ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. НАУКИ О ЖИЗНИ, 2023, том 510, с. 316–321

УДК 576.53

ВЫЯВЛЕНИЕ ЗНАЧИМЫХ РНК-СВЯЗЫВАЮЩИХ БЕЛКОВ В ПРОЦЕССЕ СПЛАЙСИНГА CD44 С ПОМОЩЬЮ АЛГОРИТМА УСИЛЕННОЙ БЕТА-РЕГРЕССИИ

© 2023 г. В. О. Новосад[1,2,*]

Представлено академиком РАН С.И. Колесниковым
Поступило 15.12.2022 г.
После доработки 01.02.2023 г.
Принято к публикации 02.02.2023 г.

Экспрессия РНК-связывающих белков и их взаимодействие со сплайсируемой пре-мРНК являются ключевым фактором в определении итогового профиля изоформ. Трансмембранный белок CD44 участвует в дифференцировании, инвазии, подвижности, росте и выживании опухолевых клеток, а также является общепринятым маркером раковых стволовых клеток и эпителиально-мезенхимального перехода. При этом функции изоформ этого белка значительно различаются. В настоящей работе разработан основанный на алгоритме усиленной бета-регрессии метод определения значимых в процессе сплайсинга РНК-связывающих белков с помощью моделирования соотношения изоформ. Применение данного метода к анализу сплайсинга CD44 в клетках колоректального рака выявило 20 значимых РНК-связывающих белков. Многие из них ранее были показаны как регуляторы ЭМП, однако впервые представлены как потенциальные факторы сплайсинга CD44.

Ключевые слова: альтернативный сплайсинг, CD44, KPP, усиленная бета-регрессия, РНК-связывающие белки, РНК секвенирование, TCGA

Список сокращений: KPP – колоректальный рак, medAPE – медиана абсолютного процента ошибки

DOI: 10.31857/S2686738922601023, **EDN:** QFUIKM

ВВЕДЕНИЕ

Альтернативный сплайсинг – это процесс регуляции транскрипции, посредством которого один ген может кодировать несколько белков со схожими структурами, но выполняющих разные функции. Данный процесс является одним из основных факторов белкового разнообразия и, по последним подсчетам, ему подвергается более 95% кодирующих белки генов [1].

Одним из основных факторов регулирования альтернативного сплайсинга является взаимодействие РНК-связывающих белков с определенными нуклеотидными последовательностями на сплайсируемой пре-мРНК [2, 3]. А именно, при связывании с пре-мРНК, некоторые белки могут стимулировать или ингибировать процесс сборки сплайсосомы на рядом стоящих сайтах сплайсинга. Как следствие, репрезентативность таких белков в окрестности некоторого экзона сплайсируемой пре-мРНК приводит к вырезанию или сохранению этого экзона в процессе альтернативного сплайсинга.

В настоящее время биоинформатический анализ процесса сплайсинга разделен на следующие направления. Во-первых, общий анализ "кода сплайсинга" – различного набора правил сплайсинга, включающего в себя паттерны сплайсинга первичной последовательности транскрипта, мутации регуляторных последовательностей, вероятности включения экзонов, а также новые, потенциальные паттерны сплайсинга [4–6]. Во-вторых, другая группа методов основана исключительно на анализе нуклеотидных последовательностей, с целью классификации сайтов сплайсинга или же предсказания силы связывания белковых регуляторов сплайсинга с определенными нуклеотидными последовательностями РНК [7, 8]. Еще одним направлением является предсказание соотношения кассетных экзонов. Основная идея в изучении данной задачи заключается в использо-

[1] *Факультет биологии и биотехнологии, Национальный исследовательский университет "Высшая школа экономики", Москва, Россия*
[2] *Федеральное государственное бюджетное учреждение науки Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова Российской академии наук, Москва, Россия*
**e-mail: vnovosad@hse.ru*

вании набора характеристик некоторой окрестности нуклеотидной последовательности рассматриваемого экзона в качестве входных данных для алгоритмов машинного обучения [6, 7, 9].

Однако большинство существующих методов либо не включают в себя анализ РНК-связывающих белков, либо основаны на нелинейных зависимостях, либо включают в себя огромное множество различных признаков, что делает данные подходы непригодными для оценки значимости действия белков – факторов сплайсинга.

В настоящем исследовании разработан метод поиска значимых РНК-связывающих белков, вовлеченных в регуляцию альтернативного сплайсинга CD44 в клетках колоректального рака (КРР). CD44 представляет собой трансмембранный гликопротеин, участвующий в различных функциях как нормальных, так и опухолевых клеток [10]. Также было показано, что изоформы белка CD44 имеют различную роль в развитии опухолевых клеток, а уровень их экспрессии может ипользоваться в качестве маркера тяжести заболевания [10].

МАТЕРИАЛЫ И МЕТОДЫ

Таблицы массового секвенирования РНК первичных опухолей ($n = 270$) проекта TCGA-COAD были загружены с сайта Broad GDAC Firehose (https://gdac.broadinstitute.org) в формате матриц количества считываний.

Для нормализации матриц считываний РНК секвенирования в таблицы в логарифмической шкале TMM-FPKM использовался алгоритм усеченного среднего из М-значений (TMM), реализованный в пакете edgeR v3.30.3 [11].

Список РНК-связывающих белков и их последовательностей связывания был загружен из баз данных Attract [12] и SpliceAid-F [13]. РНК-связывающие белки, имеющие максимальную экспрессию среди всего набора данных ниже 1 в log2(FPKM) шкале, были исключены из анализа как низкоэкспрессированные.

Для поиска значимых РНК-связывающих белков в сплайсинге CD44 были рассмотрены изоформы 3 и 4 этого белка, как наиболее экспрессируемые изоформы в клетках колоректального рака человека [14]. Обозначение изоформ приведено в соответствии с номенклатурой NCBI (https://www.ncbi.nlm.nih.gov/gene/960).

Основная часть анализа заключалась в использовании алгоритма усиленной бета-регрессии для оценки коэффициентов пропорциональности между экспрессией РНК-связывающих белков и долей рассматриваемой изоформы. В качестве метрик качества полученных линейных моделей использовали показатель коэффициента корреляции Пирсона (аналог классического R^2 для случая предсказания ограниченной переменной), коэффициент R^2, а также медиану абсолютного процента ошибки (medAPE).

Данные массового РНК секвенирования были случайным образом разделены на обучающую и валидационную выборку в соотношении 3 к 1.

АЛГОРИТМ

Бета регрессия – является частным случаем обобщенных линейных моделей и используется для моделирования ограниченных значений (например, вероятностей, от 0 до 1). В отличие от классической линейной регрессии, предполагается, что итоговая переменная y может иметь любое распределение из экспоненциального семейства, а некоторое преобразование математического ожидания выражается через линейную комбинацию рассматриваемых признаков:

$$\mu_i = E(y_i), g(\mu_i) = \theta_i = \sum_{j=1}^{m} c_j x_{ij}, \quad (1)$$

где $\{\{x_{ij}\}_{j=1}^{m}, y_i\}_{i=1}^{n}$ – выборка данных.

Двупараметрическая случайная величина имеет экспоненциальное распределение, если ее плотность может быть представлена следующим образом:

$$p(y|\theta,\varphi) = \exp((\eta_1(\theta,\varphi)f_1(y) + \eta_2(\theta,\varphi)\cdot \cdot f_2(y) - A(\theta,\varphi))h(y), \quad (2)$$

Случайная величина имеет бета распределение, если ее плотность:

$$p(y|\alpha,\ \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

где $E(y) = \frac{\alpha}{(\alpha+\beta)}$, Γ – гамма функция.

Введя перепараметризацию вида $\mu = \frac{\alpha}{(\alpha+\beta)}$, $\varphi = \alpha + \beta$, получаем:

$$p(y|\mu,\ \varphi) = \frac{\Gamma(\varphi)}{\Gamma(\mu\varphi)\Gamma((1-\mu)\varphi)} y^{\mu\varphi-1}(1-y)^{(1-\mu)\varphi-1},$$

где $E(y) = \mu$.

Тогда плотность может быть выражена в “экспоненциальной форме“ (2), где:

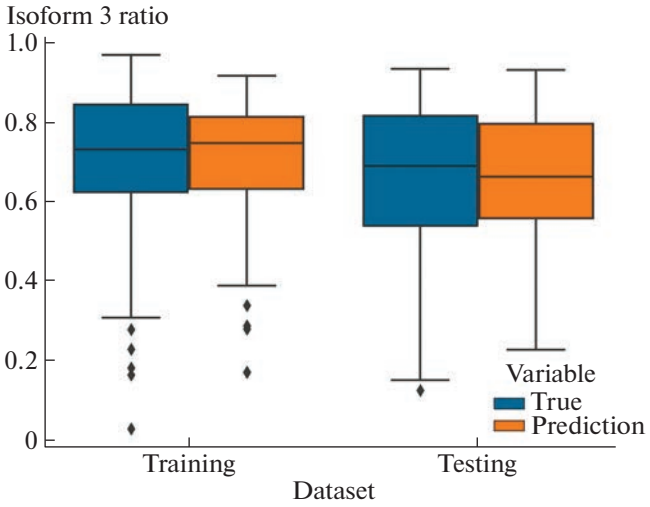


Рис. 1. Распределение настоящей и предсказанной на основе экспрессии отобранных РНК-связывающих белков доли изоформы 3 белка CD44 на обучающих и валидационных данных. Синий цвет соответствует распределению настоящей доле изоформы. Оранжевый цвет соответствует предсказанной доле изоформы.

$$\eta_1(\mu,\ \varphi) = \mu\varphi - 1,$$
$$f_1(y) = \log(y),$$
$$\eta_2(\mu,\ \varphi) = (1-\ \mu)\varphi - 1,$$
$$f_2(y) = \log(1-y),$$
$$A(\mu,\ \varphi) = \log\left(\frac{\Gamma(\mu\varphi)\Gamma((1-\ \mu)\varphi)}{\Gamma(\varphi)}\right),$$
$$h(y) = \frac{1}{y(1-y)}.$$

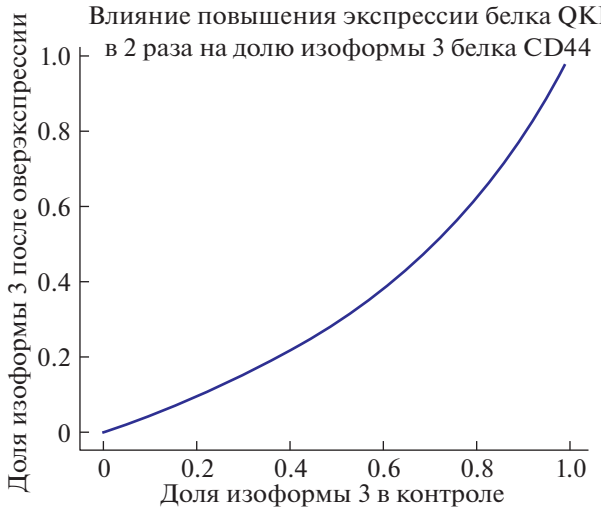
В случае бета-распределения, в роли связывающей функции $g(1)$ может быть использовано так называемое logit преобразование:

$$\theta_i = g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{j=1}^{m} c_j x_{ij}.$$

Логарифм функции правдоподобия бета-распределения для имеющейся выборки выглядит следующим образом:

$$\log(L) = \sum_{i=1}^{n} \log(p(y_i|\mu_i,\ \varphi) =$$
$$= \sum_{i=1}^{n} (\log(\Gamma(\varphi)) - \log(\Gamma(\mu_i\varphi)) -$$
$$- \log(\Gamma((1-\mu_i)\varphi)) + (\mu_i\varphi - 1)\log y_i +$$
$$+ ((1-\mu_i)\varphi - 1)\log(1-y_i).$$

Тогда, градиент логарифма функции правдоподобия бета-распределения по приближаемому параметру $\text{logit}(\mu_i)$ в наших обозначениях выглядит как:



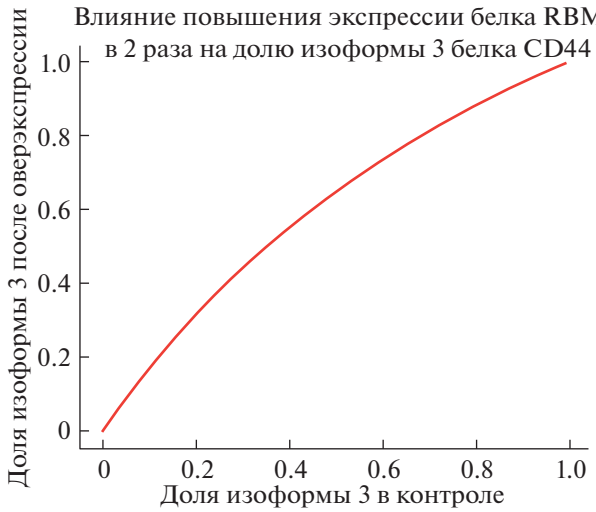


Рис. 2. Пример влияния моделируемого повышения в два раза экспрессии рассматриваемого РНК-связывающего белка (QKI/RBMX) на долю изоформы 3 белка CD44. По оси X отложена изначальная доля изоформы 3 в контрольных клетках. По оси У отложена доля изоформы 3 в клетках после оверэкспрессии рассматриваемого белка-регулятора.

$$\frac{\partial(\log(L))}{\partial\,\text{logit}(\mu_i)} = \mu_i(1-\mu_i)\varphi \times$$
$$\times\left(\frac{\Gamma'((1-\mu_i)\varphi)}{\Gamma((1-\mu_i)\varphi)} - \frac{\Gamma'(\mu_i\varphi)}{\Gamma(\mu_i\varphi)} + \log\left(\frac{y_i}{1-y_i}\right)\right).$$

В классическом случае коэффициенты c_j (1) могут быть найдены максимизацией логарифма правдоподобия модели с помощью метода градиентного спуска:

$$c_j(t+1) = c_j(t) + \gamma\sum_{i=1}^{n} \frac{\partial(\log(L))}{\partial logit(\mu_i)} x_{ij},$$

где t – номер итерации, а γ – параметр величины шага.

Таблица 1. Весовые коэффициенты вклада экспрессии РНК-связывающих белков в определение доли изоформы 3 белка CD44. Положительный/отрицательный коэффициент означает положительную/отрицательную зависимость между экспрессией соответствующего белка и долей изоформы 3 белка CD44. Абсолютное значение коэффициента определяет силу вклада изменения экспрессии соответствующего белка в изменение доли изоформы 3 белка CD44

Белок	Коэффициент
QKI	–0.89
RBMX	0.61
RBM8A	–0.34
ESRP1	0.27
TARDBP	–0.26
AKAP1	0.26
CELF1	0.23
GRSF1	0.22
PHAX	0.21
A1CF	–0.17
SRSF1	–0.17
SRSF4	0.17
RBM25	0.15
RBM14	–0.14
RBM38	–0.14
ACO1	0.13
PCBP2	0.13
OAS1	–0.12
SUPV3L1	–0.1
KHDRBS3	0.09

Усиленная бета-регрессия – это оптимизированный метод классической бета-регрессии, разработанный с целью отбора переменных и обучения регуляризованной бета регрессии с “штрафным” слагаемым. Аналогично недавно разработанным алгоритмам градиентного бустинга [15], усиленная бета регрессия основана на покомпонентном повышении градиента. Основная идея данного подхода состоит в том, чтобы оценить моделируемую переменную как линейную комбинацию так называемых базовых моделей, каждая из которых обучается предсказывать градиент вектора логарифма правдоподобия по отношению к прогнозу модели на предыдущем шаге. Таким образом, к линейной комбинации прогнозирования, полученной на шаге t, добавляется базовая модель с наименьшим показателем потерь и, как следствие, индивидуальная переменная с наилучшей предсказательной способностью. В итоге, модель, полученная на итерации t, выглядит следующим образом:

$$h(t)(x_k) = \gamma + \gamma_1 c_{i_1} x_{ki_1} + \gamma_2 c_{i_2} x_{ki_2} + \ldots + \gamma_{t-1} c_{i_{t-1}} x_{ki_{t-1}},$$

– logit-преобразованное предсказание на k-м объекте, где c_{i_j} и x_{i_j} – коэффициент и признак, полученные при обучении базовой модели на j итерации, а γ_j – параметры величины шага, подобранные максимизацией функции правдоподобия:

$$\gamma_j = \arg\max_{\gamma} L\left(\mu = \begin{Bmatrix} \text{logit}^{-1}(h(j-1)(x_1) + \gamma c_{i_j} x_{1i_j}), \\ \ldots, \text{logit}^{-1}(h(j-1)(x_n) + c_{i_j} x_{ni_j}) \end{Bmatrix} \varphi\right).$$

Основным параметром этого алгоритма является количество итераций t. Помимо точности модели, количество итераций напрямую связано с количеством отобранных признаков и, таким образом, может рассматриваться как параметр регуляризации.

Реализацию алгоритма усиленной бета-регрессии производили на языке программирования Python 3 с использованием стандартных библиотек numpy, pandas, scipy и sklearn.

Для поиска значимых РНК-связывающих белков представленная модель обучалась до тех пор, пока значение функции правдоподобия не стабилизировалось: изменение значения правдоподобия на обучающей выборке за последние 100 итераций обучения не превышало 1%. Итоговая производительность модели оценивалась на валидационной выборке.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

В данном исследовании мы использовали собственную реализацию полного алгоритма усиленной бета-регрессии на языке программирования python (основанной на идее библиотеки gamboostLSS языка программирования R [16], с дополнительной оптимизацией параметра величины шага итерации). Алгоритм доступен в виде класса BoostedBetaReg в публичном репозитории github (https://github.com/NovosadVictor/BoostedBetaRegression).

Мы применили алгоритм усиленной бета-регрессии для выявления значимых факторов сплайсинга CD44 в клетках колоректального рака. На основе данных массового РНК секвенирования из проекта TCGA-COAD мы построили линейную модель предсказания доли уровня экспрессии изоформы 3 по отношению к сумме уровней экспрессии изоформ 3 и 4.

В качестве признаков модели были использованы РНК-связывающие белки, имеющие хотя бы одну РНК-последовательность связывания, точно лежащую на нуклеотидной последовательности гена CD44.

В результате удалось построить модель, основанную на уровнях экспрессии 20 РНК-связывающих белков (см. табл. 1), с достаточно высокими показателями точности на валидационной выборке: коэф. корреляции Пирсона = 0.84, R^2 = 0.54, medAPE = 0.08 (медиана ошибки предсказаний около 8%). Распределение предсказаний на обучающей и валидационной выборках представлено на рис. 1.

Важно отметить, что, исходя из линейности построенной модели, абсолютное значение полученных весовых коэффициентов может интерпретироваться как "важность" соответствующего РНК-связывающего белка. В частности, чем выше абсолютное значение коэффициента, тем сильнее изменение экспрессии рассматриваемого белка влияет на изменение доли изоформы (см. рис. 2). Интересно, что наиболее значимые найденные РНК-связывающие белки (с максимальным абсолютным значением коэффициента) – QKI, RBMX, ESRP1 и другие – ранее были показаны как участники регуляции ЭМП [17–20]. Таким образом, использование предложенного метода позволет открывать новые белковые регуляторы процесса альтернативного сплайсинга интересующего гена.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Исследование выполнено при поддержке гранта Министерства науки и высшего образования Российской Федерации (соглашение № 075-15-2021-1049).

СПИСОК ЛИТЕРАТУРЫ

1. *Pan Q., Shai O., Lee L.J., Frey B.J., Blencowe B.J.,* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing // Nat. Genet. 2008. V. 40. P. 1413–1415.
2. *Wang Z., Burge C.B.*, Splicing regulation: From a parts list of regulatory elements to an integrated splicing code, // RNA. 2008. V. 14. P. 802–813.
3. *Wang Z., Xiao X., Van Nostrand E., Burge C.B.,* General and Specific Functions of Exonic Splicing Silencers in Splicing Control // Mol. Cell. 2006. V. 23. P. 61–70.
4. *Xiong H.Y., Barash Y., Frey B.J.,* Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context // Bioinformatics. 2011. V. 27. P. 2554–2562.
5. *Hartmann B., Valcárcel J.*, Decrypting the genome's alternative messages // Curr. Opin. Cell Biol. 2009. V. 21. P. 377–386.
6. *Xiong H.Y., Alipanahi B., Lee L.J., Bretschneider H., Merico D., Yuen R.K.C., Hua Y., Gueroussov S., Najafabadi H.S., Hughes T.R., Morris Q., Barash Y., Krainer A.R., Jojic N., Scherer S.W., Blencowe B.J., Frey B.J.*, The human splicing code reveals new insights into the genetic determinants of disease // Science. 2015. № 80. P. 347.
7. *Barash Y., Calarco J.A., Gao W., Pan Q., Wang X., Shai O., Blencowe B.J., Frey B.J.*, Deciphering the splicing code // Nature. 2010. V. 465. P. 53–59.
8. *Cereda M., Pozzoli U., Rot G., Juvan P., Schweitzer A., Clark T., Ule J.,* RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing, // Genome Biol. 2014. V. 15. P. R20.
9. *Leung M.K.K., Xiong H.Y., Lee L.J., Frey B.J.*, Deep learning of the tissue-regulated splicing code, // Bioinformatics. 2014. V. 3. P. i121–i129.
10. *Xu H., Niu M., Yuan X., Wu K., Liu A.*, CD44 as a tumor biomarker and therapeutic target., // Exp. Hematol. Oncol. 2020. V. 9. P. 36.
11. *Robinson M.D., McCarthy D.J., Smyth G.K.,* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data., // Bioinformatics. 2010. V. 26. P. 139–40.
12. *Giudice G., Sánchez-Cabo F., Torroja C., Lara-Pezzi E.*, ATtRACT–a database of RNA-binding proteins and associated motifs, Database. 2016 (2016) baw035.
13. *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. V. 41. P. D125–D131.
14. *Novosad V.O., Polikanova I.S., Tonevitsky E.A., Mal'tseva D.V.*, Expression of CD44 isoforms in human colorectal cancer patient samples and cell lines, // Cell Technol. Biol. Med. 2022. V. 1. P. 49–54.
15. *Bühlmann P., Hothorn T.,* Boosting Algorithms: Regularization, Prediction and Model Fitting, Stat. Sci. 2007. V. 22.
16. *Hofner B., Mayr A., Schmid M.,* gamboostLSS : An R Package for Model Building and Variable Selection in the GAMLSS Framework, J. Stat. Softw. 2016. V. 74.
17. *Kim E.J., Kim J.S., Lee S., Lee H., Yoon J., Hong J.H., Chun S.H., Sun D.S., Won H.S., Hong S.A., Kang K., Jo J.Y., Choi M., Shin D.H., Ahn Y., Ko Y.H.,* QKI, a miR-200 target gene, suppresses epithelial-to-mesenchymal transition and tumor growth, // Int. J. Cancer. 2019. V. 145. P. 1585–1595.
18. *Liang R., Zhang J., Liu Z., Liu Z., Li Q., Luo X., Li Y., Ye J., Lin Y.*, Mechanism and Molecular Network of RBM8A-Mediated Regulation of Oxaliplatin Resistance in Hepatocellular Carcinoma, // Front. Oncol. 2021. V. 10.
19. *Harvey S.E., Xu Y., Lin X., Gao X.D., Qiu Y., Ahn J., Xiao X., Cheng C.*, Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT, // RNA. 2018. V. 24. P. 1326–1338.
20. *Xie C., Zhou M., Lin J., Wu Z., Ding S., Luo J., Zhan Z., Cai Y., Xue S., Song Y.,* EEF1D Promotes Glioma Proliferation, Migration, and Invasion through EMT and PI3K/Akt Pathway, // Biomed Res. Int. 2020 (2020) 1–12.

IDENTIFICATION OF SIGNIFICANT RNA-BINDING PROTEINS IN THE PROCESS OF CD44 SPLICING USING THE BOOSTED BETA REGRESSION ALGORITHM

V. O. Novosad[a,b,#]

[a] *Faculty of Biology and Biotechnology, National Research University Higher School of Economics, Moscow, Russian Federation*
[b] *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russian Federation*
[#] *e-mail: vnovosad@hse.ru*

Presented by Academician of the RAS S.I. Kolesnikov

The expression of RNA-binding proteins and their interaction with the spliced pre-mRNA are the key factors in determining the final isoform profile. Transmembrane protein CD44 is involved in differentiation, invasion, motility, growth and survival of tumor cells, and is also a commonly accepted marker of cancer stem cells and epithelial-mesenchymal transition. However, the functions of the isoforms of this protein differ significantly. In this paper, we developed a method based on the boosted beta regression algorithm for identification of the significant RNA-binding proteins in the splicing process by modeling the isoform ratio. The application of this method to the analysis of CD44 splicing in colorectal cancer cells revealed 20 significant RNA-binding proteins. Many of them were previously shown as EMT regulators, but for the first time presented as potential CD44 splicing factors.

Keywords: alternative splicing, CD44, CRC, boosted beta regression, RNA-binding proteins, RNA-seq, TCGA