Математическая биология и биоинформатика
2015. Т. 10. № 1. С. 54–59. doi: 10.17537/2015.10.54

========== **БИОИНФОРМАТИКА** ==========

УДК: 577.214.5

Стемовые мультиплеты: новый подход к описанию третичных мотивов РНК

Баулин Е.Ф.[*1,2,3], **Ройтберг М.А.**[**1,2,4]

[1]*Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия*
[2]*Факультет компьютерных наук, Национальный исследовательский университет «Высшая школа экономики», Москва, 101000, Россия*
[3]*Учебный центр математической биологии, Пущинский государственный естественно-научный институт, Пущино, Московская область, 142290, Россия*
[4]*Московский физико-технический институт, Долгопрудный, Московская область, 141700, Россия.*

Аннотация. В работе предлагается подход к описанию сложных фрагментов третичной структуры РНК, основанный на понятии стемовых мультиплетов, т. е. элементов структуры, образованных двумя и более стемами. С помощью базы данных структур РНК (http://server3.lpm.org.ru/urs/) составлен список всех стемовых мультиплетов, встречающихся в базе данных PDB.

Ключевые слова: *пространственная структура РНК, мультиплет, стем, псевдоузел, линк, граф.*

ВВЕДЕНИЕ

Исследование пространственных структур РНК, а также их функций является одним из основных направлений современной молекулярной биологии. Развитие данной области затруднено отсутствием единого языка, позволяющего описывать произвольные структуры РНК, в том числе и псевдоузловые структуры. На данный момент таким общепринятым языком для классических, т. е. не содержащих псевдоузлов, структур является модель Цукера-Мэтьюза-Тернера (Nearest Neighbor Model, NNM, [1, 2]). Согласно этой модели вторичная структура представляется в виде графа, вершины которого соответствуют нуклеотидам, а ребра – водородным и ковалентным связям между нуклеотидами; в этом графе выделяются элементарные циклы (в терминологии NNM–петли, loops)[3]. Часто графы вторичных структур представляются в виде дуговых диаграмм [4]. Наиболее распространенные базы данных вторичных структур РНК [5, 6] основаны на модели NNM, из-за чего в них не рассматриваются более сложные структуры, содержащие, например, псевдоузлы.

На данный момент существует несколько подходов к описанию сложных вторичных структур. В работе [7] предлагается представление псевдоузлов в виде стандартного графа канонической вторичной структуры с дополнительными ребрами, отвечающими псевдоузлам. Главный недостаток такого подхода заключается в том, что равноправные спирали (например, образующие псевдоузлы H-типа) зачастую изображаются по-разному, что негативно отражается на наглядности и затрудняет работу исследователей. Кроме того, такой способ непригоден для изображения достаточно сложных структур, см., например, рис. 2 ниже. В работе [8] описаны би-

[*]baulin@lpm.org.ru
[**]mroytberg@lpm.org.ru

дуговые диаграммы, являющиеся более наглядным способом представления дуговых диаграмм, содержащих пересекающиеся дуги. Это представление подходит для описания псевдоузлов 2-го уровня (не более двух взаимопересекающихся дуг), однако реальные пространственные структуры РНК содержат и более сложные псевдоузлы. В работе [9] предлагается оригинальный подход к описанию вторичных структур РНК в виде графов, соответствующих её элементам (петлям, нитям, спиралям), однако, вследствие недостаточной детализации данный подход не был широко распространен. Наиболее актуальным на наш взгляд является топологический подход к классификации псевдоузлов, описанный в работах [10–12].

Описание третичных структур РНК также представляет некоторые трудности. На данный момент наиболее распространены два подхода. В работе [13] предлагается нотация неканонических спариваний между основаниями в РНК, которая накладывается на стандартный граф вторичной структуры. В работе представлены изображения наиболее типичных третичных мотивов (S-turn, E-loop). В работе [14] представлена база данных третичных мотивов РНК. Алгоритм поиска в данной работе основан на геометрии остовных атомов РНК (по умолчанию атомы C1'). В обеих работах описаны базы данных, в этих базах проведена аннотация и классификация третичных мотивов. Недостаток данных работ состоит в том, что не рассматриваются неканонические спаривания внутри стемов и между стемами.

В данной работе мы предлагаем подход к описанию третичной структуры РНК, основанный на стемовых мультиплетах (stem multiplet). *Стем* (stem) – это элемент структуры РНК, образованный двумя фрагментами цепи РНК, между которыми установлены водородные связи. При этом в стемах допускаются и неканонические спаривания. *Стемовый мультиплет* – это набор обобщенных стемов, которые связаны друг с другом водородными связями, либо содержат общие нуклеотиды. Каждый стемовый мультиплет описывается графом, вершины которого соответствуют обобщенным стемам, а ребра описывают связи между этими стемами.

Такой подход позволяет провести классификацию сложных элементов пространственной структуры РНК. В частности, с помощью ранее разработанной нами базы URSDB (http://server3.lpm.org.ru/urs/) мы выявили все типы стемовых мультиплетов, которые встречаются в базе PDB [15].

МАТЕРИАЛЫ И МЕТОДЫ

Терминология

Молекулу РНК мы будем представлять, как последовательность нуклеотидов, иначе говоря, как символьную последовательность в алфавите {A, C, G, U}. Каждый нуклеотид в молекуле имеет свой номер от 1 до L, где L – длина последовательности.

Спаривание – это пара нуклеотидов (i, j), где $i < j$, которые образуют водородные связи. При этом допускаются не только связи между комплементарными нуклеотидами и G-U связи, но и неканонические связи, см. [13].

Стем – это последовательность пар (i,j), $(i + 1, j – 1)$,..., $(i + k, j – k)$, такая что:

1) $k \geq 1$;

2) $i + k<j – k$;

3) все пары $(i + x, j – x)$, где $x = 0, ..., k$, образуют спаривания.

Фрагменты $[i, i+k]$ и $[j–k, j]$ называются *крыльями* стема.

Стем называется *каноническим,* если все его спаривания – это спаривания по Уотсон-Крику (Watson-Crick basepairs, WC) либо G-U спаривания (wobble base pairs, WB).

Таким образом, стем содержит не менее двух идущих подряд спариваний. В то же время, анализ содержимого PDB показал, что в структурах РНК есть значительное

количество "одиночных" спариваний, не входящих ни в один стем. Такие спаривания будем называть *линками.*

Будем говорить, что линк (*i, j*) *связывает* два обобщенных стема *S*1,*S*2, если *i* принадлежит одному из крыльев стема *S*1, а *j* принадлежит одному из крыльев стема *S*2. Два стема *S*1, *S*2 имеют *общий* нуклеотид, если существует нуклеотид, принадлежащий как крылу стема S1, так и крылу стема S2. *Стемовый мультиплет* (SM) – это связный граф (*V*,*E*), такой что:

1) *V* – множество стемов;
2) *E* = *EC*+*EL*, где:
   а) *EC* – это множество всех таких пар (*S*1,*S*2), что *S*1 и *S*2 имеют общий нуклеотид;
   б) *EL* – это множество всех таких пар (*S*1,*S*2), что *S*1 и *S*2 связаны линком.

Ребра стемовых мультиплетов обладают двумя характеристиками - количеством общих нуклеотидов и количеством связывающих линков.

Исходные данные

Для анализа были отобраны все документы пространственных структур РНК из банка данных PDB [15], в том числе и РНК-белковые комплексы. Разметка спариваний производилась с помощью программы DSSR из пакета 3DNA [16]. Данная программа была выбрана среди аналогичных программ исходя из того, что она оперирует геометрическими характеристиками спариваний (например, скручивание, подъем, наклон и др.), что позволяет отслеживать адекватность произведенной разметки.

Разметка стемовых мультиплетов была внедрена в разработанную ранее базу данных структур РНК URSDB (http://server3.lpm.org.ru/urs/). Для этого был написан отдельный программный модуль на языке Python, вошедший в число расширений оригинального программного пакета. Разметка была реализована в виде 3 таблиц: таблица стемовых мультиплетов, таблица связывающих линков и таблица общих нуклеотидов. Таблицы линков и нуклеотидов содержат исчерпывающий набор параметров, необходимый для детального изучения отдельных случаев.

Таблица стемовых мультиплетов, помимо стандартных характеристик графа хранит оригинальную строку-инвариант, которая позволяет сравнивать и классифицировать графы, не производя дополнительных вычислений. На данный момент, инвариант не учитывает характеристики ребер, в будущем это будет исправлено.

РЕЗУЛЬТАТЫ

Таблица 1. Распределение встречаемости стемовых мультиплетов по числу содержащихся в них стемов

Число стемов в мультплете	Число мультиплетов
2	18072
3	5023
4	1469
5	774
6	329
7	191
8–12(всего)	106
Всего	**25964**

Мы выявили все стемовые мультиплеты, содержащиеся в базе PDB (version 3.30), см. таб. 1. На данный момент (30.11.2014) в базе данных PDB выявлено 25964 стемовых мультиплета. Как и следовало ожидать, большинство из них состоят из двух

обобщенных стемов, соединенных либо (1) связывающими линками (9140), либо (2) общими нуклеотидами (5989), либо (3) и тем и другим (2943). SM из первой группы, как правило, отвечают дефекту двойной спирали (рис. 1,А), SM из второй группы (особенно если общих нуклеотидов больше одного) отвечают граничным случаям, когда нельзя точно выяснить, с каким из двух соседей связан тот или иной нуклеотид (рис. 1,Б). Особый интерес представляют SM из третьей группы – для них типично наличие триплета на границе между обобщенными стемами (рис. 1,В), причем это могут быть как и удаленные по цепи стемы, так и идущие подряд друг за другом.

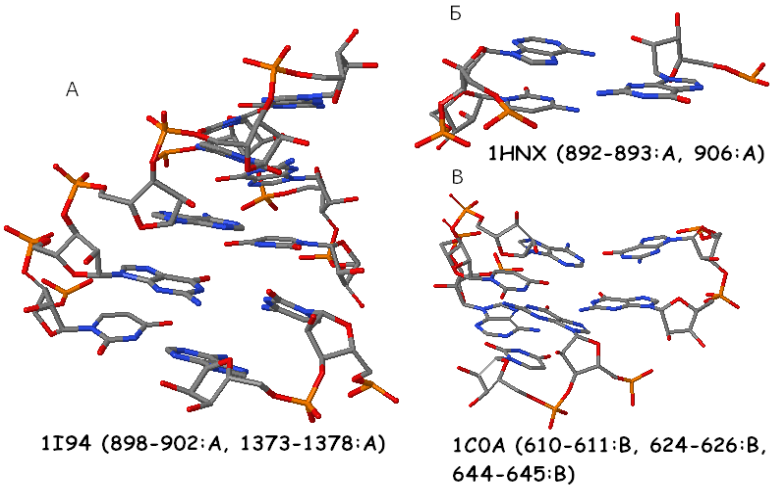

Рис. 1. Фрагменты стемовых мультиплетов, состоящих из двух обобщенных стемов: А. дефект двойной спирали. Б. альтернативное спаривание. В. триплет на границе стемов. Рисунок создан с помощью программы JSmol [17].

SM, содержащие более 7 обобщенных стемов достаточно редки (в сумме около100) и нетипичны, однако сами по себе весьма интересны. Самым большим SM является мультиплет из документа 3J12, он состоит из 12 стемов и содержит 15 нуклеотидов, принадлежащих одновременно двум стемам, и 7 связывающих стемы линков. Стоит отметить, что этот SM содержит фрагмент из 4 стемов (см. рис. 2), который встречается в примерно 150 структурах. При этом лишь в одном случае (документ 3J11) этот 4-стемовый фрагмент является полным мультиплетом; в остальных случаях он является частью более сложного мультиплета. Данный фрагмент будет нами изучен более детально в будущем.

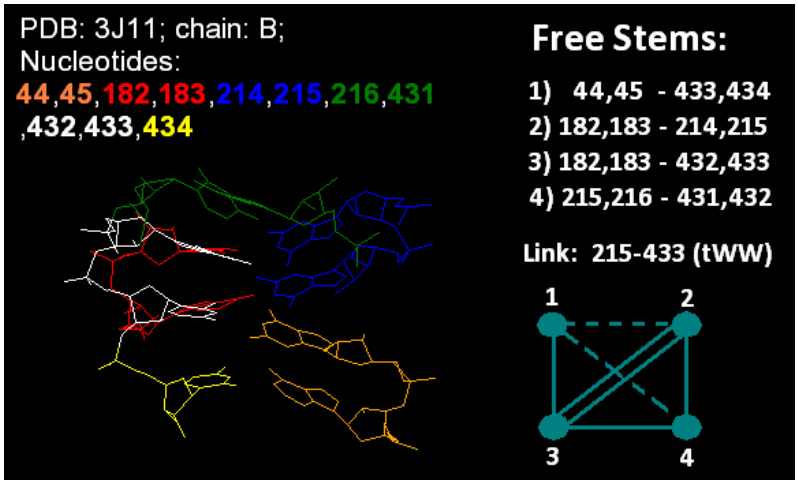

Рис. 2. Стемовый мультиплет, соответствующий полному графу из четырех вершин. Мультиплет отвечает сближению четырех удаленных по цепи фрагментов. На графе связывающие линки

показаны пунктиром, общий нуклеотид соответствует непрерывной линии. Источник: 3J11. Рисунок создан с помощью программы JSmol [17].

ЗАКЛЮЧЕНИЕ

Нами предложен новый подход к описанию сложных пространственных структур РНК, позволяющий описывать плотно уложенные фрагменты цепей. На основе предложенных определений была произведена разметка стемовых мультиплетов в пространственных структурах РНК из банка PDB. Важным отличием нашего подхода является учет неканонических связей внутри стемов. Анализ PDB показал, что в структурах РНК содержится около 176 000 стемов, из которых менее 66 000 – канонические стемы. При этом около 32 000 стемов являются «сильно неканоническими», то есть не содержат двух идущих подряд канонических спариваний.

Около 70% всех стемовых мультиплетов содержат только два обобщенных стема. В этих мультиплетах можно выделить несколько типичных видов связей между стемами: (1) только связывающие линки, (2) только общие нуклеотиды и (3) общие нуклеотиды, связанные линками. Анализ мультиплетов, содержащих 7 и более обобщенных стемов, показал, что хотя их число невелико, и все они различны, тем не менее существуют типичные подграфы, подобные тому, что представлен на рисунке 2. В будущем мы усовершенствуем используемые методы и проведем работу по выявлению консервативных фрагментов имеющихся графов и их классификации.

В то же время в ходе работы выявились и недостатки предложенного подхода. Так, он дает избыточно сложное описание для структур, образованных двумя участками РНК, в которых один нуклеотид одного участка образует водородные связи с двумя нуклеотидами другого участка. Для подобных структур, по-видимому, более адекватно представление, основанное на взаимодействии не стемов, а отдельных фрагментов РНК(«крыльев»). Мы планируем разработать такое представление и провести классификацию возможных структур РНК, основанную на этом представлении.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 12-04-00944 и № 14-01-93106).

СПИСОК ЛИТЕРАТУРЫ

1. Zuker M., Mathews D.H., Turner D.H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: *RNA Biochemistry and Biotechnology*. Eds. J. Barciszewski, B.F.C. Clark. Boston: Kluwer Academic Publishers, 1999. P. 11–43. (NATO ASI Series).
2. Xia T., SantaLucia J.Jr., Burkard M.E., Kierzek R., Schroeder S.J., Jiao X., Cox C., Turner D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*. 1998. V. 37. P. 4735.
3. Darty K., Denise A., Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*. 2009. T. 25. № 15. P. 1974.
4. Lai D., Proctor J.R., Zhu J.Y.A., Meyer I.M. R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Research*. 2012. V. 40. № 12. P. e95. doi: 10.1093/nar/gks241.
5. Vanegas P.L., Hudson G.A., Davis A.R., Kelly S.C., Kirkpatrick C.C., Znosko B.M. RNA CoSSMos: Characterization of Secondary Structure Motifs–a searchable database of secondary structure motifs in RNA three-dimensional structures. *Nucleic Acids Research*. 2012. V. 40. P. D439–D444. doi: 10.1093/nar/gkr943.
6. Popenda M., Szachniuk M., Blazewicz M., Wasik S., Burke E.K., Blazewicz J., Adamiak R.W. RNA FRABASE 2.0: an advanced web-accessible database with the

capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*. 2010. T. 11. № 1. P. 231.
7. Han K., Lee Y., Kim W. PseudoViewer: automatic visualization of RNA pseudoknots. *Bioinformatics*. 2002. T. 18. № 1. P. S321–S328.
8. Haslinger C., Stadler P.F. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bulletin of Mathematical Biology*. 1999. T. 61. № 3. P. 437–467.
9. Gan H.H., Pasquali S., Schlick T. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Research*. 2003. T. 31. № 11. P. 2926–2943.
10. Rødland E.A. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *Journal of Computational Biology*. 2006. T. 13. № 6. P. 1197–1213.
11. Bon M., Vernizzi G., Orland H., Zee A. Topological classification of RNA structures. *Journal of Molecular Biology*. 2008. T. 379. № 4. P. 900–911.
12. Reidys C.M., Huang F.W., Andersen J.E., Penner R.C., Stadler P.F., Nebel M.E. Topology and prediction of RNA pseudoknots. *Bioinformatics*. 2011. T. 27. № 8. P. 1076–1085.
13. Leontis N.B., Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA*. 2001. T. 7. № 4. P. 499–512.
14. Chojnowski G., Waleń T., Bujnicki J.M. RNA Bricks–a database of RNA 3D motifs and their interactions. *Nucleic Acids Research*. 2014. T. 42. № D1. P. D123–D131.
15. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank. *Nucleic Acids Research*. 2000. T. 28. № 1. P. 235–242.
16. Lu X.J., Olson W.K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*. 2008. T. 3. № 7. P. 1213–1227.
17. Hanson R.M., Prilusky J., Renjian Z., Nakane T., Sussman J.L. JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Israel Journal of Chemistry*. 2013. T. 53. № 3–4. P. 207–216.

Материал поступил в редакцию 28.11.2014, опубликован 06.02.2015.