \in \{1, \dots, n\}$. Определим новую микрочиповую игру v_{Sl} следующим образом:

1. Для j, такого что $v^j(S) = 1$,

$$v_{Sl}^j(T) = \begin{cases} 1, T \supseteq S \cup \{l\}, \\ 0, \text{иначе}. \end{cases}$$

2. В противном случае, $v_{Sl}^j = v^j$.

Тогда решение F удовлетворяет свойству средних потерь (СП) на $\mathcal{M}^N$, если для каждого v, v_{Sl}, как указано выше,

$$\frac{1}{S}\sum_{i \in S}[F_i(v) - F_i(v_{Sl})] = F_i(v_{Sl}) - F_i(v).$$

С другой стороны, F удовлетворяет свойству полной потери (ПП), если

$$\sum_{i \in S}[F_i(v) - F_i(v_{Sl})] = F_l(v_{Sl}) - F_l(v).$$

Заметим, что свойства СП и ПП касаются влияния добавления гена к партнерству в микрочиповой игре v. Интерпретируем аналогичные аксиомы, которые представлены в [13]. Постоянная полная (соответственно средняя) потеря здесь постулирует, что полная (соответственно средняя) потеря генов в партнерстве S равна полному (соответственно среднему) приросту гена l добавленного к S.

Рассмотрим игру $v \in \mathcal{M}^N$ и пусть M - порождающая матрица. Пусть l будет нулевым геном в v, а ген $k \neq l$. Рассмотрим матрицу M^{lk}, определенную с помощью ее строк:

$$m_{i\cdot}^{lk} = m_{i\cdot}, \text{если } i \neq l, m_{l\cdot}^{lk} = m_{k\cdot}.$$

Игра v_{lk} называется игрой, связанной с матрицей M^{lk}.

Свойство 9. Решение F удовлетворяет свойству парной согласованности (ПС) на $\mathcal{M}^N$, если для каждых v, v_{lk}, как указано выше,

$$F_k(v) = F_i(v_{lk}) + F_k(v_{lk}).$$

То есть, в новой игре нулевой ген l удаляется, и влияние ненулевого гена k «удваивается». Если в предыдущей игре v ген l не имел влияния, то теперь сила гена k распределяется между генами k и l. Следовательно, замена его другим игроком не повлияет на общую сумму сил генов. С другой стороны, чтобы сила отдельных генов, играющих роль в двух играх не менялась, сила гена k, которая была в предыдущей игре, должна быть разделена между k и l.

Теорема 2. Существует один и только один индекс: $\varphi : \mathcal{M}^N \to \mathbb{R}^N$, удовлетворяющий свойствам НГ, С, РР, ИС, ПС. Это индекс Банзафа.

Доказательство теоремы представлено в [11]

Тем самым, можно утверждать, что кооперативная теория игр, включая вектор Шепли и индекс Банзафа, могут применятся в контексте микрочиповых игр.

Глава 3. Результаты

Рассмотрим применение теории кооперативных игр для анализа следующих данных экспрессии генов. В 2014 году группа ученых из Японии собрала данные микрочипов из тканей лобной и височной коры головного мозга, а также гиппокампа у пациентов с наличием болезни Альцгеймера или другой сосудистой деменции. Данные по экспрессии генов можно посмотреть в [14]. Эти данные представлены в трех видах: в виде тепловой карты, где значения экспрессии генов выражены различными оттенками зеленого и розового: ярко-зеленый цвет соответствует низкому уровню экспрессии гена, ярко-розовый – высокому уровню; в виде графиков и в виде текстового файла с числовыми значениями экспрессии. Будем использовать третий вариант представления данных по экспрессии генов, текстовый файл. Поставим задачу ранжирования генов по их силе при болезни Альцгеймера по отношению к другой деменции с помощью вектора Шепли и индекса Банзафа.

Перед применением метода, основанного на теории кооперативных игр, необходимо представить данные в виде матрицы, где по строкам расположены гены, а по столбцам – образцы исследований. Всего мы имеем 79 образцов, из которых 32 – с болезнью Альцгеймера. Также в файле содержится 5209 строк, из которых мы используем только те 4033 строки, в которых содержится уникальная информация о генах. Для построения булевой матрицы экспрессии используем наивный дискриминантный метод. Код программы по работе с данными написаны на языке Python 3 (Прил. 1).

Основной алгоритм следующий: на вход подается матрица данных экспрессии генов, далее с помощью дискриминантного метода строится булевая матрица экспрессии, причем отдельно рассматриваются данные, взятые из разных частей головного мозга: лобной коры, височной коры и гиппокампа. Затем выявляются опоры столбцов, рассматриваются только те

коалиции, которые содержат полностью хотя бы одну из этих опор, считаются для них характеристические функции и строится решение в виде вектора Шепли и индекса Банзафа. Выводом программы является файл, содержащий гены в порядке убывания их индекса значимости.

Рассмотрим МЭС $E = \langle N, S_R, S_D, A^{S_R}, A^{S_D} \rangle$. За образцы с интересующим нас сценарием взяты образцы тканей с болезнью Альцгеймера, за «здоровые» – образцы с другой деменцией. Мощность множества генов $|N| = 4033$, мощности множеств образцов с заболеванием Альцгеймера и со «здоровыми» тканями, соответственно, равны $|S_D| = 32$ и $|S_R| = 47$.

Рассмотрим первые 10 генов по убыванию индекса значимости, найденного с помощью вектора Шепли (табл. 1) и индекса Банзафа (табл. 2):

Ген	**Значение Шепли** ($* \mathbf{10^3}$)
RYR1	*1,42125*
HYDIN2	*1,3214*
FAHD2B	*1,31450*
GALNT15	*1,27645*
ECHDC2	*1,17365*
PLEC	*1,16275*
PFDN5	*1,14733*
ZC3H11A	*1,13735*
NDUFS6	*1,1145*
WDR49	*1,08939*

Таблица 1. Первые 10 генов по индексу значимости, посчитанному с помощью вектора Шепли

Ген	**Индекс Банзафа** ($* \mathbf{10^{56}}$)
METTL3	*2.04115*
NDUFS6	*2.04115*
AK130448	*2.04115*
CCDC88C	*2.04115*
TRPV1	*2.04115*
DNHD1	*2.04115*
HYDIN2	*2.04115*
KIF5B	*2.04115*
THBS2	*2.04115*
ZNF621	*2.04115*

Таблица 2. Первые 10 генов по индексу значимости, посчитанному с помощью индекса Банзафа

Глобально данная разница в значениях индекса значимости, посчитанных разными способами, представлена на графике (рис. 1). Под рангом гена здесь понимается номер, который имеет этот ген при упорядочивании всех генов по убыванию величины индекса значимости, вычисленного с помощью вектора Шепли.

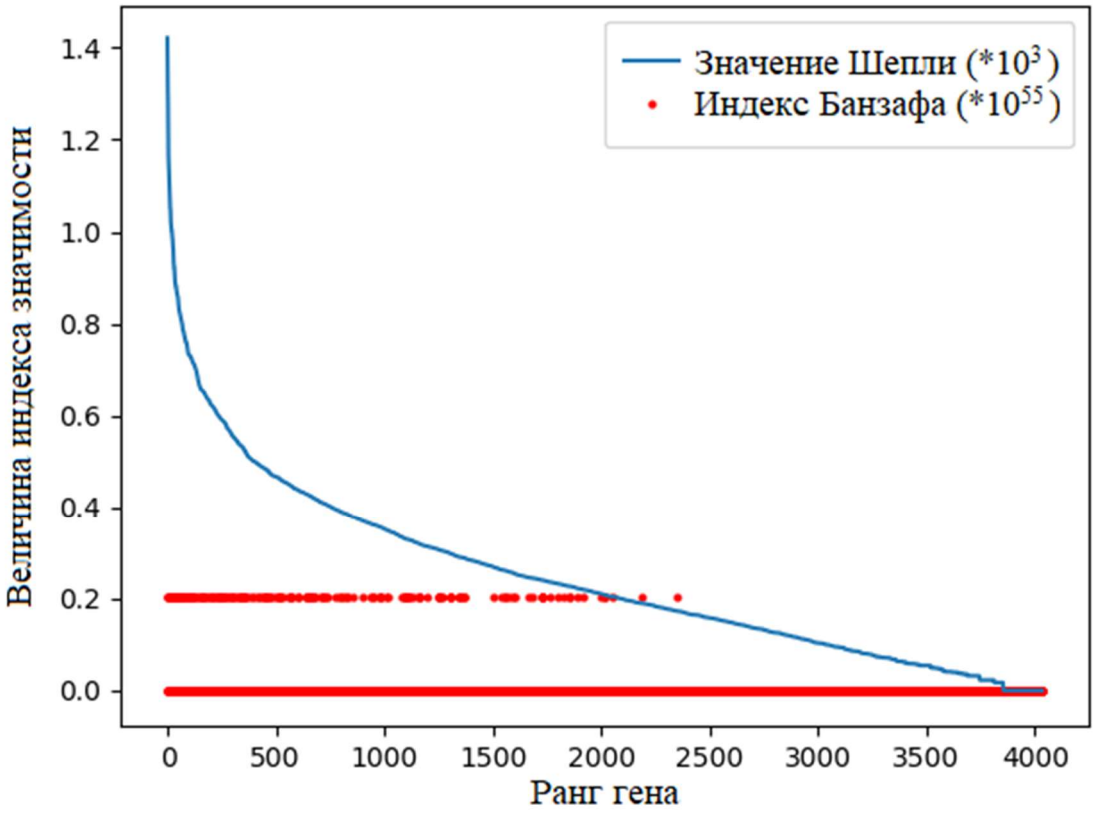

Рисунок 1. График изменения индекса значимости с уменьшением ранга гена

Причиной разницы в результатах может являться величина, стоящая в знаменателе формулы (2), по которой вычисляется индекс Банзафа. Так, если при вычислении индекса значимости гена с помощью вектора Шепли (1) делим коэффициент единогласия на $|S|$, т. е. на мощность множества генов, входящих в коалицию, то при использовании индекса Банзафа делим на $2^{|S|-1}$, из-за чего появляются трудности при относительно большом количестве генов.

Рассмотрим еще три МЭС с тем же множеством образцов тканей, но разным количеством генов. Для этого возьмем группы генов, равные по мощности, с наибольшими и наименьшими рангами, а также гены из середины списка.

На графиках представлены ситуации при анализе 100 (рис. 2), 30 (рис. 3) и 15 (рис. 4) генов.

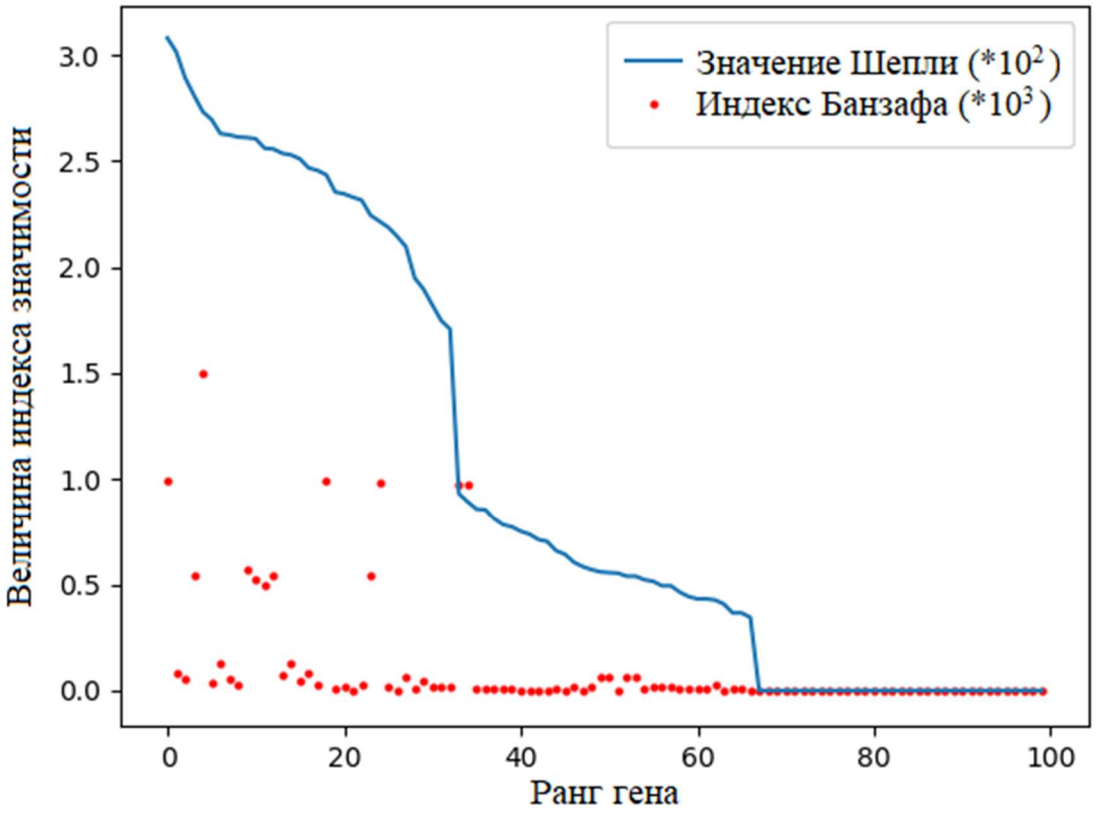

Рисунок 2. График изменения индекса значимости с уменьшением ранга гена для 100 генов

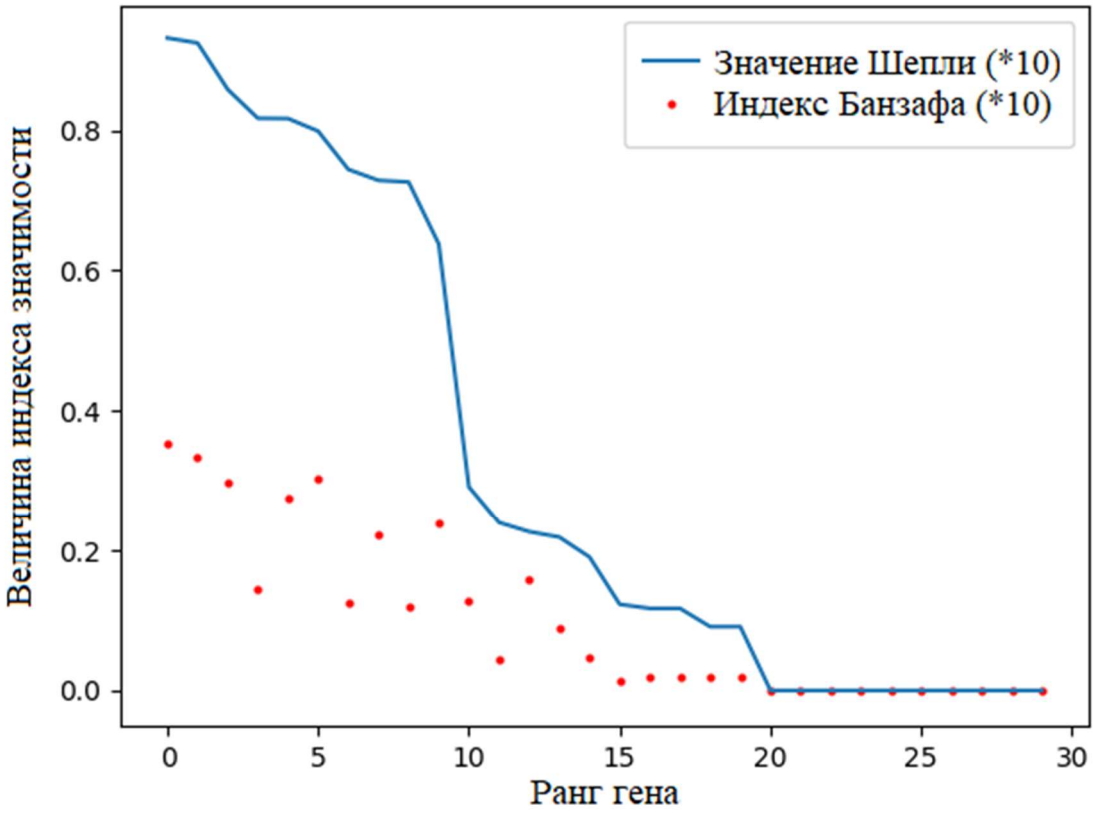

Рисунок 3. График изменения индекса значимости с уменьшением ранга гена для 30 генов

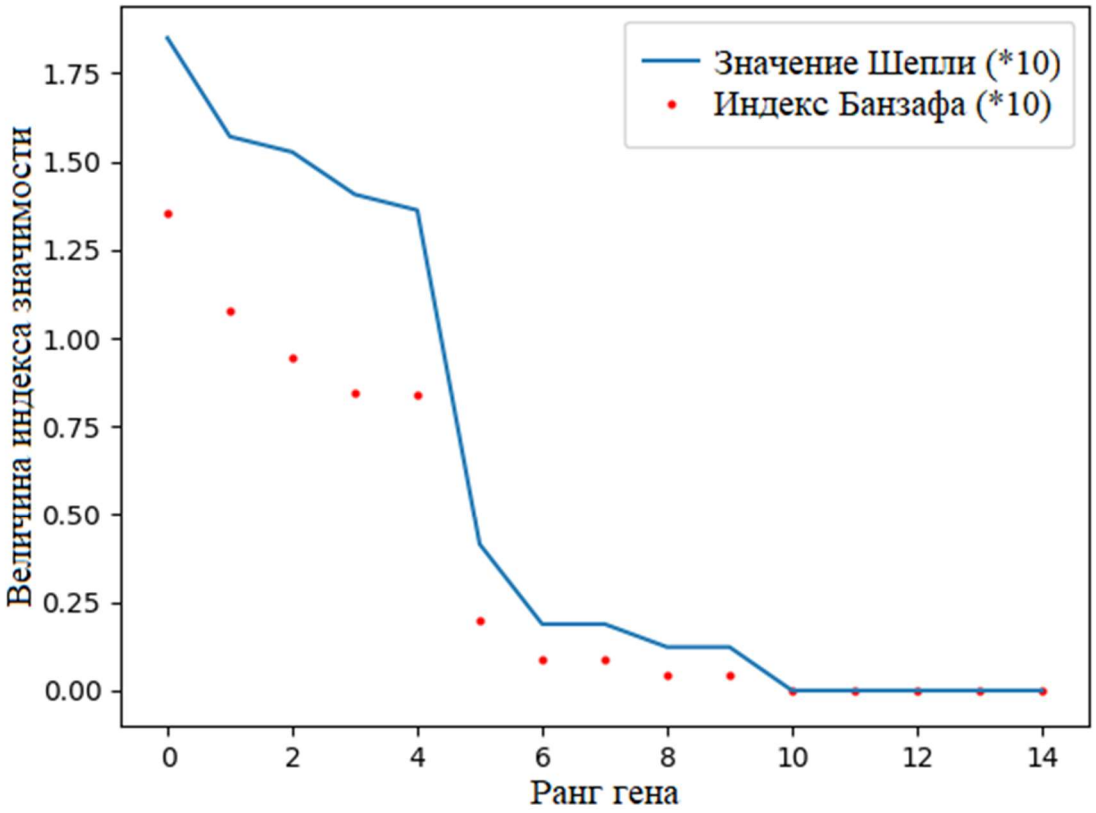

Рисунок 4. График изменения индекса значимости с уменьшением ранга гена для 15 генов

Можно заметить, что при уменьшении количества рассматриваемых генов уменьшается разница в ранжировании генов. Причем, в большинстве случаев, меняется только порядок генов в пределах взятых групп.

Доказано, что у мужчин и женщин болезнь Альцгеймера протекает по-разному. Рассмотрим две МЭС $E' = \langle N, S'_R, S'_D, A^{S'_R}, A^{S'_D} \rangle$ и $E'' = \langle N, S''_R, S''_D, A^{S''_R}, A^{S''_D} \rangle$. В первом случае имеем 22 «здоровых» и 15 тканей с болезнью Альцгеймера, во втором случае – 25 и 17 образцов, соответственно. В таблице 3 представлены первые 10 позиций по индексу значимости, посчитанного с помощью вектора Шепли для этих двух случаев.

Мужчины		**Женщины**	
Ген	**Значение Шепли** $(* \mathbf{10^3})$	**Ген**	**Значение Шепли** $(* \mathbf{10^3})$
ACAA1	*0,87529*	ZC3H11A	*1,40275*
HNRNPA3P1	*0,79922*	CCDC190	*1,39448*
MMAB	*0,79355*	ALDH2	*1,19962*
TDH	*0,7922*	AF074983	*1,10729*
NINL	*0,78509*	RHBDD1	*1,08732*
TDRD3	*0,78163*	ARHGEF10	*1,07735*
WDR49	*0,77295*	AK023372	*1,02686*
ZNF185	*0,77011*	ZNF761	*0,99062*
PFDN5	*0,76712*	ADGRA3	*0,98517*
ZNF844	*0,76267*	RYR1	*0,98269*

Таблица 3. Первые 10 генов по индексу значимости, посчитанному с помощью вектора Шепли отдельно для мужчин и для женщин

Построим также графики, показывающий разницу между индексами значимости, посчитанными разными способами для мужчин (рис. 5) и женщин (рис. 6).

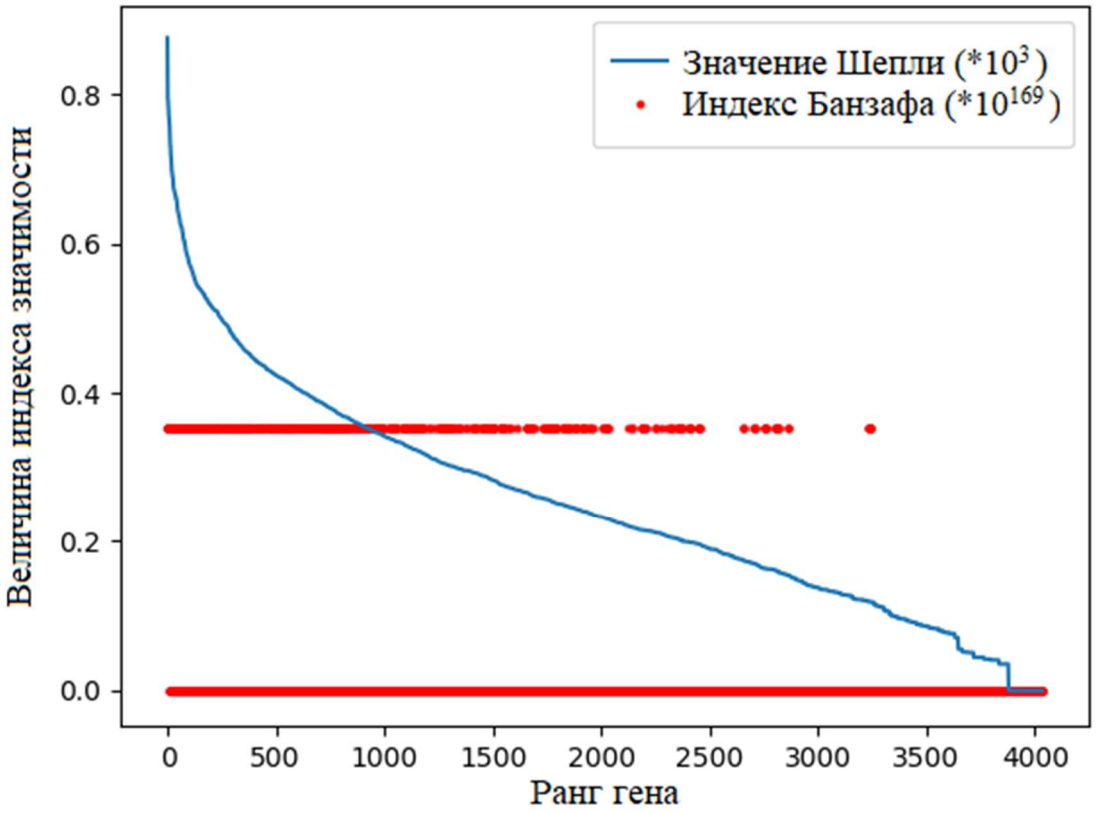

Рисунок 5. График изменения индекса значимости с уменьшением ранга гена у мужчин

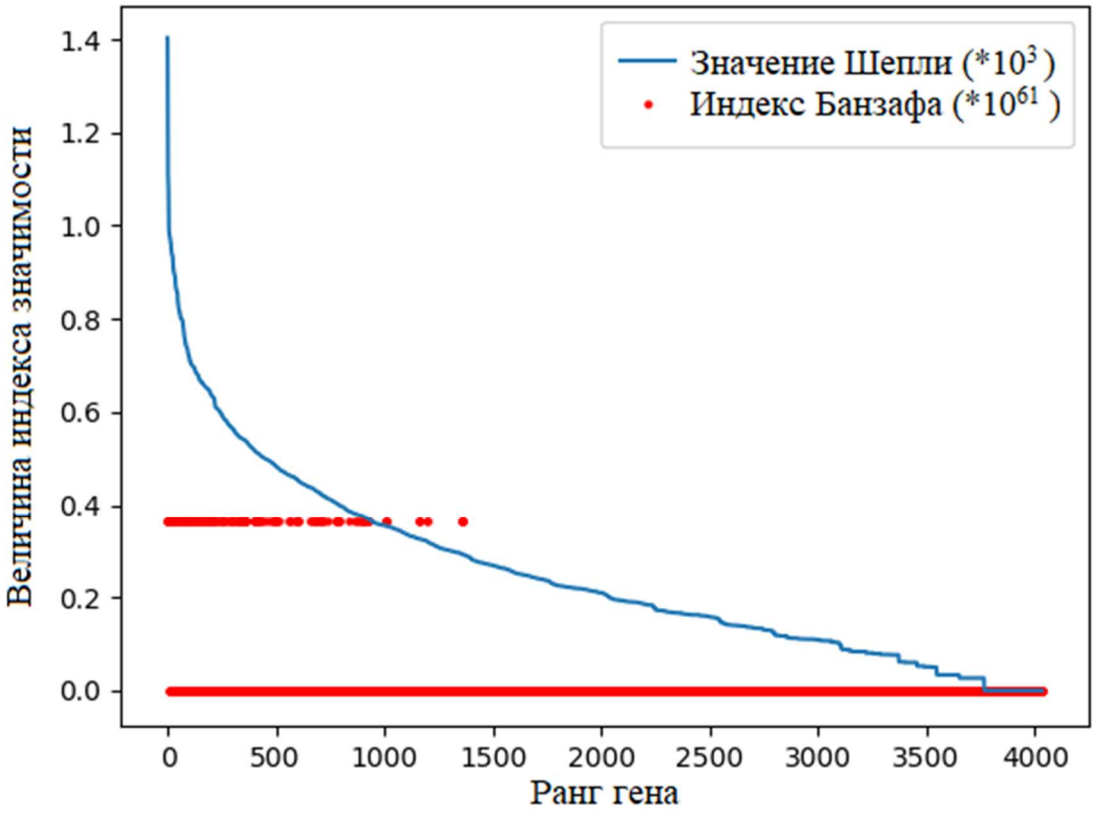

Рисунок 6. График изменения индекса значимости с уменьшением ранга гена у женщин

Многие полученные результаты подтверждены современными научными публикациями: так, например, влияние гена RYR1 на процесс болезни Альцгеймера описано в [15], влияние гена ALDH2 - в [16], роль шаперонов в патогенезе болезни Альцгеймера, в комплекс которых входит ген PFDN5, описана в [17]. Также выявлены гены, которые могут играть ключевую роль при патогенезе этой болезни, но еще не рассмотренные в научной литературе.

Выводы

До применения теории кооперативных игр математические методы, используемые для извлечения информации из экспрессии генов, можно было разделить на три группы: статистические методы, используемые для идентификации генов, регулируемых различными интересующими условиями; неконтролируемые методы анализа, используемые в качестве метода для идентификации групп генов или образцов с похожим поведением; инструменты прогнозирования класса, используемые для классификации образцов по известным категориям морфологии, известных биологических особенностей, клинических результатов и т. д. в соответствии с паттернами экспрессии генов. Данный подход имеет, по крайней мере, два отличия от классических: во-первых, используемый класс коалиционных игр, предоставляет эффективную возможность описать связь между глобальной экспрессией каждой коалиции генов и представляющим интерес биологическим состоянием и, следовательно, включить в последовательный анализ все возможные связи взаимодействия генов, связанные с биологическим состоянием; во-вторых, подход основан на идее применения структур решений к микрочиповым играм и на сильной связи между теорией игр и управляемыми свойствами.

Обычно интерпретация результатов, полученных с помощью классических статистических методов, сильно зависит от теоретической модели, используемой для анализа, или от сильных предположений об эталонной совокупности, из которой собираются образцы. Для применения подхода, основанного на теории кооперативных игр, необходимы только слабые предположения о популяции, и то, что четко обозначено априори, является границей для правдоподобной интерпретации результатов.

Результатом использования рассмотренного в этой работе подхода является распределение вектора решения, примененного к микрочиповой игре. Этот результат интерпретируется с использованием базовых свойств,

которые должны быть удовлетворены путем нахождения индекса значимости каждого гена в взаимосвязи экспрессии коалиций и генетического заболевания. Данный подход представляет интерес для генетического анализа, который все еще является относительно новой темой для исследований и предлагает новый математический подход для данной области.

Заключение

В данной работе было рассмотрено приложение теории кооперативных игр к анализу экспрессии генов. А именно: определены понятия микрочиповой игры, булевой матрицы экспрессии, индекса значимости генов, партнерства генов и т.д., представлена аксиоматическая характеристика решений теории кооперативных игр в контексте микрочиповых игр, рассмотрен практический пример на основе изучения влияния генов на развитие болезни Альцгеймера. Также была реализована программа по подсчету вектора Шепли и индекса Банзафа, которая позволяет наглядно показать результаты применения данного отдела теории игр к исследованию силы генов.

Список литературы

1. Moretti S., Patrone F., Bonassi S. The class of microarray games and the relevance index for genes // TOP. 2007. No 15. P. 256–280.
2. Drmanac R., Crkvenjakov R. Process for obtaining genome by hybridization and oligonucleotidic tests. // Yugoslav Patent Application YU0057087A. 1990.
3. Лысов Ю., Флорентьев В., Хорлин А., Храпко К., Шик В., Мирзабеков А. Определение нуклеотидной последовательности ДНК гибридизацией с олигонуклеотидами. Новый метод // Докл. Акад. Наук СССР. 1988. Т. 303. Стр. 1508−1511.
4. Джиоев Ю. П. Биочиповые технологии: этапы развития и инновационная стратегия в области молекулярной диагностики инфекционных заболеваний. // Бюллетень Восточно-Сибирского научного центра Сибирского отделения Российской академии медицинских наук. 2007. № 2. С. 119–123.
5. Amaratunga D., Cabrera J. Exploration and Analysis of DNA Microarray and Protein Array Data. New York: John Wiley & Sons. 2004. P. 267.
6. Baldi P., Hatfield G. W. DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling. Cambridge University Press, Cambridge. 2002. P. 230.
7. Parmigiani G., Garrett E. S., Irizarry R., Zeger S. L., eds. The analysis of gene expression data: methods and software. New York: Springer. 2003. P. 511.
8. Moretti S. Minimum cost spanning tree games and gene expression data analysis // ACM international conference proceeding series. 2006. P. 199–208.
9. Fragnelli V., Moretti S. A game theoretical approach to the classification problem in gene expression data analysis // Comput. Math. Appl. 2008. Vol. 55. No 5. P. 950–959.
10. Albino D., Scaruffi P., Moretti S., Coco S., Di Cristofano C., Cavazzana A., et al. Identification of low intratumoral gene expression heterogeneity in neuroblastic tumors by wide-genome expression analysis and game theory. // Cancer. 2008. No 113. P. 1412–1422.

11. Lucchetti R., Moretti S., Patrone F., Radrizzani P. The Shapley and Banzhaf value in microarray games. // Computers & Operations Research 37. 2010. P. 1406–1412.
12. Bower J. M., Bolouri H. Computational Modeling of Genetic and Biochemical Networks. Massachusetts University of Technology. 2001.
13. Laruelle A., Valenciano F. Shapley–Shubik and Banzhaf indices revisited. // Mathematics of Operations Research. 2001. No 26. P. 89–104.
14. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/geo/gds/analyze/analyze.cgi?ID=GDS4758.
15. Chami M., Checler F. Ryanodine receptors: dual contribution to Alzheimer disease? // Channels (Austin). 2014. P. 168–168.
16. Ma L., Lu Z. N. Role of ADH1B rs1229984 and ALDH2 rs671 gene polymorphisms in the development of Alzheimer's disease // Genet. Mol. Res. 2016. Vol. 15. No 4. P. 1–8.
17. Loke S. Y., Wong P. T., Ong W. Y. Global gene expression changes in the prefrontal cortex of rabbits with hypercholesterolemia and/or hypertension // J. Mol. Neurosci. 2007. No 102. P. 33–56.

Приложение

Приложение 1.

```
import numpy as np
import matplotlib.pyplot as plt
from itertools import combinations, chain, compress

class Gene(object):
 def __init__(self, Name, Sh, Bn, Index):
  self.Name = Name
  self.Sh = Sh
  self.Bn = Bn
  self.Index = Index

def Sh_key(gene):
  return gene.Sh
def Bn_key(gene):
  return gene.Bn

fileN = open('UniqNumbers.txt','r')
uniq = []
for line in fileN:
 uniq.append(line[0:-1])

file = open('NamesOfGenes.txt','r')
names = []
for i, line in enumerate(file):
 if str(i) in uniq:
  names.append([])
  st = ''
```

```
 for x in line:
  if (x == '\t' or x == '\n') and x != '':
   names[-1].append(st)
   st = ''
  else:
   st += x

fileExp = open('RightAllGenes.txt', 'r')
exp = []
for i, line in enumerate(fileExp):
 if str(i) in uniq:
  exp.append([])
  st = ''
  for x in line:
   if x == ' ' or x == '\n':
    if st != '':
     exp[-1].append(float(st))
    st = ''
   else:
    st += x
exp = np.array(exp)

min = []
max = []
bexp = []

for i in range(len(exp)):
 min.append([])
 max.append([])
 bexp.append([])
```

```
min[i].append(exp[i,32])
max[i].append(exp[i,32])
for j in range(33,42):
 if exp[i,j] < min[i][0]:
  min[i][0] = exp[i,j]
 elif exp[i,j] > max[i][0]:
  max[i][0] = exp[i,j]
min[i].append(exp[i,42])
max[i].append(exp[i,42])
for j in range(43,61):
 if exp[i,j] < min[i][1]:
  min[i][1] = exp[i,j]
 elif exp[i,j] > max[i][1]:
  max[i][1] = exp[i,j]
min[i].append(exp[i,61])
max[i].append(exp[i,61])
for j in range(62,79):
 if exp[i,j] < min[i][2]:
  min[i][2] = exp[i,j]
 elif exp[i,j] > max[i][2]:
  max[i][2] = exp[i,j]

for j in range(7):
 if exp[i,j] < min[i][0] or exp[i,j] > max[i][0]:
  bexp[i].append(1)
 else:
  bexp[i].append(0)

for j in range(7,17):
```

```
 if exp[i,j] < min[i][1] or exp[i,j] > max[i][1]:
  bexp[i].append(1)
 else:
  bexp[i].append(0)

for j in range(17,32):
 if exp[i,j] < min[i][2] or exp[i,j] > max[i][2]:
  bexp[i].append(1)
 else:
  bexp[i].append(0)

sum = 0

for j in range(len(bexp[i])):
  sum += bexp[i][j]
 print(bexp[i], sum)

bexp = np.array(bexp)
bexpT = bexp.transpose()
s=[]

for i in range(len(bexpT)):
 s.append([])
 for j in range(len(bexpT[i])):
  if bexpT[i,j]:
   s[i].append(j+1)
 print(s[i])

sh = np.zeros(len(bexp))
bn = np.zeros(len(bexp))
```

```
for i in range(len(s)):
 for j in s[i]:
  sh[j-1] += 1/len(s[i])
  bn[j-1] += 1/2**(len(s[i])-1)

sh = sh/len(bexpT)
bn = bn/len(bexpT)

genes = []
for i in range(len(names)):
 genes.append(Gene(names[i], sh[i], bn[i], i))

sort_sh = sorted(genes, key = Sh_key, reverse = True)
sort_bn = sorted(genes, key = Bn_key, reverse = True)

fileTop = open('TopOfAllGenes.txt', 'w')
top = []
top.append('Rang\tName of Gene\tShapley Value\t\tName of Gene\tBanzhaf
Value\n')

for i in range(len(names)):
 top.append(str(i+1) + '.\t' + sort_sh[i].Name[0] + '\tSh = ' +
str(sort_sh[i].Sh) + '\t\t' + sort_bn[i].Name[0] + '\tBn = ' +
str(sort_bn[i].Bn)+'\n')
fileTop.writelines(top)

for i in range(30):
 print(i,'.', sort_sh[i].Name[0], '\t\t: Sh =', sort_sh[i].Sh,'\t',
sort_bn[i].Name[0], '\t\t: Bn =', sort_bn[i].Bn)
```

```
s = []
for i in range(5):
 s.append(sort_sh[i].Index)
 s.append(sort_sh[len(genes)//2-2+i].Index)
 s.append(sort_sh[-i-1].Index)

file30 = open('UniqNum15.txt','w')
for i in range(len(s)):
 file30.writelines(str(s[i])+'\n')

s = []
for i in range(10):
 s.append(sort_sh[i].Index)
 s.append(sort_sh[len(genes)//2-5+i].Index)
 s.append(sort_sh[-i-1].Index)

file30 = open('UniqNum30.txt','w')
for i in range(len(s)):
 file30.writelines(str(s[i])+'\n')

s = []
for i in range(33):
 s.append(sort_sh[i].Index)
 s.append(sort_sh[len(genes)//2-16+i].Index)
 s.append(sort_sh[-i-1].Index)
s.append(sort_sh[len(genes)//2+17].Index)

file30 = open('UniqNum100.txt','w')
for i in range(len(s)):
```

```
        file30.writelines(str(s[i])+'\n')

        X = range(len(genes))
        Y1 = []
        Y2 = []
        for i in range(len(genes)):
         Y1.append(sort_sh[i].Sh*pow(10,3))
         Y2.append(sort_sh[i].Bn*pow(10,55))

        plt.xlabel('number')
        plt.ylabel('value')
        plt.plot(X,Y1)
        plt.scatter(X,Y2, marker='.', linewidth = 0.05, color = 'red')
        plt.legend(('Shapley (*10^3)','Banzaf (*10^55)'))
        plt.show()


        print(sh)
        print(bn)
print(np.sum(sh), np.sum(bn))
```