ПАТТЕРН ОПРЕДЕЛЕННЫХ ДИНУКЛЕОТИДОВ микроРНК АРАБИДОПСИСА СВЯЗАН С УРОВНЕМ ИХ СОДЕРЖАНИЯ В РАСТЕНИИ

В.Г. Левицкий, И.В. Хомичева, Н.А. Омельянчук, М.П. Пономаренко, Н.А. Колчанов

Институт цитологии и генетики СО РАН, Новосибирск, Россия,
e-mail: khomicheva@bionet.nsc.ru

Проведен анализ связи между контекстными характеристиками последовательностей зрелых микроРНК (миРНК) *Arabidopsis thaliana* и экспериментальными данными по содержанию миРНК в основных тканях растения. Установлены две контекстные закономерности. Взаимное присутствие динуклеотидов UG в позициях 17–19 и CA в позициях 19–21 относительно 5′ конца миРНК характерно для миРНК с высоким уровнем содержания в тканях стебля, а их отсутствие – с низким. Для тканей стручков присутствие динуклеотида UG в позициях с 6-ой по 12-ую при отсутствии динуклеотида CC в позициях с 15-ой по 21-ую связано с высоким уровнем содержания миРНК в тканях, а обратное событие – с низким его уровнем.

Введение

МикроРНК – одноцепочечные некодирующие РНК длиной около 20–24 нуклеотидов, которые комплементарно или частично комплементарно связываются с мРНК эукариот и приводят к ингибированию трансляции белка (Olsen, Ambros, 1999; Chen, 2004) или разрушению мРНК (Llave *et al*., 2002; Yekta *et al*., 2004). Значимость миРНК в процессах развития проявляется в фенотипических аномалиях мутантов по генам миРНК или их биогенеза (Palatnik *et al*., 2003; Chen, 2004; Baulcombe, 2004). К настоящему времени в геноме *A. thaliana* обнаружено более ста генов миРНК (Griffiths-Jones, 2004).

В процессе биогенеза из первичного транскрипта гена миРНК растений ферментом группы Дайсер (РНКаза III типа) вырезается предшественник с характерной шпилечной структурой, из которого этим же ферментом вырезается фрагмент стеблевой части (миРНК-миРНК* дуплекс, Doi *et al*., 2003; Henderson *et al.*, 2006). Далее зрелая миРНК включается в состав специального мультибелкового комплекса RISC (RNA-induced silencing complexes) (Bartel, 2004; Tang, 2005), который распознает мРНК мишень.

Показано, что целый ряд свойств последовательностей миРНК, таких, как термодинамические свойства миРНК-миРНК* дуплекса (Khvorova *et al*., 2003), частоты определенных нуклеотидов, сила водородных связей (Dezulian, 2005) являются район-специфичными. Структура миРНК является блочной, т. е. существует паттерн отдельно расположенных контекстных сигналов.

Вопрос о контекстной специфичности отдельных районов миРНК до сих пор детально не изучен. Это вызвано как относительной слабостью контекстных сигналов, так и ограниченностью числа известных миРНК и наличием среди них существенной доли гомологов. Метод консенсусов обеспечивает выравнивание последовательностей миРНК близкородственных семейств (таких, как 159 и 319, 165 и 166 и т. д.), выравнивание миРНК разных семейств возможно только с привлечением информации о вторичной структуре (Griffiths-Jones *et al*., 2003). До настоящего времени не было выявлено каких-либо общих контекстных характеристик нуклеотидных последовательностей зрелых миРНК разных семейств. Существующие мето-

ды распознавания миРНК основаны на использовании априорной информации о биологической модели: структура миРНК предшественника, комплементарность миРНК к району связывания в мРНК мишени, феномен консервативности миРНК (Bengert, Dandekar, 2005).

На основании множества свидетельств о важности миРНК в регуляции трансляции была выдвинута гипотеза об общей роли «кода миРНК» и «кода транскрипционных факторов» в определении тканеспецифичной экспрессии генов (Hobert, 2004). По-видимому, применение специальных средств анализа ДНК сайтов связывания транскрипционных факторов к анализу миРНК способно выявить скрытые, неизвестные к настоящему времени, контекстные характеристики миРНК.

Для анализа миРНК мы использовали метод SiteGA, успешно примененный ранее для распознавания сайтов связывания транскрипционных факторов (Левицкий и др., 2006, в печати). Основным достоинством этого подхода является рассмотрение взаимных зависимостей частот встреч динуклеотидов в различных районах нуклеотидных последовательностей. В нашей работе проведен анализ последовательностей зрелых миРНК и экспериментальных данных по содержанию ряда миРНК в разных тканях *A. thaliana*. Выявлена зависимость между совместной встречаемостью пар локально позиционированных динуклеотидов (ЛПД) и содержанием миРНК в отдельных тканях.

МАТЕРИАЛЫ И МЕТОДЫ

Выборки, использованные в анализе

Последовательности зрелых миРНК *A. thaliana* были извлечены из базы данных microRNA Registry, http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/browse.pl (Griffiths-Jones, 2004). Нами использованы только миРНК длиной 21 нт, которые, согласно номенклатуре Rfam, образуют 37 семейств гомологичных последовательностей. Нами были объединены миРНК близкородственных семейств: 156 и 157, 165 и 166, 170 и 171, таким образом, рассматривалось всего 34 семейства, содержащих 57 последовательностей.

Из 57 миРНК, описанных выше, для 17 известны экспериментальные данные по содержанию миРНК в основных органах *A. thaliana* (Axtell, Bartel, 2005) (в*ыборка экспериментальных данных*). Нами проанализировано 18 экспериментов для 7 органов: соцветия (4 эксперимента), стебли (2), стручки (2), стеблевые листья (2), розеточные листья (2), проростки – (короткий день – 2, длинный день – 2), корни (2).

Для преодоления гетерогенности исходных данных, которая заключается в том, что последовательности внутри одного семейства высокогомологичны, а численности семейств варьируют, нами использована следующая итеративная процедура обучения метода SiteGA. Было проведено 100 итераций обучения (*выборки обучения*), при этом каждый раз в обучение метода входило по одной случайно выбираемой последовательности из каждого семейства миРНК. Таким образом, всякий раз в обучение входило только по одному представителю каждого семейства.

Метод SiteGA

Задачу поиска ЛПД в пределах миРНК решает генетический алгоритм (ГА), использующий популяцию особей, представляющих собой наборы локальнопозиционированных динуклеотидов (ЛПД). Каждый ЛПД особи характеризуется положением $[a; b]$ в пределах всей миРНК $[A; B]$, а также типом d_j динуклеотида ($j = 1, \ldots, 16$). Для границ a и b возможной локализации динуклеотида используются позиции его первого нуклеотида, при этом всегда выполняется условие $A \le a \le b \le B - 1$. Работа ГА начинается с того, что для каждой особи популяции случайным образом задаются типы и положения всех ЛПД. Затем ГА итеративно производит циклы мутаций и рекомбинаций. Мутация меняет положение или тип одного ЛПД особи. Рекомбинация осуществляет обмен ЛПД между двумя разными особями. При этом в ГА максимизируемым параметром приспособленности особи является расстояние R^2 Махаланобиса, рассчитываемое по следующей формуле:

$$R^2 = \sum_{k=1}^{N}\sum_{n=1}^{N}\{[f_n^{(2)} - f_n^{(1)}]\times \times S_{n,k}^{-1}\times[f_k^{(2)} - f_k^{(1)}]\}. \quad (1)$$

Здесь N – общее число ЛПД в текущем наборе, $f_n^{(1)}$ и $f_n^{(2)}$ – средние частоты n-го ЛПД по выборкам миРНК и случайных последовательностей (полученных перемешиванием природных) соответственно; S^{-1} – обратная матрица объединённой ковариационной матрицы, которая равна сумме ковариационных матриц для выборок природных и случайных последовательностей по частотам ЛПД. Результатом работы ГА является конкретный набор ЛПД.

Применение метода SiteGA

Согласно вышеописанному методу SiteGA, выявлялись значимые по критерию Стьюдента положительные и отрицательные корреляции между частотами ЛПД зрелых миРНК (*выборка экспериментальных данных*). Положительная корреляция между частотами двух ЛПД соответствует наиболее вероятному присутствию или отсутствию обоих динуклеотидов по сравнению со случайными последовательностями (нами использованы последовательности нулевой марковости). При отрицательной корреляции с большей вероятностью наблюдается присутствие только одного из динуклеотидов пары. Таким образом, положительная и отрицательная корреляции между двумя ЛПД выделяют в выборке две контрастные группы последовательностей, в которых чаще всего присутствуют/отсутствуют оба динуклеотида пары, либо только один из динуклеотидов. Этот признак разбиения на две группы с различными свойствами контекста был использован нами для анализа экспериментальных данных (*выборка экспериментальных данных*) по содержанию миРНК в различных органах. Для каждой итерации обучения по набору полученных ранее ЛПД нами подсчитаны значимые корреляции между ЛПД. В итоге пары динуклеотидов, соответствующие наиболее частым среди ста итераций обучения корреляциям, использованы в качестве свойств контекста для сравнения с экспериментальными данными.

Применение критерия Фишера

Для установления зависимостей между уровнем содержания миРНК в различных тканях *A. thaliana* и полученными разбиениями на две группы с различными свойствами контекста был применён точный критерий Фишера для четырехпольных таблиц. Пример составления четырехпольной таблицы по экспериментальным данным для зависимости «содержание миРНК в соцветиях (эксперимент 2) – корреляция динуклеотидов [6; 12] UG и [15; 20] CC» приведен в табл. 1.

Для каждой итерации проведен подсчет значимых по критерию Фишера зависимостей «уровень содержания миРНК – значимая по критерию Стьюдента корреляция частот ЛПД», рассчитанных для *выборки экспериментальных данных*.

Таблица 1

Пример составления четырехпольной таблицы по экспериментальным данным (содержание миРНК в стеблях, эксперимент (2)) и корреляции динуклеотидов* [6; 12] UG и [15; 20] CC. Значимость по критерию Фишера $p < 0{,}05$

Группы миРНК по состоянию пары локально позиционированных динуклеотидов	Число миРНК с содержанием	
	высоким	низким
присутствие только [6; 12] UG	3	2
присутствие только [15; 20] CC	0	9

* Приведенные границы соответствуют пределам локализации 1-го нуклеотида динуклеотида.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Метод SiteGA был построен в соответствии с моделью, описанной выше. На основании оценок точности распознавания миРНК нами найдено оптимальное число характеристик $N = 28$, которое и было использовано далее в ходе 100 итераций обучения метода SiteGA (*выборки обучения*). Сравнение с экспериментальными данными произведено согласно представленному выше критерию Фишера.

Две наиболее достоверные зависимости между корреляциями частот ЛПД и содержанием миРНК в различных тканях *A. thaliana* приведены в табл. 2.

Первая зависимость представлена наиболее часто встречающейся положительной корреляцией между динуклеотидами [17; 18] UG и [19; 20] CA, эксперимент: стручки (2). Присутствие этих динуклеотидов соответствует высокому уровню содержания миРНК, а их отсутствие – низкому. Вторая зависимость характеризуется отрицательной корреляцией динуклеотидов [6; 12] UG и [15; 20] CC, эксперимент: стебли (2). Присутствие динуклеотида [6; 12] UG при отсутствии динуклеотида [15; 20] CC связано с высоким уровнем содержания миРНК в тканях, а обратное событие – с низким. Для других тканей сравнимого результата получено не было.

Нами также исследован вопрос выявления общих контекстных характеристик последовательностей миРНК без их отношения к данным по содержанию миРНК. Частота появления значимых корреляций на *выборке обучения* (37 последовательностей, см. выше) представлена в табл. 3.

Таким образом, в пяти наиболее значимых корреляциях принимают участие ЛПД (табл. 3, выделено жирным), которые выявлены нами как связанные с уровнем содержания миРНК в тканях стеблей и стручков (табл. 2). При этом ЛПД [19; 20] CA обнаружен нами в составе двух корреляций, а каждый из ЛПД [15; 20] CC, [6; 12] UG, [17; 18] UG – в составе одной. Это еще раз подтверждает, что контекстные характеристики, выявленные нами в ходе сравнения с экспериментальными данными, имеют важное функциональное значение для миРНК.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (гранты № 03-04-48469-а, 05-04-49111-а, 03-07-90181-в), Междисциплинарным интеграционным проектом фундаментальных исследований СО РАН № 119 «Генные сети: теоретический анализ, компьютерное моделирование и экспериментальное конструирование», проектом «Компьютерное моделирование и экспериментальное конструирование генных сетей» Программы Президиума РАН по моле-

Таблица 2

Представление двух наиболее достоверных классов зависимостей между корреляциями локально позиционированных динуклеотидов и содержанием миРНК в различных органах

Класс*	Название органа, номер эксперимента	Локализации динуклеотидов и знак коэффициента корреляции	Частота обнаружения зависимости при 100 итерациях	Значимость, × 10E-2	
				зависимости по критерию Фишера	коэффициента корреляции по критерию Стьюдента
I	стебли (2)	**[17;18] UG [19;20] CC** +	96	2,7	2,2
II	стручки (2)	**[6;12] UG [15;20] CC** – [7;12] UG [15;20] CC – * [6;12] UG [16;20] CC – *	32 4 2	4,9 4,9 4,9	4,1 4,6 4,1

* Указаны также корреляции, имеющие незначительные отличия в локализации динуклеотидов. Корреляции приводятся в порядке частоты их обнаружения, начиная с самой частой (выделена жирным).

Таблица 3

Наиболее часто обнаруживаемые значимые корреляции,
полученные в процессе обучения метода (*выборка обучения* – 37 последовательностей)

Корреляция	Знак	Значимость, коэффициента корреляции по критерию Стьюдента, ×10-2	Частота обнаружения корреляции при 100 итерациях
[4; 8] CC, **[15; 20] CC**	+	1,7	51
[1; 1] UU, **[19; 20] CA**	–	2,0	50
[17; 18] UG, [17; 19] UC	–	0,13	37
[19; 20] CA, [19; 19] CU	–	3,9	36
[6; 12] UG, [13; 14] CA	–	1,2	34
[4; 5] UA, [1; 1] UC	+	3,9	29
[11; 12] UG, [13; 14] CA	–	2,9	28
[2; 7] AA, [11; 17] CG	+	2,4	28
[17; 19] UC, [3; 8] CG	+	2,9	27
[4; 6] GC, [19; 19] CU	+	2,0	27

кулярно-физико-химической биологии (10.4), CDRF (Y2-B-08-02).

Работа частично поддержана Госконтрактом с Федеральным агентством по науке и инновациям «Идентификация перспективных мишеней действия новых лекарственных препаратов на основе реконструкции генных сетей» приоритетного направления «Живые системы».

Литература

Левицкий В.Г., Игнатьева Е.В., Ананько Е.А. и др. Распознавание сайтов связывания транскрипционных факторов с помощью метода SiteGA // Биофизика. 2006. Т. 51. Вып. 4. (в печати).

Axtell M.J., Bartel D.P. Antiquity of microRNAs and their targets in land plants // Plant Cell. 2005. V. 17. № 6. P. 1658–1673.

Bartel D. MicroRNAs: Genomics, biogenesis, mechanism, and function // Cell. 2004. V. 116. P. 281–297.

Baulcombe D. RNA silencing in plants // Nature. 2004. V. 431. P. 356–363.

Bengert P., Dandekar T. Current efforts in the analysis of RNAi and RNAi target genes // Brief Bioinform. 2005. V. 6. № 1. P. 72–85.

Chen, X. A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development // Science. 2004. V. 303. P. 2022–2025.

Dezulian T., Palatnik J.F., Huson D., Weigel D. Conservation and divergence of microRNA families in plants // Genome Biology. 2005. V. 6. P. 13.

Doi N., Zenno S., Ueda R. *et al*. Short-interfering-RNA-mediated gene silencing in mammalian cells requires Dicer and eIF2C translation initiation factors // Curr. Biol. 2003. V. 13. P. 41–46.

Griffiths-Jones S. The microRNA registry // Nucl. Acids Res. 2004. V. 32. P. 109–111.

Griffiths-Jones S., Bateman A., Marshall M. *et al*. Rfam: an RNA family database // Nucl. Acids Res. 2003. V. 31. № 1. P. 439–441.

Henderson I.R., Zhang X., Lu C. *et al*. Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning // Nature Genet. 2006. V. 38. P. 721–725.

Hobert O. Common logic of transcription factor and microRNA action // Trends Biochem. Sci. 2004. V. 29. P. 462–468.

Khvorova A., Reynolds A., Jayasena S.D. Functional siRNAs and miRNAs exhibit strand bias // Cell. 2003. V. 115. № 2. P. 209–216.

Llave C., Kasschau K., Rector M., Carrington J. Endogenous and silencing-associated small RNAs in plants // Plant Cell. 2002. V. 14. P. 1605–1619.

Olsen P., Ambros V. The lin-4 regulatory RNA controls developmental timing in *C. elegans* by blocking LIN-14 protein synthesis after the initiation of translation // Dev. Biol. 1999. V. 216. P. 671–680.

Palatnik J.F., Allen E., Wu X. *et al*. Control of leaf morphogenesis by microRNAs // Nature. 2003. V. 425. P. 257–263.

Tang G. siRNA and miRNA: an insight into RISCs // Trends Biochem. Sci. 2005. V. 30. № 2. P. 106–14.

Yekta S., Shih I., Bartel D. MicroRNA-directed cleavage of HOXB8 mRNA // Science. 2004. V. 304. P. 594–596.

THE PATTERN OF SOME DINUCLEOTIDES OF ARABIDOPSIS miRNA AND THEIR CONTENT IN THE PLANT

V.G. Levitsky, I.V. Khomicheva, N.A. Omelianchuk, M.P. Ponomarenko, N.A. Kolchanov

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia,
e-mail: khomicheva@bionet.nsc.ru

Summary

MicroRNAs (miRNAs) are small RNA that interact with target mRNAs causing cognate mRNA degradation or translation repression, play an important regulatory role in animals and plants. Discovery of specific miRNA features in the light of experimental data on miRNA abundance allows to predict its tissue-specific expression pattern. We revealed that mutual occurrence of dinucleotides UG in positions from 17 to 19 and CA in positions from 19 to 21 (relative to 5′ end of *Arabidopsis thaliana* miRNA) corresponds to the high accumulation level of miRNAs in stems whereas the absence of both dinucleotides at the same locations corresponds to the low accumulation level. The presence of dinucleotide UG in positions from 6 to 12 together with absence of dinucleotide CC in positions from 15 to 21 corresponds to the high accumulation level of miRNAs in siliques whereas the opposite event to the low level of accumulation.