БИОФИЗИКА, 2017, том 62, вып. 1, с. 47–52

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 575. 112:004

10% КОНСЕРВАТИВНЫХ САЙТОВ микроРНК В ПОЗВОНОЧНЫХ НЕПРАВИЛЬНО ВЫРАВНЕНЫ

© 2017 г. К.А. Просвиров*, А.А. Миронов* **, Р.А. Солдатов* **

**Факультет биоинженерии и биоинформатики Московского государственного университета имени М.В. Ломоносова, 119234, Москва, Ленинские горы, 1/73;*

***Институт проблем передачи информации, 127051, Москва, Большой Каретный пер., 19*

E-mail: prosvirov.k@gmail.com

Поступила в редакцию 16.06.16 г.
После доработки 10.10.16 г.

МикроРНК – малые эндогенные некодирующие РНК, ответственные за ингибирование экспрессии белков на уровне трансляции посредством комплементарного взаимодействия со специфичным сайтом в мРНК. Большинство функциональных сайтов связывания (второй–седьмой нуклеотиды микроРНК), которые расположены в 3′-нетранслируемой области, консервативны. Когда дело доходит до предсказания этих сайтов, используется сравнительная геномика и ее базовый инструмент – множественное выравнивание. Однако множественные выравнивания накапливают ошибки из-за сильного расхождения видов. Кроме того, в процессе эволюции сайты связывания могут перемещаться вдоль последовательности. В работе оценена доля консервативных сайтов связывания микроРНК, которые не могут быть предсказаны существующими методами. Введено понятие L-консервативности – сайт называется L-консервативным, если все виды в выравнивании имеют его внутри окна шириной L нуклеотидов. Получено значительное увеличение количества дополнительных сайтов без потери чувствительности. Сделана оценка прироста количества потенциальных сайтов связывания для видов с разной дивергенцией.

Ключевые слова: микроРНК, сравнительная геномика, множественные выравнивания.

МикроРНК – обширный класс малых некодирующих РНК длиной ~ 22 нуклеотида. Они пост-транскрипционно репрессируют мРНК через комплементарное связывание преимущественно с 3′-нетранслируемой областью (3′НТО) [1,2]. МикроРНК вовлечены во множество биологических процессов, включая клеточный рост, рак, метаболизм и дифференцировку [3,4]. В биогенезе микроРНК у животных задействовано множество белков [5,6]. Первичный транскрипт микроРНК разрезается в ядре РНКазой III типа Drosha, высвобождая шпильку, в которой 3′-конец, как правило, на два нуклеотида длиннее 5′-конца – такая структура называется пре-микроРНК [7]. Пре-микроРНК затем транспортируется в цитоплазму белка-транспортера exportin5 и ГТФазы Ran, где в дальнейшем разрезается второй РНКазой III типа Dicer, высвобождая ~ 22-нуклеотидный дуплекс. Одна цепь дуплекса загружается в белок Аргонавт, образуя с другими белками комплекс RISC. Более представленный продукт – микроРНК, а другая цепь – микроРНК* [8,9]. Далее микроРНК служит матрицей комплексу RISC для репрессии мРНК. Для узнавания мишени критична комплементарность первых восьми нуклеотидов микроРНК. Также для увеличения эффективности регуляции существуют такие особенности, как 3′-дополнительное связывание, локальная доступность сайта в структуре мРНК, позиция сайта внутри 3′НТО [10,11].

Определение мишеней микроРНК является важным фактором для понимания регуляции белковой экспрессии на уровне мРНК. Однако в связи с плохим пониманием главных свойств, отличающих правильные мишени микроРНК, это по-прежнему является сложной биоинформатической задачей [12]. Консервативность широко применяется для поиска функциональных элементов. Сравнительный геномный анализ опирается на множественные выравнивания последовательностей, которые могут содержать ошибки, например, выравнивание человек–мышь содержит 13% неправильно выравненных нуклеотидов [13–15]. Слабо дивергировавшие виды дают мало информации о консервативности, в то время как далекие геномы сталки-

Сокращение: 3′НТО – 3′-нетранслируемая область.

Консервативный октамер (L = 0)	L-консервативный октамер (L = 9)	L-консервативный октамер (L = 15)
AUGA**AGACAGAA**ACU	AUGA**AGACAGAA**ACU	AUGA**AGACAGAA**ACUAAG-CAC-AACU
ACGA**AGACAGAA**GCU	ACGA**AGACAGA-A**CU	ACGAAGACAG-AGCUA**AGACAGAA**ACU
AUGG**AGACAGAA**AUU	AUGG**AGACAGAA**AUU	AUGG**AGACAGAA**AUUAA--CAG-AACU
AUGA**AGACAGAA**ACU	AUGA**AGACAGA-A**CU	AUGAAGACAG-AACUA**AGACAGAA**ACU
AUGG**AGACAGAA**AUU	AUGG**AGACAGAA**AUU	AUGG**AGACAGAA**AUUTAGAC--AAACU
AUGA**AGACAGAA**ACU	AUGA**AGACAGAA**ACU	AUGA**AGACAGAA**ACUA**AGACAGAA**ACU

Рис. 1. Консервативные сайты (подчеркнуты) в множественном геномном выравнивании.

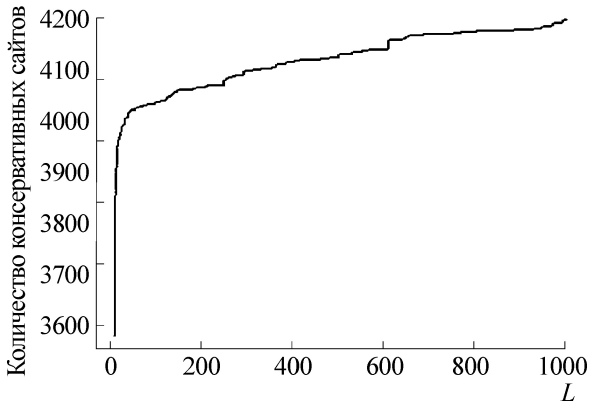

Рис. 2. Прирост количества консервативных сайтов в зависимости от размера рамки.

ваются с проблемой ошибок в выравнивании. Локальные ошибки выравнивания возникают по причине того, что алгоритм неспособен правильно расставить пропуски в олигонуклеотидной последовательности, состоящий из оснований одного вида [16,17]. Как следствие, такие ошибки ведут к неправильному расположению сайту и его потере при предсказании функциональных элементов. Мы вводим понятие L-консервативности. Сайт называется L-консервативным, если все последовательности в выравнивании имеют его внутри окна ширины L. Этот подход позволил нам идентифицировать большее количество консервативных сайтов микроРНК. Также мы наблюдали недопредставленность консервативных сайтов в быстро эволюционирующих 3′НТО, что объясняется плохим выравниванием, а также миграцией сайтов при эволюции [18].

МАТЕРИАЛЫ И МЕТОДЫ

Данные. Консервативные микроРНК и выравнивания 3′НТО были скачаны из базы данных TargetScan. Были получены подвыравнивания, содержащие следующие виды: *Homo sapiens*, *Mus musculus*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Otolemur garnetti*, *Macaca mulatta*, *Felix catus*, *Canis llupis*. Все микроРНК были кластеризованы в семейства на основе общего seed-региона, а в качестве первого нуклеотида для консенсуса был выбран наиболее часто встречающийся нуклеотид первой позиции в семействе [19–21]. Таким образом, получилось 83 семейства и, следовательно, 83 консенсусных последовательности.

Подход. Некорректные выравнивания ортологичных k-меров могут возникать из-за неправильной расстановки гэпов, также возможна миграция сайтов узнавания в процессе эволюции [20,21]. Мы определяем понятие L-консервативности следующим образом: сайт называется L-консервативным, если во множественном выравнивании в окне размера L для каждой последовательности содержится данный сайт, причем сайт может попадать на участки с делециями – мы учитываем только значащие символы в выравнивании. На рис. 1 представлены варианты консервативных и L-консервативных октамеров при разных значениях L.

РЕЗУЛЬТАТЫ

Дополнительные сайты найдены при увеличении рамки. Алгоритм был применен к набору данных 3′НТО и глубоко консервативных микроРНК для потенциального нахождения консервативных и L-консервативных сайтов. В качестве общей проверки всего пайплайна был подсчитано отношение консервативных сайтов и контрольных на рамке длиною в восемь нуклеотидов. Отношение оказалось равным 5,62, что полностью соответствует ранним данным [22]. Это подтверждает тот факт, что алгоритм не содержит ошибки. Был отмечен прирост дополнительно найденных сайтов при увеличении длины допустимой рамки, до выхода на плато происходило насыщение на 10% относительно количества консервативных сайтов для рамки L = 8 (рис. 2). Понятно, что сильный

локальный прирост до очень плавного прироста и является следствием ошибок выравнивания, в то же время слабый прирост при больших L отражает частоту появления случайных октамеров в 3′НТО. Для проведения контроля были сгенерированы случайные октамеры с использование бернуллиевой случайной величины. Так как контроль был построен случайно, без сохранения нуклеотидного/динуклеотидного состава, не проверялось совпадение/перекрывание с уже существующими сайтами в 3′НТО. В качестве оцениваемой величины было взято отношение найденных сайтов к контрольным сайтам, таким образом, видно, что на начальных длинах рамок найденных сайтов больше, чем случайно сгенерированных, что говорит о том, что локальные ошибки выравнивания возникают в небольшом окне (рис. 3а). Поэтому мы предполагаем, что найденные сайты не случайны, а действительно представляют функциональные элементы.

Включение невыравненных сайтов не приводит к потере в чувствительности. Для грубой оценки метода был применен параметр, именуемый «чувствительностью». «Чувствительность» – способность метода правильно определять консервативные сайты. Мы преобразовали его в «относительное изменение чувствительности» согласно уравнению (1), в котором «чувствительность», «сигнал» и «контроль» – есть функции от длины рамки L [23]:

$$RSC(L) = \\ = \frac{Signal(L) - Signal(L = 8) - Background(L)}{Signal(L)}, \quad (1)$$

где $RSC(L)$ – относительная чувствительность метода при длине рамки L, $Signal\ (L)$ – количество найденных консервативных сайтов при использовании окна длиною L, $Background\ (L)$ – количество случайных сайтов при длине рамки L. График относительного изменения чувствительности представлен на рис. 3б.

Дивергенция – одна из основных причин ошибок выравнивания сайтов. В данном исследовании мы определяем дивергенцию следующим образом [24]:

– В выравнивании выбираем лишь позиции, содержащие буквы в каждой строке.

– Считаем долю замен для каждой пары последовательностей.

– Берем среднее и обозначаем за дивергенцию:

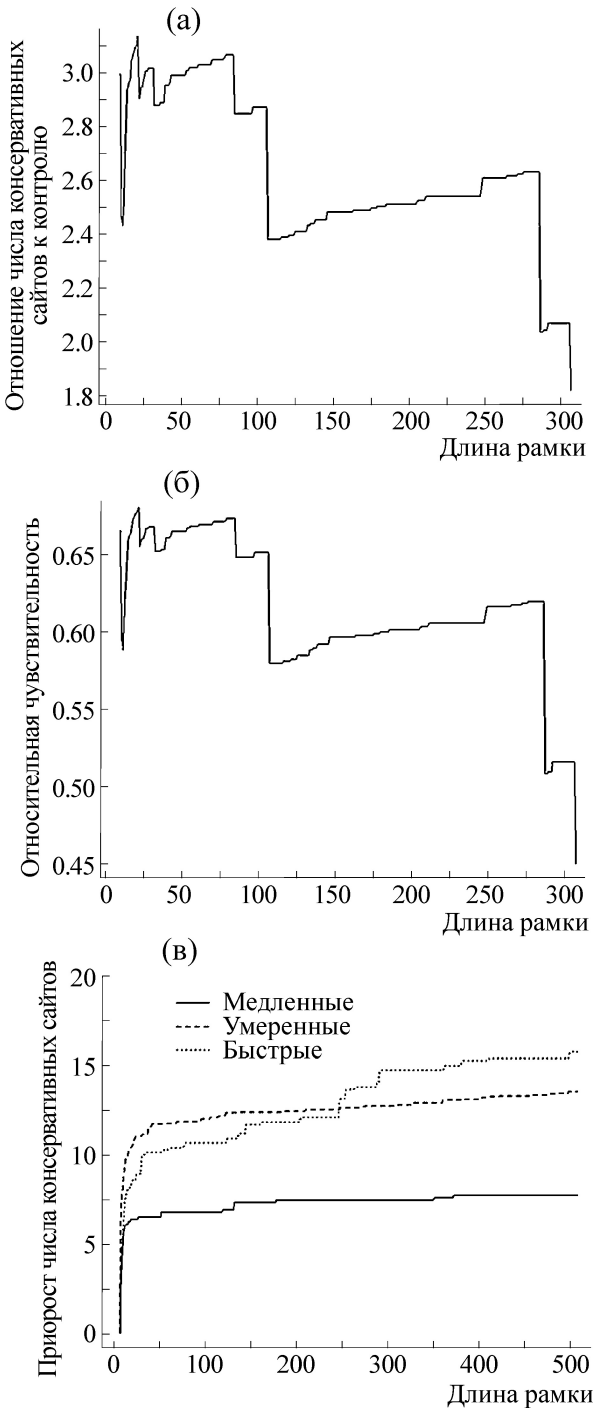

Рис. 3. (а) – Зависимость отношения консервативных сайтов к контролю от длины рамки, (б) – зависимость относительного изменения чувствительности от длины рамки, (в) – зависимость процента консервативных сайтов для трех групп с разной скоростью дивергенции.

$$Div = \frac{\sum_{i=0}^{n-1} \sum_{i \neq j}^{q} \delta(seq_i x_j, seq_{i+1} x_j)}{n},$$

Таблица 1. Примеры регулируемых генов, имеющих предсказанные сайты

МикроРНК	Ген	log2 изменения экспрессии	Рамка
miR-1	SS18	–0,05	9
miR-1	SRGAP2	0,20	9
miR-1	SS18	–0,76	9
miR-16	PPM1A	0,37	11
miR-16	SMAD7	–1,88	9
miR-16	PPM1A	–1,14	11
miR-16	SMAD7	–0,42	9
miR-16	TBL1XR1	0,63	9
miR-16	PPM1A	–1,22	11
miR-16	SMAD7	–0,12	9
miR-16	TBL1XR1	0,20	9
miR-30	FAM91A1	2,84	9
miR-30	LRRC17	1,30	9
miR-30	NUS1	1,06	9
miR-30	UBN1	0,27	9
miR-30	ZNRF1	–1,83	9
miR-128	NF1	0,60	10
miR-133	PRDM16	–0,19	9
miR-133	RB1CC1	–0,91	9
miR-148	MTMR12	0,83	9
miR-181a	RAD21	–2,31	9
miR-181a	RSBN1	–0,36	11
miR-181a	TCERG1	–1,07	9
miR-181a	RSBN1	0,29	11
miR-181a	TCERG1	–0,40	9
miR-181a	ABCB7	–1,29	47

где n – количество организмов, q – количество пар, w – длина выравнивания, $\delta(x, y) = \begin{cases} 1, \text{если } x = y \\ 0, \text{если } x \neq y. \end{cases}$

Мы разделили все последовательности по трем корзинам в зависимости от уровня дивергенции [25,26]. Для каждой корзины была дополнительно подсчитана средняя длина последовательностей, которая и объясняет дивергенцию в каждой группе. Чем длиннее последовательность, тем более вероятно, что там произойдут замены, которые лишь усилят расхождение последовательностей. Была обнаружена прямая зависимость между дивергенцией и количество дополнительно найденных сайтов (рис. 3в).

Существует регуляция генов с плохо выравненными сайтами. Для проверки наличия биологического эффекта были использованы данные по трансфекции/сверхэкспрессии. Для этого мы взяли данные экспериментов по 26 микроРНК и применили наш метод [27]. В табл. 1 приведены примеры генов с плохо выравненными сайтами микроРНК в 3′НТО, для которых существует биологический эффект. Изменение уровня экспрессии указано в виде отрицательного десятичного логарифма. Видно, что сверхэкспрессия этих микроРНК действительно снижает экспрессию данных генов. Оказалось, что реальные сайты связывания, которые считались не консервативными, являются на самом деле консервативными, но теряются при наивном определении консервативности. Статистическое подтверждение этого факта представлено в табл. 2, где с использованием критерия Вилкоксона были подсчитаны p-value для каждого эксперимента по трансфекции для пары мишени/контроль, а также суммарно по всем микроРНК. В табл. 1 в основном представлены гены, для которых рамка сайта невелика, от 9 до 11, что скорее связано с плохим выравни-

Таблица 2. Статистическая значимость для экспериментов по оверэкспрессии микроРНК

miR-1_-_24_Hour_-_hela_(Lim)	9,70562e–06
miR-124_-_24_Hour_-_hela_(Lim)	5,01659e–16
miR-7_-_24_Hour_-_hela_(Grimson)	1,145787e–05
miR-9_-_24_Hour_-_hela_(Grimson)	1,489342e–19
miR-122_-_24_Hour_-_hela_(Grimson)	0,0001350548
miR-128_-_24_Hour_-_hela_(Grimson)	1,215177e–20
miR-132_-_24_Hour_-_hela_(Grimson)	1,019474e–10
miR-133_-_24_Hour_-_hela_(Grimson)	5,013692e–10
miR-148_-_24_Hour_-_hela_(Grimson)	3,600954e–07
miR-1_-_24_Hour_-_hela_(Bartel)	4,601052e–08
miR-124_-_24_Hour_-_hela_(Bartel)	4,199761e–23
miR-30_-_32_Hour_-_hela_(Selbach)	4,021801e–11
miR-16_-_32_Hour_-_hela_(Selbach)	0,0190532
miR-1_-_32_Hour_-_hela_(Selbach)	8,292086e–16
Все мишени против всех контролей	1,591368e–125

ванием. В выборке также присутствует ген ABCB7, для которого событием является смещение сайта на 47 нуклеотидов. Возможным механизмом такого перемещения могут являться мутации, которые возникают внутри сайта, но одновременно где-то в 3′НТО появляются компенсаторные мутации, которые восстанавливают нужный для регуляции сайт, но уже в другой позиции вдоль 3′НТО.

ДИСКУССИЯ

Биологические последовательности, особенно полинуклеотидные, действительно зачастую плохо выравниваются. Но это не есть биологический эффект, это скорее недостаток современных алгоритмов. Нами был предложен метод, который основывается на расширении понятия консервативности, который позволяет учитывать, во-первых, возможные ошибки выравнивания, возникающие в малых окнах и полинуклеотидных последовательностях, и, во-вторых, возможные эволюционные события. Второе объясняется тем, что эволюционное расхождение геномов нередко приводит к тому, что на фоне сохранения регуляторной последовательности она может перемещаться по геному, исчезая в одних позициях и возникая в других.

Эффективность нашего подхода можно наблюдать по введенной нами величине – «чувствительности», которая вводилась не совсем стандартным образом, но, несмотря на это, она не падает на достаточно коротких рамках, которых вполне хватит для поиска консервативных сайтов регуляции методами сравнительной геномики. Статистический анализ и сравнение со случайными последовательностями показали, что мы можем находить новые сайты, которые, как показал анализ данных по сверхэкспрессии, действительно являются функциональными элементами, при этом величина смещения сайтов составляет около 10 нуклеотидов, что может быть обусловлено скорее ошибками выравнивания.

ВЫВОДЫ

В работе было введено ослабленное определение консервативности сайтов связывания – L-консервативность. Это позволило найти до 10% новых консервативных сайтов. При использовании окон, близких к $L = 9$, чувствительность метода повышается.

Регуляция многих генов с сайтами в рамках 9–11 нуклеотидов подтверждена данными дифференциальной экспрессии при трансфекции микроРНК.

Ошибки выравнивания появляются из-за различных дивергенций и неспособности алгоритмов выравнивания правильно расставить нуклеотиды в полинуклеотидных последовательностях.

Количество неправильно выравненных сайтов увеличивается при увеличении дивергенции генов.

Исследование выполнено при финансовой поддержке Российского научного фонда (грант № 14-14-00088).

СПИСОК ЛИТЕРАТУРЫ

1. N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel, Science **294** (5543), 858 (2001).
2. I. Alvarez-Garcia and E. A. Miska, Development **132** (21), 4653 (2005).
3. G. Mullokandov, B. D. Brown, et al., Nature Methods **9**, 840 (2012).
4. A. M. Ardekani and M. M. Naeini, Avicenna J. Med. Biotechnol. **2** (4), 161 (2010).
5. Y. Lee, V. N. Kim, et al., Nature **425** (6956), 415 (2003).
6. E. Ladewiq, E. C. Lai, et al., Genome Res. **22** (9), 1634 (2012).
7. J. E. Park and V. N. Kim, Nature **475**, 201 (2011).
8. K. Okamura, E. C. Lai, et al., Nat. Struct. Mol. Biol. **15** (4), 354 (2008).
9. L. Guo and Z. Lu, PLoS ONE **5** (6), e11387 (2010).
10. R. C. Friedman, D. P. Bartel, et al., Genome Res. **19** (1), 92 (2009). doi: 10.1101/gr.82701.108.
11. D. P. Bartel, et al., Mol Cell, **38** (6), 789 (2010); doi: 10.1016/j.molcel.2010. 6.5.
12. P. M. Szabo, Z. Tömböl, V. Molnár, et al., Curr. Bioinformatics **5** (1), 81 (2010).
13. R. C. Friedman and C. B. Burge, Mehods Mol. Biol. **1097**, 457 (2014).
14. M. Kellis, E. S. Lander, et at., J. Comput. Biol. **11** (2–3), 319 (2004).
15. M. Brudno and I. Dubchak, Genome Res. **14** (4), 685 (2004).
16. G. Lunter, J. Hein, et al., Genome Res. **18** (2), 298 (2008).
17. D. A. Pollard, M. B. Eisen, et al., BMC Bioinformatics **7**, 376 (2006).
18. R. Satija, J. Hein, and G. A. Lunter, Bioinformatics **26** (17), 2116 (2010).
19. R. Satija, L. Pachter, and J. Hein, Bioinformatics **24** (10), 1236 (2008).
20. Q. Zhang, J. Pell, R. Canino-Koning, et al., PLoS ONE **9** (7), e101271 (2014).
21. B. P. Lewis, C. B. Burge, and D. P. Bartel, Cell **120** (1), 15 (2005).
22. P. Benjamin and D. P. Bartel, Cell **120** (1), 15 (2005).
23. D. M. Hamby, Environ Monit Assess. **32** (2), 135 (1994). doi: 10.1007/BF00547132.
24. G. M. Cooper, et al., Genome Res. **13**, 813 (2003).
25. G. Jordan and N. Goldman, Mol. Biol. Evol. **29** (4), 1125 (2012).
26. M. Spivakov, J. Akhtar, P. Kheradpour, et al., Genome Biol. **13** (9), R49 (2012).
27. A. Khan and D. S. Marks, Nature Biotechnol. **27**, 549 (2009), Published online: 24 May 2009, corrected online: 8 July 2009.

10% of Conserved miRNA-Binding Sites in Vertebrates Are Misaligned

K.A. Prosvirov*, A.A. Mironov* **, and R.A. Soldatov* **

**Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskiye Gory 1/73, Moscow, 119234 Russia*

***Institute for Information Transmisson Problems, Russian Academy of Sciences, Bolshoy Karetnyi per. 19, Moscow, 127051 Russia*

MiRNAs are small endogenous noncoding RNAs that are responsible for repression of protein expression at the level of translation of mRNA through complementary binding with the specific site. Most of the functional binding sites (2^{nd} to 7^{th} nc of miRNA), positioned within the 3' untranslated region, are conserved. When it comes to prediction of these sites, comparative genomics is widely used alongside with its basic tool – multiple sequence alignment. However, multiple sequence alignments are prone to accumulate errors due to the huge divergence of species. Besides, during evolution binding sites can migrate along the sequence. The aim of this work is to estimate the fraction of conserved miRNA-binding sites, which cannot be predicted with the contemporary tools because of these phenomena. We introduce the term of L-conserved sites. Site is considered L-conserved if each sequence in the alignment has it within the window of length L. We observed the significant increase of the additional found conserved sites without the loss in sensitivity. This increase was also compared with the divergence of species.

Key words: miRNA, comparative genomics, multiple sequence alignment