**COMPUTER RESEARCH AND MODELING**
**2026 VOL. 18 NO. 3 P. e697–e714**
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. e697–e714.

© 2026 Alexey A. Pirogov
This work is licensed under the Creative Commons Attribution-NoDerivs 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by-nd/3.0/
or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ**
**И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 3 С. e697–e714**
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

© 2026 Алексей Алексеевич Пирогов
Статья доступна по лицензии Creative Commons Attribution-NoDerivs 3.0 Unported License.
Чтобы получить текст лицензии, посетите веб-сайт http://creativecommons.org/licenses/by-nd/3.0/
или отправьте письмо в Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

## Introduction

Alternative splicing is a key mechanism of post-transcriptional gene regulation in eukaryotes. Genomic DNA contains two types of sequence elements: coding regions, known as exons, and non-coding intervening regions, known as introns. During pre-mRNA processing, introns are removed and exons are joined to form a mature mRNA molecule. In alternative splicing, this process can proceed in different ways: individual exons may be included in or excluded from the final transcript, allowing a single gene to encode multiple protein isoforms with distinct structural and functional properties [Marasco, Kornblihtt, 2023]. Dysregulation of alternative splicing is closely associated with a broad range of diseases, especially cancer [Li et al., 2021].

RNA-binding proteins (splicing factors) play a central role in the regulation of alternative splicing: they bind to pre-mRNA and influence splice-site selection. Their activity relies on the recognition of specific cis-regulatory elements, including splicing enhancers and silencers, followed by modulation of spliceosome assembly [Black, 2003; Fu, Ares, 2014]. Because several RNA-binding proteins may act jointly on the same transcript in a context-dependent manner, predicting splicing outcomes directly from the genomic sequence remains difficult [Fu, Ares, 2014]. The identification of relevant regulators for a given gene therefore remains an important and methodologically challenging problem [Li et al., 2021].

Among alternatively spliced genes, CD44 is of particular interest in cancer biology. It encodes a transmembrane glycoprotein whose isoforms differ in the extracellular domain: ten variable exons (v1–v10) can be included or skipped in different combinations, and exon v1 is known to be absent from human CD44 transcripts [Mishra et al., 2019]. CD44 alternative splicing is frequently disrupted in cancer and produces isoforms with tissue-specific properties and distinct effects on tumor progression [Prochazka et al., 2014]. The standard isoform, CD44s, contains no variable exons and functions as a hyaluronic acid receptor and growth factor co-receptor, integrating microenvironmental signals and regulating cell adhesion, migration, proliferation, and survival [Yan et al., 2015; Hassn Mesrati et al., 2021]. Variant isoforms, collectively referred to as CD44v, have been identified as markers of cancer stem cells and are associated with chemoresistance and metastasis [Yan et al., 2015].

Identifying RNA-binding proteins that regulate CD44 splicing is therefore important both biologically and clinically. From a basic research perspective, the regulatory landscape of CD44 splicing remains incompletely understood: the inclusion of variable exons in different combinations is governed by the joint action of several splicing factors, producing a complex context-dependent system whose mechanisms in cancer are still largely unclear. From a translational perspective, the regulators identified in such analyses may become therapeutic targets: a directed shift in isoform balance, in particular from CD44v variants toward the standard CD44s isoform, could potentially weaken the cancer stem cell phenotype, reduce chemoresistance, and limit metastatic potential [Yan et al., 2015; Mishra et al., 2019]. A natural way to analyze isoform expression is to normalize each isoform by total gene expression; the resulting proportions represent the relative contribution of each isoform and lie in the interval (0, 1). The task can then be stated as follows: given the expression levels of RNA-binding proteins, estimate a linear relationship that quantifies the influence of each splicing factor on the proportion of a particular isoform. Because of the bounded nature of the response variable, standard linear models are not appropriate here [Kieschnick, McCullough, 2003]. A more natural choice is

beta regression [Ferrari, Cribari-Neto, 2004], which assumes a beta-distributed response and explicitly accounts for the bounded support of proportions.

This approach was applied to CD44 splicing in colorectal cancer cells in [Novosad, 2023]: beta regression was used to identify important RNA-binding proteins, with the target variable defined as the proportion of the CD44v8–10 isoform relative to the sum of CD44v8–10 and CD44s, the two most highly expressed isoforms in this tissue type.

The present study also focuses on CD44 in colorectal cancer cells, but extends the analysis to a broader set of CD44 isoforms. Two strategies for identifying significant splicing factors are proposed and compared: a *one-versus-all* scheme, in which a separate beta regression model is fitted independently for each isoform, and an *isoform tree* method, based on a hierarchical decomposition by variable exons. In addition, a new implementation of beta regression is introduced. Unlike the classical statistical formulation [Ferrari, Cribari-Neto, 2004], the model is implemented as a PyTorch-based neural network regressor [Paszke et al., 2019] with built-in Elastic Net regularization [Zou, Hastie, 2005] for feature selection. The implementation is validated on synthetic data and then applied to the identification of CD44 splicing regulators, including several new candidate factors.

## Theoretical model of beta regression

We first recall the basic definitions used in the model.

**Definition 1.** A random variable $\xi$ is said to follow a beta distribution with parameters $\alpha$, $\beta$ (denoted by $\xi \sim \mathcal{B}(\alpha, \beta)$) if its probability density function is given by

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

where $\mathcal{B}(\alpha, \beta) = \int\limits_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ is the beta function;

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

is the indicator function.

Note that for $\xi \sim \mathcal{B}(\alpha, \beta)$ the expectation and the variance are, respectively,

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

Let $y_1, \ldots, y_N \in (0,\,1)$ be the observed values of the target (predicted) variable. In our model it is assumed that $y_i \sim \mathcal{B}(\alpha_i, \beta_i)$ for all $i \in \{1, \ldots, N\}$, where $N$ is the number of observations.

**Definition 2.** For a sample $\xi_1, \ldots, \xi_n$, the likelihood function $L(x_1, \ldots, x_n)$, denoted below by $L(\overline{x})$, is defined as

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

where $f_{\xi_i}(x)$ denotes the density, or probability mass function in the discrete case, of the random variable $\xi_i$.

It is often more convenient to work with $\ln L_1(\overline{x})$ rather than $L_1(\overline{x})$ itself. Since the logarithm is strictly increasing, this transformation preserves the extrema of the likelihood. The log-likelihood function for the beta regression model is therefore

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

Below, $-L(\overline{x})$ is used as the loss function.

Using $\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, where $\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}\,dx$ is the gamma function [Abramowitz, Stegun, 1964], the likelihood can be rewritten as

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

We now introduce the reparameterization commonly used in beta regression [Ferrari, Cribari-Neto, 2004; Vasconcellos, Cribari-Neto, 2005].

For convenience and computational simplicity, beta regression is usually expressed in terms of distributional moments, such as the mean and variance, rather than directly in terms of $\alpha$ and $\beta$ [Smithson, Verkuilen, 2006]. We therefore reparameterize the distribution.

Let $\mu_i = \frac{\alpha_i}{\alpha_i+\beta_i}$ and $\phi_i = \alpha_i + \beta_i$. The corresponding moments are

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

Under this substitution, the likelihood becomes

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

The beta regression model estimates both parameters of the reparameterized distribution: $\mu_i$ (the mean) and $\phi_i$ (the precision). We use the following notation: $X \in \mathbb{R}^{N\times m}$ is the feature matrix for the mean ($m$ features and $N$ observations); $\omega \in \mathbb{R}^m$ is the corresponding weight vector; $Z \in \mathbb{R}^{N\times l}$ is the feature matrix for the precision parameter ($l$ features, not necessarily equal to $m$, and $N$ observations); and $\eta \in \mathbb{R}^l$ is the corresponding weight vector.

It is assumed that

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

where $g_1 : (0, 1) \to \mathbb{R}$ is a strictly monotone differentiable link function. The precision parameter can either be modeled analogously through a strictly monotone differentiable link function $g_2 : (0, +\infty) \to \to \mathbb{R}$ or treated as constant across observations, $\phi_i \equiv \phi = \text{const}$ [Cribari-Neto, Zeileis, 2010]. In this model, $g_1(x) = \text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$, so $g_1^{-1}(x) = \sigma(x) = \frac{1}{1+e^{-x}}$.

The regression problem is then to find $\omega$ and $\eta$ such that

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

## Implementation and justification of its correctness

In the practical implementation, we set $X = Z$ and use $g_2(x) = \ln(e^x - 1)$ as the link function for the precision parameter, so that $g_2^{-1}(x) = \text{softplus}(x) = \ln(1 + e^x)$. This choice has two advantages. First, it satisfies the conditions imposed on $g_2(x)$ above. Second, it belongs to the class $C^\infty$, which improves numerical stability when computing the gradients used to minimize the loss function [Glorot et al., 2011].

The implementation was written in Python 3.12.4 using PyTorch 2.3.1 [Paszke et al., 2019]. Model training proceeds as follows. The feature matrix $X$ is first normalized [LeCun et al., 1998], and the model weights are then optimized by minimizing the loss evaluated on the observed responses $\overline{y}$. Optimization is performed using the ADAM (Adaptive Moment Estimation) optimizer [Kingma, Ba, 2015]. At each iteration $t$, ADAM computes two exponentially smoothed gradient statistics, the first moment $m_t$ and the second moment $v_t$:

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

Bias-corrected estimates are then introduced:

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

The weights are updated according to

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

where $\theta \in \{\omega, \eta\}$ denotes the current weight vector, $lr$ is the learning rate, $\varepsilon = 10^{-8}$ is a numerical stability constant, $\delta_1$ and $\delta_2$ are the exponential decay rates for the moments, and $\nabla_\theta$ is the gradient with respect to the weights. Let $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ denote the digamma function [Abramowitz, Stegun, 1964]. The gradient components of the beta-regression likelihood are then

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

where the $i$-th components of the auxiliary vectors are

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

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

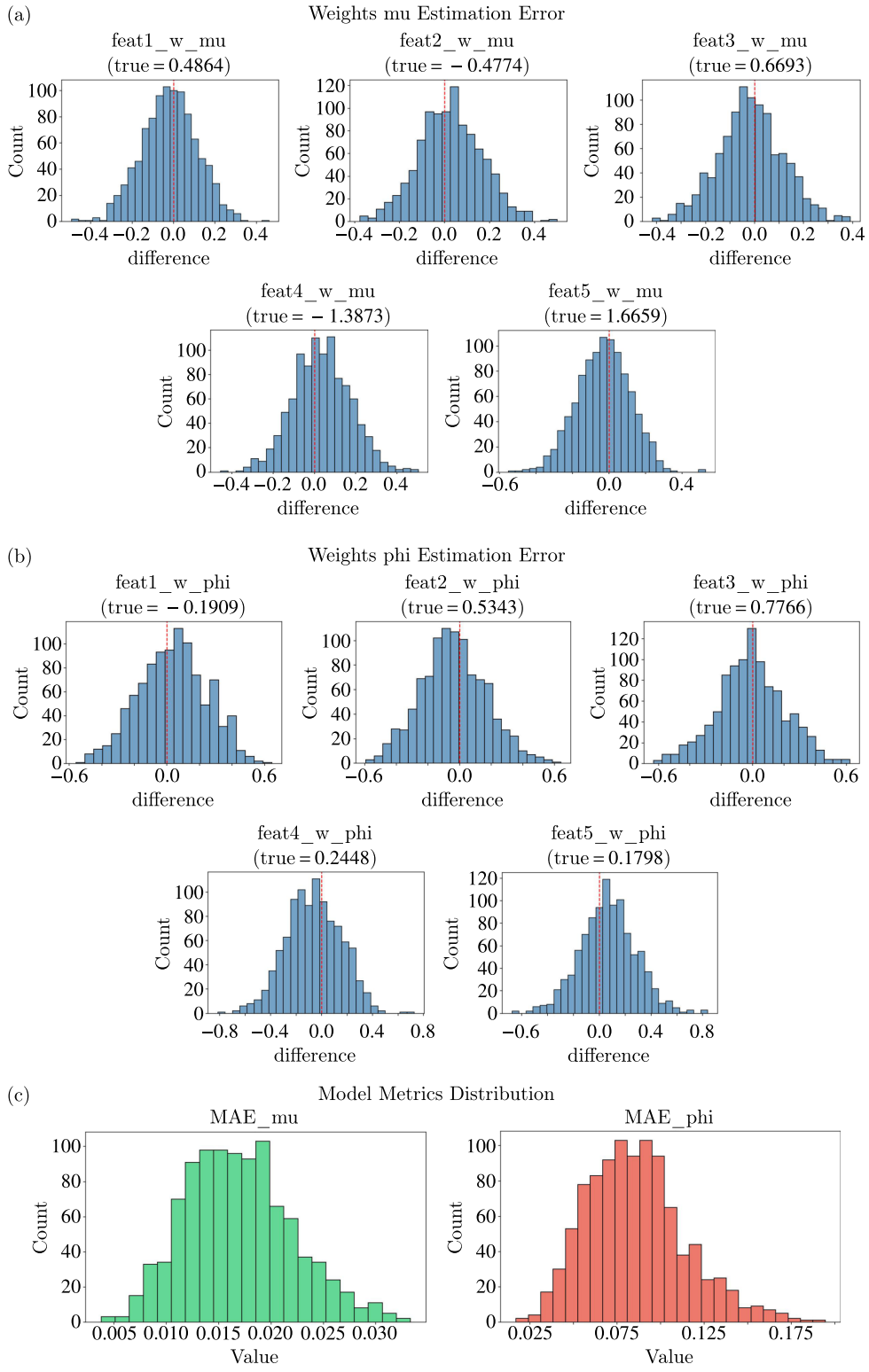


Figure 1. Histograms of model parameter recovery errors on synthetic data. Panels (a) and (b) show the deviations of the estimated weights from their true values (shown above the corresponding histograms) for a set of five features. The horizontal axis shows the deviation of the estimated parameter, and the vertical axis shows the number of models in the corresponding bin. Panel (c) shows the distributions of the mean absolute errors for the mean parameter $\mu$ (left) and the precision parameter $\phi$ (right). The horizontal axis shows the MAE for the corresponding recovered parameter, and the vertical axis shows the number of models in the corresponding bin

Training and prediction were validated using synthetic tests. A series of 1000 simulations was performed with fixed weight values generated in advance from the standard normal distribution. In each simulation, the columns of the feature matrix were generated independently from the uniform distribution on [0, 1], with 1000 observations per matrix. The true values of the mean and precision parameters were then computed from the generated data:

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

The target variable was then generated as

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

The model was trained on these data, after which we computed the deviations of the recovered weights from their true values and the mean absolute deviations of the predicted mean and precision. The distributions of these deviations across all simulations are shown as histograms in Fig. 1.

As shown in Fig. 1, the deviation distributions for both sets of weights are concentrated near zero, indicating no pronounced bias in the estimates. The mean absolute deviation of the predictions for $\mu$ is concentrated around 0.01–0.025, whereas for $\phi$ it lies between 0.05 and 0.15, which is acceptable given the range of the target variables.

## Materials and methods

To improve interpretability and select significant RNA-binding proteins, regularization was applied to the mean-model weights. Specifically, penalty terms corresponding to $l_1$- and $l_2$-regularization were added to the loss function, yielding Elastic Net regularization [Zou, Hastie, 2005]:

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

where $\gamma_1$ and $\gamma_2$ are the regularization hyperparameters.

The final loss function is therefore

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

Expression data for CD44 splice variants and potential splicing factors in colorectal cancer cells were obtained from the open TCGA PanCancer repository [Weinstein et al., 2013]. Reference transcript sequences, together with their genomic coordinates and exon composition, were obtained from GENCODE 21 [Frankish et al., 2021]. The list of RNA-binding proteins and their motifs was downloaded from the Attract [Giudice et al., 2016] and SpliceAid-F [Giulietti et al., 2013] databases. RNA-binding motifs provided in RNA orientation were converted into the corresponding DNA sequences using the reverse-complement operation. Motif occurrences in transcripts were identified by exact substring matching, with no mismatches allowed. The presence of at least one motif occurrence

was used as the threshold for selecting potentially significant factors. All features varied across the sample; constant and near-zero features were absent from the feature matrix and therefore required no additional filtering. In total, 331 patients and 161 potential splicing factors were selected.

Among the 38 annotated isoforms, we retained only those with median expression proportions of at least 0.03. The excluded transcripts carry little regulatory information because their expression proportions remain close to zero throughout the sample; in such cases, the beta regression model degenerates into a trivial constant predictor and cannot reveal meaningful dependencies on splicing-factor expression. This filtering left 5 splice variants: CD44s, CD44v8–9, CD44v3–10, CD44v8–10, and CD44v6–10.

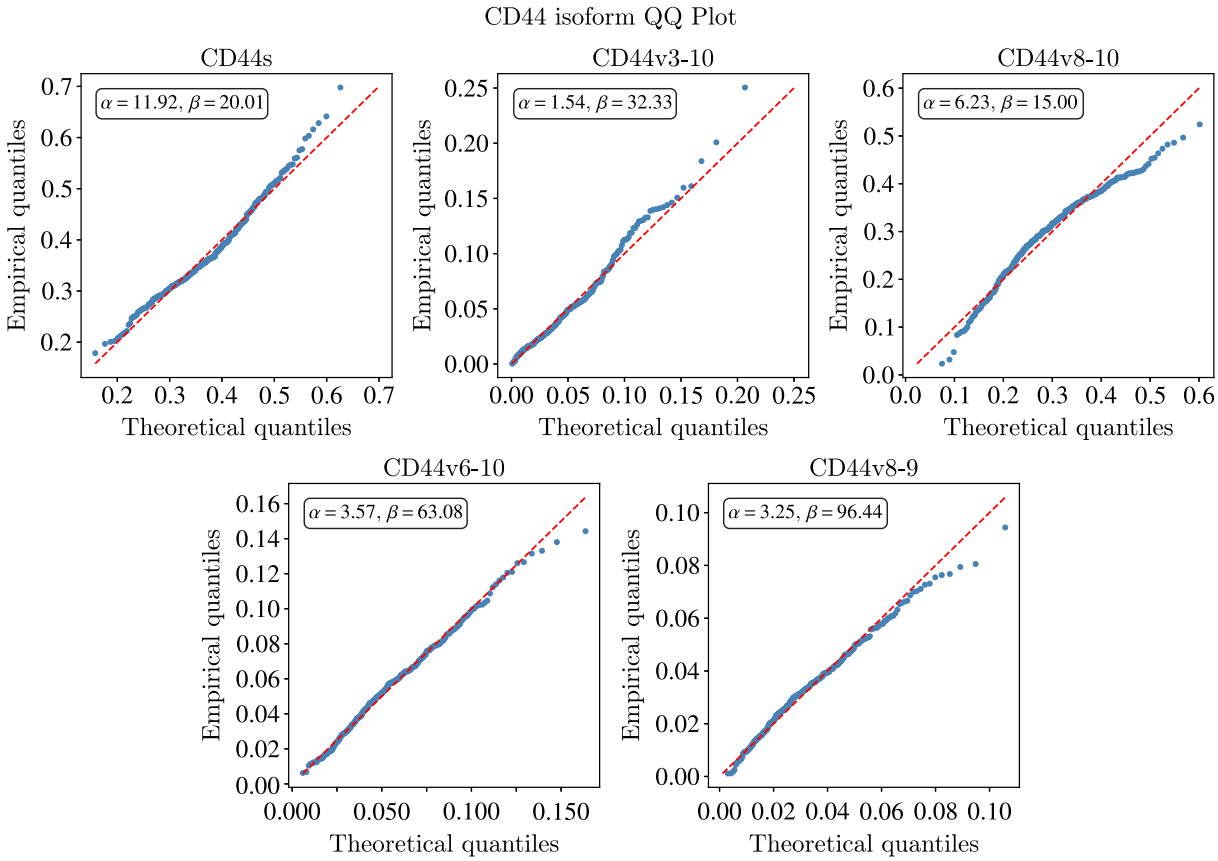


Figure 2. Quantile – quantile plots for the selected CD44 isoforms. The horizontal axis shows the quantiles of the theoretical beta distribution, with parameters $\alpha$ and $\beta$ fitted by maximum likelihood and reported in the legend of each plot; the vertical axis shows the quantiles of the empirical isoform-expression distribution

To assess whether beta regression is appropriate for modeling the proportions, quantile – quantile plots were constructed for each transcript. The empirical quantiles of the observed proportions were compared with the theoretical quantiles of a beta distribution whose parameters were estimated by maximum likelihood. Most points in the isoform QQ plots lie close to the bisector of the first quadrant, indicating agreement with the beta distribution (Fig. 2).

Two formulations of the regression problem were considered for identifying significant splicing factors.

The first approach follows a *one versus all* scheme: for each of the five splice variants, a separate beta regression model predicts the proportion of that transcript relative to total gene expression. The feature set includes only those splicing factors whose motifs are present in the sequence of the transcript under consideration. This formulation treats the regulation of each splice variant independently, making it possible to identify factors that specifically affect a given isoform.

The second approach uses an *isoform tree*. An exon is defined as variable if it is not present in every isoform within the group under consideration; for example, exon 10 is variable for CD44v8–10 and CD44v8–9. Based on this definition, a binary tree is constructed recursively. The root contains the full set of transcripts under consideration. At each step, the first variable exon in genomic order is selected and used to split the transcripts in the current node into two subsets: those containing this exon (left child) and those not containing it (right child). Each node is assigned its own set of splicing factors whose motifs occur on the corresponding variable exon. The procedure is repeated until each leaf corresponds to a single transcript. At each internal node, a separate beta regression model is trained to predict the conditional probability of inclusion of the corresponding variable exon. The final probability of each isoform is then obtained as the product of the conditional probabilities along the path from the corresponding leaf to the root. Formally, let $\pi(t)$ be the set of internal nodes on the path from the root to the leaf corresponding to isoform $t$; let $\widehat{\mu}_v$ be the mean predicted by the beta regression model at node $v$ (the conditional probability of inclusion of the corresponding variable exon); and let $b(t, v) \in \{0, 1\}$ indicate whether isoform $t$ contains the exon associated with node $v$. The final probability of isoform $t$ is

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

In contrast to the first approach, this method explicitly identifies factors associated with inclusion of a particular variable exon rather than with the isoform as a whole.

The two approaches were compared by evaluating model quality. The sample was split into training and validation sets containing 264 and 67 patients, respectively. Mean absolute error and Pearson correlation were used as quality metrics. Hyperparameters for each approach were selected with priority given to maximizing Pearson correlation while also accounting for mean absolute error. To reduce the search space and save computational resources, optimization was performed in two stages. First, during preliminary tuning, local sensitivity to the learning rate was analyzed: the parameter was varied to identify critical boundaries beyond which the quality metrics deteriorated sharply. After the stability region had been identified, grid search was applied. The learning rate ($lr$) was searched over the interval from $10^{-2}$ to $10^{-1}$ with a step of $10^{-2}$, and all combinations of regularization coefficients $\gamma_1$ and $\gamma_2$ from the set $\left\{10^{-3}, 10^{-2}, 10^{-1}, 1\right\}$ were considered. The resulting hyperparameters were $lr = 4.5 \cdot 10^{-2}$, $\gamma_1 = 10^{-2}$, and $\gamma_2 = 10^{-2}$ for the first approach, and $lr = 5 \cdot 10^{-2}$, $\gamma_1 = 10^{-2}$, and $\gamma_2 = 10^{-2}$ for the second.

Proteins were subsequently ranked by the absolute value of their predicted weights, and the top 15 were selected. To statistically verify the most important features, the significance of associations between the expression of the splicing factors used to predict isoform proportions and the observed values of these proportions was assessed using Spearman's rank correlation coefficient.

The null hypothesis was the absence of a monotonic dependence between the expression of a splicing factor and the expression level of the corresponding isoform ($H_0$: $\rho = 0$), and the alternative hypothesis was the presence of a statistically significant monotonic association ($H_1$: $\rho \neq 0$). For each "splicing factor – isoform" pair the Spearman correlation coefficient was computed, after which the obtained $p$-values were adjusted for multiple hypothesis testing by the Benjamini – Hochberg method.

The proposed beta regression implementation was also compared with two implementations previously described in the literature. The first was the R implementation proposed in [Cribari-Neto, Zeileis, 2010], run using R version 4.6.0 and the Betareg library version 3.2.4 with default training parameters. The second was the Python implementation described in [Novosad, 2023], trained for 200 epochs. The implementation proposed in this work was used with the parameters specified above. For simplicity, all three regressions were applied within the one-versus-all setting. Four quantities were recorded for comparison: Pearson correlation on the training and validation sets, MAE on the validation set, and training time. To combine these metrics, the following functional was computed separately for each isoform:

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

where $P_{val}$ and $P_{train}$ are the Pearson correlations on the validation and training sets, respectively, $MAE_{val}$ is the validation MAE, and $T$ is the training time. This functional was chosen for several reasons. First, the main objective is to detect relationships between features and the target variable, so validation Pearson correlation is particularly important. Second, the model should be stable: a high training correlation accompanied by a drop on the validation set indicates overfitting. The second factor in the numerator captures this stability requirement. The third factor captures the accuracy of probability recovery: the larger the validation MAE, the smaller the metric. The logarithm in the denominator accounts for training time while keeping values that differ by only a few seconds close to one another. The correction inside the logarithm ensures that $\log_{10} > 1$, preventing the functional from diverging. By construction, larger values of this functional correspond to better model performance with respect to the selected criteria.

## Results and discussion

Table 1 presents the results of testing the three implementations on different CD44 isoforms.

After computing the functional introduced above and averaging it over isoforms, we obtained $F_R = 0.349$, $F_{Novosad} = 0.244$, and $F_{neural} = 0.559$. The results in Table 1 and the functional values indicate that the neural network implementation of beta regression, although in some cases inferior to the other implementations in terms of $MAE_{val}$, is overall better suited to the problem considered here.

When comparing approaches for identifying important RNA-binding proteins, one must account not only for the different forms of their target variables (normalized probabilities versus independent estimates), but also for the different ways in which their feature spaces are constructed.

To exclude the possibility that differences in prediction quality were caused solely by normalization in the tree model (where isoform proportions are constrained to sum to one), we analyzed

Table 1. Comparison of three beta regression implementations (R, Novosad, and neural) for CD44 isoforms. For each isoform, the table reports prediction quality on the validation set ($MAE_{val}$), Pearson correlation coefficients on the training and validation sets ($P_{train}$ and $P_{test}$), and model training time (time (s)), reflecting the computational complexity of the approaches. The R and Novosad models correspond to the literature methods discussed above, whereas neural denotes the neural network implementation of beta regression developed in this work

| Isoform | Model | $MAE_{val}$ | $P_{train}$ | $P_{val}$ | Time (s) |
|---|---|---|---|---|---|
| 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 |

the predictive ability of the tree at the level of its individual, unnormalized branching nodes. The results showed that the hierarchical approach does not capture the biological signal stably in the data under study: despite satisfactory training-set performance, validation Pearson correlations for the intermediate nodes, which predict conditional exon-inclusion probabilities, drop to near-zero or negative values (from −0.15 to 0.02).

This substantial loss of generalization can be explained by architectural limitations of the tree model. By construction, each internal node uses as features only those splicing factors whose motifs are located on the corresponding variable exon. This rigid narrowing of the feature space discards the broader regulatory context. Splicing is a complex process in which exon inclusion can be influenced by factors binding to neighboring introns or other transcript regions. Because the feature sets are local and limited, the node-level models overfit the small amount of available data and lose predictive power.

As a result, errors accumulate when the final prediction is assembled. Because the final probability of each isoform is computed as the product of conditional probabilities along the path from the root to the leaf, poor unnormalized node-level predictions are compounded. This led to very low metrics for the tree leaves: the validation Pearson correlation averaged only 0.12 (from −0.04 to 0.38 across individual splice variants), and the mean absolute error (MAE) was 0.37.

In contrast, the *one versus all* approach models each splice variant independently and uses factors whose motifs occur anywhere in the full transcript sequence as features. Preserving the broader biological context and avoiding the rigid constraint of hierarchical normalization yielded substantially higher accuracy. As shown in Table 1, Pearson correlations for the OVA model range from 0.58 to 0.80, while MAE values decrease to 0.037–0.085.

Given its clear mathematical and biological advantages, the "one-versus-all" method was selected for the final analysis. Complete lists of the top 15 features for each isoform, together with the results of statistical significance testing, are provided in the Appendix.

Analysis of the selected feature sets identified RNA-binding proteins involved in the regulation of CD44 splicing. For all features discussed below, the association between factor expression and isoform proportion is statistically significant ($p_{adj} < 0.05$). For ease of interpretation, two quantities are reported in the text: the contribution of the feature to prediction in the multivariate model (weight $w$) and Spearman's rank correlation coefficient $\rho$. For some features, the sign of the predicted model weight may differ from the sign of Spearman's correlation coefficient. This is expected: $\rho$ reflects the isolated (marginal) association between a factor and an isoform, whereas the weights $w$ are obtained from a multivariate model and account for splicing-factor co-expression (multicollinearity), interactions, and the mutually exclusive nature of isoform proportions.

The selected regulators are discussed below.

The most notable example is *HNRNPL* ($w = 0.064$, $\rho = 0.470$), the leading positive predictor of the CD44v8–9 isoform. HNRNPL is a canonical regulator of CD44 alternative splicing: it binds to CA-rich intronic elements flanking the variable exons and promotes their inclusion [Hui et al., 2005; Heiner et al., 2010]. For the same isoform, *YBX1* had a negative weight ($w = -0.035$, $\rho = 0.205$), which at first glance appears to contradict the known role of YBX1 as an activator of exon v4 inclusion [Stickeler et al., 2001]. However, this contradiction is resolved when the exon composition is taken into account: CD44v6–10 does not contain v4, so high YBX1 activity competitively redistributes splicing in favor of v4-containing isoforms, reducing the proportion of CD44v.

For the CD44v8–10 and CD44v8–9 isoforms, *MBNL1* acts as a negative regulator, with weights and correlations $w = -0.064$, $\rho = -0.321$ and $w = -0.016$, $\rho = -0.325$, respectively. This is also biologically consistent: MBNL1 is known to suppress EMT-associated alternative splicing in colorectal cancer, where EMT (epithelial – mesenchymal transition) is associated with the acquisition of motility and invasiveness [Oltean, Bates, 2014; Navvabi et al., 2021]. In differentiated epithelial cells, MBNL1 suppresses the inclusion of variable exons, whereas its decrease during EMT allows CD44 variant isoforms to accumulate. Finally, the largest contribution to the regulation of isoforms with an extended variable region (CD44v3–10 and CD44v6–10, with weights and correlations $w = 0.108$, $\rho = 0.436$ and $w = 0.061$, $\rho = 0.433$, respectively) is made by *HNRNPF*. HNRNPF is known to bind G-quadruplex structures in CD44 pre-mRNA and promote variable-exon inclusion during EMT, activating the transition from the standard CD44s isoform to variant isoforms [Huang et al., 2017].

In addition to previously established splicing factors, the model identified regulators whose roles in CD44 splicing are not yet fully understood. One example is *QKI*, selected for CD44v8–10 ($w = -0.078$, $\rho = -0.225$) and, to a lesser extent, for CD44v6–10 ($w = -0.024$, $\rho = -0.173$). QKI has been associated with tumor suppression, although its involvement in CD44 splicing remains insufficiently studied [Maltseva, Tonevitsky, 2023; Zhu et al., 2024]. In addition to QKI, *TARDBP* (TDP-43) was detected for the CD44v8–10 splice variant ($w = -0.057$, $\rho = 0.184$). TARDBP is known to regulate alternative splicing in colorectal cancer [Ma et al., 2021] and to suppress variable exons in CD44 splicing in breast cancer [Guo et al., 2022].

Among potentially novel regulators, three splicing factors are particularly noteworthy.

The first candidate is *NUDT21*, selected by the model as a regulator of CD44s ($w = -0.061$, $\rho = -0.572$) and CD44v8–9 ($w = -0.018$, $\rho = 0.169$). NUDT21 controls the choice between proximal and distal polyadenylation sites and has been characterized as an inhibitor of glioblastoma growth [Masamha et al., 2014]. The negative weights for both isoforms suggest that NUDT21 activity shifts the balance of CD44 transcripts toward isoforms with a short variable region (v2–v5); however, this hypothesis requires experimental testing and direct analysis of APA events at the CD44 locus.

The second potential regulator is *AGO2* ($w = -0.028$, $\rho = -0.534$ for CD44s). Argonaute-2 is the central effector of the RISC complex [Hammond et al., 2001] and is traditionally regarded as a mediator of siRNA-dependent silencing. At the same time, AGO2 has been shown to be capable of regulating alternative splicing co-transcriptionally [Alló et al., 2009]. Its negative association with CD44s suggests that AGO2-dependent activity may promote switching from the standard CD44 isoform to variant isoforms. As in the case of NUDT21, the specific target exons and the mechanism of this switch require experimental verification.

The third candidate is *ACO1* ($w = 0.034$, $\rho = 0.126$ for CD44v8–10). ACO1 (cytosolic aconitase 1, IRP1) is a bifunctional protein: when iron levels are sufficient, it functions as a Krebs cycle enzyme, whereas under iron deficiency it switches to an RNA-binding mode and regulates mRNA stability and translation through iron-responsive elements [Rouault, 2006]. ACO1 overexpression has been reported in rectal tumors [Choi et al., 2011], but its involvement in alternative splicing regulation has not been described previously. Its positive association with CD44v8–10 suggests the existence of an "iron metabolic status – CD44 splicing" axis, making ACO1 the most interesting of the three candidates for further experimental study.

First, NUDT21 and AGO2 are established regulators of RNA metabolism, so their potential link to splicing, although not verified for CD44, is methodologically expected. ACO1, by contrast, has not previously been considered in this context. Second, switching of ACO1 between enzymatic and RNA-binding modes is determined by intracellular iron levels, which are systemically disrupted in the tumor microenvironment [Torti, Torti, 2013]. This creates a plausible basis for metabolically dependent dynamic regulation of the CD44 isoform composition.

## Conclusion

This work proposes an approach to alternative splicing analysis based on modeling isoform proportions with beta regression implemented as a neural network regressor that jointly estimates the mean and precision parameters. The key idea is to treat relative isoform expression levels as realizations of a beta distribution whose parameters depend on RNA-binding protein expression. Compared with classical statistical tools, such as standard beta regression packages in R, and existing specialized approaches, the proposed neural network architecture provides high computational stability on multidimensional expression data. This makes it possible to estimate the context-dependent systemic contribution of each regulator while accounting for mutual influence and co-expression.

The proposed implementation was validated on synthetic data, demonstrating the absence of systematic bias and acceptable accuracy in recovering the distribution parameters. The model was then applied to TCGA data for CD44 in colorectal cancer cells to compare two problem formulations: the "one-versus-all" scheme, in which a separate model is fitted independently for each isoform, and the

isoform tree method, based on hierarchical decomposition of transcripts by variable exons followed by modeling of their conditional inclusion probabilities.

The "one-versus-all" scheme substantially outperformed the tree approach in prediction quality and was therefore used for the subsequent analysis of splicing-factor significance.

The results were statistically verified and are consistent with known biological data: previously described regulators of CD44 splicing (HNRNPL, MBNL1, HNRNPF, and YBX1) were found among the significant features. At the same time, the model identified several potentially novel candidates, including ACO1, NUDT21, and AGO2. ACO1 is of particular interest because it led to a new hypothesis linking intracellular iron metabolism to regulation of the CD44 isoform composition.

Future work will extend the analysis to other tumor types in order to identify common and tissue-specific regulators of CD44 splicing. Comparative studies across cancer tissues will help assess the generality of the identified patterns and reveal context-dependent splicing factors. This approach should clarify mechanisms of alternative splicing regulation in different oncological models and help identify potential therapeutic targets.

## Acknowledgements

The author thanks his scientific advisor, Alexey Vladimirovich Galatenko, and Anton Pavlovich Zhiyanov for methodological recommendations and fruitful discussion of the results. The author is also deeply grateful to the Editor-in-Chief of the journal, Alexey Ivanovich Lobanov, and to the reviewer for their careful attention to the manuscript and valuable comments that helped improve the paper.

## 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