Предсказание вторичной структуры РНК методом машинного обучения*

Нестерова И. С., Попова М.
irina.nesterova@phystech.edu
[1]МФТИ(ГУ)

Вторичная структура РНК является важной особенностью, которая определяет функциональные свойства РНК. Поэтому очень важно уметь её предсказывать. Все классические методы определения основываются на физическом методе минимизации свободной энергии, которые зависят от экспериментальных данных. Мы же используем новые методы машинного обучения, а именно нейронные сети для автоматизации работы. В нашей работе мы планируем сделать модель нейронной сети, которая на основании двух свойств – SMILES и графа будет классифицировать молекулу.

Ключевые слова: *РНК, машинное обучение, вторичная структура.*

Введение

В этой работе мы представим модель предсказания свойств молекул. Свойство может быть как бинарное - задача классификации, так и вещественное - задача регрессии. Каждая молекула будет описываться двумя разными способами: первый - это строка символов SMILES, второй - это молекулярный граф. Модель представляет из себя нейронную сеть с двумя входами - SMILES и граф.

Ещё в прошлом десятилетии методы предсказания структуры РНК основывались на минимизации потенциала свободной энергии, например [1]. Но, к сожалению, эти методы оказались не достаточно надёжными, что сподвигнуло к созданию новых методов, одни из которых основываюся на машинном обучении, а именно на нейронных сетях.

Ранее уже были созданы модели, которые предсказывали свойства по SMILES, например [2] для предсказания устойчивого вирусологического ответа на наличие гепатита С, и по графам, например [3] для регуляции экспресии генов, отдельно. Мы объеденим эти два параметра и проверим улучшит ли это результат.

РНК взаимодействия являются фундаментальными для клеточной регуляции [4]. Поэтому существует множество методов расчёта их взаимодействия, например [5]. Более того, в нескольких исследованиях подчеркивалось участие молекул РНК в начале и прогрессировании заболеваний человека, включая неврологические расстройства [6]. Бактериальные небольшие РНК имеют широкий спектр регуляторных функций, начиная от экологического зондирования и патогенеза [7]. Можно подвести итог, что РНК учавствует во многих важных биологических процессах.

Постановка задачи

В нашей задаче мы рассматриваем выборку молекул $M = \{m_i \mid i = 1, \ldots, n\}$. Молекула описывается своим молекулярным графом, а также строкой символов SMILES. Описание молекулярного графа представляет из себя матрицу смежности графа и матрицу признаков вершин. Для $m_i \in M : A_i$ – матрица смежности, $A_i \in \mathbb{R}^{|V_{m_i}| \times |V_{m_i}|}$, где V_{m_i} – множество вершин(атомов) молекулы $m_i \in M$. F_i – матрица признаков вершин $F_i \in \mathbb{R}^{|V_{m_i}| \times d}$, где d – количество признаков у вершин. Примерами признаков могут служить физические

Научный руководитель: Стрижов В. В. Задачу поставил: Попова М. О. Консультант: Никитин Ф. О.

и химические свойства атомов, такие как тип атома, валентность, заряд и т.д. Каждая молекула из M обладает свойством $Y = \{y_1, y_2, \ldots, y_n\},\ y_i \in \mathbb{R}$. Примером свойств могут служить растворимость, токсичность, биоактивность, энергия атомизации и т.д. Рассмотрим множество параметрических моделей из класса нейронных сетей $f(y, m, w) \in \mathcal{F}$ – множество моделей, $w \in W$ – множество возможных параметров, $\mathcal{F} = \{f_i, m \to f_i(m)\}$. Требуется решить задачу предсказания y_i для молекулы m_i. Будем рассматривать эту задачу как задачу регрессии, т.к $y_i \in \mathbb{R}$, считая что $y_i \sim \mathcal{N}(\bar{y}, \sigma_i)$, где $\mathcal{N}$ - нормальное распределение. Введем функционал качества $l(y, m, w) = (y - f(m, w))^2$. Тогда задача нахождения оптимального набора параметров модели f может быть записана в виде:$\hat{W} = = \underset{w \in W}{\operatorname{argmin}} \sum_{(y,m) \in (Y,M)} l(y, m, w)$

Заключение

Здесь будет заключение

Литература

[1] Michael Zuker, David H Mathews, and Douglas H Turner. Algorithms and thermodynamics for rna secondary structure prediction: a practical guide. In *RNA biochemistry and biotechnology*, pages 11–43. Springer, 1999.

[2] M Martinot-Peignoux, L Comanor, JM Minor, MP Ripault, B-N Pham, N Boyer, C Castelnau, N Giuily, D Hendricks, and Patrick Marcellin. Accurate model predicting sustained response at week 4 of therapy with pegylated interferon with ribavirin in patients with chronic hepatitis c. *Journal of viral hepatitis*, 13(10):701–707, 2006.

[3] Yongmei Ji, Xing Xu, and Gary D Stormo. A graph theoretical approach for predicting common rna secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20(10):1591–1602, 2004.

[4] Ahmad M Khalil and John L Rinn. Rna–protein interactions in human health and disease. In *Seminars in cell & developmental biology*, volume 22, pages 359–365. Elsevier, 2011.

[5] Federico Agostini, Andreas Zanzoni, Petr Klus, Domenica Marchese, Davide Cirillo, and Gian Gaetano Tartaglia. cat rapid omics: a web server for large-scale prediction of protein–rna interactions. *Bioinformatics*, 29(22):2928–2930, 2013.

[6] Rory Johnson, Wendy Noble, Gian Gaetano Tartaglia, and Noel J Buckley. Neurodegeneration as an rna disorder. *Progress in neurobiology*, 99(3):293–315, 2012.

[7] Jayavel Sridhar and Paramasamy Gunasekaran. Computational small rna prediction in bacteria. *Bioinformatics and biology insights*, 7:BBI–S11213, 2013.