КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 3 С. 697–714
DOI: 10.20537/2076-7633-2026-18-3-697-714

Ки&М

СПЕЦИАЛЬНЫЙ ВЫПУСК

УДК: 004.032.26, 57.087.1

Применение бета-регрессии в задаче альтернативного сплайсинга гена CD44

А. А. Пирогов

НИУ ВШЭ,
Россия, 101000, г. Москва, ул. Мясницкая, д. 20

E-mail: aapirogov@hse.ru

Получено 09.04.2026, после доработки — 08.06.2026.
Принято к публикации 08.06.2026.

Изоформный состав гена CD44 тесно связан с прогрессированием злокачественных опухолей и формированием раковых стволовых клеток. Хотя механизм альтернативного сплайсинга обеспечивает выборочное включение вариабельных экзонов этого гена, общий регуляторный ландшафт данного процесса до сих пор остается малоизученным.

В настоящей работе предложен новый вычислительный подход к анализу регуляции сплайсинга, основанный на моделировании долей экспрессии транскриптов с помощью аппарата бета-регрессии. Автор реализовал математическую модель в виде глубокого нейросетевого регрессора, который совместно оценивает параметры вероятностного распределения и применяет регуляризацию методом эластичной сети для отбора наиболее значимых биологических признаков. Разработанный метод успешно применен для идентификации рибонуклеиново-связывающих белков, выполняющих роль факторов сплайсинга и напрямую управляющих выбором сайтов сборки гена CD44 в клетках колоректального рака. В построенных регрессионных моделях целевыми переменными выступают строго нормированные доли экспрессии изоформ, а в качестве признаков используются уровни экспрессии потенциальных регуляторов.

В ходе исследования детально сопоставлены две схемы постановки задачи машинного обучения. Первая схема реализует классический подход «один против всех», предполагающий построение отдельной изолированной модели для каждой рассматриваемой изоформы. Второй подход представляет собой метод дерева изоформ, базирующийся на последовательном иерархическом разбиении транскриптов по наличию вариабельных экзонов с независимым подбором регуляторных белков на каждом этапе алгоритма. Вычислительная корректность программной реализации была предварительно подтверждена на искусственно сгенерированных данных: модель продемонстрировала высокую точность восстановления параметров при полном отсутствии систематического смещения оценок. Последующий сравнительный анализ на клинических данных выявил явное преимущество схемы «один против всех» с точки зрения итогового качества и стабильности предсказаний.

Биологический анализ результатов не только подтвердил участие известных регуляторов гена CD44, но и позволил выявить новые потенциальные факторы регуляции, среди которых выделяются белки ACO1, NUDT21 и AGO2. Для фактора ACO1 автор сформулировал оригинальную гипотезу, напрямую связывающую внутриклеточный метаболизм железа и регуляцию изоформного состава гена CD44. Полученные выводы существенно расширяют текущее понимание молекулярных механизмов онкогенеза.

Ключевые слова: бета-регрессия, машинное обучение, сплайсинг, CD44

COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 3 P. 697–714
DOI: 10.20537/2076-7633-2026-18-3-697-714

Ки&М

SPECIAL ISSUE

UDC: 004.032.26, 57.087.1

Application of beta regression to the CD44 alternative splicing problem

A. A. Pirogov

HSE University,
20 Myasnitskaya st., Moscow, 101000, Russia

E-mail: aapirogov@hse.ru

Received 09.04.2026, after completion — 08.06.2026.
Accepted for publication 08.06.2026.

Aberrant alternative splicing of the CD44 gene drives colorectal cancer progression and facilitates the emergence of cancer stem cells. Although biomedical research recognizes this transmembrane glycoprotein as a major catalyst of malignancy, deciphering its multi-isoform regulatory networks remains a complex analytical challenge. To address this knowledge gap, this study presents a machine learning framework designed to decode these biological mechanisms. The author constructed a neural network regressor based on beta regression to model bounded isoform proportions. This computational architecture jointly estimates both the mean and the precision parameters of the underlying probability distribution. Furthermore, the system employs elastic net regularization to perform quantitative feature selection from high-dimensional molecular expression data.

The investigation evaluates the proposed framework using gene expression profiles from colorectal cancer patients. The primary objective involves identifying specific ribonucleic acid-binding proteins acting as regulatory splicing factors. The experimental design contrasts two distinct mathematical modeling strategies. The first configuration incorporates an independent ”one-vs-all” approach that treats each transcript variant as an isolated regression target. The second formulation utilizes a structured ”isoform tree” method that directly mirrors hierarchical exon inclusion relationships. Validation experiments on synthetically generated datasets confirmed the mathematical integrity of the network. The model recovered true distribution parameters with precision and exhibited no systematic bias. Comprehensive empirical comparisons subsequently demonstrated that the independent ”one-vs-all” layout consistently outperforms the hierarchical tree configuration in predictive stability and accuracy.

The computational analysis maps the regulatory landscape of the CD44 gene. The framework validates several established splicing factors while uncovering new candidate proteins, including ACO1, NUDT21, and AGO2. Based on these statistical associations, the paper introduces a biological hypothesis. This concept functionally connects intracellular iron metabolism via the ACO1 protein with the shifting balance of CD44 variants. These discoveries provide deeper insights into oncogenic splicing regulation. Ultimately, they highlight molecular targets for future therapeutic interventions aimed at suppressing the cancer stem cell phenotype.

Keywords: beta regression, machine learning, splicing, CD44

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 3, pp. 697–714 (Russian).

Введение

Альтернативный сплайсинг является одним из ключевых механизмов посттранскрипционной регуляции генной экспрессии у эукариот. Геномная ДНК содержит в себе два типа последовательностей: кодирующие, называемые экзонами, и некодирующие вставки, называемые интронами. В ходе процессинга пре-мРНК интроны вырезаются, а экзоны соединяются в зрелую молекулу мРНК. При альтернативном сплайсинге этот процесс не является детерминированным: часть экзонов может включаться в итоговую последовательность или исключаться из нее, что позволяет одному гену кодировать множество белковых изоформ с различными структурными и функциональными свойствами [Marasco, Kornblihtt, 2023]. Известно, что нарушения регуляции альтернативного сплайсинга тесно связаны с развитием широкого спектра заболеваний, прежде всего онкологических [Li et al., 2021].

Ключевую роль в регуляции альтернативного сплайсинга играют РНК-связывающие белки (сплайсинг-факторы) — молекулы, связывающиеся с пре-мРНК и управляющие выбором сайтов сплайсинга. Механизм их действия основан на распознавании специфических цис-регуляторных элементов — сплайсинговых энхансеров и сайленсеров — и последующей модуляции сборки сплайсосомы [Black, 2003; Fu, Ares, 2014]. Совместное действие нескольких РНК-связывающих белков на один транскрипт носит контекстно зависимый характер, что существенно затрудняет предсказание результата сплайсинга по последовательности генома [Fu, Ares, 2014]. Идентификация значимых регуляторов для конкретного гена остается и по сей день актуальной и методически сложной задачей [Li et al., 2021].

Среди генов с альтернативным сплайсингом ген CD44 занимает особое место в онкологическом контексте. CD44 кодирует трансмембранный гликопротеин, члены семейства которого различаются по экстраклеточному домену: десять вариабельных экзонов (обозначаемых v1–v10) могут включаться или исключаться в различных комбинациях, причем известно, что экзон v1 отсутствует в сплайсинге человека [Mishra et al., 2019]. Альтернативный сплайсинг CD44, нередко нарушенный при онкологических заболеваниях, порождает изоформы со свойствами, различающимися в зависимости от типа ткани и по-разному влияющими на прогрессию опухоли [Prochazka et al., 2014]. Стандартная изоформа CD44s, не имеющая в себе никаких вариабельных экзонов, функционирует как рецептор гиалуроновой кислоты и корецептор факторов роста, интегрируя сигналы микроокружения и регулируя адгезию, миграцию, пролиферацию и выживаемость клеток [Yan et al., 2015; Hassn Mesrati et al., 2021]. Вариантные изоформы CD44v, в свою очередь, идентифицированы как маркеры раковых стволовых клеток и связаны с химиорезистентностью и метастазированием [Yan et al., 2015].

Таким образом, идентификация РНК-связывающих белков, регулирующих сплайсинг гена CD44, представляет одновременно фундаментальный и прикладной интерес. С фундаментальной точки зрения регуляторный ландшафт сплайсинга CD44 остается недостаточно изученным: совместное действие нескольких сплайсинг-факторов, определяющее включение вариабельных экзонов в различных сочетаниях, формирует сложную контекстно зависимую систему, механизмы которой в онкологическом контексте во многом не выяснены. В прикладном отношении идентифицированные регуляторы могут стать мишенями для терапевтического вмешательства: направленное смещение баланса изоформ (в частности, от вариантных изоформ CD44v к стандартной CD44s) потенциально способно ослабить фенотип раковых стволовых клеток, снизить химиорезистентность и ограничить метастатический потенциал опухоли [Yan et al., 2015; Mishra et al., 2019]. Естественным способом работы с экспрессиями изоформ является их нормировка на суммарную экспрессию гена: полученные пропорции отражают относительный вклад каждой изоформы и принимают значения в интервале (0, 1). Задача тогда формулируется следующим образом: по уровням экспрессии РНК-связывающих белков восстановить линейную зависимость,

позволяющую оценить степень влияния каждого сплайсинг-фактора на долю конкретной изоформы. В силу природы целевой переменной стандартные линейные модели здесь неприменимы [Kieschnick, McCullough, 2003]: для моделирования пропорций естественно использовать бета-регрессию [Ferrari, Cribari-Neto, 2004], предполагающую бета-распределение отклика и явно учитывающую ограниченность его области значений.

В работе [Novosad, 2023] этот подход был реализован применительно к сплайсингу CD44 в клетках колоректального рака: в качестве целевой переменной рассматривалась доля изоформы CD44v8–10 относительно суммы изоформ CD44v8–10 и CD44s как наиболее экспрессируемых в данном типе ткани, а бета-регрессия использовалась для определения важных РНК-связывающих белков.

В настоящей работе CD44 также рассматривается в клетках колоректального рака, однако анализ расширен на больший набор изоформ этого гена. Для выявления значимых сплайсинг-факторов предложены и сопоставлены два подхода: схема *один против всех*, в которой для каждой изоформы независимо строится отдельная модель бета-регрессии, и метод *дерева изоформ*, основанный на иерархической декомпозиции по вариабельным экзонам. Помимо этого, предлагается новая реализация бета-регрессии: в отличие от классической статистической постановки [Ferrari, Cribari-Neto, 2004] модель реализована как нейросетевой регрессор на базе PyTorch [Paszke et al., 2019] со встроенной регуляризацией Elastic Net [Zou, Hastie, 2005] для отбора значимых признаков. Корректность реализации верифицирована на синтетически сгенерированных данных, после чего модель применена к задаче идентификации регуляторов сплайсинга CD44 с выделением новых потенциальных кандидатов.

Теоретическая модель бета-регрессии

Прежде чем переходить непосредственно к теоретической составляющей модели, напомним базовые определения, связанные с ней.

Определение 1. Говорим, что случайная величина ξ задана функцией распределения бета с параметрами α, β (обозначается как $\xi \sim \mathcal{B}(\alpha, \beta)$), если плотность распределения случайной величины имеет вид

$$p_{\xi}(x) = \frac{1}{\mathcal{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \cdot I_{x\in[0,1]}, \tag{1}$$

где $\mathcal{B}(\alpha, \beta) = \int\limits_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ — бета-функция;

$$I_{x\in[0,1]} = \begin{cases} 1, & x \in [0, 1], \\ 0, & \text{если иначе}, \end{cases}$$

— индикаторная функция.

Отметим, что для $\xi \sim \mathcal{B}(\alpha, \beta)$ математическое ожидание и дисперсия соответственно равны

$$E\xi = \frac{\alpha}{\alpha+\beta}, \quad D\xi = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Пусть $y_1, \ldots, y_N \in (0, 1)$ — наблюдаемые значения целевой (предсказываемой) переменной. В нашей модели предполагается, что для всех $i \in \{1, \ldots, N\}$ выполняется $y_i \sim \mathcal{B}(\alpha_i, \beta_i)$, где N — количество наблюдений.

Определение 2. Функцией правдоподобия $L(x_1, \ldots, x_n)$ (в дальнейшем будем использовать обозначение $L(\overline{x})$) для выборки $\xi_1, \ldots, \xi_n$ называется функция, задаваемая следующим образом:

$$L_1(\overline{x}) = \prod_{i=1}^{n} f_{\xi_i}(x_i),$$

где $f_{\xi_i}(x)$ — плотность или закон распределения случайной величины ξ_i.

Зачастую вместо $L_1(\overline{x})$ рассматривают $\ln L_1(\overline{x})$, так как в силу строгого возрастания логарифма $\ln f(x)$ имеет те же точки минимума/максимума, что и $f(x)$. Пользуясь введенными определениями, получаем, что логарифмическая функция правдоподобия в модели бета-регрессии имеет следующий вид:

$$L(\overline{x}) = \ln L_1(\overline{x}) = \sum_{i=1}^{N}[-\ln(\mathcal{B}(\alpha_i, \beta_i)) + (\alpha_i - 1)\ln(x_i) + (\beta_i - 1)\ln(1 - x_i)]. \tag{2}$$

В дальнейшем будем называть $-L(\overline{x})$ функцией потерь.

Используя факт, что $\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, где $\Gamma(n) = \int\limits_0^\infty x^{n-1}e^{-x}\,dx$ — гамма-функция [Abramowitz, Stegun, 1964], функцию правдоподобия можно переписать в следующем виде:

$$L(\overline{x}) = \sum_{i=1}^{N}[\ln\Gamma(\alpha_i + \beta_i) - \ln\Gamma(\alpha_i) - \ln\Gamma(\beta_i) + (\alpha_i - 1)\ln(x_i) + (\beta_i - 1)\ln(1 - x_i)]. \tag{3}$$

Переходя непосредственно к описанию самой модели бета-регрессии, стоит отметить преобразование, стандартно используемое при работе с моделью [Ferrari, Cribari-Neto, 2004; Vasconcellos, Cribari-Neto, 2005].

В этой задаче регрессии стандартно принято восстанавливать не сами параметры, а моменты распределений, такие как математическое ожидание и дисперсия, ввиду удобности и упрощения вычислений [Smithson, Verkuilen, 2006]. Поэтому мы перепараметризуем распределение.

Пусть $\mu_i = \frac{\alpha_i}{\alpha_i+\beta_i}$ и $\phi_i = \alpha_i + \beta_i$. Тем самым моменты распределения будут иметь вид

$$E(y) = \mu, \quad D(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

При использовании данной замены функция правдоподобия примет вид

$$L(\overline{x}) = \sum_{i=1}^{N}[\ln\Gamma(\phi_i) - \ln\Gamma(\mu_i\phi_i) - \ln\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i - 1)\ln(x_i) + (1-\mu_i)\phi_i\ln(1-x_i)]. \tag{4}$$

Модель бета-регрессии одновременно оценивает оба параметра перепараметризованного распределения: μ_i (среднее) и ϕ_i (точность). Введем обозначения: $X \in \mathbb{R}^{N\times m}$ — матрица признаков для среднего (m признаков, N наблюдений); $\omega \in \mathbb{R}^m$ — вектор весов X; $Z \in \mathbb{R}^{N\times l}$ — матрица признаков для параметра точности (l признаков; m не обязательно равно l; N наблюдений); $\eta \in \mathbb{R}^l$ — вектор весов Z.

Предполагается, что

$$g_1(\mu_i) = \sum_{j=1}^{m}\omega_j X_{ij}, \quad g_2(\phi_i) = \sum_{k=1}^{l}\eta_k Z_{ik}, \tag{5}$$

где g_1 — строго монотонная дифференцируемая функция связи, $g_1\colon (0, 1) \to \mathbb{R}$. Для параметра точности возможны два случая: ϕ_i либо моделируется аналогично через строго монотонную дифференцируемую функцию связи $g_2\colon (0, +\infty) \to \mathbb{R}$, либо полагается постоянным для всех наблюдений, $\phi_i \equiv \phi = \text{const}$ [Cribari-Neto, Zeileis, 2010]. В рамках этой модели $g_1(x) = \text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$, тогда $g_1^{-1}(x) = \sigma(x) = \frac{1}{1+e^{-x}}$.

Тогда задачу регрессии можно сформулировать следующим образом: найти ω и η такие, что

$$\omega,\, \eta = \arg\min_{\omega', \eta'}(-L(\overline{x})). \tag{6}$$

Реализация и обоснование ее корректности

В рамках практической реализации модели предполагается, что $X = Z$, а также в качестве функции преобразования для параметра точности возьмем $g_2(x) = \ln(e^x - 1)$, то есть такую, что $g_2^{-1}(x) = \text{softplus}(x) = \ln(1 + e^x)$. Такой выбор обусловлен сразу двумя соображениями. Во-первых, эта функция удовлетворяет всем условиям, наложенным на функцию $g_2(x)$, упомянутым ранее. Во-вторых, данная функция принадлежит классу C^∞, что обеспечивает численную устойчивость при вычислении градиентов, необходимых при минимизации функции потерь [Glorot et al., 2011].

Реализация выполнена на языке программирования Python версии 3.12.4 с использованием библиотеки Pytorch версии 2.3.1 [Paszke et al., 2019]. Обучение модели содержательно устроено следующим образом: сначала матрица признаков X нормализуется [LeCun et al., 1998], после чего происходит оптимизация функции потерь с поправкой на целевую переменную $\overline{y}$. Она производится с помощью оптимизатора ADAM (Adaptive Moment Estimation) [Kingma, Ba, 2015]. Суть метода заключается в том, что на каждой итерации t оптимизатор вычисляет две экспоненциально сглаженные статистики градиента — среднее m_t и дисперсию v_t:

$$m_t = \delta_1 m_{t-1} + (1 - \delta_1)\nabla_\theta(-L(\overline{x})), \quad v_t = \delta_2 v_{t-1} + (1 - \delta_2)(\nabla_\theta(-L(\overline{x})))^2,$$

после чего вводятся оценки с поправкой на смещение:

$$\widehat{m}_t = \frac{m_t}{1 - \delta_1^t}, \quad \widehat{v}_t = \frac{v_t}{1 - \delta_2^t},$$

и веса обновляются по правилу

$$\theta_t = \theta_{t-1} - \frac{lr}{\sqrt{\widehat{v}_t} + \varepsilon}\widehat{m}_t,$$

где $\theta \in \{\omega, \eta\}$ — текущие веса, lr — шаг обучения (learning rate), $\varepsilon = 10^{-8}$ — константа численной устойчивости, δ_1, δ_2 — экспоненциальные скорости затухания для моментов, ∇_θ — оператор градиента по весам. Обозначим через $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ дигамма-функцию [Abramowitz, Stegun, 1964]. Тогда компоненты градиентов функции правдоподобия в бета-регрессии имеют вид

$$\nabla_{\omega_j} L = \frac{\partial L}{\partial w_j} = \sum_{i=1}^{N} d_i^{(\mu)} X_{ij}, \quad \nabla_{\eta_j} L = \frac{\partial L}{\partial \eta_j} = \sum_{i=1}^{N} d_i^{(\phi)} X_{ij}, \quad j = 1, \ldots, m,$$

где i-е компоненты вспомогательных векторов равны

$$d_i^{(\mu)} = \phi_i\mu_i(1 - \mu_i)\left[\psi((1 - \mu_i)\phi_i) - \psi(\mu_i\phi_i) + \ln\frac{y_i}{1 - y_i}\right],$$

$$d_i^{(\phi)} = \sigma\left(\sum_{k=1}^{N}\eta_k X_{ik}\right)\cdot[\psi(\phi_i) - \mu_i\psi(\mu_i\phi_i) - (1 - \mu_i)\psi((1 - \mu_i)\phi_i) + \mu_i\ln y_i + (1 - \mu_i)\ln(1 - y_i)].$$

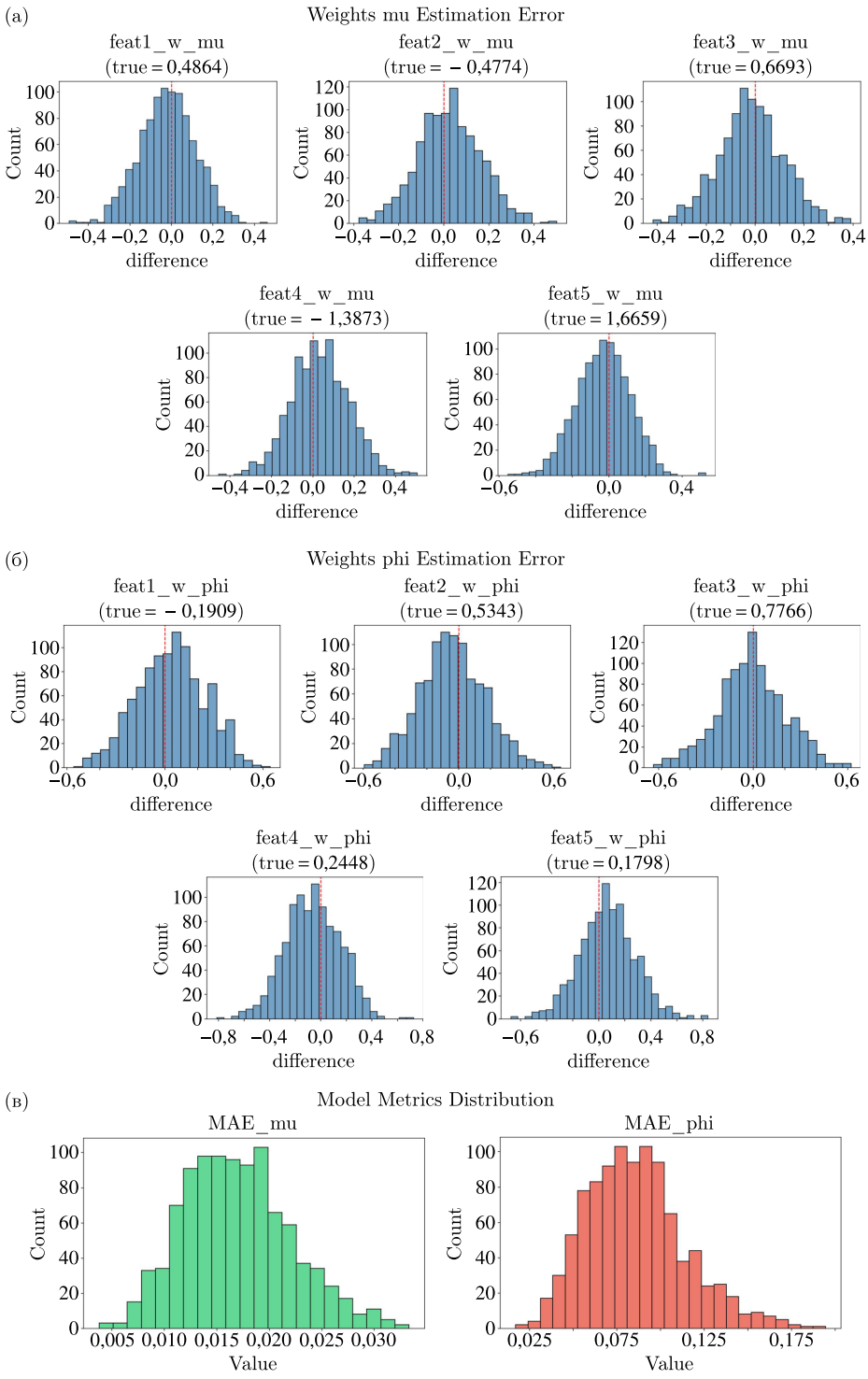

Рис. 1. Гистограммы распределения ошибок восстановления параметров модели на синтетических данных. На панелях (а), (б) представлены распределения отклонений оцененных весов от их истинных значений (указанных над соответствующими гистограммами) для набора из пяти признаков. По горизонтальной оси отложены значения отклонений оцененного параметра, в то время как по вертикальной оси — число моделей, для которых значение отклонения попадает в соответствующий биновый интервал. На панели (в) представлены распределения среднеабсолютных ошибок для параметров среднего μ (левый график) и точности ϕ (правый график). По горизонтальной оси отложены значения метрики MAE для соответствующего восстанавливаемого параметра, тогда как по вертикальной оси указано число моделей, для которых значение MAE попадает в указанный биновый интервал

Проверка корректности обучения и предсказания модели проводилась на синтетических тестах. Для этого была проведена серия из 1000 симуляций с одними и теми же значениями весов, предварительно сгенерированных из стандартного нормального распределения. В каждой симуляции столбцы матрицы признаков были независимо сгенерированы из равномерного на отрезке [0, 1] распределения, число наблюдений в каждой матрице равно 1000. По сгенерированным данным вычислялись истинные значения параметров среднего и точности:

$$\mu_i = \sigma\left(\sum_{j=1}^{m} \omega_j X_{ij}\right), \quad \phi_i = \text{softplus}\left(\sum_{j=1}^{m} \omega_j X_{ij}\right),$$

после чего целевая переменная генерировалась как

$$y_i \sim \mathcal{B}(\mu_i\phi_i, (1-\mu_i)\phi_i).$$

Затем модель обучалась на этих данных и вычислялись отклонение восстановленных весов от истинных, а также среднеабсолютное отклонение предсказаний среднего и точности. Распределение отклонений по всем симуляциям визуализировалось в виде гистограмм (см. рис. 1).

Из представленных на рис. 1 результатов следует, что для обоих наборов весов распределения отклонений сосредоточены вблизи нуля, что говорит нам об отсутствии ярко выраженной смещенности получаемых оценок. Среднеабсолютное отклонение предсказаний μ сосредоточено в районе 0,01–0,025, в то время как для ϕ — в промежутке от 0,05 до 0,15, что является приемлемым результатом, принимая во внимание область значений целевых переменных.

Материалы и методы

В модель регрессии была введена регуляризация весовых коэффициентов среднего с целью повышения интерпретируемости и отбора значимых РНК-связывающих белков. В частности, к функции потерь были добавлены штрафные слагаемые, соответствующие l_1- и l_2-регуляризации, то есть добавлена регуляризация методом Elastic Net [Zou, Hastie, 2005]:

$$\gamma_1 \sum_{i=1}^{n} |\omega_i|, \quad \gamma_2 \sum_{i=1}^{n} \omega_i^2,$$

где γ_1 и γ_2 — гиперпараметры регуляризации.

В результате итоговая функция потерь принимает вид

$$-L(\overline{x}) + \gamma_1 \sum_{i=1}^{n} |\omega_i| + \gamma_2 \sum_{i=1}^{n} \omega_i^2.$$

Данные об уровнях экспрессии сплайс-вариантов гена CD44 и потенциальных сплайсинг-факторов на клетках колоректального рака были получены из открытого репозитория TCGA PanCancer [Weinstein et al., 2013], референсные последовательности транскриптов, а также их геномные координаты и экзонный состав — из базы данных GENCODE 21 [Frankish et al., 2021]. Список РНК-связывающих белков и их мотивов был скачан из баз данных Attract [Giudice et al., 2016] и SpliceAid-F [Giulietti et al., 2013]. Мотивы связывания РНК-связывающих белков, представленные в РНК-ориентации, были преобразованы в соответствующие ДНК-последовательности с использованием операции обратного комплементарного преобразования. Поиск вхождений мотивов в транскриптах осуществлялся методом точного сопоставления подстрок без допуска несовпадений. В качестве порогового значения для отбора потенциально значимых факторов

использовалось наличие хотя бы одного вхождения мотива. Все признаки демонстрировали вариативность по выборке; константные или околонулевые признаки в матрице признаков отсутствовали и дополнительно не фильтровались. Всего были отобраны 331 пациент и 161 потенциальный сплайсинг-фактор.

Из 38 аннотированных изоформ были отобраны изоформы, медианные доли экспрессий которых не ниже 0,03, так как иные транскрипты несут минимальную регуляторную информацию: их доля экспрессии стабильно близка к нулю во всей выборке, вследствие чего модель бета-регрессии вырождается в тривиальное константное предсказание и не способна выявить значимые зависимости от экспрессии сплайсинг-факторов. В результате осталось 5 сплайс-вариантов: CD44s, CD44v8–9, CD44v3–10, CD44v8–10, CD44v6–10.

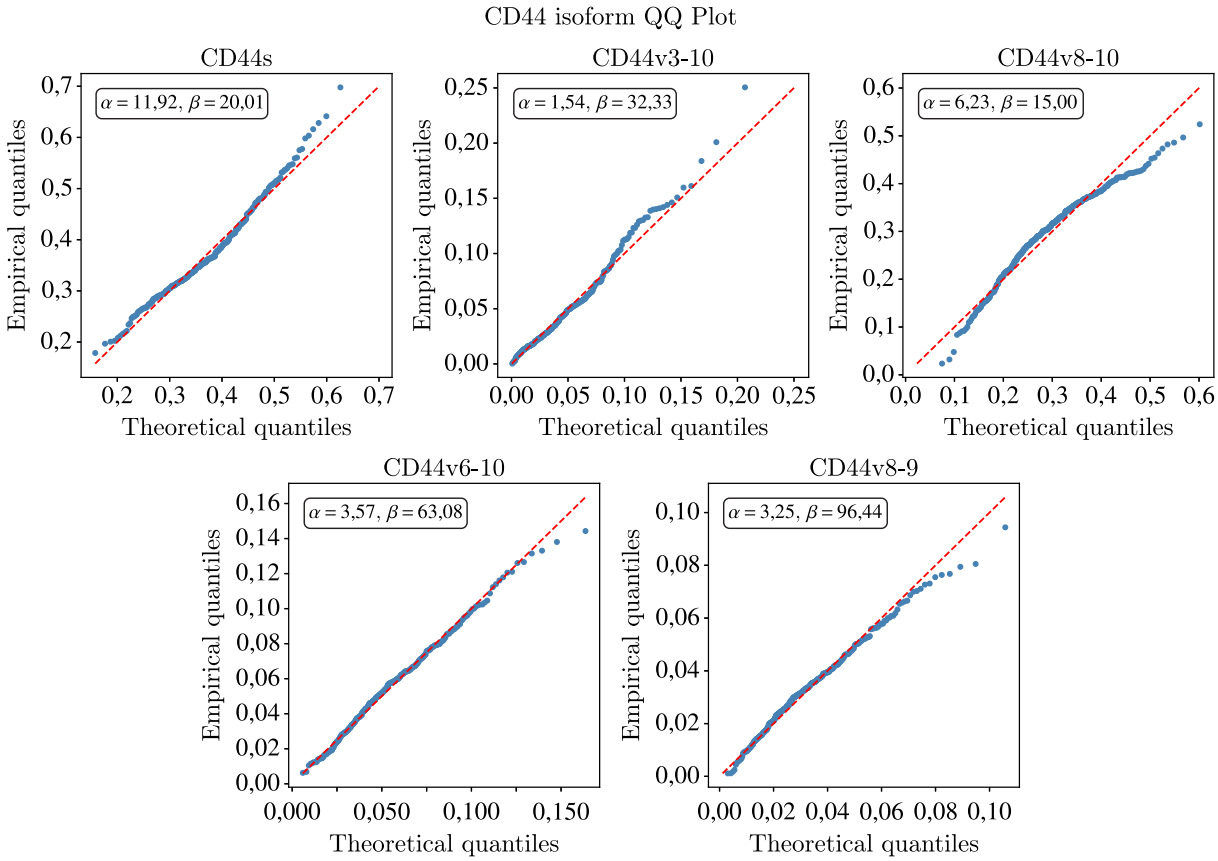

Рис. 2. Квантиль-квантиль-графики для отобранных изоформ гена CD44. По горизонтальной оси отложены значения квантилей теоретического бета-распределения с параметрами α и β, подобранными методом максимального правдоподобия и указанными в легенде для каждого соответствующего рисунка; по вертикальной оси указаны значения квантилей фактического распределения экспрессий изоформ

Для обоснования применимости бета-регрессии к моделированию долей для каждого из транскриптов были построены квантиль-квантиль-графики: эмпирические квантили наблюдаемых долей сопоставлялись с теоретическими квантилями бета-распределения, параметры которого оценивались методом максимального правдоподобия. Близость большинства точек изоформ к биссектрисе первого квадранта свидетельствует о соответствии данных бета-распределению (рис. 2).

Для выявления значимых сплайсинг-факторов было рассмотрено два подхода, различающихся способом постановки регрессионной задачи.

Первый подход реализует схему сравнения *one versus all*: для каждого из пяти сплайс-вариантов в отдельности строится модель бета-регрессии, предсказывающая долю данного транскрипта относительно суммарной экспрессии гена. В качестве признаков используются только те сплайсинг-факторы, мотивы которых присутствуют в последовательности рассматриваемого транскрипта. Такой подход позволяет рассматривать регуляцию каждого сплайс-варианта независимо от остальных, что дает возможность выявить факторы, специфично влияющие именно на данную изоформу.

Второй подход основан на методе, называемом *деревом изоформ*. Условимся называть экзон вариабельным, если он присутствует не во всех экзонных составах рассматриваемой группы изоформ; например, для CD44v8-10 и CD44v8-9 таким вариабельным экзоном выступает экзон 10. На основе этого определения строится бинарное дерево по следующему рекурсивному правилу: в корне располагается полный набор рассматриваемых транскриптов; на каждом шаге выбирается первый по порядку расположения (номеру) вариабельный экзон, разделяющий транскрипты текущего узла на два подмножества — содержащие данный экзон (левый потомок) и не содержащие его (правый потомок). Каждому узлу сопоставляется свой набор сплайсинг-факторов, чьи мотивы присутствуют на соответствующем вариабельном экзоне. Процедура повторяется рекурсивно до тех пор, пока каждый лист не будет соответствовать отдельному транскрипту. На каждом внутреннем узле дерева обучается отдельная модель бета-регрессии, предсказывающая условную вероятность включения соответствующего вариабельного экзона. Таким образом, итоговая вероятность каждой изоформы восстанавливается как произведение условных вероятностей вдоль пути от соответствующего листа к корню. Формально: пусть $\pi(t)$ — множество внутренних узлов дерева на пути от корня до листа, соответствующего изоформе t; $\widehat{\mu}_v$ — предсказанное моделью бета-регрессии среднее в узле v (условная вероятность включения соответствующего вариабельного экзона); $b(t, v) \in \{0, 1\}$ — индикатор того, содержит ли изоформа t экзон узла v. Тогда итоговая вероятность изоформы t имеет вид

$$P(t) = \prod_{v \in \pi(t)} \widehat{\mu}_v^{b(t, v)} (1 - \widehat{\mu}_v)^{1 - b(t, v)}. \tag{7}$$

В отличие от первого подхода данный метод явно позволяет идентифицировать факторы, специфичные для включения конкретного вариабельного экзона, а не изоформы в целом.

Для сопоставления эффективности описанных подходов была проведена оценка качества моделей. Выборка предварительно делилась на обучающую и валидационную, размеры которых составляли 264 и 67 пациентов соответственно. Метриками качества были выбраны среднеабсолютная ошибка и корреляция Пирсона. По каждому подходу подбор гиперпараметров осуществлялся с приоритетом максимизации корреляции Пирсона при одновременном учете среднеабсолютной ошибки. Процедура оптимизации выполнялась в два этапа для сужения пространства поиска и экономии вычислительных ресурсов. Сначала, на этапе предварительной настройки, проводился локальный анализ чувствительности модели к темпу обучения: путем варьирования параметра определялись критические границы, за пределами которых метрики качества скачкообразно ухудшались. После выявления зоны стабильности применялся решетчатый поиск. Для темпа обучения (lr) перебор осуществлялся в интервале от 10^{-2} до 10^{-1} с шагом 10^{-2}, а для коэффициентов регуляризации γ_1 и γ_2 рассматривались все возможные комбинации значений из множества $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Таким образом, для первого подхода гиперпараметры равны $lr = 4{,}5 \cdot 10^{-2}$, $\gamma_1 = 10^{-2}$, $\gamma_2 = 10^{-2}$, в то время как для второго — $lr = 5 \cdot 10^{-2}$, $\gamma_1 = 10^{-2}$, $\gamma_2 = 10^{-2}$.

В дальнейшем также в топ-15 отбирались белки по абсолютному значению предсказанных весов. Вместе с этим для статистической верификации полученных наиболее важных признаков была проведена оценка значимости ассоциаций между экспрессией сплайсинг-факторов, ис-

пользованных в модели для прогнозирования долей изоформ, и наблюдаемыми значениями этих долей с использованием коэффициента ранговой корреляции Спирмена.

Нулевая гипотеза состояла в отсутствии монотонной зависимости между экспрессией сплайсинг-фактора и уровнем экспрессии соответствующей изоформы (H_0: $\rho = 0$), альтернативная гипотеза — в наличии статистически значимой монотонной ассоциации (H_1: $\rho \neq 0$). Для каждой пары «сплайсинг-фактор – изоформа» вычислялся коэффициент корреляции Спирмена, после чего полученные p-значения корректировались с учетом множественного тестирования гипотез методом Бенджамини – Хохберга.

Дополнительно для сопоставления эффективности предложенной реализации бета-регрессии было проведено ее сравнение с двумя реализациями, ранее представленными в литературе. Первая литературная модель бета-регрессии — имплементация в R, предложенная в [Cribari-Neto, Zeileis, 2010]. Для ее запуска использовались R версии 4.6.0 и библиотека Betareg версии 3.2.4, параметры обучения были взяты стандартными. Второй литературной версией была выбрана модель, описанная в работе [Novosad, 2023], реализованная на языке программирования Python, с числом эпох на обучение, равным 200. Предложенная же в текущей работе реализация была взята с параметрами, указанными раннее. Для простоты эксперимента данные регрессии были применены к модели one-versus-all. При их сравнении фиксировалось четыре параметра: значения корреляции Пирсона на обучающей и валидационной выборках, значение метрики MAE на валидационной выборке, а также время обучения. Для сравнения моделей по этим метрикам вводится функционал, который высчитывается отдельно для каждой изоформы:

$$F = \frac{P_{val} \cdot (1 - (P_{train} - P_{val})) \cdot (1 - MAE_{val})}{\log_{10}(T + 10)},$$

где P_{val}, P_{train} — значения корреляции Пирсона на валидационной и обучающей выборках соответственно, MAE_{val} — значение метрики MAE на валидационной выборке, T — время. Выбор такого функционала обусловлен сразу несколькими причинами. Для начала первоочередной задачей выступает выявление связей между признаками и целевой переменной, поэтому важную роль играет значение корреляции Пирсона на валидационной выборке. Помимо этого, необходима также стабильность: если при обучении наблюдается высокое значение корреляции, в то время как на тесте оно падает, то это свидетельствует о переобучении данной модели. Именно за описание стабильности и отвечает второй множитель в числителе. Третий множитель отвечает за точность восстановления вероятностей: чем сильнее среднеабсолютное отклонение на валидационной выборке, тем меньше значение метрики. Логарифм в знаменателе выбран для того, чтобы учесть, как долго происходит процесс обучения модели. При этом значения, отличающиеся друг от друга на несколько секунд, должны иметь достаточно близкие значения функционала, что и гарантирует логарифм. Поправка в логарифме нужна для того, чтобы гарантировать $\log_{10} > 1$, тем самым не давая функционалу уходить на бесконечность. Непосредственно из вида функционала видно, что чем лучше модель по указанным параметрам, тем больше значение.

Результаты и обсуждение

В таблице 1 представлены результаты тестирования трех имплементаций по разным изоформам CD44.

Тогда, вычисляя значения предложенного раннее функционала и усредняя его значения по изоформам, мы получим $F_R = 0{,}349$, $F_{Novosad} = 0{,}244$, $F_{neural} = 0{,}559$. Из результатов, представленных в таблице 1, и значений функционала видно, что нейросетевая реализация бета-регрессии, хоть и уступает в некоторых случаях по значениям метрики MAE_{val} другим имплементациям, в совокупности является более удачным подходом в рамках поставленной задачи.

Таблица 1. Сравнение трех реализаций бета-регрессии (R, Novosad и neural) на уровне изоформ гена CD44. Для каждой изоформы представлены метрики качества предсказания на валидационной выборке (MAE_{val}), коэффициенты корреляции Пирсона на обучающей и валидационной выборках (P_{train} и P_{test}), а также время обучения моделей (время (с)), отражающее вычислительную сложность подходов. Модели R и Novosad соответствуют ранее упомянутым литературным методам, тогда как neural является разработанной в рамках текущей работы нейросетевой реализацией бета-регрессии

Изоформа	Модель	MAE_{val}	P_{train}	P_{val}	Время (с)
CD44s	R	0,0507	0,9236	0,7167	0,135
	Novosad	0,0418	0,8485	0,7994	120,826
	neural	0,0510	0,7889	0,7539	1,210
CD44v3-10	R	0,0324	0,8501	0,2453	0,116
	Novosad	0,0231	0,7429	0,5594	94,992
	neural	0,0835	0,6546	0,5893	0,458
CD44v8-10	R	0,0446	0,9005	0,7455	0,099
	Novosad	0,0438	0,8316	0,7471	73,147
	neural	0,0533	0,8607	0,8005	0,522
CD44v6-10	R	0,0167	0,7719	0,6151	0,061
	Novosad	0,0166	0,7615	0,6814	174,166
	neural	0,0848	0,6793	0,6258	0,392
CD44v8-9	R	0,0128	0,7289	0,4736	0,069
	Novosad	0,0117	0,7108	0,5399	189,791
	neural	0,0370	0,8354	0,7857	0,434

При сравнении подходов к выявлению важных РНК-связывающих белков необходимо учитывать фундаментальное различие не только в форматах их целевых переменных (нормированные вероятности против независимых оценок), но и в механизмах формирования пространства признаков.

Чтобы исключить вероятность того, что разница в качестве предсказаний вызвана исключительно эффектом нормировки в древесной модели (где сумма долей изоформ сводится к единице), мы проанализировали предсказательную способность дерева на уровне его индивидуальных ненормированных узлов ветвления. Результаты показали, что иерархический подход изначально не способен стабильно улавливать биологический сигнал на исследуемых данных: несмотря на удовлетворительные показатели на обучающей выборке, на валидационной выборке корреляция Пирсона для промежуточных узлов (предсказывающих условную вероятность включения экзона) падает до околонулевых или отрицательных значений (в диапазоне от −0,15 до 0,02).

Столь существенное падение обобщающей способности объясняется архитектурными ограничениями древесной модели. Согласно логике построения дерева на каждом внутреннем узле в качестве признаков используются исключительно те сплайсинг-факторы, мотивы которых локализованы на соответствующем вариабельном экзоне. Такое жесткое сужение признакового пространства приводит к потере глобального контекста регуляции. Сплайсинг является сложным процессом, где на включение экзона могут влиять факторы, связывающиеся с соседними интронами или другими участками транскрипта. Локальная ограниченность признаков приводит к тому, что модели в узлах дерева переобучаются на малом объеме данных и теряют предсказательную силу.

Как следствие, при сборке финального прогноза происходит лавинообразное накопление ошибки. Поскольку итоговая вероятность каждой изоформы вычисляется как произведение условных вероятностей вдоль пути от корня к листу, низкое качество ненормированных прогнозов в узлах мультиплицируется. Это отразилось в крайне низких метриках для листьев дерева: валидационная корреляция Пирсона составила в среднем лишь 0,12 (от −0,04 до 0,38 для отдельных сплайс-вариантов), а средняя абсолютная ошибка (MAE) — 0,37.

В противовес этому подход *one versus all* позволяет рассматривать регуляцию каждого сплайс-варианта независимо, используя в качестве признаков факторы, мотивы которых присутствуют в последовательности транскрипта целиком. Сохранение полного биологического контекста и снятие жесткого ограничения на иерархическую нормировку позволили достичь принципиально иного уровня точности. По результатам, отраженным в таблице 1, значения корреляции Пирсона для OVA-модели лежат в диапазоне от 0,58 до 0,80, а MAE снизилась до значений от 0,037 до 0,085.

Опираясь на убедительное превосходство метода «один против всех» как математически, так и биологически, именно он был выбран для итогового анализа. Полные списки топ-15 признаков для каждой изоформы, а также результаты проверки статистической значимости полученных результатов приведены в приложении.

Анализ полученных наборов важных признаков позволил идентифицировать РНК-связывающие белки, вовлеченные в регуляцию сплайсинга гена CD44. Важно подчеркнуть, что для всех обсуждаемых далее признаков связь между экспрессией фактора и долей изоформы является статистически достоверной ($p_{adj} < 0{,}05$). Для удобства интерпретации в тексте приводятся два показателя: вклад признака в предсказание многомерной модели (вес w) и коэффициент ранговой корреляции Спирмена ρ. Стоит отметить, что для некоторых признаков знак предсказанного веса в модели может не совпадать со знаком коэффициента корреляции Спирмена. Это ожидаемый эффект, обусловленный тем, что ρ отражает изолированную (маржинальную) связь между фактором и изоформой, в то время как веса w извлекаются из многомерной модели и учитывают коэкспрессию сплайсинг-факторов (мультиколлинеарность), их взаимодействия и взаимоисключающий характер долей изоформ.

Более детально среди признаков были отобраны следующие регуляторы.

Наиболее весомым примером является *HNRNPL* ($w = 0{,}064$, $\rho = 0{,}470$), ведущий положительный предиктор изоформы CD44v8–9. HNRNPL является каноническим регулятором альтернативного сплайсинга CD44: он связывается с CA-богатыми интронными элементами, фланкирующими вариабельные экзоны, и способствует их включению [Hui et al., 2005; Heiner et al., 2010]. Для этой же изоформы получилось, что *YBX1* несет отрицательный вес ($w = -0{,}035$, $\rho = 0{,}205$), что на первый взгляд противоречит известной роли YBX1 как активатора включения экзона v4 [Stickeler et al., 2001]. Однако это противоречие снимается при учете экзонного состава: CD44v6–10 не содержит v4, поэтому высокая активность YBX1 конкурентно перераспределяет сплайсинг в пользу v4-содержащих изоформ, снижая долю CD44v.

Для изоформ CD44v8–10 и CD44v8–9 белок *MBNL1* выступает отрицательным регулятором (с весами и корреляциями $w = -0{,}064$, $\rho = -0{,}321$ и $w = -0{,}016$, $\rho = -0{,}325$ соответственно), что также сходится с биологией: MBNL1 известен как супрессор ЭМП-ассоциированного (эпителиально-мезенхимальный переход, приобретение подвижности и инвазивности) альтернативного сплайсинга в колоректальном раке [Oltean, Bates, 2014; Navvabi et al., 2021]. В дифференцированных эпителиальных клетках он подавляет включение вариабельных экзонов, тогда как их снижение при ЭМП позволяет CD44-вариантным изоформам накапливаться. Наконец, наибольший вклад в регуляцию изоформ с расширенной вариабельной частью (CD44v3–10 и CD44v6–10; со значениями весов и корреляции $w = 0{,}108$, $\rho = 0{,}436$ и $w = 0{,}061$, $\rho = 0{,}433$ соответственно) вносит белок *HNRNPF*. Известно, что HNRNPF связывается с G-квадруплексными структурами в пре-мРНК CD44 и способствует включению вариабельных экзонов при ЭМП, активируя переход от стандартной изоформы CD44s к вариантным [Huang et al., 2017].

Помимо доказанных раннее сплайсинг-факторов, модель также смогла выявить и те регуляторы, чьи функции в вопросе сплайсинга CD44 еще не до конца изучены. К таким, например, относится белок *QKI*, отобранный для CD44v8–10 ($w = -0{,}078$, $\rho = -0{,}225$) и в меньшей степени

для CD44v6–10 (w = −0,024, ρ = −0,173), роль которого связывают с супрессией онкозаболеваний, однако его участие в сплайсинге CD44 остается недостаточно изученным [Maltseva, Tonevitsky, 2023; Zhu et al., 2024]. Помимо QKI, еще был обнаружен *TARDBP* (TDP-43) для сплайс-варианта CD44v8–10 (w = −0,057, ρ = 0,184), который известен своей регуляцией альтернативного сплайсинга в рамках колоректального рака [Ma et al., 2021], а также подавлением вариабельных экзонов в сплайсинге CD44 в раке груди [Guo et al., 2022].

Говоря о потенциально новых регуляторах, здесь можно выделить три сплайсинг-фактора.

Первым кандидатом выступает белок *NUDT21*, отобранный моделью как регулятор CD44s (w = −0,061, ρ = −0,572) и CD44v8–9 (w = −0,018, ρ = 0,169). NUDT21 контролирует выбор между проксимальным и дистальным сайтом полиаденилирования и охарактеризован как ингибитор роста глиобластомы [Masamha et al., 2014]. Отрицательные веса для обеих изоформ позволяют предположить, что активность данного белка смещает баланс транскриптов CD44 в пользу изоформ с коротким вариабельным регионом (v2–v5); однако эта гипотеза требует экспериментальной проверки и прямого анализа APA-событий в локусе CD44.

Вторым потенциальным регулятором выступает *AGO2* (w = −0,028, ρ = −0,534 для CD44s). Argonaute-2 является центральным эффектором RISC-комплекса [Hammond et al., 2001] и традиционно рассматривается как медиатор миРНК-зависимого сайленсинга. Вместе с тем показано, что AGO2 способен регулировать альтернативный сплайсинг котранскрипционно [Alló et al., 2009]. Его отрицательная ассоциация с CD44s открывает гипотезу о том, что AGO2-зависимая активность способствует переключению со стандартной на вариантные изоформы CD44; конкретные экзоны-мишени и механизм такого переключения, аналогично NUDT21, требуют экспериментальной верификации.

Третьим кандидатом является *ACO1* (w = 0,034, ρ = 0,126 для CD44v8–10). ACO1 (цитозольная аконитаза 1, IRP1) — бифункциональный белок: при достаточном уровне железа он функционирует как фермент цикла Кребса, тогда как при его дефиците переключается в режим РНК-связывающего белка, регулируя стабильность и трансляцию мРНК через железо-чувствительные элементы [Rouault, 2006]. Гиперэкспрессия ACO1 зафиксирована в опухолях прямой кишки [Choi et al., 2011], однако его причастность к регуляции альтернативного сплайсинга ранее не описывалась. Положительная ассоциация с CD44v8–10 позволяет предположить существование оси «метаболический статус железа – сплайсинг CD44», что представляет наибольший интерес для дальнейшего экспериментального изучения среди всех трех кандидатов по совокупности причин.

Во-первых, NUDT21 и AGO2 являются признанными регуляторами РНК-метаболизма, и их связь со сплайсингом, пусть и не верифицированная для CD44, методологически ожидаема; ACO1 же никогда не рассматривался в данном контексте. Во-вторых, переключение ACO1 между ферментативным и РНК-связывающим режимами определяется уровнем внутриклеточного железа, который в опухолевом микроокружении системно нарушен [Torti, Torti, 2013], что создает предпосылки для зависящей от метаболического контекста динамической регуляции изоформного состава CD44.

Заключение

В настоящей работе предложен подход к анализу альтернативного сплайсинга, основанный на моделировании долей изоформ с помощью бета-регрессии, реализованной в виде нейросетевого регрессора с совместной оценкой параметров среднего и точности. Ключевая идея метода заключается в том, что относительные уровни экспрессии изоформ рассматриваются как реализации бета-распределения, параметры которого зависят от экспрессии РНК-связывающих белков. В отличие от классических статистических инструментов (таких как стандартные пакеты бета-регрессии среды R) и существующих профильных подходов предложенная нейросетевая

архитектура обеспечивает высокую вычислительную стабильность на многомерных данных экспрессии. Это позволяет эффективно оценивать контекстно зависимый системный вклад каждого регулятора с учетом их взаимного влияния и коэкспрессии.

Корректность предложенной реализации была подтверждена на синтетических данных, где показаны отсутствие систематического смещения оценок и приемлемая точность восстановления параметров распределения. Применение модели к данным TCGA для гена CD44 в клетках колоректального рака позволило сопоставить два подхода к постановке задачи: схему «один против всех», в которой для каждой изоформы независимо строится отдельная модель, и метод дерева изоформ, основанный на иерархической декомпозиции транскриптов по вариабельным экзонам с последующим моделированием условных вероятностей их включения.

Установлено, что схема «один против всех» существенно превосходит древесный подход по качеству предсказания. Это позволило использовать данный подход для последующего анализа значимости сплайсинг-факторов.

Полученные результаты строго верифицированы статистически и согласуются с известными биологическими данными: среди значимых признаков обнаружены ранее описанные регуляторы сплайсинга CD44 (HNRNPL, MBNL1, HNRNPF, YBX1). Одновременно модель выявила ряд потенциально новых кандидатов, включая ACO1, NUDT21 и AGO2. Наибольший интерес представляет белок ACO1, для которого впервые выдвинута гипотеза о связи между внутриклеточным метаболизмом железа и регуляцией изоформного состава CD44.

В дальнейшем планируется расширение анализа на другие типы опухолей с целью выявления общих и специфических регуляторов сплайсинга CD44. Предусматривается проведение сравнительных исследований между различными раковыми тканями для оценки универсальности выявленных закономерностей и идентификации контекстно зависимых сплайсинг-факторов. Такой подход позволит уточнить механизмы регуляции альтернативного сплайсинга в различных онкологических моделях и выявить потенциальные терапевтические мишени.

Благодарности

Автор выражает благодарность своему научному руководителю Алексею Владимировичу Галатенко, а также Антону Павловичу Жиянову за методологические рекомендации и плодотворное обсуждение результатов работы. Автор также глубоко признателен главному редактору журнала Алексею Ивановичу Лобанову и рецензенту за внимательное отношение к рукописи и ценные замечания, способствовавшие улучшению статьи.

Список литературы (References)

Abramowitz M., Stegun I. A. (eds.) Handbook of mathematical functions with formulas, graphs, and mathematical tables. — Washington, D.C.: National Bureau of Standards, 1964. — 1046 p.

Alló M., Buggiano V., Fededa J. P., Petrillo E., Schor I., de la Mata M., Agirre E., Plass M., Eyras E., Elela S. A., Klinck R., Chabot B., Kornblihtt A. R. Control of alternative splicing through siRNA-mediated transcriptional gene silencing // Nat. Struct. Mol. Biol. — 2009. — Vol. 16, No. 7. — P. 717–724. — DOI: 10.1038/nsmb.1620

Black D. L. Mechanisms of alternative pre-messenger RNA splicing // Annu. Rev. Biochem. — 2003. — Vol. 72. — P. 291–336. — DOI: 10.1146/annurev.biochem.72.121801.161720

Choi S. Y., Jang J. H., Kim K. R. Analysis of differentially expressed genes in human rectal carcinoma using suppression subtractive hybridization // Clin. Exp. Med. — 2011. — Vol. 11, No. 4. — P. 219–226. — DOI: 10.1007/s10238-010-0130-5

Cribari-Neto F., Zeileis A. Beta regression in R // J. Stat. Softw. — 2010. — Vol. 34, No. 2. — P. 1–24. — DOI: 10.18637/jss.v034.i02

Ferrari S. L. P., Cribari-Neto F. Beta regression for modelling rates and proportions // J. Appl. Stat. — 2004. — Vol. 31, No. 7. — P. 799–815. — DOI: 10.1080/0266476042000214501

Frankish A., Diekhans M., Jungreis I., Lagarde J., Loveland J. E., Mudge J. M., Sisu C., Wright J. C., Armstrong J., Barnes I., Berry A., Bignell A., Boix C., Carbonell Sala S., Cunningham F., Di Domenico T., Donaldson S., Fiddes I. T., García Girón C., Gonzalez J. M., Grego T., Hardy M., Hourlier T., Howe K. L., Hunt T., Izuogu O. G., Johnson R., Martin F. J., Martínez L., Mohanan S., Muir P., Navarro F. C. P., Parker A., Pei B., Pozo F., Riera F. C., Ruffier M., Schmitt B. M., Stapleton E., Suner M.-M., Sycheva I., Uszczynska-Ratajczak B., Wolf M. Y., Xu J., Yang Y. T., Yates A., Zerbino D., Zhang Y., Choudhary J. S., Gerstein M., Guigó R., Hubbard T. J. P., Kellis M., Paten B., Tress M. L., Flicek P. GENCODE 2021 // Nucleic Acids Res. — 2021. — Vol. 49, No. D1. — P. D916–D923. — DOI: 10.1093/nar/gkaa1087

Fu X. D., Ares M. Context-dependent control of alternative splicing by RNA-binding proteins // Nat. Rev. Genet. — 2014. — Vol. 15, No. 10. — P. 689–701. — DOI: 10.1038/nrg3778

Giudice G., Sánchez-Cabo F., Torroja C., Lara-Pezzi E. ATtRACT — a database of RNA-binding proteins and associated motifs // Database. — 2016. — Vol. 2016. — P. baw035. — DOI: 10.1093/database/baw035

Giulietti M., Piva F., D'Antonio M., D'Onorio De Meo P., Paoletti D., Castrignanò T., D'Erchia A. M., Picardi E., Zambelli F., Principato G., Pavesi G., Pesole G. SpliceAid-F: a database of human splicing factors and their RNA-binding sites // Nucleic Acids Res. — 2013. — Vol. 41. — P. D125–D131. — DOI: 10.1093/nar/gks997

Glorot X., Bordes A., Bengio Y. Deep sparse rectifier neural networks // Proc. 14th Int. Conf. Artif. Intell. Stat. (AISTATS). — 2011. — Vol. 15. — P. 315–323.

Guo L., Ke H., Zhang H., Zou L., Yang Q., Lu X., Zhao L., Jiao B. TDP43 promotes stemness of breast cancer stem cells through CD44 variant splicing isoforms // Cell Death Dis. — 2022. — Vol. 13, No. 5. — P. 428. — DOI: 10.1038/s41419-022-04867-w

Hammond S. M., Boettcher S., Caudy A. A., Kobayashi R., Hannon G. J. Argonaute2, a link between genetic and biochemical analyses of RNAi // Science. — 2001. — Vol. 293, No. 5532. — P. 1146–1150. — DOI: 10.1126/science.1064023

Hassn Mesrati M., Syafruddin S. E., Mohtar M. A., Syahir A. CD44: a multifunctional mediator of cancer progression // Biomolecules. — 2021. — Vol. 11, No. 12. — P. 1850. — DOI: 10.3390/biom11121850

Heiner M., Hui J., Schreiner S., Hung L. H., Bindereif A. HnRNP L-mediated regulation of mammalian alternative splicing by interference with splice site recognition // RNA Biol. — 2010. — Vol. 7, No. 1. — P. 56–64. — DOI: 10.4161/rna.7.1.10402

Huang H., Zhang J., Harvey S. E., Hu X., Cheng C. RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF // Genes Dev. — 2017. — Vol. 31, No. 22. — P. 2296–2309. — DOI: 10.1101/gad.305862.117

Hui J., Hung L.-H., Heiner M., Schreiner S., Neumüller N., Reither G., Haas S. A., Bindereif A. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing // EMBO J. — 2005. — Vol. 24, No. 11. — P. 1988–1998. — DOI: 10.1038/sj.emboj.7600677

Kieschnick R., McCullough B. D. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions // Stat. Model. — 2003. — Vol. 3, No. 3. — P. 193–213. — DOI: 10.1191/1471082X03st053oa

Kingma D. P., Ba J. Adam: a method for stochastic optimization // 3rd Int. Conf. Learn. Representations (ICLR 2015). — San Diego, 2015. — DOI: arXiv:1412.6980

LeCun Y., Bottou L., Orr G. B., Müller K.-R. Efficient BackProp // Neural Networks: Tricks of the Trade / ed. by G. B. Orr, K.-R. Müller. — Springer, 1998. — P. 9–50. — DOI: 10.1007/3-540-49430-8_2

Li J., Pan T., Chen L., Wang Q., Chang Z., Zhou W., Li X., Xu G., Li X., Li Y., Zhang Y. Alternative splicing perturbation landscape identifies RNA binding proteins as potential therapeutic targets in cancer // Mol. Ther. Nucleic Acids. — 2021. — Vol. 24. — P. 792–806. — DOI: 10.1016/j.omtn.2021.04.005

Ma X., Ying Y., Xie H., Liu X., Wang X., Li J. The regulatory role of RNA metabolism regulator TDP-43 in human cancer // Front. Oncol. — 2021. — Vol. 11. — P. 755096. — DOI: 10.3389/fonc.2021.755096

Maltseva D., Tonevitsky A. RNA-binding proteins regulating the CD44 alternative splicing // Front. Mol. Biosci. — 2023. — Vol. 10. — P. 1326148. — DOI: 10.3389/fmolb.2023.1326148

Marasco L. E., Kornblihtt A. R. The physiology of alternative splicing // Nat. Rev. Mol. Cell Biol. — 2023. — Vol. 24. — P. 242–254. — DOI: 10.1038/s41580-022-00545-z

Masamha C. P., Xia Z., Yang J., Albrecht T. R., Li M., Shyu A.-B., Li W., Wagner E. J. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression // Nature. — 2014. — Vol. 510, No. 7505. — P. 412–416. — DOI: 10.1038/nature13261

Mishra M. N., Chandavarkar V., Sharma R., Bhargava D. Structure, function and role of CD44 in neoplasia // J. Oral Maxillofac. Pathol. — 2019. — Vol. 23, No. 2. — P. 267–272. — DOI: 10.4103/jomfp.JOMFP_246_18

Navvabi N., Kolikova P., Hosek P., Zitricky F., Navvabi A., Vycital O., Bruha J., Palek R., Rosendorf J., Liska V., Pitule P. Altered expression of MBNL family of alternative splicing factors in colorectal cancer // Cancer Genomics Proteomics. — 2021. — Vol. 18, No. 3. — P. 295–306. — DOI: 10.21873/cgp.20260

Novosad V. Identification of significant RNA-binding proteins in the process of CD44 splicing using the boosted beta regression algorithm // Dokl. Biochem. Biophys. — 2023. — Vol. 510, No. 1. — P. 99–103. — DOI: 10.1134/S1607672923700199

Oltean S., Bates D. O. Hallmarks of alternative splicing in cancer // Oncogene. — 2014. — Vol. 33, No. 46. — P. 5311–5318. — DOI: 10.1038/onc.2013.533

Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., Köpf A., Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. PyTorch: an imperative style, high-performance deep learning library // Adv. Neural Inf. Process. Syst. — 2019. — Vol. 32. — P. 8024–8035. — DOI: arXiv:1912.01703

Prochazka L., Tesarik R., Turanek J. Regulation of alternative splicing of CD44 in cancer // Cell. Signal. — 2014. — Vol. 26, No. 10. — P. 2234–2239. — DOI: 10.1016/j.cellsig.2014.07.011

Rouault T. A. The role of iron regulatory proteins in mammalian iron homeostasis and disease // Nat. Chem. Biol. — 2006. — Vol. 2, No. 8. — P. 406–414. — DOI: 10.1038/nchembio807

Smithson M., Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables // Psychol. Methods. — 2006. — Vol. 11, No. 1. — P. 54–71. — DOI: 10.1037/1082-989X.11.1.54

Stickeler E., Fraser S. D., Honig A., Chen A. L., Berget S. M., Cooper T. A. The RNA binding protein YB-1 binds A/C-rich exon enhancers and stimulates splicing of the CD44 alternative exon v4 // EMBO J. — 2001. — Vol. 20, No. 14. — P. 3821–3830. — DOI: 10.1093/emboj/20.14.3821

Torti S. V., Torti F. M. Iron and cancer: more ore to be mined // Nat. Rev. Cancer. — 2013. — Vol. 13, No. 5. — P. 342–355. — DOI: 10.1038/nrc3495

Vasconcellos K. L. P., Cribari-Neto F. Improved maximum likelihood estimation in a new class of beta regression models // Braz. J. Probab. Stat. — 2005. — Vol. 19, No. 1. — P. 13–31.

Weinstein J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project // Nat. Genet. — 2013. — Vol. 45, No. 10. — P. 1113–1120. — DOI: 10.1038/ng.2764

Yan Y., Zuo X., Wei D. Concise review: emerging role of CD44 in cancer stem cells: a promising biomarker and therapeutic target // Stem Cells Transl. Med. — 2015. — Vol. 4, No. 9. — P. 1033–1043. — DOI: 10.5966/sctm.2015-0048

Zhu W., Yang W., Sun G., Huang J. RNA-binding protein quaking: a multifunctional regulator in tumour progression // Ann. Med. — 2024. — Vol. 57, No. 1. — P. 2443046. — DOI: 10.1080/07853890.2024.2443046

Zou H., Hastie T. Regularization and variable selection via the elastic net // J. R. Stat. Soc. B. — 2005. — Vol. 67, No. 2. — P. 301–320. — DOI: 10.1111/j.1467-9868.2005.00503.x