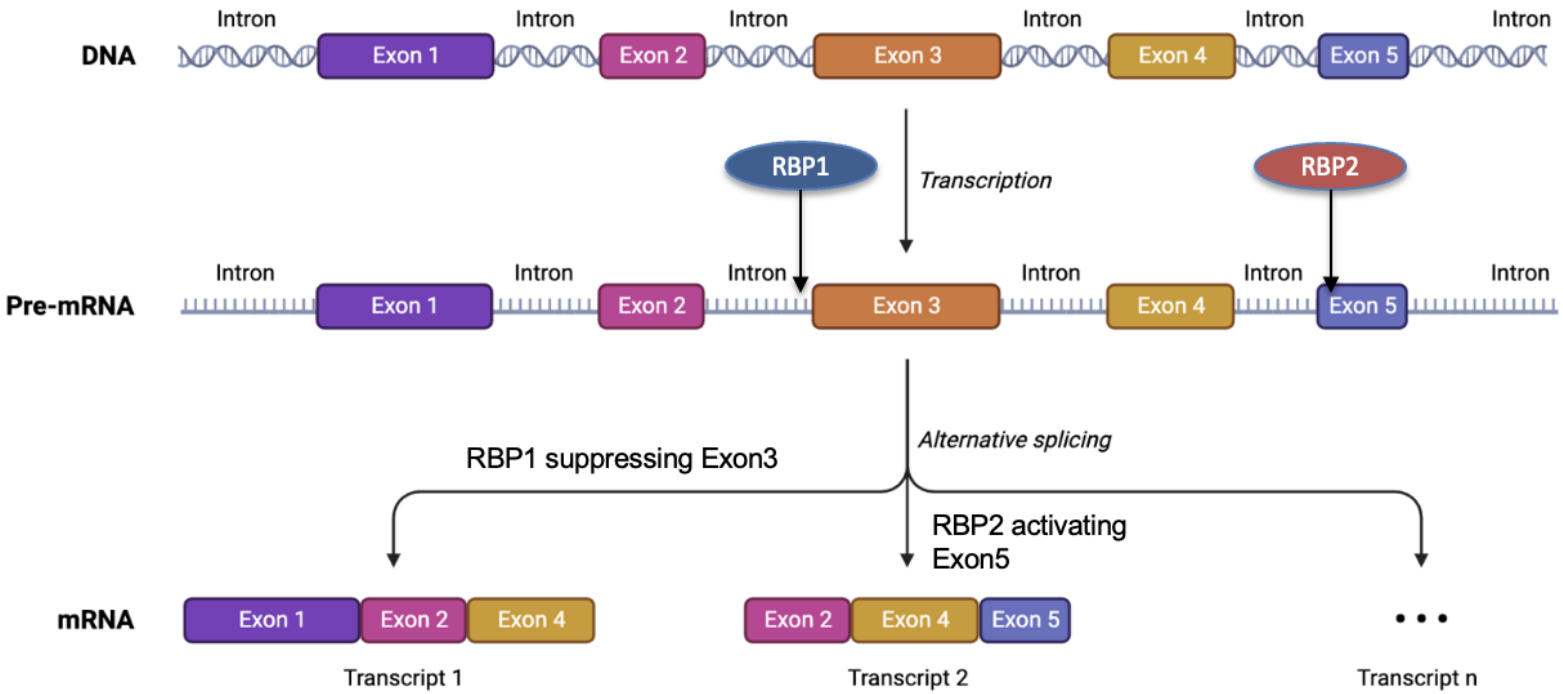
Выявление важных РНК-связывающих белков в процессе сплайсинга гена CD44

Пирогов А.А., Галатенко А.В., Жиянов А.П.,
Староверов В.М.

МГУ им. М.В. Ломоносова, НИУ ВШЭ

2025

Постановка задачи



Применение

- Появление неканонических изоформ в значительном количестве может быть характерным для некоторых заболеваний, например, рак[1].

[1] Y. Zhang, J. Qian, C. Gu. et al, *Alternative splicing and cancer: a systematic review*

Применение

- Появление неканонических изоформ в значительном количестве может быть характерным для некоторых заболеваний, например, рак[1].
- В частности, известно, что изменение экспрессии изоформ гликопротеина CD44 тесно связано с колоректальным раком[2]

[1] Y. Zhang, J. Qian, C. Gu. et al, *Alternative splicing and cancer: a systematic review*

[2] M.P. Raigorodskaya, V.O. Novosad, S.A Tonevitskaya et al. *Expression of CD44 Isoforms in Human Colorectal Cancer Cell Lines*

Предыдущие результаты

В статье [3] представлена модель "дерева изоформ". Эта модель классифицирует изоформы в зависимости от того, содержат ли они определенный экзон. Затем она анализирует данные об РНК-связывающих белках, которые связываются с этим экзоном.

[3] V.O. Novosad, *Identification of Significant RNA-Binding Proteins in the Process of CD44 Splicing Using the Boosted Beta Regression Algorithm*

Предыдущие результаты

В статье [3] представлена модель "дерева изоформ". Эта модель классифицирует изоформы в зависимости от того, содержат ли они определенный экзон. Затем она анализирует данные об РНК-связывающих белках, которые связываются с этим экзоном.
На основе этой информации модель бета-регрессии предсказывает вероятность того, что экзон будет включен в группу изоформ. Важной особенностью модели является ее способность определять, какие именно RBP сильнее всего влияют на это предсказание.

[3] V.O. Novosad, *Identification of Significant RNA-Binding Proteins in the Process of CD44 Splicing Using the Boosted Beta Regression Algorithm*

Предыдущие результаты

В статье [3] представлена модель "дерева изоформ". Эта модель классифицирует изоформы в зависимости от того, содержат ли они определенный экзон. Затем она анализирует данные об РНК-связывающих белках, которые связываются с этим экзоном.
На основе этой информации модель бета-регрессии предсказывает вероятность того, что экзон будет включен в группу изоформ. Важной особенностью модели является ее способность определять, какие именно RBP сильнее всего влияют на это предсказание.
Из недостатков подхода можно отметить технические проблемы с моделью, а также сильная потеря качества предсказания из-за ограничений, связанных с предположениями.

---

[3] V.O. Novosad, *Identification of Significant RNA-Binding Proteins in the Process of CD44 Splicing Using the Boosted Beta Regression Algorithm*

Текущая реализация

В нашей реализации мы использовали данные с TCGA Pan-Cancer (PANCAN). Помимо этого, вместо дерева изоформ реализуется принцип "один против всех" – выбирается самый экспрессированный сплайс-вариант (CD44–201), берутся все садящиеся на него РНК-связывающие белки (127 штук) и строится обучение модели.

Текущая реализация

В нашей реализации мы использовали данные с TCGA Pan-Cancer (PANCAN). Помимо этого, вместо дерева изоформ реализуется принцип "один против всех" – выбирается самый экспрессированный сплайс-вариант (CD44–201), берутся все садящиеся на него РНК-связывающие белки (127 штук) и строится обучение модели.
В ходе нашей работы было рассмотрено 4 различные регрессии со встроенными в них методами подбора важных признаков (регуляризации):

Текущая реализация

В нашей реализации мы использовали данные с TCGA Pan-Cancer (PANCAN). Помимо этого, вместо дерева изоформ реализуется принцип "один против всех" – выбирается самый экспрессированный сплайс-вариант (CD44–201), берутся все садящиеся на него РНК-связывающие белки (127 штук) и строится обучение модели.
В ходе нашей работы было рассмотрено 4 различные регрессии со встроенными в них методами подбора важных признаков (регуляризации):

- Стандартная линейная регрессия с преобразованием $\sigma(x) = \frac{1}{1+e^{-x}}$.

Текущая реализация

В нашей реализации мы использовали данные с TCGA Pan-Cancer (PANCAN). Помимо этого, вместо дерева изоформ реализуется принцип "один против всех" – выбирается самый экспрессированный сплайс-вариант (CD44–201), берутся все садящиеся на него РНК-связывающие белки (127 штук) и строится обучение модели.
В ходе нашей работы было рассмотрено 4 различные регрессии со встроенными в них методами подбора важных признаков (регуляризации):

- Стандартная линейная регрессия с преобразованием $\sigma(x) = \frac{1}{1+e^{-x}}$.
- Логистическая регрессия

Текущая реализация

В нашей реализации мы использовали данные с TCGA Pan-Cancer (PANCAN). Помимо этого, вместо дерева изоформ реализуется принцип "один против всех" – выбирается самый экспрессированный сплайс-вариант (CD44–201), берутся все садящиеся на него РНК-связывающие белки (127 штук) и строится обучение модели.
В ходе нашей работы было рассмотрено 4 различные регрессии со встроенными в них методами подбора важных признаков (регуляризации):

- Стандартная линейная регрессия с преобразованием $\sigma(x) = \frac{1}{1+e^{-x}}$.
- Логистическая регрессия
- "Линейная" и "неслинейная" бета-регрессии.

Для сравнения между ними и определения качества предсказаний были выбраны метрики MAE (mean absolute error) и корреляция Пирсона.

Результаты

На текущий момент модели показывают следующие результаты по метрикам:

Результаты

На текущий момент модели показывают следующие результаты по метрикам:

- Линейная регрессия:
  train: MAE: 0.19, pearson: 0.003
  test: MAE: 0.19 , pearson: 0.001

Результаты

На текущий момент модели показывают следующие результаты по метрикам:

- Линейная регрессия:
  train: MAE: 0.19, pearson: 0.003
  test: MAE: 0.19 , pearson: 0.001
- Лог. регрессия:
  train: MAE: 0.48, pearson: 0.002
  test: MAE: 0.46, pearson: 0.001

Результаты

На текущий момент модели показывают следующие результаты по метрикам:

- Линейная регрессия:
  train: MAE: 0.19, pearson: 0.003
  test: MAE: 0.19 , pearson: 0.001
- Лог. регрессия:
  train: MAE: 0.48, pearson: 0.002
  test: MAE: 0.46, pearson: 0.001
- Линейная бета-регрессия:
  train: MAE: 0.14, pearson: 0.13
  test: MAE: 0.14, pearson: 0.01

Результаты

На текущий момент модели показывают следующие результаты по метрикам:

- Линейная регрессия:
  train: MAE: 0.19, pearson: 0.003
  test: MAE: 0.19 , pearson: 0.001
- Лог. регрессия:
  train: MAE: 0.48, pearson: 0.002
  test: MAE: 0.46, pearson: 0.001
- Линейная бета-регрессия:
  train: MAE: 0.14, pearson: 0.13
  test: MAE: 0.14, pearson: 0.01
- Нелинейная бета-регрессия:
  train: MAE: 0.06, pearson: 0.89
  test: MAE: 0.16, pearson: 0.03

Даже при условии такого качества среди важных признаков встречается белок QKI с положительной корреляцией, что соответствует исследованиям. [4]

[4] D.V. Maltseva, A.G. Tonevitsky, *RNA-binding proteins regulating the CD44 alternative splicing*

Дальнейшие планы исследования

Первоочередная задача – улучшение качества моделей на данных для более достоверного выявления важных признаков.

Дальнейшие планы исследования

Первоочередная задача – улучшение качества моделей на данных для более достоверного выявления важных признаков. В качестве таких методов уже были попытки встроить в процесс обучения кросс-валидацию (K-folds), а также сделать автоматический подбор гиперпараметров с помощью библиотеки Optuna. Существенного улучшения модели не получили.

Дальнейшие планы исследования

Первоочередная задача – улучшение качества моделей на данных для более достоверного выявления важных признаков. В качестве таких методов уже были попытки встроить в процесс обучения кросс-валидацию (K-folds), а также сделать автоматический подбор гиперпараметров с помощью библиотеки Optuna. Существенного улучшения модели не получили.

После улучшения качества приоритет будет отдан определению наиболее подходящей модели, а также последующему отбору важных признаков и исследованию их биологической значимости.

Литература

- Y. Zhang, J. Qian, C. Gu. et al, *Alternative splicing and cancer: a systematic review*, Signal Transduction and Targeted Therapy, Vol. 6, 78, 2021.
- M.P. Raigorodskaya, V.O. Novosad, S.A Tonevitskaya et al. *Expression of CD44 Isoforms in Human Colorectal Cancer Cell Lines*, Appl Biochem Microbiol vol 58, pp 992–996, 2022.
- V.O. Novosad, *Identification of Significant RNA-Binding Proteins in the Process of CD44 Splicing Using the Boosted Beta Regression Algorithm*, Doklady Biochemistry and Biophysics, vol. 510, no. 1, pp. 99–103, 2023.
- D.V. Maltseva, A.G. Tonevitsky, *RNA-binding proteins regulating the CD44 alternative splicing, Front. Mol. Biosci.,vol 10, 2023.*

Спасибо за внимание!