Доклады Международной конференции "Математическая биология и биоинформатика". Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e61. doi: 10.17537/icmbb18.35

Предсказание сайтов связывания ионов магния с РНК методами машинного обучения

Баулин Е.Ф.[1, 2], Тихонова П.О.[3], Ройтберг М.А.[1, 2, 3]

[1]*ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН*
[2]*Московский физико-технический институт (Государственный университет)*
[3]*Национальный исследовательский университет «Высшая школа экономики»*

baulin@lpm.org.ru

Пространственная структура РНК значительно зависит от её окружения, состоящего преимущественно из ионов различных металлов. Ионы магния играют важную роль в функционировании молекул РНК, в отличие от ионов натрия и калия, которые, как правило, компенсируют общий отрицательный заряд структуры. Современные экспериментальные методы разрешения структур макромолекул, такие как рентгеновская кристаллография, часто не могут точно определить местоположение малых молекул, в том числе ионов металла. В данной работе был использован алгоритм машинного обучения «случайный лес» для предсказания сайтов связывания ионов магния с РНК по её пространственной структуре. Результаты работы алгоритма были сопоставлены с результатами работы двух существующих вычислительных сервисов, решающих аналогичную задачу. Несмотря на то, что представленная модель показала немного лучшие результаты в сравнении с существующими подходами, средняя точность предсказаний оказалась недостаточной для решения поставленной задачи. В работе показано, что на данный момент не существует универсального вычислительного сервиса, способного с приемлемой точностью предсказывать сайты связывания ионов магния для произвольной пространственной структуры РНК.

Ключевые слова: *структура РНК, ионы металла, сайт связывания, машинное обучение.*

Machine learning for Mg2+-binding sites prediction in RNA structures

Baulin E.[1, 2], Tikhonova P.[3], Roytberg M.[1, 2, 3]

[1]*IMPB RAS – Branch of KIAM RAS*
[2]*Moscow Institute of Physics and Technology (State University)*
[3]*National Research University Higher School of Economics*

RNA spatial structure highly depends on metal ions environment. Magnesium ions are crucial for RNA functional properties in contrast to sodium and potassium ions that mainly compensate the negative charge of RNA. Modern experimental methods for resolving macromolecular structures, such as X-ray crystallography, are often unable to detect small ligands such as metal ions. In this work we used a machine learning to predict locations of Mg2+-binding sites using RNA spatial structure information. We compared our model with two existing computational services aimed to solve the same problem. Although our model has shown slightly better results compared to existing tools, however, its overall accuracy has been less than satisfactory. All in all, we show that the existing benchmarking approaches require improvement to be able to predict locations of Mg2+-binding sites for arbitrary RNA structures with reasonable accuracy.

Key words: RNA structure, metal ions, binding site, machine learning.

1. Введение

Ионы металлов играют важную роль в формировании пространственной структуры РНК, её стабилизации, а также во взаимодействии РНК с другими молекулами [1–3]. Взаимодействие ион–РНК осуществляется двумя способами – неспецифическое (диффузионное) и специфическое связывание. При диффузионном связывании полностью гидратированный (т.е. окруженный молекулами воды) ион подходит к гидратированной поверхности РНК и нейтрализует её отрицательный заряд. Взаимодействие осуществляется через два слоя молекул воды, непосредственного контакта не происходит. В свою очередь специфическое связывание происходит либо путём пересечения водных оболочек (у иона и РНК есть «общие» молекулы воды), либо путём образования прямых

Proceedings of the International Conference "Mathematical Biology and Bioinformatics". *Ed. V.D. Lakhno. Vol. 7. Pushchino: IMPB RAS, 2018. Paper No. e61. doi: 10.17537/icmbb18.35*

координационных связей между ионом и атомами РНК.

В отличие от моновалентных ионов, таких как Na+ и K+, которые связываются с РНК в основном диффузионно [4], специфически связанные ионы Mg2+ могут образовывать структурные мотивы, необходимые для функционирования молекул РНК [5].

Известно, что ионы предпочитают подходить либо к фосфатам, либо к атомам большой бороздки в спиралях [6]. Это обусловлено тем, что в этих местах концентрируется отрицательный заряд.

Большое число работ было проведено для анализа паттернов связывания ионов с отдельными нуклеотидами и даже с отдельными основаниями [7]. Известно, что ион может связываться с любым атомом нуклеотида, в том числе с углеродом [8]. Известны также наиболее предпочтительные для связывания атомы [8].

Тем не менее, в настоящий момент не существует надежного способа определить, с каким участком данной молекулы РНК будет связан ион (говорят: где находится *сайт связывания* иона). Каждый из существующих подходов имеет свои недостатки. Существует три основных подхода для определения сайтов связывания.

Экспериментальные подходы можно разделить на два типа: методы разрешения структур макромолекул (ядерно-магнитный резонанс, рентгеновская кристаллография и др.), которые также позволяют определять окружение структуры; и так называемые rescue-методы [9, 10], которые позволяют определять местоположение и функциональную значимость специфически связанных ионов. Методы разрешения структур зачастую не позволяют точно определять координаты малых молекул, в том числе ионов металлов [11], что определяет актуальность задачи предсказания их сайтов связывания. В свою очередь rescue-методы являются дорогостоящими, как по стоимости, так и по времени их работы.

На данный момент существует ряд подходов, основанных на компьютерном имитационном моделировании (молекулярная динамика, броуновская динамика). Так, в работе [12] описана модель броуновской динамики диффузии ионов металлов для предсказания сайтов связывания, а в работе [13] используются уравнения Пуассона–Больцмана для определения вероятных областей расположения ионов. Среди недостатков данных методов можно отметить необходимость больших вычислительных мощностей и недостаточно надежные данные для расчета параметров взаимодействия.

Существуют также вычислительные методы определения сайтов связывания ионов путём минимизации энергии связывания на основе подобранных статистических потенциалов взаимодействия.

На данный момент существует два вычислительных сервиса, предсказывающих сайты связывания ионов магния по пространственной структуре РНК. Сервис FEATURE [14] использует U-критерий Манна–Уитни для определения значимости структурных признаков РНК, которые используются для классификации точек пространства на два класса – содержащие и несодержащие ион магния. Сервис MetalIonRNA [15] основан на вычислении анизотропного статистического потенциала для пар ковалентно-связанных атомов РНК. Вывод о наличии сайта связывания делается на основе суммы значений перекрывающихся потенциалов.

В данной работе описан метод предсказания сайтов связывания ионов магния с РНК, основанный на алгоритме машинного обучения «случайный лес». Проведён сравнительный анализ работы представленной модели и сервисов FEATURE и MetalIonRNA. Насколько нам известно, классические методы машинного обучения к решению данной задачи не применялись.

2. Материалы и методы

2.1. Постановка задачи

Будем решать задачу классификации фрагментов РНК на два типа – связанные с ионом магния (класс 1) и несвязанные (класс 0). Поскольку по координатам невозможно достоверно определить специфичность связывания с ионом, класс 1 будем определять по расстоянию от фрагмента до ближайшего иона магния. Фрагмент принадлежит к классу 1, если в радиусе X Å находится хотя бы один ион магния. Рассмотрим несколько вариантов задачи. Фрагментом РНК в разных вариантах будем считать нуклеотид, фрагмент нуклеотида (фосфат / рибоза / основание) или O/N атом РНК. Значения радиусов X выберем равными 3 Å, 5 Å или 7 Å. Поскольку вариант с радиусом в 3 Å ограничен прямыми контактами с ионом магния, будем также рассматривать вариант задачи с расстоянием от 3 Å до 7 Å, в котором учитываются только контакты через молекулы воды.

2.2. Обучающая выборка

Источником данных была выбрана база URSDB [16], включающая все РНК-содержащие документы банка PDB [17] с аннотированными структурными элементами РНК. Всего URSDB включает около 5000 структур РНК, содержащих около 195000 ионов магния. Поскольку в PDB присутствует множество повторов, для создания обучающей выборки использовалось неизбыточное подмножество структур РНК [18], т.е. для каждого организма и типа РНК была выбрана одна структура в качестве представителя класса. Представитель класса определялся двумя разными способами – структура с наилучшим разрешением (код minresol) или структура с максимальным числом ионов

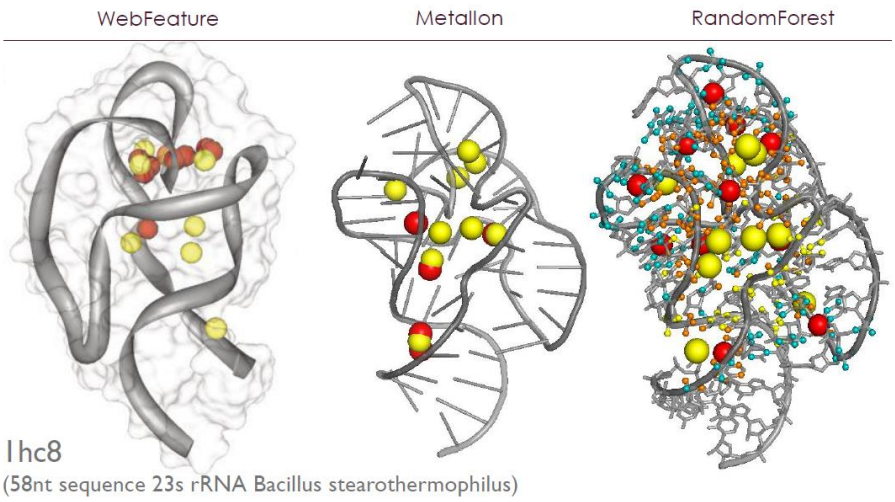

Рис. Результаты работы алгоритмов WebFeature, MetalIon и RandomForest. Реальные ионы магния представлены большими жёлтыми шарами, предсказанные – большими красными. Маленькими шариками выделены результаты классификации атомов РНК алгоритмом RandomForest – верно отнесённые к классу 1 (рыжий цвет), ошибочно отнесённые к классу 1 (голубой цвет), ошибочно отнесённые к классу 0 (жёлтый цвет).

магния (код maxmg). Таким образом, было составлено 24 выборки (два варианта выбора представителя, три варианта выбора элемента выборки, четыре варианта значения радиуса), каждая из которых содержала около 330 структур РНК. Доля элементов класса 1 составляла от 5 % до 20 %, в зависимости от выбранного радиуса и элемента выборки.

Каждый элемент выборки содержал от 362 до 383 (в зависимости от варианта задачи) значений признаков, которые можно разделить на 7 групп:

1) общие характеристики структуры – длина цепи, присутствие белка, разрешение структуры;
2) торсионные углы;
3) тип основания;
4) данные о спариваниях оснований;
5) данные о вторичной структуре;
6) вид фрагмента нуклеотида или атома;
7) связь с магнием (1 или 0);

Признаки групп (2–5) включали данные, как о выбранном нуклеотиде, так и о его соседях (по 2 нуклеотида в каждую сторону).

2.3. Подготовка данных

По результатам предварительных экспериментов из 24 выборок была определена выборка для дальнейшей работы – minresol_A7. В данной выборке представителем класса структур является структура с наилучшим разрешением, элементом выборки является O/N атом РНК, присутствие иона магния определяется в радиусе 7 Å.

Среди более чем 300 признаков были выбраны наиболее информативные, дисперсия которых превышала заданный порог. Пропущенные значения (например, данные о соседях для крайних нуклеотидов цепи) во всех случаях заменялись нулевыми значениями. Среди признаков, описывающих значения торсионных углов, были выявлены 6 групп скоррелированных значений, из каждой группы выбирался единственный признак. Всего в итоговой модели рассматривалось 170 признаков.

2.4. Описание модели

Для решения поставленной задачи был выбран алгоритм машинного обучения «случайный лес» (реализация RandomForest из Python-библиотеки scikit-learn [19]), как наиболее устойчивый к переобучению. Обучающая выборка балансировалась автоматически методом under-sampling. Параметры алгоритма были определены в ходе тестовых запусков. Итоговая модель включала следующие значения основных параметров: Max_depth = 26, Min_samples_leaf = 20, Max_features = 0.7. В процессе кросс-валидации было выставлено правило, запрещающее разным элементам выборки, принадлежащим одной структуре РНК, попадать одновременно в тестовую и обучающую части выборки.

Репозиторий с исходными кодами и результатами экспериментов см. https://pollytikhonova.github.io/coursework/.

3. Результаты

Результаты работы описанной модели были сопоставлены с результатами сервисов FEATURE и MetalIonRNA. Поскольку онлайн-версия сервиса FEATURE в настоящее время недоступна, сравнения проводились на 12 структурах, описанных в работе [14]. Для 8 структур из 12 результаты нашей модели незначительно превосходили результаты существующих сервисов, однако разница составляла 5–10 % и находилась в рамках погрешности. Стоит отметить, что в отличие от алгоритмов FEATURE и MetalIonRNA наша модель не предсказывает точного местоположения ионов магния, поэтому для сравнительного анализа

результатов была реализована программа аппроксимации координат предсказанных ионов магния с помощью алгоритма кластеризации K-средних. На рисунке представлены результаты предсказаний трёх алгоритмов на примере структуры с PDB-кодом 1hc8. Визуализация результата FEATURE взята из работы [14].

Стоит отметить, что, несмотря на достаточное качество полученных результатов для выбранных 12 структур, средние результаты работы описанной модели на всей выборке не позволяют сделать вывод об успешном решении поставленной задачи. Так, значение F-меры для всей выборки не превышает значения 0.3. Также, Сервис MetalIonRNA для произвольных структур РНК, не включая структуры из работы [14], показывал точность не более 60 % по количеству предсказанных ионов магния от числа ионов магния, содержащихся в структуре.

Таким образом, на данный момент не существует сервиса, способного с приемлемой точностью предсказывать сайты связывания ионов магния для произвольных пространственных структур РНК.

4. Благодарности

Работа выполнена при поддержке гранта РФФИ 16-04-01640.

5. Список литературы

1. Tan Z., Zhang W., Shi Y., Wang F. RNA folding: structure prediction, folding kinetics and ion electrostatics. In: *Advance in Structural Bioinformatics*. Dordrecht: Springer, 2015. P. 143–183. doi: 10.1007/978-94-017-9245-5_11.
2. Saunders A.M., DeRose V.J. Beyond Mg2+: functional interactions between RNA and transition metals. *Current opinion in chemical biology*. 2016. V. 31. P. 153–159. doi: 10.1016/j.cbpa.2016.02.015.
3. Bowman J.C., Lenz T.K., Hud N.V., Williams L.D. Cations in charge: magnesium ions in RNA folding and catalysis. *Current opinion in structural biology*. 2012. V. 22. № 3. P. 262. doi: 10.1016/j.sbi.2012.04.006.
4. Draper D.E. A guide to ions and RNA structure. *RNA* 2004. V. 10. № 3. P. 335–343. doi: 10.1261/rna.5205404.
5. Pyle A. Metal ions in the structure and function of RNA. *JBIC Journal of Biological Inorganic Chemistry*. 2002. V. 7. № 7–8. P. 679–690. doi: 10.1007/s00775-002-0387-6.
6. Cate J.H., Doudna J.A. Metal-binding sites in the major groove of a large ribozyme domain. *Structure*. 1996. V. 4. № 10. P. 1221–1229. doi: 10.1016/S0969-2126(96)00129-3.
7. Cerda B.A., Wesdemiotis C. Li+, Na+, and K+ binding to the DNA and RNA nucleobases. Bond energies and attachment sites from the dissociation of metal ion-bound heterodimers. *Journal of the American Chemical Society*. 1996. V. 118. № 47. P. 11884–11892. doi: 10.1021/ja9613421.
8. Lippert B. Multiplicity of metal ion binding patterns to nucleobases. *Coordination Chemistry Reviews*. 2000. V. 200. P. 487–516. doi: 10.1016/S0010-8545(00)00260-5.
9. Hougland J.L., Kravchuk A.V., Herschlag D., Piccirilli J.A. Functional identification of catalytic metal ion binding sites within RNA. *PLoS biology* 2005. V. 3. № 9. P. *e*277. doi: 10.1371/journal.pbio.0030277.
10. Frederiksen J.K., Li N.S., Das R., Herschlag D., Piccirilli J.A. Metal-ion rescue revisited: biochemical detection of site-bound metal ions important for RNA folding. *RNA*. 2012. V. 18. № 6. P. 1123–1141. doi: 10.1261/rna.028738.111.
11. Klein D.J., Moore P.B., Steitz T.A. The contribution of metal ions to the structural stability of the large ribosomal subunit. *RNA*. 2004. V. 10. № 9. P. 1366–1379. doi: 10.1261/rna.7390804.
12. Hermann T., Westhof E. Exploration of metal ion binding sites in RNA folds by Brownian-dynamics simulations. *Structure*. 1998. V. 6. № 10. P. 1303–1314. doi: 10.1016/S0969-2126(98)00130-0.
13. Misra V.K., Draper D.E. Mg2+ binding to tRNA revisited: the nonlinear Poisson-Boltzmann model. *Journal of molecular biology*. 2000. V. 299. № 3. P. 813–825. doi: 10.1006/jmbi.2000.3769.
14. Banatao D.R., Altman R.B., Klein T.E. Microenvironment analysis and identification of magnesium binding sites in RNA. *Nucleic acids research*. 2003. V. 31. № 15. P. 4450–4460. doi: 10.1093/nar/gkg471.
15. Philips A., Milanowska K., Lach G., Boniecki M., Rother K., Bujnicki J.M. MetalionRNA: computational predictor of metal-binding sites in RNA structures. *Bioinformatics*. 2011. V. 28. № 2. P. 198–205. doi: 10.1093/bioinformatics/btr636.
16. Baulin E., Yacovlev V., Khachko D., Spirin S., Roytberg M. URS DataBase: universe of RNA structures and their motifs. *Database*. 2016. doi: 10.1093/database/baw085.
17. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank, 1999–. *International Tables for Crystallography*. 2006. V. F. P. 675–684. doi: 10.1107/97809553602060000722.
18. Leontis N.B., Zirbel C.L. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: *RNA 3D structure analysis and prediction.* Berlin, Heidelberg: Springer, 2012. P. 281–298. doi: 10.1007/978-3-642-25740-7_13.
19. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J. Scikit-learn: Machine learning in Python. *Journal of machine learning research*. 2011. V. 12. P. 2825–2830. doi: 10.1016/j.patcog.2011.04.006.