Научная статья
УДК 519.6

В.И. Мартьянов
Байкальский государственный университет,
Иркутский национальный исследовательский технический университет,
Иркутский государственный университет,
г. Иркутск, Российская Федерация

Теоретико-множественные модели данных в задаче расчета вторичных структур РНК

Аннотация. Задачи расчета вторичных структур нуклеотидных последовательностей РНК являются важными частными случаями комбинаторных задач высокой сложности (NP трудные задачи), которые можно решать логико-эвристическими методами. Кроме того, известна возможность применения программного обеспечения решения задач расчета вторичных структур нуклеотидных последовательностей РНК для филогенетического анализа (восстановление всех мутаций от предка к потомку).

В настоящей работе предлагаются более общие средства представления данных и более мощные алгоритмы расчета вторичных структур нуклеотидных последовательностей РНК, чем в предыдущих работах автора. Кроме того, отметим, что программное обеспечение имеет большие возможности адаптации для решения задач филогенетического анализа.

Вычислительные эксперименты, результаты которых приводятся в данной работе, проводились на исходных данных, взятых с сайта www.ncbi.nlm.nih.gov. Кроме того, данные также брались из банков нуклеотидных последовательностей GenBank (США) и EMBL (Европа).

Адаптация разработанных методов расчета вторичных структур проводилась вначале на поиске полного спектра совершенных повторов в генетических текстах (геномы Bradyrhizobium japonicum (BA0000040_GR), Streptomyces avermitilis (BA0000030_GR)). В результате чего был создана программа, настраиваемая по многим параметрам, включая работу в алфавите, расширенным консенсусными символами (подробнее во введении).

Ключевые слова. Вторичные структуры нуклеотидных последовательностей РНК, филогенетический анализ, NP трудные задачи, удовлетворение ограничениям.

Информация о статье. Дата поступления: 14 декабря 2022 г.

Original article

V.I. Martyanov
Baikal State University,
Irkutsk National Research Technical University,
Irkutsk State University,
Irkutsk, Russian Federation

Set-Theoretical Data Models in the Problem of Calculating Secondary Structures of RNA

Abstract. The tasks of counting the secondary structures of nucleotide sequences are a special case of combinatorial problems of high complexity (NP-problems) that can be solved by logic-heuristic methods. In addition, it is possible

DOI: 10.17150/2713-1734.2022.4(4).343-357

to use special software programs to solve the calculation of secondary structures of RNA nucleotide sequences for phylogenetic analysis (recovery of all mutations from an ancestor to the next one).

In this paper, we propose more general means of data presentation and more powerful algorithms for calculating the secondary structures of RNA nucleotide sequences than in previous works of the author. In addition, we note that the software has great flexibility to solve problems of phylogenetic analysis. Computational experiments, the results of which are presented in this work, were carried out on the primary data taken from the website www.ncbi.nlm.nih.gov. In addition, data were also taken from the GenBank (USA) and EMBL (Europe) nucleotide sequence banks.

The adaptation of the developed methods for calculating secondary structures was first carried out by searching for the full range of perfect repeats in genetic texts (the genomes of Bradyrhizobium japonicum (BA0000040_GR), Streptomyces avermitilis (BA0000030_GR)). As a result, a program was created that is customizable in many ways, including working in the alphabet, with extended consensus symbols (for more details, see the introduction).

Keywords. Secondary structures of RNA nucleotide sequences, phylogenetic analysis, NP difficult tasks, satisfying constraints.

Article info. Received 14 December, 2022.

Введение

В настоящей работе предлагаются модификации логико-эвристического подхода для представления данных и алгоритмов решения NP-трудных задач, что проверено для расчета вторичных структур нуклеотидных последовательностей РНК [1; 2]. Кроме того, проверена возможность адаптации для решения задач филогенетического анализа [3].

В генетическом тексте микроРНК (кандидаты на создание вторичных структур нуклеотидных последовательностей РНК) являются несовершенными палиндромными повторами протяженности 19–26 нуклеотидов, причем промежуток между повторными областями (раздел) должен быть от 8 до 120 нуклеотидов. Структура раздела и сами повторы должны удовлетворять ряду специальных условий [4, с. 11511; 5, с. 119]:

– наличие в микроРНК нуклеотидов G и C не меньше 35 % и не больше 65% соответственно;

– минимум числа связанных нуклеотидов у гена микроРНК и комплементарного участка — 16, максимум числа не связанных нуклеотидов — 5;

– микроРНК имеет не больше 6 последовательно идущих совпадающих оснований подряд;

– число подряд следующих неспаренных нуклеотидов не более 2;

– минимум свободной энергии петли (шпилечной структуры), частью которой является микроРНК, – менее чем 55 ккал/моль;

– минимум свободной энергии микроРНК — не более чем 29 ккал/моль;

– число не имеющих пары (выпяченных) или неспаренных асимметрично оснований – не более 2 на микроРНК. Таким образом, комплементарная область и микроРНК не должны отличаться числом нуклеотидов более, чем на 2;

– промежуточный район (связующий микроРНК и комплементарную область) содержит короткие (2,3 нк и более) палиндромные повторы, упорядоченные в зеркальном порядке.

Молекулярная эволюция геномов [6] связана с изменениями генотипа, которые называются мутациями. Различают следующие виды мутаций [7]:

– точечные мутации, которые сводятся к заменам одного нуклеотида в последовательности ДНК;

– транслокации — перенос гена в другое место генома;

– делеции — удаление участка гена;

– инверсии — поворот участка гена на 180 градусов;

– дупликации — удвоение участка гена.

Полный филогенетический анализ (восстановление всех мутаций от предка к потомку), как расчетная задача, как минимум *NP*-трудная проблема [8–10], а это требует вариативности подходов, эффективных при тех или иных условиях. Отметим также, что автоматизация филогенетического анализа [3, с. 27] может строится на алгоритмах расчета вторичных структур РНК [1, с. 78], как решение комбинаторных задач высокой сложности.

Логико-эвристические методы решения комбинаторных задач высокой сложности [2, с. 470] будут рассматриваться как преобразования исходной (инициальной) многоосновной модели (*для данной вычислительной проблемы это БД отношений (реляционная), где домены являются основными множествами, а реляционные таблицы — это предикаты на основных множествах*) в конечную (финальную), где выполнены необходимые условия (ограничения).

Для данной расчетной задачи будет найдено конечное множество преобразований и тем самым можно будет искать перебором нужную структуру в сгенерированном пространстве многоосновных моделей, полученных из начальных, применением той или иной последовательности преобразований.

Предлагаемые методы исходной точкой оценки сложности имеют проверку принадлежности кортежа $(a_1, a_2, \ldots, a_n)$ отношениям $H_1, H_2, \ldots, H_k$, которые определены на конечных множествах $A_1, A_2, \ldots, A_n$. Для такой постановки соответствующий результат работы [11] может быть сформулирован следующим образом.

Теорема А. *Верхняя граница сложности проверки принадлежности кортежа* $(a_1, a_2, \ldots, a_n)$ *отношениям* $H_1, H_2, \ldots, H_k$ *не превышает* $O(n+k)$.

DOI: 10.17150/2713-1734.2022.4(4).343-357

Формализация описания кандидатов на создание вторичных структур нуклеотидных последовательностей РНК

Составляющими **элементами** кандидатов на создание вторичных структур нуклеотидных последовательностей РНК будем считать **фрагменты стеблей** и **петли** [4, с. 11511; 5, с. 119]. Каждый элемент является словом в алфавите нуклеотидов *{A, T, C, G}* и консенсусных символов (Таблица 1).

Таблица 1

Консенсусные символы

Y = *C или T*	D = *A или G или T*
K = *G или T*	H = *A или C или T*
S = *G или C*	V = *A или G или C*
M = *A или C*	B = *G или C или T*
R = *A или G*	N = *A или G или C или T*
W = *A или T*	

Таким образом, одним из основных множеств, рассматриваемых многоосновных моделей, будет алфавит

$$\boldsymbol{\Theta} = \{A, T, C, G, Y, K, S, M, R, W, D, H, V, B, N\}.$$

Кандидаты на создание вторичных структур состоят из последовательностей элементов (слов в алфавите $\boldsymbol{\Theta}$). **Стебли**, как части кандидатов, являются последовательностями элементов, которые образуют цепь. Каждая **петля** состоит из только одного элемента.

Для элемента основной количественной характеристикой является общее количество нуклеотидов. Соответственно, основной количественной характеристикой стебля является сумма нуклеотидов всех элементов стебля.

Замечание 1. *Предполагаем, что элементы имеют начало и конец, причем связанные элементы стебля соединяются концом первого элемента с началом следующего.*

Математической формализацией для кандидатов на создание вторичных структур нуклеотидных последовательностей РНК будут многоосновные модели вида

$$\boldsymbol{MM = < \Theta, Arcc, Rrel, V; Ssect, Aang, R >}, \quad (1)$$

где совокупность $\boldsymbol{\Theta}$ — алфавит из нуклеотидов *{A, T, C, G}* и консенсусных символов; **Arcc** — система элементов; **Rrel** — система связей элементов; **V** — система натуральных чисел (представляет длину элементов и энергию в некоторой шкале); отображение ***Ssect:*** **Arcc** → **V**, т.е. указывает числовые параметры данных; отображение ***Aang:*** **Rrel** → **V**, т.е. определяет численность элементов; трехместный предикат ***R*** соединяет связь элементов **Rrel** с соответствующими объектами, т.е. ***R*** — отношение на множествах **Rrel, Arcc, Arcc**.

Замечание 2. *Разработанное программное обеспечение работает только с ограниченными совокупностями, что достигается рассмотрением ограниченных систем* ***Arcc, Rrel***, а *также установкой о выполнении шкалы изменения численных параметров элементов и связей элементов, т.е. совокупность* ***V*** *ограничена ненулевым числом, согласованным с параметрами убывания значения, и максимальным, определенным максимумом на размер кандидатов на создание вторичных структур нуклеотидных последовательностей РНК.*

Постановка задачи расчета вторичных структур РНК

Общий подход к решению NP-трудных задач логико-эвристическими методами [2] интерпретируется как перестройка первично определенной (исходной) многосоставной математической модели [12; 13] (1):

$$MM_{ini} = < \Theta, Arcc, Rrel, \mathbf{V}; Ssect, Aang, R >, \qquad (2)$$

где совокупность Θ — алфавит из нуклеотидов {A, T, C, G} и консенсусных символов; **Arcc** — система элементов; **Rrel** — система связей элементов; **V** — система натуральных чисел (представляет длину элементов и энергию в некоторой шкале); отображение ***Ssect:*** **Arcc → V**, т.е. указывает числовые параметры данных; отображение ***Aang:*** **Rrel → V**, т.е. определяет численность элементов; трехместный предикат ***R*** соединяет связь элементов **Rrel** с соответствующими объектами, т.е. ***R*** — отношение на множествах **Rrel, Arcc, Arcc**, MM_{fin}, где выполнены рестрикции $ST_1, \ldots, ST_w$.

Такую перестройку логично принимать допустимым (не оптимальным) управлением из динамического программирования [14], где $ST_1, \ldots, ST_w$ — фазовые рестрикции.

При решении (расчете вторичных структур) модель (*2*) преобразуется в базу данных, поиск перестройки (последовательность шагов) для получения необходимой модели, удовлетворяющей рестрикциям, становится NP-трудной задачей [15].

Для изучаемых (и решаемых) вопросов (поиск кандидатов на создание вторичных структур нуклеотидных последовательностей РНК) полная методика (схема) решения NP-трудных задач здесь не востребована (и не может быть применена полностью на данном этапе работы с точки зрения логико-эвристических методов) из-за невозможности рабочих определений конкретных задач.

Остановимся на рассмотрении вычислительной емкости действий по реализации рестрикций $ST_1, \ldots, ST_w$ на многоосновных алгебраических системах типа (1) с проверкой свободы быстроты выполнения рестрикций от их количества.

DOI: 10.17150/2713-1734.2022.4(4).343-357

Установление правильности рестрикций требует выполнимости вложения кандидатов на создание вторичных структур нуклеотидных последовательностей РНК (образцов) в исследуемую последовательность нуклеотидов и консенсусных символов (1).

Определим правила представления анализируемой последовательности нуклеотидов и консенсусных символов для рассматриваемой схемы установки удовлетворения рестрикций.

Образующими частями вторичных структур и анализируемой последовательности нуклеотидов и консенсусных символов являются **элементы** и **связи элементов**, цифровым заданием которых являются число малых степеней увеличения у **элементов** (**связей элементов**) — числовых характеристик, имеющие **низшие и высшие** уровни.

Следовательно, отображение ***Ssect:*** **Arcc → V** становится отображением ***Ssect:*** **Arcc → V x V**, аналогично, отображение ***Aang:*** **Rrel → V** становится отображением ***Aang:*** **Rrel → V x V**.

Замечание 3. *Наименьшие и наибольшие значения отображений для рассматриваемой предметной области мало существенны в плане вычислительной сложности, но крайне важны для рестрикций, где 2 или 3 лишних нуклеотида иногда определяют существование кандидата на создание вторичных структур нуклеотидных последовательностей РНК.*

Пусть многоосновные модели:

$$\begin{aligned} R_1 &= <\Theta, \text{Arcc}_1, \text{Rrel}_1, \text{V}; \text{Ssect}, \text{Aang}, R>; \\ R_2 &= <\Theta, \text{Arcc}_2, \text{Rrel}_2, \text{V}; \text{Ssect}, \text{Aang}, R>; \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ R_m &= <\Theta, \text{Arcc}_m, \text{Rrel}_m, \text{V}; \text{Ssect}, \text{Aang}, R> \end{aligned} \qquad (3)$$

задают искомые образцы в анализируемой последовательности нуклеотидов и консенсусных символов (1).

Анализ последовательности нуклеотидов и консенсусных символов выполняет определение ***BCEX*** однозначных отображений $\{\xi_{i,j}\}$ моделей R_1, R_2, …, R_m в модель M (1) [7, 8], т.е. эквивалентные вложения $\xi_{i,j}$: $R_i \rightarrow M$ состоят из однозначных вложений

$$\xi_{i,j}: Arcc_i \rightarrow Arcc; \quad \xi_{i,j}: Rrel_i \rightarrow Rrel, \qquad (4)$$

таких, что:

а) если $\xi_{i,j}(Ar) = Arr$, где $Ar \in Arcc_i$, $Arr \in Arcc$, $Ssect(Ar) = (v_1, v_2)$,

$$Ssect(Arr) = (v_3, v_4), \text{ то } v_1 <= v_3 <= v_4 <= v_2;$$

б) если $\xi_{i,j}(Re) = Rrel$, где $Re \in Rrel_i$, $Rrel \in Rrel$, $Aang(Re) = (v_1, v_2)$,

$$Aang\ (Rrel) = (v_3, v_4),\quad \text{то}\quad v_1 <= v_3 <= v_4 <= v_2\ ;$$

в) если R (Re, Ar_1, Ar_2), где Re $\in Rrel_i$, $Ar_1 \in Ar_i$, $Ar_2 \in Arcc_i$, то

$$R\ (\xi_{i,j}\ (\text{Re}), \xi_{i,j}\ (Ar_1), \xi_{i,j}\ (Ar_2)).$$

Оценка сложности анализа последовательности нуклеотидов и консенсусных символов

Для облегчения понимания идеи доказательства основного результата рассмотрим доказательство ***теоремы А.*** Основой его является представление декартова произведения конечных множеств A_1 x A_2 x … x A_n в древовидной форме ***Tree*** [11].

Определим более точно конечные множества $A_1 = \{a_{1,1}, a_{1,2}, ..., a_{1,m1}\}$,

$A_2 = \{a_{2,1}, a_{2,2}, ..., a_{2,m2}\}$, …, $A_n = \{a_{n,1}, a_{n,2}, ..., a_{n,mn}\}$ (Схема 1).

Схема 1

Этажи дерева *Tree*

№ этажа	Значения вершин дерева												
0	Корень дерева												
1	$a_{1,1}$				$a_{1,2}$				…	$a_{1,m1}$			
2	$a_{2,1}$	$a_{2,2}$	…	$a_{2,m2}$	$a_{2,1}$	$a_{2,2}$	…	$a_{2,m2}$	…	$a_{2,1}$	$a_{2,2}$	…	$a_{2,m2}$
…	…								…	…			
n	$a_{n,1}$	$a_{n,2}$	…	$a_{n,mn}$	$a_{n,1}$	$a_{n,2}$	…	$a_{n,mn}$	…	$a_{n,1}$	$a_{n,2}$		$a_{n,mn}$

Понятно, что Схема 1 является ***универсумом*** для любых таблиц реляционной БД с доменами $A_1, A_2, …, A_n$. Т.е. представление отношения H в таблице состоит в пометке вершин **n**-го этажа, если путь от корня дерева до этой вершины **n**-го этажа дает кортеж из отношения H.

Проверка принадлежности кортежа $(a_1, a_2, …, a_n)$, где $a_1 \in A_1$, $a_2 \in A_2$, …, $a_n \in A_n$, т.е. отношению H производится за **n** шагов в Схеме 1 (этот процесс в дальнейшем будем называть ***интерпретацией*** кортежа $(a_1, a_2, …, a_n)$ на дереве ***Tree***). Действительно, a_1 позиционируется на 1-ом этаже за 1 шаг, a_2 позиционируется на 2-ом этаже за 1 шаг и так далее, a_n позиционируется на **n**-ом этаже за 1 шаг, где и определяется принадлежность кортежа $(a_1, a_2, …, a_n)$ отношению H.

Таким образом, проверка осуществляется за **n** шагов. Для доказательства ***теоремы А*** достаточно пометить вершины **n**-го этажа на принадлежность отношениям $H_1, H_2, …, H_k$. Тогда в результате ***интерпретации*** кортежа $(a_1, a_2, …, a_n)$ за **n** шагов на дереве ***Tree*** получим вершину **n**-го этажа, пометки которой покажут принадлежность (или непринадлежность) отношениям $H_1, H_2, …, H_k$.

DOI: 10.17150/2713-1734.2022.4(4).343-357

Оценка ***O(n+k)***, а не ***O(n)***, получается из-за необходимости пройти по списку отметок **n**-го этажа, что и дает добавку ***O(k)***.

Замечание 4. *Результат* ***теоремы A*** *типичный, так называемый обмен памяти на эффективность* [5]. *Конечно, задание декартового произведения деревом увеличивает необходимую зону памяти, но скорость выполнения операций предельно ускоряется. Следует отметить также, что на практике универсум* Схема 1 *не строится, а строится только его часть, состоящая из кортежей отношений* H_1, H_2, …, H_k. *Вообще говоря, это замедляет скорость интерпретации, но незначительно не более, чем на* ***ln(m)****, где* $m = max\{m1, m2, \ldots, mn\}$.

Прежде, чем перейти к изложению основного результата, определим ***универсум*** для анализируемой последовательности нуклеотидов и консенсусных символов, имеющих не более **n** элементов, и **k** вариантов элементов и связей элементов, т.е. множество **V** имеет **k** элементов, а минимальный сектор элемента или угол пересечения элементов равен (360 / **k**) градусов.

Пусть $\mathbf{Arcc}_1$, $\mathbf{Arcc}_2$, …., $\mathbf{Arcc}_n$ — множества элемент *всех характеристик*, т.е.

$$\mathbf{Arcc}_1 = \{ ar_{1,1}, ar_{1,2}, \ldots, ar_{1,k}\};$$
$$Arcc_2 = \{ ar_{2,1}, ar_{2,2}, \ldots, ar_{2,k}\};$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..$$
$$Arcc_n = \{ ar_{n,1}, ar_{n,2}, \ldots, ar_{n,k}\},$$

где $Ssect(ar_{i,j}) = (j, j)$.

Далее, пусть $Rrel_1$, $Rrel_2$, ..., $Rrel_{n-1}$ — множества связей элементов *всех характеристик*, т.е.

$$Rrel_1 = \{ re_{1,1}, re_{1,2}, \ldots, re_{1,k}\};$$
$$Rrel_2 = \{ re_{2,1}, re_{2,2}, \ldots, re_{2,k}\};$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\ldots..$$
$$Rrel_{n-1} = \{ re_{n-1,1}, re_{n-1,2}, \ldots, re_{n-1,k}\},$$

где $Aang(re_{i,j}) = (j, j)$.

Дерево ***TreeImage*** (универсум Схема 2 для всех анализируемых последовательностей нуклеотидов и консенсусных символов, имеющих не более **n** элементов, и **k** вариантов элементов и связей этих элементов) строится по аналогии дерева ***Tree*** для декартова произведения

$$\mathbf{Arcc}_1 \times Rrel_1 \times \mathbf{Arcc}_2 \times Rrel_2 \times \ldots \times Rrel_{n-1} \times \mathbf{Arcc}_n$$

Соглашения по представлению элементов Схемы 2 следующие:

– для элемента ${}_{\theta}\alpha_{\xi,\chi}$ — число θ является *позицией* на этаже схемы (№ клетки в строке); $\alpha_{\xi,\chi}$ является χ-ым элементом из множества $Arcc_{\xi}$ или $Rrel_{\xi}$;

Схема 2

Этажи дерева *TreeImage*

№ этажа	Значения вершин дерева												
0	Корень дерева												
1	$_1ar_{1,1}$				$_2ar_{1,2}$				…	$_{m1+k-1}ar_{1,k}$			
2	$_1re_{1,1}$	$_2re_{1,2}$	…	$_kre_{1,k}$	$_{k+1}re_{1,1}$	$_{k+2}re_{1,2}$	…	$_{2k}re_{1,k}$	…	$_{m2}re_{1,1}$	$_{m2+1}re_{1,2}$	…	$_{m2+k-1}re_{1,k}$
…	…								…	…			
2n-1	$_1ar_{n,1}$	$_2ar_{n,2}$	…	$_kar_{n,k}$	$_{k+1}ar_{n,1}$	$_{k+2}ar_{n,2}$	…	$_{2k}ar_{n,k}$	…	$_tar_{n,1}$	$_{t+1}ar_{n,2}$	…	$_{t+k1}ar_{n,k}$

– числа mi, где i — номер этажа, равны $k^i - k$; число $t = k^{(2n-1)} - k$ (отметим, что данные числа имеют чисто технический характер и уменьшают громоздкость выражений, стоящих в конце строк Схемы 2).

Интуитивно понятно, что Схема 2 содержит все анализируемые последовательности нуклеотидов и консенсусных символов, имеющих не более **n** элементов, и **k** вариантов элементов и связей элементов. Структуру дерева на Схеме 2 будем задавать отношениями ***Par(x, y)*** и ***Brot(x, y)***.

Отношение ***Par(x, y)*** задает отношение «*родитель-потомок*», например, $Par(_2ar_{i,j}, _{k+w}re_{1,2})$, где $1 <= w <= k$. Отношение ***Par(x, y)*** связывает элементы, расположенные на соседних этажах, и может быть определено строго математически, а именно, $Par(_\theta\alpha_{\xi,\chi}, _\gamma\beta_{\zeta,\eta})$ тогда и только тогда, когда

– $\xi = \zeta$ или $\xi + 1 = \zeta$;

– $(\theta-1)*k <= \zeta <= (\theta-1)*k + k - 1$.

Отношение ***Brot (x, y)*** задает отношение «*быть братом*», например,

$Brot(_{k+1}re_{1,1}, _{k+2}re_{1,1})$. Отношение ***Brot(x, y)*** связывает элементы, расположенные на одном этаже и связанные с одним элементом верхнего этажа отношением «*родитель-потомок*», и может быть определено строго математически, а именно, $Brot(_\theta\alpha_{\xi,\chi}, _\gamma\beta_{\zeta,\eta})$ тогда и только тогда, когда

– $\alpha = \beta$;

– $\xi = \zeta$;

– $\theta < \zeta$ и $\zeta - \theta < k$, а также $[\theta / k] > 0$, где операция $[]$ остаток от деления.

Интерпретация ξ произвольной связной анализируемой последовательности нуклеотидов и консенсусных символов (1) (в дальнейшем термин «последовательность «последовательности, имеющие не более **n** элементов, и **k** вариантов элементов и связей элементов», если, конечно, не оговорено противное), где

$$Arcc = \{ ar_1, ar_2, ..., ar_w \}, \; w <= n, \; Rrel = \{ re_1, re_2, ..., re_t \} \qquad (5)$$

DOI: 10.17150/2713-1734.2022.4(4).343-357

на дереве ***TreeImage*** производится по следующей схеме, заданной по индукции.

Основание индукции. Пусть $i = 1$ и $Ssect(ar_1) = (j, j)$ и $RRrel_1 = \{ re_i \mid R(re_t, Ar_1, Ar_2) \}$, где $Ar_1 = ar_1$ или $Ar_2 = ar_1\}$, $Arr_1 = \{ aar_i \mid R(re_t, Ar_1, Ar_2) \}$, где $re_i \in RRrel_1$, $Ar_1 = ar_1$ и $Ar_2 = aar_i$ или $Ar_2 = ar_1$ и $Ar_1 = aar_i\}$. Тогда полагаем

$\xi(ar_1) = {}_{j}ar_{1,j}$, $\xi(re_t) = {}_{(j-1)*k+v}re_{1,v}$, где $re_i \in RRrel_1$, $Aang(re_t) = Aang(re_{1,v}) = v$. Если $aar_i \in Arr_1$, то $\xi(aar_t) = {}_{(d-1)*k+e}ar_{2,e}$, где d — <u>позиция</u> элемента $\xi(re_t)$ (т.е. $d = (j-1)*k+v$), $Ssect(aar_t) = (e, e)$.

Отметим, что $Par(\xi(ar_1), \xi(re_t)), Par(\xi(re_t), \xi(aar_t))$, а также для любых aar_i и aar_j из Arr_1 выполняется $Brot(\xi(aar_i), \xi(aar_j))$.

Интерпретация ξ продолжается для множества элементов $Arcc_1 = Arcc \setminus (\{ar_1\} \cup Arr_1)$, множества связей элементов $Rrel_1 = Rrel \setminus RRrel_1$.

<u>Замечание 5.</u> *Важнейшим моментом построения отношений* $Par(x, y)$ и $Brot(x, y)$ *на Схеме 2 является их <u>конструктивизм</u> (эффективная вычислимость за один шаг), и это свойство сохраняется при построении интерпретации* ξ*, как для основания индукции (так и для индукционного шага, что будет показано ниже).*

Индукционный шаг. Пусть после i-го шага получены непустые множества элементов $Arr_i = \{ar_{i1}, ar_{i2}, \ldots, ar_{iw}\}$, $Arcc_i = Arcc_{i-1} \setminus Arr_i$, множества связей элементов $RRrel_i$, $Rrel_i = Rrel_{i-1} \setminus RRrel_i$, причем по аналогии с основанием индукции элементы из множества $\xi(Arr_i)$ располагаются на $2 * i + 1$ этаже Схемы 2, связи элементов из множества $\xi(RRrel_i)$ располагаются на $2 * i$ этаже Схемы 2.

По алгоритму основания индукции будем проводить построение для каждого элемента $ar_\alpha \in Arr_i$ таким образом, что ar_α не принадлежит объединению $\{ar_1\} \cup Arr_1 \cup \ldots \cup Arr_{i-1}$.

Пусть $Ssect(ar_\alpha) = (j, j)$ и $RRrel_{i+1\alpha} = \{ re_s \mid R(re_s, Ar_1, Ar_2) \}$, где $re_s \in Rrel_i$, а также $Ar_1 = ar_\alpha$ или $Ar_2 = ar_\alpha\}$, $Arr_{i+1\alpha} = \{ aar_u \mid R(re_s, Ar_1, Ar_2) \}$, где $Ar_1 = ar_\alpha$ и $Ar_2 = aar_u$ или $Ar_2 = ar_\alpha$ и $Ar_1 = aar_u\}$. Пусть $\xi(ar_\alpha) = {}_{\beta}ar_{i+1,j}$. Тогда $\xi(re_s) = {}_{(\beta-1)*k+v}re_{i+1,v}$, где $re_s \in RRrel_{i+1\alpha}$, $Aang(re_s) = Aang({}_{(\beta-1)*k+v}re_{i+1,v}) = v$.

Если $aar_u \in Arr_{i+1\alpha}$, то $\xi(aar_u) = {}_{(d-1)*k+e}ar_{2,e}$, где d — <u>позиция</u> элемента $\xi(re_s)$ (т.е. $d = (\beta-1)*k+v$), $Ssect(aar_u) = (e, e)$.

Отметим, что $Par(\xi(ar_\alpha), \xi(re_s)), Par(\xi(re_s), \xi(aar_u))$, а также для любых aar_i и aar_j из $Arr_{i+1\alpha}$ выполняется $Brot(\xi(aar_i), \xi(aar_j))$.

Полагаем множество элементов Arr_{i+1} равным объединению всех $Arr_{i+1\alpha}$, где ar_α произвольный элемент из множества Arr_i такая, что ar_α не принадлежит объединению $\{ar_1\} \cup Arr_1 \cup \ldots \cup Arr_{i-1}$, также множество связей элементов $RRrel_{i+1}$ полагаем равным объединению всех $RRrel_{i+1\alpha}$, соответствующие произвольным элементам ar_α из множества Arr_i (смотри, выше).

Интерпретация ξ продолжается для множества элементов $Arcc_{i+1} = Arcc_i \setminus Arr_i$, множества связей элементов $Rrel_{i+1} = Rrel_i \setminus RRrel_{i+1}$.

Так как по условию интерпретируемая последовательность нуклеотидов и консенсусных символов (5) является связной, то процесс интерпретации ξ будет закончен не более, чем за **w** шагов индукции, где **w** — количество элементов.

Отметим, что соответствие при построении интерпретации ξ для любого элемента или связи элементов анализируемой последовательности нуклеотидов и консенсусных символов (5) *производится за один шаг*, так как «связывание», соответствующего элемента Схемы 2 производится вычислением одной арифметической формулы. Таким образом, доказана

Лемма 1. *Верхняя граница сложности построения интерпретации ξ для связной последовательности нуклеотидов и консенсусных символов (1) не превышает* $O(w + t)$, *где* $w(t)$ *— количество элементов (соответственно, связей элементов).*

Теорема 1. *Пусть каждая из моделей* R_1, R_2, …, R_m *(3) имеет не более* **n** *элементов и представляет связное анализируемой последовательности нуклеотидов и консенсусных символов. Тогда анализ связной последовательности нуклеотидов и консенсусных символов (1) имеет верхнюю границу сложности не превышающую* $O(((w + t)*w) + m)$, *где* $w(t)$ *— количество элементов (соответственно, связей элементов) анализируемой последовательности нуклеотидов и консенсусных символов (1), причем множества элементов и связей элементов представлены выражениями (5).*

Доказательство. Построим интерпретации всех моделей R_1, R_2, …, R_m на универсуме Схема 2 для всех анализируемой последовательности нуклеотидов и консенсусных символов, имеющих не более **n** элементов, и **k** вариантов элементов и связей элементов (сложность этой процедуры, конечно, не входит в оценку доказываемой теоремы).

Далее, пометим все вершины Схемы 2 номерами моделей, чьи элементы соответствуют этим вершинам. Каждой многоосновной модели R_i сопоставим пару чисел (a_i, b_i), где a_i — количество помеченных вершин Схемы 2, соответствующих элементам, $\mathbf{b_i}$ — количество помеченных вершин Схемы 2, соответствующих связям элементов (конечно, помеченных номером i).

Построим совокупность интерпретаций $\xi_1, \xi_2, \ldots, \xi_w$ на Схеме 2 (с помеченными вершинами), которые отличаются выбором первого элемента для *основания индукции*. А именно, интерпретация ξ_1 начинается традиционно с элемента ar_1, интерпретация ξ_2 начинается с элемента ar_2 и так далее. Последняя интерпретация ξ_w начинается, соответственно, с элемента ar_w.

Введем для каждой интерпретации ξ_i множество пар

DOI: 10.17150/2713-1734.2022.4(4).343-357

$$(a_{i1}, b_{i1}), (a_{i2}, b_{i2}), \ldots, (a_{im}, b_{im}), \quad (6)$$

где a_{ij} (b_{ij}) — количество помеченных j вершин Схемы 2, соответствующих элементам (соответственно, связям элементов), полученных для интерпретации ξ_i. Если пара (a_{ij}, b_{ij}) равна паре (a_j, b_j), то, таким образом, найдено изоморфное вложение j-го последовательности нуклеотидов и консенсусных символов (образца) в анализируемую последовательность нуклеотидов и консенсусных символов (1).

В силу ***леммы 1***, построение каждого отображения ξ_i требует не более $w + t$ шагов и, таким образом, верхняя граница сложности поиска всех изоморфных вложений не более $O(((w + t)*w) + m)$, где «добавка» $O(m)$ возникает из-за необходимости сравнивать пары (6) с парами (a_j, b_j). Теорема доказана.

Замечание 6. *Следует отметить также, что на практике универсум* Схема 2 *не строится, а строится только его часть, состоящая из элементов и связей элементов моделей* $R_1, R_2, \ldots, R_m$ *(3). Вообще говоря, это замедляет скорость интерпретации, но незначительно, не более, чем на* $\ln(k)$.

Заключение

Алгоритм подбора кандидатов на создание вторичных структур нуклеотидных последовательностей РНК (при выполнении специальных условий 1–6) отрабатывался на геноме *Avermitilis*. В табл. 2 приведено определенное количество кандидатов.

Таблица 2

№	Имя хромосомы	Количество кандидатов
1	**Complite-chromosome1-chain1**	**23 фрагмента**
2	**Complite-chromosome2-chain1**	**51 фрагмент**
3	**Complite-chromosome3-chain1**	**43 фрагмента**
4	**Complite-chromosome4-chain1**	**45 фрагментов**

Алгоритм (программа) работал около 4 ч на обычном компьютере без использования параллельных вычислений.

Программа была проверена для поиска кандидатов на создание вторичных структур нуклеотидных последовательностей РНК на генетическом тексте *Homo sapiens*. При этом к специальным ограничениям генома *Avermitilis* устанавливались два добавочных требования:

– число не имеющих пары (выпяченных) или неспаренных асимметрично оснований – не более 2 на микроРНК. Таким образом, комплементарная область и микроРНК не должны отличаться числом нуклеотидов более чем на 2;

– промежуточный район (связующий микроРНК и комплементарную область) содержит короткие (2,3 нк и более) палиндромные повторы, упорядоченные в зеркальном порядке.

Было найдено количество кандидатов на создание вторичных структур нуклеотидных последовательностей РНК (табл. 3). Алгоритм (программа) работал около 9 ч на обычном компьютере без использования параллельных вычислений

В настоящей работе предлагаются более общие средства представления данных и более мощные алгоритмы расчета вторичных структур нуклеотидных последовательностей РНК, чем в работе [1]. Кроме того, программное обеспечение имеет бóльшие возможности адаптации для решения задач филогенетического анализа (восстановление всех мутаций от предка к потомку) по сравнению с методами, рассмотренными в статье [3].

Таблица3

№	Имя хромосомы	Количество кандидатов
1	**Complite-chromosome1-chain1**	**492 последовательности**
2	**Complite-chromosome2-chain1**	**416 последовательностей**
3	**Complite-chromosome3-chain1**	**262 последовательности**
4	**Complite-chromosome4-chain1**	**185 последовательностей**
5	**Complite-chromosome5-chain1**	**237 последовательностей**
6	**Complite-chromosome6-chain1**	**223 последовательности**
7	**Complite-chromosome7-chain1**	**315 последовательностей**
8	**Complite-chromosome8-chain1**	**156 последовательностей**
9	**Complite-chromosome9-chain1**	**217 последовательностей**
10	**Complite-chromosome10-chain1**	**212 последовательности**
11	**Complite-chromosome11-chain1**	**197 последовательностей**
12	**Complite-chromosome12-chain1**	**285 последовательностей**
13	**Complite-chromosome13-chain1**	**92 последовательности**
14	**Complite-chromosome14-chain1**	**193 последовательности**
15	**Complite-chromosome15-chain1**	**173 последовательности**
16	**Complite-chromosome16-chain1**	**233 последовательности**
17	**Complite-chromosome17-chain1**	**293 последовательности**
18	**Complite-chromosome18-chain1**	**77 последовательностей**
19	**Complite-chromosome19-chain1**	**362 последовательности**
20	**Complite-chromosome20-chain1**	**143 последовательности**
21	**Complite-chromosome X-chain1**	**219 последовательностей**
22	**Complite-chromosome Y-chain1**	**24 последовательности**

Список использованной литературы

1. Архипов В.В. Логико-эвристические методы поиска вторичных структур РНК / В.В. Архипов, Ю.М. Константинов, В.И. Мартьянов. – EDN LCZSRX // Современные технологии. Системный анализ. Моделирование. – 2010. – № 1. – С. 162–167.

2. Мартьянов В.И. Логико-эвристические методы сетевого планирования и распознавание ситуаций / В.И. Мартьянов // Проблемы управления и моделирования в сложных системах : труды III Междунар. конф., Самара, 4 сент. 2001. – Самара, 2001. – С. 469–473.

3. Мартьянов В.И. Применение логико-эвристических методов для некоторых задач филогенетического анализа / В.И. Мартьянов, М.Л. Скуматов. – EDN OEFVVT // Современные технологии. Системный анализ. Моделирование. – 2011. – № 2. – С. 27 31.

4. Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes / E. Bonnet, J. Wuyts, P. Rouze, Y. Peer // Proceedings of the National Academy of Sciences. – 2004. – Vol. 101, no. 31. – P. 11511–11516.

5. Lindow M. Computational evidence for hundreds of non-conserved plant microRNAs / M. Lindow, A. Krogh // BMC Genomics. – 2005. – Vol. 6. – P. 119.

6. Fujii T. Predator–Prey Molecular Ecosystems / T. Fujii, Y. Rondelez. – DOI 10.1021/nn3043572 / ACS Nano. – 2013. – Vol. 7, iss. 1. – P. 27–34.

7. Афонников Д.А. Молекулярная эволюция белков / Д.А. Афонников // Информационная биология. – URL: www.bionet.nsc.ru.

8. Maier D. The complexity of some problems on subsequences and supersequences / D. Maier // Journal of the Assocmtton for Computing Machinery. – 1977. – Vol. 25, no. 2. – P. 322–336.

9. Wagner R.A. On the complexity of the extended string-to-string correction problem / R.A. Wagner. – DOI 10.1145/800116.803771 // STOC '75: Proceedings of the seventh annual ACM symposium on Theory of computing. – 1975. – P. 218–223.

10. Lipsky W.Jr. Two NP-complete problems Rrelated information retrieval / W.Jr. Lipsky // Fundamentals of Computation Theory : International Conference on Fundamentals of Computation Theory. – Berlin, 1977, – P. 123–154.

11. Кнут Д. Искусство программирования для ЭВМ. Сортировка и поиск / Д. Кнут. – Москва : Мир, 1978. – Т. 3. – 848 с.

12. Мальцев А.И. Алгебраические системы / А.И. Мальцев. – Москва : Наука, 1967. – 324 с.

13. Пинус А.Г. Вопросы разрешимости расширенных теорий / А.Г. Пинус. – EDN UUGABH // Успехи математических наук. – 1978. – Т. 33, № 2. – С. 49–84.

14. Беллман Р. Динамическое программирование / Р. Беллман. – Москва : Изд-во иностр. лит., 1960. – 400 с.

15. Обзор приложений логико-эвристических методов решения комбинаторных задач высокой сложности / В.И. Мартьянов, В.В. Архипов, М.Д. Каташевцев, Д.В. Пахомов. – EDN NRBKXP // Современные технологии. Системный анализ. Моделирование. – 2010. – № 4. – С. 61–67.

References

1. Arkhipov V.V., Konstantinov Yu.M., Mart'yanov V.I. Logical-Heuristic Methods of Searching for Secondary RNA Structures. *Sovremennye tekhnologii. Sistemnyi analiz. Modelirovanie = Modern Technologies. System Analysis. Modeling*, 2010, no. 1, pp. 162–167. (In Russian). EDN: LCZSRX.

2. Mart'yanov V.I. Logical-Heuristic Methods of Network Planning and Recognition of Situations. *Complex systems: control and modeling problems*. Proceedings of the 3rd International Conference*, Samara, September 4, 2001.* Samara, 2001, pp. 469–473. (In Russian).

3. Mart'yanov V.I., Skumatov M.L. Adaptation of the Logical-Heuristic Approaches for Some Problems of the Phylogenetic Analysis. Sovremennye tekhnologii. Sistemnyi analiz. Modelirovanie = Modern Technologies. System Analysis. Modeling, 2011, no. 2, pp. 27–31. (In Russian). EDN: OEFVVT.

4. Bonnet E., Wuyts J., Rouze P., Peer Y. Detection of 91 Potential Conserved Plant Micrornas in Arabidopsis Thaliana and Oryza Sativa Identifies Important Target Genes. *Proceedings of the National Academy of Sciences*, 2004, vol. 101, no. 31. pp. 11511–11516.

5. Lindow M., Krogh A. Computational Evidence for Hundreds of Non-Conserved Plant Micrornas. *BMC Genomics,* 2005, vol. 6, pp. 119.

6. Fujii T., Rondelez Y. Predator–Prey Molecular Ecosystems. *ACS Nano*, 2013, vol. 7, iss. 1. pp, 27–34. DOI: 10.1021/nn3043572.

7. Afonnikov D.A. Molecular Evolution of Proteins. *Information Biology*. Available at: www.bionet.nsc.ru. (In Russian).

8. Maier D. The Complexity of Some Problems on Subsequences and Super Sequences. *Journal of the Assocmtton for Computing Machinery,* 1977, vol. 25, no. 2, pp. 322–336.

9. Wagner R.A. On the complexity of the extended string-to-string correction problem. *STOC '75: Proceedings of the Seventh Annual ACM Symposium on Theory of Computing*, 1975, pp. 218–223. DOI: 10.1145/800116.803771.

10. Lipsky W.Jr. Two NP-complete Problems Rrelated Information Retrieval. *Fundamentals of Computation Theory. International Conference on Fundamentals of Computation Theory,* Berlin, 1977, pp. 123–154.

11. Knut D. The Art of Computer Programming. Vol. 3: Sorting and Searching. 1973. 782 p. (Russ. ed.: Knut D. *The Art of Computer Programming*. Moscow, Mir Publ., vol.3, 1978. 848 p.).

12. Maltsev A.I. *Algebraic systems*. Moscow, Nauka Publ., 1967. 324 p.

13. Pinus A.G. Questions of Solvability of Extended Theories. *Uspekhi matematicheskikh nauk = Successes of Mathematical Sciences*, 1978, vol. 33, no. 2, pp. 49–84. (In Russian). EDN: UUGABH.

14. Bellman R. *Dynamic Programming*. New Jersey, 1957. 342 p. (Russ. ed.: Bellman R. *Dynamic Programming*. Moscow, Inostrannaya literatura Publ., 1960. 400 p.).

15. Mart'yanov V.I., Arkhipov V.V., Katashevtsev M.D., Pakhomov D.V. Logic-Heuristic Methods for Solving Combinatorial Problems of High Complexity Applications Preview. *Sovremennye tekhnologii. Sistemnyi analiz. Modelirovanie = Modern Technologies. System Analysis. Modeling*, 2010, no. 4, pp. 205–211. (In Russian). EDN: NRBKXP.

Информация об авторе

Мартьянов Владимир Иванович — доктор физико-математических наук, старший научный сотрудник, Байкальский государственный университет; профессор, Иркутский национальный исследовательский технический университет; профессор, Иркутский государственный университет, г. Иркутск, Российская Федерация, e-mail: martvliv@mail.ru.

Information about the Author

Vladimir I. Martyanov — Doctor of Physical and Mathematical Sciences, Senior Researcher, Baikal State University; Professor, Irkutsk National Research Technical University; Professor, Irkutsk State University, Irkutsk, Russian Federation, e-mail: martvliv@mail.ru.

Для цитирования

Мартьянов В.И. Теоретико-множественные модели данных в задаче расчета вторичных структур РНК. — DOI 10.17150/2713-1734.2022.4(4).343-357 // System Analysis & Mathematical Modeling. — 2022. — Т. 4, № 4. — С. 343–357.

For Citation

Martyanov V.I. Set-Theoretical Data Models in the Problem of Calculating Secondary Structures of RNA. *System Analysis & Mathematical Modeling*, 2022, vol. 4, no. 4, pp. 343–357. (In Russian). DOI: 10.17150/2713-1734.2022.4(4).343-357.