

# Поиск элементов структур РНК

Е.Ф.Баулин<sup>(1,2)</sup>, С.А.Спирин<sup>(3)</sup>, М.А.Ройтберг<sup>(2)</sup>

Пушкинский Естественно-Научный Институт, Пушкино<sup>(1)</sup>, Институт Математических Проблем Биологии РАН, Москва<sup>(2)</sup>, Институт Физико-Химической Биологии им. А.Н. Белозерского МГУ, Москва<sup>(3)</sup>

[baulin@lpm.org.ru](mailto:baulin@lpm.org.ru)

We present a database URS (the Universe of RNA Structures) of RNA structures and their elements (e.g. base pairs, loops, stems, pseudoknots) and a search engine allowing one to search such elements and obtain statistics on them. The search engine is available on the web at <http://server3.lpm.org.ru/urs/>

## 1 База данных

### 1.1 Материалы и методы

В качестве исходных данных используются документы из Protein Data Bank (PDB,[1]). На данный момент в базе данных содержится информация о 2988 РНК-содержащих структурах. Поскольку отдельный документ из PDB может содержать несколько моделей одной структуры, все документы были предварительно разделены на отдельные модели (суммарное число моделей – 9437).

Для разметки спариваний между основаниями в РНК была использована программа DSSR из пакета X3DNA [5]. Преимущество данной программы перед аналогами состоит в том, что она является наиболее современной, а также предоставляет наиболее детальную информацию о вторичной структуре РНК.

Для обработки исходных данных используется оригинальный пакет программ, производящий разметку элементов вторичной структуры РНК, таких как нити, крылья, петли, спирали и псевдоузлы.

### 1.2 Возможности

На данный момент база данных URS DB состоит из 51 таблицы. Все таблицы можно условно разделить на 4 группы, а именно: таблицы исходных данных PDB (документы, цепи, основания, атомы), таблицы результатов работы DSSR (спаривания между основаниями, динуклеотидные шаги, спирали, петли, третичные мотивы), таблицы разметки элементов вторичной структуры (петли, псевдоузлы, элементарные закрытые участки, мультиплеты и др.), вспомогательные таблицы. Таблицы второй группы (за исключением таблицы спариваний) в работе веб-интерфейса не участвуют, а хранятся для решения технических задач, в частности, для анализа работы программы DSSR и сравнительного анализа. Таблицы четвертой группы постоянно меняются и

обновляются в соответствии с ведущейся экспериментальной работой. На данный момент при работе с веб-интерфейсом среди этих таблиц доступна только таблица спиральных мультиплетов.

### 1.3 Особенности

Для разметки элементов вторичной структуры РНК используется оригинальный подход, позволяющий описывать как классические структуры, так и структуры, содержащие псевдоузлы. Подход основан на обобщении определения петли, предложенном в работах Мэтьюза-Тернера [7]. Как и в случае Мэтьюза-Тернера, каждая петля соответствует спирали («внешняя спираль» петли) и, наоборот, каждая спираль является внешней спиралью некоторой петли. Главное отличие от определения Мэтьюза-Тернера состоит в том, что петля может содержать «крылья» спиралей, такие спирали образуют псевдоузел с внешней спиралью петли. Это позволяет применять данный подход к псевдоузловым структурам. В случае классических структур РНК петли по Мэтьюзу-Тернеру и по предложенному подходу идентичны. Более подробно см. нашу работу [8].

Для описания и классификации псевдоузлов используется метод, схожий с предложенными в работах [2,3,6]. Основой метода является разложение вторичной структуры на элементарные закрытые участки (ЭЗУ). Каждому псевдоузловому ЭЗУ ставится в соответствие т.н. сигнатура крыльев (пример: H-узел = abAB, kissing hairpins = abAcBC).

Одним из главных отличий описанной базы данных является совмещение детальной информации о вторичной структуре РНК (включая псевдоузлы) с данными о водородных связях между РНК и белком. На данный момент не существует баз, обладающих обоими указанными свойствами.

## 2 Веб-интерфейс

### 2.1 Общая информация

Разработанный интерфейс позволяет пользователю формировать выборку интересующих его РНК-содержащих структур, после чего собирать статистику по имеющимся в них структурным элементам. Кроме того пользователю доступны

индивидуальные структуры и элементы с подробными данными для детального анализа.

## 2.2 Формирование выборки структур

Формирование выборки структур происходит согласно запросу к базе данных; запрос состоит из конъюнкции нескольких дизъюнкций элементарных условий. Элементарные условия, доступные пользователю на данный момент, состоят из 4 групп: (1) общая информация о PDB документе, (2) информация о содержащихся макромолекулах, (3) шаблоны структурных элементов, а также (4) информация о содержащихся спариваниях между основаниями в РНК и между атомами РНК и белка. Результатом запроса является список структур, удовлетворяющих введенным параметрам.

## 2.3 Страница структуры

Для индивидуального анализа интересующей пользователя структуры имеется окно структуры, доступное по клику на ID структуры в списке результатов поиска структур. Данное окно содержит детальную информацию об имеющихся в структуре цепях, спариваниях, петлях, спиралах и псевдоузлах. В будущем этот список будет пополняться. Также в окно встроено 3D-визуализатор молекул Jsmol [4].

## 2.4 Статистика и поиск элементов

После формирования выборки документов пользователь имеет возможность собирать статистику структурных элементов, содержащихся в выбранных структурах. На данный момент в качестве таких элементов доступны цепи, спаривания, петли, спирали, псевдоузлы, мультиплеты, а также РНК-белковые взаимодействия. По каждому виду элемента доступны фильтры, позволяющие ограничивать элементы необходимыми параметрами, а также полные списки найденных элементов с возможностью вывода детальной информации об отдельном элементе по клику.

## 3 Результаты

Был проведен анализ имеющихся в базе псевдоузлов. Все псевдоузлы были разбиты на классы согласно предложенному методу. Каждому классу псевдоузлов соответствует определенная сигнатура крыльев. Количественный анализ показал, что наиболее распространены H-узлы (сигнатура abAB, 2762 случаев) и kissing hairpins (сигнатура abAcBC, 526 случаев). Всего в базе имеется 26 различных классов псевдоузлов. В таблице 1 указаны сигнатуры классов и количество реализаций этих сигнатур в базе данных.

Таблица 1

Сигнатура	Количество
abAB	2762
abAcBC	526

abAcdeDfgFEChBHiGI	289
abAcdeBEfDFC	16
abcdCADB	15
abAcBdCD	13
abAcdeDfgFEChijIkBKHIGLJ	12
abAcdeCeBEfDF	11
abAcdefEghGFDiBIjHJC	9
abAcdeBDeCE	8
abcBdefECDAF	7
abcABC	4
abcdCefAFDEB	3
abcBdeDCAFEF	2
abcBdeDAEC	2
abcdeDfgEGhFBCAH	2
abAcdeDfBFCE	1
abAcdeDfBFgEGC	1
abcdCABeDE	1
abcACdBD	1
abcdeDBCAfghiEIGHF	1
abcBdefEghGFDijkJICLIhMhMAhNoNKO	1
abcBdefgFhiHGEjCjKIADL	1
abcBdeAEfgDGChFH	1
abcdCeBEAFDF	1
abcdBCAD	1

## Список литературы

- [1] Berman H. M. et al. // *Nucleic acids research*. – 2000. – Т. 28. – №. 1. – С. 235-242.
- [2] Bon M. et al. Topological classification of RNA structures // *Journal of molecular biology*. – 2008. – Т. 379. – №. 4. – С. 900-911.
- [3] Condon A. et al. Classifying RNA pseudoknotted structures // *Theoretical Computer Science*. – 2004. – Т. 320. – №. 1. – С. 35-50.
- [4] Hanson R. M. et al. JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia // *Israel Journal of Chemistry*. – 2013. – Т. 53. – №. 3-4. – С. 207-216.
- [5] Lu X. J., Olson W. K., Bussemaker H. J. // *Nucleic acids research*. – 2010. – Т. 38. – №. 14. – С. 4868-4876.
- [6] Reidys C. M. et al. Topology and prediction of RNA pseudoknots // *Bioinformatics*. – 2011. – Т. 27. – №. 8. – С. 1076-1085.
- [7] Zuker M., Mathews D. H., Turner D. H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide // *RNA biochemistry and biotechnology*. – Springer Netherlands, 1999. – С. 11-43.
- [8] Баулин Е. Ф., Астахова Т. В., Ройтберг М. А. Классификация элементов вторичной структуры РНК // *Математическая биология и биоинформатика*. – 2012. – Т. 7. – №. 2. – С. 567-571.