СОДЕРЖАНИЕ микроРНК В *ARABIDOPSIS THALIANA* КОРРЕЛИРУЕТ С ВСТРЕЧАЕМОСТЬЮ ТЕТРАНУКЛЕОТИДОВ WRHW И DRYD

М.П. Пономаренко, Н.А. Омельянчук, А.В. Катохин, Н.А. Колчанов

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

МикроРНК (миРНК) – это недавно открытые короткие РНК длиной 20–24 основания, которые транскрибируются РНК-полимеразой II, не кодируют белки и регулируют экспрессию генов. Анализ с помощью системы ACTIVITY экспериментальных данных о содержании 26 зрелых миРНК в 7 органах *Arabidopsis thaliana* показал, что содержание миРНК в этих органах коррелирует со встречаемостью тетрануклеотидов WRHW и DRYD в последовательностях миРНК. Построенная линейная регрессия неизвестных количественных величин содержания произвольной миРНК в органах *A. thaliana* по встречаемости тетрануклеотидов WRHW и DRYD в известной последовательности этой миРНК дала статистически достоверные предсказания на независимых контрольных данных. Таким образом, нами впервые обнаружены конкретные особенности нуклеотидного контекста последовательностей миРНК, которые могут влиять на биологические функции этих молекул – на их способность накапливаться в органах *A. thaliana*.

Введение

МикроРНК (миРНК) – это эндогенные РНК длиной 20–24 основания, которые комплементарно связываются с матричной РНК (мРНК), что приводит к ингибированию трансляции или к разрушению этой мРНК. МикроРНК образуются путем созревания РНК-предшественников, транскрибированных РНК-полимеразой II с особых генов и свернутых в структуру «шпильки» (Lee, Ambros, 2001). Содержание различных миРНК в *Arabidopsis thaliana* варьирует от нескольких молекул до 50 тыс. молекул на клетку (Bartel, 2004). Высокое содержание миРНК в клетках может определяться как высоким уровнем транскрипции их генов, так и низкими темпами их распада. Факторы нуклеотидного контекста, определяющие стабильность миРНК, до сих пор не были обнаружены.

Для коротких интерференционных РНК (киРНК), очень похожих на миРНК по механизму подавления экспрессии генов, такие факторы нуклеотидного контекста были недавно найдены (Khvorova *et al.*, 2003). Очевидно, что поиск закономерностей влияния нуклеотидного контекста миРНК на их функционирование является необходимым для адекватного понимания регуляторного воздействия миРНК на экспрессию генов и для дизайна миРНК-подобных посттранскрипционных регуляторов для генов-мишеней в *A. thaliana*.

В этой работе мы исследовали с помощью компьютерной системы ACTIVITY (Ponomarenko *et al.*, 1997) экспериментальные данные о содержании зрелых миРНК в соцветии, стебле, стручке, проростке, корне, стеблевом и розеточном листьях *A. thaliana* из работы (Axtell, Bartel, 2005). В результате мы впервые обнаружили факторы нуклеотидного контекста миРНК, которые определяют их стабильность: высокое содержание миРНК в органах *A. thaliana* коррелирует с высокой встречаемостью тетрануклеотидов WRHW и DRYD в нуклеотидной последовательности этой миРНК. Построена линейная регрессия неизвестных количественных величин содержания произвольной миРНК в *A. thaliana* по количественным величинам встречаемости тетрануклеотидов

WRHW и DRYD в нуклеотидной последовательности этой миРНК. Показано, что эта регрессия дает достоверные предсказания на независимых контрольных данных ($\alpha < 0{,}05$).

Материалы и методы

Мы исследовали экспериментальные данные из работы (Axtell, Bartel, 2005), которые приведены в табл. 1. Всего было 26 зрелых миРНК. Их нуклеотидные последовательности имели длину от 20 до 24 оснований (здесь и далее: $E_k = e_{k,1} \ldots e_{ki} \ldots e_{k,L} = 20$, где $e_{ki} \in \{A, U, G, C\}$ – код нуклеотида согласно номенклатуре IUB-IUPAC, $1 \le i \le L = 20$ – номер позиции в последовательности миРНК, $1 \le k \le 26$ – порядковый номер миРНК). Содержание этих миРНК в органах *A. thaliana*, выраженное в логарифмических единицах, [миРНК]k, было от -0,543 до 7,774.

Эти величины [миРНК]k были получены нами путем усреднения (Axtell, Bartel, 2005) по четырем биологическим репликам для случаев соцветия и проростка (колонки I и VI) и по двум таким репликам в случае других органов (колонки II–V и VII). Кроме того, в колонке VIII приведены величины среднего содержания миРНК в *A. thaliana*, которые были усреднены нами по всем указанным органам и затем также исследованы с помощью ACTIVITY (Ponomarenko *et al.*, 1997).

Поскольку длины миРНК варьировали от 20 нт до 24 нт, то мы анализировали фрагменты длиной 20 нт от 5′-конца миРНК, как это показано прописными буквами в табл. 1. Поскольку до сих пор не были обнаружены факторы нуклеотидного контекста, определяющие стабильность миРНК, то с помощью ACTIVITY (Ponomarenko *et al.*, 1997) мы исследовали простейшее количественное свойство символьной нуклеотидной последовательности миРНК: встречаемость коротких слов-олигонуклеотидов $Z(m) = z_1 \ldots z_m$ фиксированной длины m от 1 нт до 4 нт, взвешенных с учетом локализации со стартом в позиции i этой последовательности:

$$X_{Z(m),F}(E) = \sum_{i=1,L-m+1} F(i) \prod_{j=1,m} \Delta(e_{i+j-1} \in z_j), \quad (1)$$

здесь: согласно номенклатуре IUB-IUPAC, $z_j \in \{A, U, G, C, W = \{A, U\}, R = \{A, G\}, M = \{A, C\}, K = \{U, G\}, Y = \{U, C\}, S = \{G, C\}, B = \{U, G, C\}, V = \{A, G, C\}, H = \{A, U, C\}, D = \{A, U, G\}, N = \{A, U, G, C\}\}$; Δ (истина) = 1, Δ (ложь) = 0; F(i) – весовая функция, моделирующая влияние Z(m) со стартом в позиции i последовательности E, на содержание миРНК в *A. thaliana* с помощью эвристического правила «чем выше F(i), тем больше Z(m) в позиции i влияет на содержание миРНК в *A. thaliana*».

Для заданной последовательности E и слова-олигонуклеотида Z(m) формула (1) суммирует веса F(i) всех позиций i, со стартом в которых этот Z(m) встретился в этой последовательности E. На рис. 1 показаны два примера весовых функций F(i). Непрерывная линия U-образной формы имеет на этом рисунке один пик в центральной части миРНК и с помощью этого пика моделирует наибольшее влияние слов-олигонуклеотидов Z(m) из центральной части миРНК на содержание этой миРНК в *A. thaliana*.

Аналогично пунктирная линия S-образной формы имеет один переход от своего минимума на 5′-конце миРНК к своему максимуму на 3′-конце миРНК и, таким образом, моделирует наибольшее влияние Z(m) в 3′-половине миРНК на содержание этой миРНК в *A. thaliana*. Всего мы проанализировали 360 весовых функций

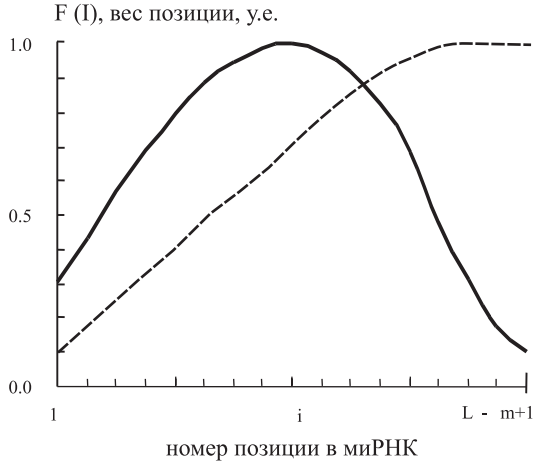

Рис. 1. Примеры весовых функций F(i), с помощью которых формула (1) моделирует положение важных олигонуклеотидов Z(m) длины m в позиции i центральной (непрерывная линия) и 3′-концевой (пунктир) частях миРНК длиной L. Всего анализируется по 180 функций F(i) каждого из этих двух типов.

Таблица 1

Содержание миРНК в органах *A. thaliana* (колонки I–VIII, логарифмические единицы) и результаты его анализа

миРНК	Последовательность миРНК	I	II	III	IV	V	VI	VII	VIII	WRHW	DRYD	Прогноз
mir158	UCCCAAAUGUAGACAAAGCA	3,590	3,889	**4,016**	**5,722**	**4,288**	5,851	**6,612**	**4,853**	1,790	2,013	5,780
mir159	UUUGGAUUGAAGGGAGCUCUa	**4,657**	5,237	4,179	5,788	5,076	6,003	5,528	**5,210**	1,300	1,399	4,329
mir160	UGCCUGGCUCCCUGUAUGCCa	3,469	**3,252**	2,085	**4,146**	**3,363**	4,678	**5,699**	**3,813**	0,685	2,434	4,881
mir161.1	UGAAAGUGACUACAUCGGGGt	4,373	4,331	3,655	4,756	**4,703**	**5,133**	**5,795**	4,678	**1,224**	**1,873**	4,852
mir161.2	UCAAUGCAUUGAAAGUGACUa	2,999	3,517	**2,427**	4,531	3,794	4,392	5,637	3,900	**1,779**	**1,634**	5,267
mir163	UUGAAGAGGACUUGGAACUUcgau	**0,722**	**1,687**	0,739	**4,642**	2,560	1,308	2,049	**1,958**	0,608	1,700	3,815
mir164	UGGAGAAGCAGGGCACGUGCa	**3,668**	**4,166**	**4,270**	**4,520**	4,187	**4,366**	**4,467**	4,235	**1,644**	**1,460**	4,861
mir165	UCGGACCAGGCUUCAUCCCCc	0,658	0,722	**0,622**	**1,126**	0,717	**0,998**	**1,441**	**0,898**	0,000	0,700	1,702
mir166	UCGGACCAGGCUUCAUUCCCc	1,494	**1,353**	**1,224**	**1,728**	**1,380**	1,522	**1,668**	**1,481**	0,000	0,700	1,702
mir167	UGAAGCUGCCAGCAUGAUCUa	4,929	2,132	**5,663**	5,639	**6,275**	**5,921**	4,134	4,956	2,226	1,172	**5,248**
mir168	UCGCUUGGUGCAGGUCGGGAa	**3,479**	**3,532**	**3,158**	4,349	**4,985**	**4,508**	**4,151**	**4,023**	1,000	1,249	3,738
mir170	UGAUUGAGCCGUGUCAAUAUc	**1,458**	**1,139**	1,153	**1,551**	1,652	**1,926**	-0,275	1,229	0,476	1,170	**2,945**
mir172	AGAAUCUUGAUGAUGCUGCAu	**6,257**	**7,769**	5,019	**7,774**	6,215	**3,966**	6,060	6,151	2,698	2,598	**7,742**
mir173	UUCGCUUGCAGAGAGAAAUCac	**0,786**	0,795	**-0,232**	0,945	**1,086**	**1,320**	0,830	0,790	1,610	0,549	**3,619**
mir390	AAGCUCAGGAGGGAUAGCGCc	**3,452**	**2,473**	**0,982**	2,888	3,557	2,718	**2,489**	**2,651**	0,434	2,128	4,149
mir394	UUGGCAUUCUGUCCACCUCC	**2,186**	3,014	1,505	2,174	1,204	2,534	1,400	2,003	0,000	0,247	**1,107**
mir396	UUCCACAGCUUUCUUGAACU	**3,184**	**4,805**	2,856	**5,048**	**5,753**	4,266	4,167	**4,297**	1,262	0,549	3,162
mir156	UGACAGAAGAGAGUGAGCAC	1,182		**2,343**	3,230	**1,959**	**6,517**	**3,277**	3,085	**0,974**	**2,081**	4,797
mir169	CAGCCAAGGAUGACUUGCCGa	**0,108**	1,548		**2,272**	2,367	**3,206**	3,159	2,110	0,000	1,611	**2,899**
mir171	UGAUUGAGCCGCGCCAAUAUc	3,215	**0,780**	2,379	1,551	**1,594**	**2,580**		**2,017**	0,476	1,170	2,945
mir398	UGUGUUCUCAGGUCACCCCUg	0.115	**0,640**	**1,710**	**3,201**	**3,522**		**1228**	1,736	**0,568**	**0,347**	1,985
mir156/157	UUGACAGAAGAUAGAGAGCAc			-0,094	1,353	1,569	3,767	**0,299**		1,890	1,021	4,608
mir162	UCGAUAAACCUCUGCAUCCAg	**0,116**			0,644	0,055				1,485	1,205	4,317
mir391	UUCGCAGGAGAGAUAGCGCCa				1,517	0,832				0,922	1,910	4,504
mir319	UUGGACUGAAGGGAGCUCCC	0,754	1,254							1,436	1,311	4,392
mir397b	UCAUUGAGUGCAUCGUUGAUg		-0,543		**0,940**	**0,216**		1,509		1,000	1,249	3,738
Коэффициент линейной корреляции, R		0,624	0,637	0,590	0,626	0,516	0,628	0,686	0,798	рис. 2а	рис. 2б	0,834
Уровень статистической значимости, α		0,0025	0,0025	0,005	0,01	0,01	0,0025	0,0005	0,00005			0,05

I – соцветие, II – стебель, III –стручок, IV – стеблевой лист, V – розеточный лист, VI – проросток, VII – корень, VIII – среднее.

F(i) этих двух типов: 180 U-образных F(i) с одним пиком (максимум или минимум) в пределах миРНК и 180 S-образных F(i) с одним переходом (возрастание или убывание). Эти весовые функции F(i) различались по ширине и по положению их единственных пиков/переходов внутри миРНК. Комбинирование всех этих 360 весовых функций со всеми возможными словами-олигонуклеотидами Z(m) длины m от 1 нт до 4 нт позволило нам изучить $360\times\{14 + 14\times14 + 14\times15\times14 + 14\times15\times15\times14\} = 17010000 \approx 2\times10^7$ различных количественных величин $X_{Z(m),\,F}$, каждая из которых может быть вычислена по формуле (1) для любой символьной нуклеотидной последовательности миРНК.

Общеизвестно, для того чтобы вывод «встречаемость $X_{Z(m),\,F}(E)$ коррелирует с содержанием [миРНК]» был статистически обоснованным, все пары величин $\{X_{Z(m),\,F}(E_k), [\text{миРНК}]_k\}$ должны отвечать требованию простой регрессии:

$$[\text{миРНК}]_{Z(m),\,F}(E_k) = a + b \times X_{Z(m),\,F}(E_k), \quad (2)$$

здесь: a, b – коэффициенты регрессии, вычисляемые стандартным способом по проверяемому набору пар вещественных чисел $\{X_{Z(m),\,F}(E_k), [\text{миРНК}]_k\}$.

Формула (2) предсказывает по последовательности E_k миРНК количественную величину $[\text{миРНК}]_{Z(m),\,F}(E_k)$ содержания этой миРНК в *A. thaliana* на основе учета встречаемости $X_{Z(m),\,F}(E_k)$ олигонуклеотидов Z(m) в этой миРНК. Условием применимости регрессии (2) является наличие достоверных соответствий между предсказанными и экспериментальными величинами, $[\text{миРНК}]_{Z(m),\,F}(E_k)$ и $[\text{миРНК}]_k$.

Для этой проверки ACTIVITY сначала формирует с помощью метода bootstrap (Hayes *et al.*, 1989) из всех анализируемых пар чисел $\{X_{Z(m),\,F}(E_k), [\text{миРНК}]_k\}$ семь поднаборов таких пар. Затем для каждого из этих семи поднаборов ACTIVITY проверяет 11 соответствий между предсказанием и экспериментом. В частности, среди этих 11 соответствий проверяются линейная, знаковая и ранговые корреляции. Таким образом, всего ACTIVITY (Ponomarenko *et al.*, 1997) проверяет $7\times11 = 77$ частных соответствий между предсказанными и экспериментальными величинами $[\text{миРНК}]_{Z(m),\,F}(E_k)$ и $[\text{миРНК}]_k$.

Проверка каждого n-го частного соответствия ($1 \le n \le 77$) заключается в оценке уровня его статистической значимости α_n, которая затем преобразуется, в терминах нечетких логик Задэ (Zadeh, 1965), в оценку обоснованности проверяемого предсказания:

$$q_n(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}]) = \begin{cases} 1, & \text{если } \alpha_n \le 0{,}01; \\ 1{,}3 - 28{,}3\alpha_n + 55{,}6\alpha_n^2, & \text{если } 0{,}1 \ge \alpha_n \ge 0{,}01; \\ -1, & \text{если } \alpha_n \ge 0{,}1. \end{cases} \quad (3)$$

Каждому достоверному соответствию ($\alpha_n < 0{,}05$) между предсказанными и экспериментальными величинами, $[\text{миРНК}]_{Z(m),\,F}(E_k)$ и $[\text{миРНК}]_k$ формула (3) приписывает положительную оценку его обоснованности $q_n((X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])$ от 0 до 1, недостоверному – отрицательную такую оценку от –1 до 0. Для каждого предсказания $\{[\text{миРНК}]_{Z(m),\,F}(E_k)\}$ система ACTIVITY (Ponomarenko *et al.*, 1997) получает всего 77 частных оценок его обоснованности $q_n(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])$, которые, согласно теории принятия решений (Fishburn, 1970), она затем усредняет в интегральную оценку обоснованности:

$$Q(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}]) = \{\Sigma_{n=1,77}\, q_n(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])\}/77. \quad (4)$$

Формула (4) дает каждому предсказанию тем большую оценку обоснованности $Q(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])$, чем больше количество достоверных соответствий между этим предсказанием и экспериментом $[\text{миРНК}]_{Z(m),\,F}(E_k)$ и $[\text{миРНК}]_k$ и чем выше их статистическая значимость α. Поэтому самая высокая положительная оценка обоснованности $Q(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])$ указывает именно тот олигонуклеотид Z(m) и именно ту весовую функцию F(i), с помощью которых по известным последовательностям

миРНК $\{E_k\}$ система ACTIVITY (Ponomarenko *et al.*, 1997) предсказывала (формулы (1) и (2)) такие величины $[\text{миРНК}]_{Z(m),\,F}(E_k)$ содержания миРНК в *A. thaliana*, для которых наблюдалось самое лучшее согласие с экспериментальными величинами $[\text{миРНК}]_k$.

РЕЗУЛЬТАТЫ

Используя формулы (1)–(4), система ACTIVITY (Po-nomarenko *et al.*, 1997) проанализировала так называемые «обучающие» поднаборы данных, объемы которых составляли 50 % всех данных (табл. 1, колонки I–VIII, жирный шрифт) и каждый из которых равномерно представлял все наблюдаемые экспериментальные величины содержания миРНК в органах *A. thaliana*. Остальные 50 % этих данных (табл. 1, нормальный шрифт) были контрольными.

Для каждой анализируемой миРНК с последовательностью E_k по формуле (1) вычислялись все 17010000 взвешенные оценки встречаемости $X_{Z(m),\,F}(E_k)$. Для каждой $X_{Z(m),\,F}(E_k)$ с помощью экспериментальных величин $\{[\text{миРНК}]_k\}$ (табл. 1, колонки I–VIII, жирный шрифт) были построены регрессии (2) для предсказания содержания миРНК $\{[\text{миРНК}]_{Z(m),\,F}(E_k)\}$ по известным последовательностям E_k соответствующих миРНК. Затем по формулам (3) и (4) были получены оценки обоснованности $Q(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])$ каждого из этих предсказаний. Всего для 8 «обучающих» поднаборов (табл. 1, колонки I–VIII, жирный шрифт) было получено $8 \times 17010000 = 136080000 \approx 10^8$ таких оценок $Q(X_{Z(m),\,F}(E) \rightarrow [\text{миРНК}])$.

В случае U-образных весовых функций наиболее обоснованной Q = 0,477 была регрессия (2) величин среднего содержания миРНК (табл. 1, колонка VIII) по величинам встречаемости $X_{WRHW,\,F1}(E)$ тетрануклеотидов WRHW, взвешенных функцией $F_1(i)$ с максимумом в центре миРНК (рис. 1, непрерывная линия). В колонке WRHW для всех исследуемых миРНК приведены величины $X_{WRHW,\,F1}(E_k)$, вычисленные по формуле (1). Соответствие между этими $X_{WRHW,\,F1}(E_k)$ и средним содержанием миРНК $\{[\text{миРНК}]_k\}_{VIII}$ показано на рис. 2а. Темными кружками на этом рисунке показаны 10 миРНК из «обучающего» поднабора (табл. 1, колонка VIII, жирный шрифт), пунктиром – регрессия (2). Аналогично светлыми кружками и непрерывной линией показаны 11 контрольных миРНК (табл. 1, колонка VIII, нормальный шрифт) и их регрессия (2). На этих независимых контрольных данных коэффициент линейной корреляции между $X_{WRHW,\,F1}(Ek)$ и средним содержанием миРНК был равен R = 0,733, что означает достоверность предсказаний ($\alpha < 0{,}025$).

В случае S-образных весовых функций самой обоснованной Q = 0,466 была регрессия (2) содержания миРНК в корнях (табл. 1, колонка VII) по встречаемости $X_{DRYD,\,F2}(E)$ тетрануклеотидов DRYD, взвешенных функцией $F_2(i)$ с линейным возрастанием весов от минимума на 5′-конце миРНК до максимума на 3′-конце миРНК (рис. 1, пунктир). В колонке DRYD приведены все величины $X_{DRYD,\,F2}(E_k)$, на рис. 2б показано соответствие между $X_{DRYD,\,F2}(E)$ и содержанием миРНК в корнях *A. thaliana* $\{[\text{миРНК}]_k\}_{VII}$. Коэффициент линейной корреляции R = 0,605 между $X_{DRYD,\,F2}(E_k)$ и содержанием миРНК в корнях *A. thaliana* $\{[\text{миРНК}]_k\}_{VIII}$ был статистически достоверным ($\alpha < 0{,}05$) на независимых контрольных данных.

Существенно, что линейная корреляция между содержанием миРНК в корнях *A. thaliana* (табл. 1, колонка VII) и средним содержанием миРНК (колонка VIII) была достоверной ($R = 0{,}910$, $\alpha < 10^{-6}$), тогда как встречаемости $X_{WRHW,\,F1}(E)$ и $X_{DRYD,\,F2}(E)$ выявленных тетрарибонуклеотидов WRHW и DRYD были независимыми ($R = 0{,}324$, $\alpha > 0{,}10$). Это означает, что тетрануклеотиды WRHW и DRYD независимо влияют на содержание миРНК в *A. thaliana*. Поэтому с помощью найденный $X_{WRHW,\,F1}(E)$ и $X_{DRYD,\,F2}(E)$ мы построили множественную линейную регрессию среднего содержания миРНК в *A. thaliana*:

$$[\text{миРНК}](E_k) = 0{,}782 + 1{,}314 \times X_{WRHW,\,F1}(E_k) + 0{,}756 \times X_{DRYD,\,F2}(E_k), \quad (5)$$

здесь: 0,782, 1,314 и 0,756 – коэффициенты множественной линейной регрессии, вычисляемые стандартным способом по набору из пар чисел $\{X_{WRHW,\,F1}(E_k), X_{DRYD,\,F2}(E_k)\}$, выделенных

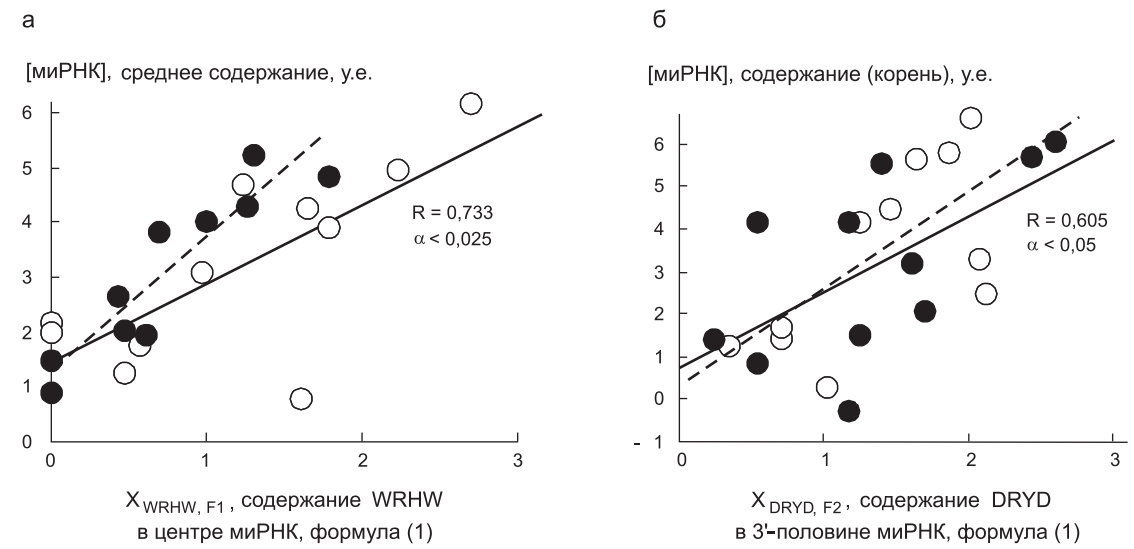

Рис. 2. Контекстные закономерности миРНК.

а – среднее содержание в миРНК в *A. thaliana* (вертикальная ось, эксперимент (Axtell, Bartel, 2005)) достоверно коррелирует со встречаемостью тетрануклеотидов WRHW, взвешенных функцией F_1 (рис. 1, непрерывная линия), в последовательности мРНК (горизонтальная ось, предсказание, настоящая работа); б – содержание миРНК в корнях *A. thaliana* (вертикальная ось, эксперимент (Axtell, Bartel, 2005)) достоверно коррелирует со встречаемостью тетрануклеотидов DRYD, взвешенных функцией F_2 (рис. 1, пунктир), в последовательности мРНК (горизонтальная ось, предсказание, настоящая работа).

Темные кружки и пунктир – обучающие данные; светлые кружки и непрерывная линия - контрольные данных; для контрольных данных приведены коэффициент линейной корреляции R и уровень его статистической значимости a.

жирным шрифтом в колонках WRHW и DRYD, и соответствующих экспериментальным величинам $\{[\text{миРНК}]_k\}_{\text{VIII}}$ (колонка VIII).

В колонке «Прогноз» для всех 26 исследуемых миРНК представлены величины [миРНК](E), предсказанные по формуле (5). В этой колонке жирным шрифтом выделены предсказания для шести миРНК, которые мы до сих пор не использовали ни для оптимизации формулы (5), ни для поиска тетрануклеотидов WRHW и DRYD (колонки WRHW и DRYD, VII и VIII: нормальный шрифт). В нижней строке колонки «Прогноз» показан коэффициент линейной корреляции R = 0,834 между этими шестью независимыми предсказаниями и соответствующими им экспериментальными величинами среднего содержания миРНК в *A. thaliana*. Эти контрольные предсказания имеют достоверное согласие с экспериментом ($\alpha < 0,05$).

Наконец, в двух нижних строках табл. 1 приведены коэффициенты линейной корреляции между предсказанными по формуле (5) и экспериментальными величинами содержания миРНК во всех семи органах *A. thaliana*. Можно видеть, что все эти корреляции являются достоверными ($\alpha < 0,01$). Это означает, что мы впервые обнаружили факторы нуклеотидного контекста миРНК, которые влияют на стабильность миРНК: высокое содержание миРНК в органах *A. thaliana* коррелирует с высокой встречаемостью WRHW и DRYD в последовательностях миРНК. Установлено также, что построенная на этой основе линейная регрессия неизвестных величин содержания миРНК по величинам встречаемости WRHW и DRYD в известных последовательностях зрелых миРНК дает достоверные предсказания на независимых контрольных данных.

Обсуждение

Представляется интересным, что с помощью формулы (5) встречаемость WRHW в последовательностях миРНК объясняет 56 ± 14 % дисперсии экспериментально наблюдаемого содержания

миРНК в органах *A. thaliana*, встречаемость DRYD объясняет 44 ± 14 % дисперсии этих экспериментальных данных. Как можно здесь видеть, $(56 - 44) \pm (14^2 + 14^2)^{1/2}\ \% = 12 \pm 19{,}8\ \%$ и $(56 + 44) \pm (14^2 + 14^2)^{1/2}\ \% = 100 \pm 19{,}8\ \%$, найденные нами WRHW и DRYD вносят, соответственно, равные и исчерпывающие вклады в дисперсию данных эксперимента (Axtell, Bartel, 2005).

Интересно также, что из всех возможных 44 = 256 тетрануклеотидов 212 – не являются WRHW или DRYD, 8 – являются только WRHW, 20 – только DRYD, 16 являются одновременно и WRHW, и DRYD. Согласно точному критерию Фишера, WRHW и DRYD достоверно коррелируют между собой **($\alpha < 10\text{-}9$)**. Действительно, WRHW и DRYD имеют консенсус WRYW. Дополнение G в крайние позиции этого консенсуса WRYW дает DRYD, выявленный в качестве важного для 3′-половины миРНК. «Консенсусный» RY этого DRYD указывает на важность цилиндрической формы A-спирали 3′-половины миРНК при формировании ее гетеродуплекса с РНК-мишенью. Аналогично дополнение A в третью позицию этого консенсуса WRYW дает WRHW, который мы нашли в качестве важного в центре миРНК. Два крайних «консенсусных» W этого WRHW указывают на функциональную важность именно слабых водородных связей (W = A + U) в центре гетеродуплекса миРНК с ее мишенью. Предсказанные выше два свойства гетеродуплекса миРНК с ее мишенью: (а) цилиндрическая формы 3′-конца A-спирали и (б) слабые водородные связи в центре согласуются с данными экспериментов (Khvorova *et al.*, 2003; Haley, Zamore, 2004), где эти свойства были показаны для так называемых малых интерференционных РНК, самых похожих на микроРНК по строению, функциям и механизмам регуляции.

В целом результаты настоящей работы позволяют заключить, что содержание миРНК в органах *A. thaliana* коррелирует со встречаемостью тетрануклеотидов WRHW и DRYD в последовательности миРНК. Эта впервые выявленная нами контекстная закономерность миРНК может быть полезна для понимания взаимосвязи между структурой и функцией миРНК и конструирования миРНК-подобных посттранскрипционных регуляторов для генов-мишеней *A. thaliana*.

Благодарности

Настоящая работа была поддержана грантом ЖС-12.3/002 № 02.434.11.3004.

Литература

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.

Fishburn P.C. Utility theory for decision making. N.Y.: John Wiley and Sons, 1970.

Haley B., Zamore P.D. Kinetic analysis of the RNAi enzyme complex // Nat. Struct. Mol. Biol. 2004. V. 11. № 7. P. 599–606.

Hayes K.G., Perl M.L., Efron B. Application of the bootstrap statistical method to the tau-decay-mode problem // Phys. Rev. D. Part. Fields. 1989. V. 39. P. 274–279.

Khvorova A., Reynolds A., Jayasena S.D. Functional siRNAs and miRNAs exhibit strand bias // Cell. 2003. V. 115. P. 209–216.

Lee R.C., Ambros V. An extensive class of small RNAs in *Caenorhabditis elegans* // Science. 2001. V. 294. P. 862–864.

Ponomarenko M.P., Kolchanova A.N., Kolchanov N.A. Generating programs for predicting the activity of functional sites // J. Comput. Biol. 1997. V. 4. P. 83–90.

Zadeh L.A. Fuzzi sets // Information and Control. 1965. V. 8. P. 338–353.

THE CONTENT OF miRNAs IN *ARABIDOPSIS THALIANA* CORRELATES WITH THE OCCURRENCE OF TETRAMERS WRHW AND DRYD

M.P. Ponomarenko, N.A. Omelyanchuk, A.V. Katokhin, N.A. Kolchanov

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

Summary

MicroRNAs (miRNAs) are short recently discovered non-protein-coding and RNAs, which regulate gene expression. Using the system ACTIVITY to study the microarray data on the content of mature miRNA in *Arabidopsis thaliana*, we found that a high content of miRNA correlates with a high occurrence of the tetranucleotides WRHW and DRYD in this miRNA sequence. It is shown that the linear regression of the unknown quantitative content of arbitrary miRNA on the basis of the known occurrences of the WRHW and DRYD within its sequence gives statistically significant predictions with independent control data. Thus we first report that the sequence of mature miRNAs may also influence their ability to accumulate in tissues.