КОРРЕЛЯЦИИ ОПЕРОННОЙ СТРУКТУРЫ С ДЛИНОЙ ГЕНОМА У 14 ВИДОВ МИКОПЛАЗМ

С.А. Лашин [1, 2], Ю.Г. Матушкин [1], Т.М. Хлебодарова [1], В.А. Лихошвай [1, 2]

[1] Учреждение Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, Новосибирск, Россия, e-mail: lashin@bionet.nsc.ru;
[2] Новосибирский государственный университет, Новосибирск, Россия

В работе анализируются геномы 14 видов *Mycoplasma* существенно различающихся размерами своих геномов. Показано, что при уменьшении длины генома у *Mycoplasma* уменьшаются плотность предсказанных сайтов терминации транскрипции (количество сайтов, нормированное на число генов), плотность предсказанных единиц транскрипции и возрастает среднее количество генов, входящих в предсказанные единицы транскрипции. Для предсказания сайтов терминации транскрипции (термотивы) разработан новый метод. В работе формулируется и обосновывается гипотеза о существовании эволюционной тенденции к укрупнению оперонных структур в процессе уменьшения размеров генома в ходе дегенеративной эволюции у *Mycoplasma*.

Ключевые слова: *Mycoplasma*, терминация транскрипции, опероны, эволюция.

Введение

Бактерии рода *Mycoplasma* – чрезвычайно полиморфные микроорганизмы, которые обладают примитивной организацией и самыми малыми геномами среди прокариот. В процессе дегенеративной эволюции они потеряли клеточную стенку и превратились в настоящих паразитов, которые 99 % времени проводят внутри клеток хозяина. Размеры их генома варьируют от 580 т.п.н. для *Mycoplasma genitalium* до 1380 т.п.н. для *M. mycoides* subsp. *mycoides* LC, т. е. различаются более чем в 2 раза (Razin *et al*., 1998). Сравнительный анализ геномов двух тесно связанных видов *M. genitalium* и *M. pneumoniae* (размер генома 816 т.п.н.) показал, что структура этих геномов существенно различается. 209 генов (точнее, открытых рамок считывания) не выявлено у *M. genitalium* в сравнении с геномом *M. pneumoniae* (Himmelreich *et al*., 1997). Анализ полностью секвенированных геномов микоплазм показал, что эти организмы не только полностью или частично утеряли гены, контролирующие метаболические пути синтеза аминокислот и структур клеточной стенки, но и претерпели значительную потерю генов, участвующих в таких процессах, как репарация и рекомбинация ДНК и клеточное деление (Razin *et al*., 1998).

Mycoplasma относятся к Firmicutes, для которых характерен ρ-независимый механизм терминации транскрипции. Поиск сайтов терминации транскрипции и анализ их распределения в геномах *Mycoplasma* позволяет прояснить некоторые аспекты формирования транскриптонов у этих организмов, что особенно интересно в свете столь значительных изменений структуры их геномов, приобретенных в ходе эволюции. Под термином «предсказанные транскриптоны» (ПТ) или «предсказанные транскрипционные единицы» (ПТЕ) мы понимаем как опероны, состоящие из цистронов, т. е. отдельных кодирующих полипептиды последовательностей, так и комплексы генов тРНК или рРНК, которые не транслируются, но транскрибируются единой РНК, а затем процессируются. В контексте работы последовательности ДНК, кодирующие полипептид, тРНК и рРНК, будем называть термином «ген».

Данная статья посвящена изучению особенностей формирования транскрипционных единиц в пределах рода *Mycoplasma*. Проана-

лизированы полногеномные нуклеотидные последовательности 14 видов *Mycoplasma*. В работе формулируется и обосновывается гипотеза о наличии эволюционной тенденции к увеличению средней длины транскрипционных единиц микоплазм при уменьшении размеров их геномов. Для подтверждения гипотезы в работе проведен анализ распределения потенциальных сайтов терминации транскрипции (тер-мотивов) с использованием нового оригинального метода их определения.

Материалы и методы

В работе использованы нуклеотидные последовательности геномов 14 видов *Mycoplasma*, характеристика которых приведена в табл. 1. Последовательности взяты из базы данных GenBank (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria).

Для предсказания сайтов терминации транскрипции использованы программа FindTerm, параметры которой были выбраны по умолчанию (http://linux1.softberry.com/berry.phtml), и программа TermPred. В программе TermPred реализован оригинальный метод поиска тер-мотивов, описанный в данной статье. Параметры метода TermPred подобраны таким образом, чтобы минимизировать отношение тер-мотивов ложных по положению к тер-мотивам истинным по положению. Истинными по положению считаются тер-мотивы, расположенные и ориентированные относительно генов таким образом, что они не вступают в смысловой конфликт с первичной оперонной разметкой генома. Обычно истинными по положению тер-мотивами являются те, которые лежат в межгенных участках, а ложными – расположенные внутри генов. Первичная оперонная разметка проведена с использованием программ FGenesB (http://linux1.softberry.com/berry.phtml) и OperonSet (Matushkin *et al.*, 2007). Для поиска тер-мотивов использовались последовательности длиной 660 нуклеотидов, лежащие в 3′-областях генов, с координатами (–60,600) относительно последнего нуклеотида гена. Статистическая обработка результатов проводилась с использованием критерия Стьюдента для коэффициента корреляции Пирсона.

Результаты и обсуждение

Микоплазмы представляют род одноклеточных организмов, в который входят как сво-

Таблица 1

Генетическая статистика для 14 видов рода *Mycoplasma*

Название вида	Количество генов*	Длина генома, п.н.	Средняя длина гена	Количество генов на 1000 п.н.
M. genitalium	519	580086	1117,70	0,89
*M. arthritidis*_158L3_1	666	820453	1231,91	0,81
*M. mobile*_163K	667	777079	1165,04	0,86
*M. hyopneumoniae*_7448	696	920079	1321,95	0,76
*M. hyopneumoniae*_J	698	897405	1285,68	0,78
*M. synoviae*_53	713	799476	1121,28	0,89
*M. hyopneumoniae*_232	727	892758	1228,00	0,81
M. pneumoniae	732	816394	1115,29	0,90
M. gallisepticum	765	996422	1302,51	0,77
*M. agalactiae*_PG2	792	877438	1107,88	0,90
M. pulmonis	814	963879	1184,13	0,84
*M. capricolum*_ATCC_27343	854	1010023	1182,70	0,85
*M. mycoides*_subsp_*mycoides*_ SC_str_PG1	1052	1211703	1151,81	0,87
M. penetrans	1069	1358633	1270,94	0,79

* В контексте изложения под геном понимается участок ДНК, кодирующий белок, рРНК или тРНК.

бодноживущие, так и паразитирующие виды. В настоящее время секвенированы геномы 14 видов микоплазм. Их геномы содержат от 0,5 до 1,4 мегабаз и от 500 до 1100 генов (табл. 1). Считается, что уменьшение генома микоплазм связано с переходом к паразитическому образу жизни. Этот процесс сопровождается утерей генов и упрощением организации метаболизма и морфологии микоплазм-паразитов. В связи с тем, что накопилось достаточное количество секвенированных геномов, появилась возможность проанализировать плотность упаковки информации в геномах микоплазм в зависимости от их длины. Естественно предположить, что по мере уменьшения длины генома будет происходить увеличение плотности генов на геном (количества генов на единицу длины, например на 1000 н.п.). Как видно из рис. 1, подобная тенденция, действительно, выявляется, однако она незначительная. По линейному тренду плотность генов возрастает при уменьшении длины генома примерно на 3 %, но корреляция этих величин является недостоверной: $r = –0{,}095$, $p = 0{,}373$.

В то же время из рис. 2 видно, что плотность предсказанных транскрипционных единиц падает при уменьшении длины генома почти на 50 %. Коэффициент корреляции между количеством генов в геномах и плотностью предсказанных транскрипционных единиц, найденных программой OperonSet, равен $r = 0{,}68$ и достоверен, $p = 0{,}004$. Для ПТЕ, выявленных программой FGenesB, коэффициент корреляции $r = 0{,}63$ и также достоверен, $p = 0{,}007$.

На рис. 3 приведены результаты анализа среднего количества генов в предсказанной транскрипционной единице в зависимости от длины генома *Mycoplasma*. Видно, что среднее количество генов в ПТЕ увеличивается практически в 2 раза при уменьшении количества генов в геномах в 2 раза. Это свидетельствует о том, что в среднем плотность ПТЕ в геноме уменьшается с уменьшением длины генома (количества генов).

Таким образом, на основе представленных результатов мы формулируем гипотезу о существовании эволюционной тенденции к укрупнению оперонных структур в процессе уменьшения размеров генома в ходе дегенеративной эволюции у микоплазм.

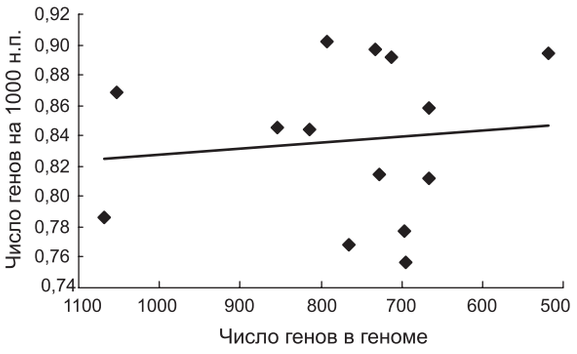

Рис. 1. Динамика изменения плотности генов (число генов/1000 н.п.) в геномах *Mycoplasma* в зависимости от их размеров.

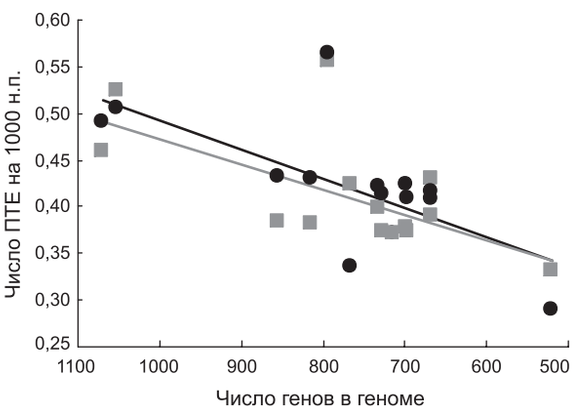

Рис. 2. Динамика изменения плотности предсказанных транскрипционных единиц (ПТЕ/1000 н.п.) в геномах *Mycoplasma* в зависимости от их размера. Черные круги и линия тренда соответствуют расчетам программы OperonSet, серые – FGenesB. Зависимости полностью сохраняются при разных параметрах соответствующих программ.

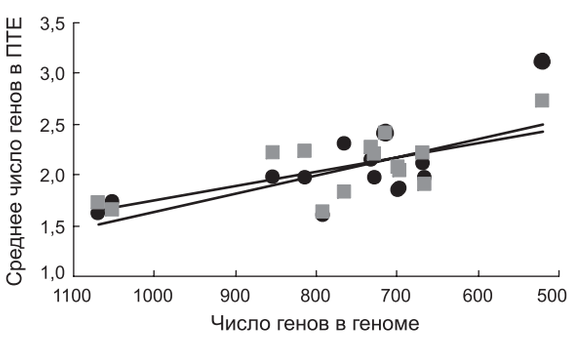

Рис. 3. Зависимость среднего числа генов в предсказанной OperonSet транскрипционной единице (ПТЕ) от размера генома *Mycoplasma*.

Для обоснования гипотезы мы предприняли дополнительный анализ распределения тер-мотивов в геномах бактерий рода *Mycoplasma*. Для анализа мы использовали два метода поиска тер-мотивов. Первый метод реализован в интернет-доступной версии программы FindTerm. Второй, TermPred, разработан в настоящей статье и описывается ниже.

Метод распознавания тер-мотивов TermPred

На рис. 4 приведена схема типичного мотива, который определяет терминацию транскрипции у *Mycoplasma* и других Firmicutes. Особенностью данного мотива является наличие GC-богатого инвертированного повтора, который в процессе синтеза РНК формирует вторичную структуру в виде шпильки, и следующего за ним Т-богатого участка (рис. 4). Появление такой структуры на 3′-конце растущей РНК приводит к запуску ρ-независимого механизма терминации транскрипции (Farnham, Platt, 1981; Lynn *et al*., 1988; Wilson, von Hippel, 1995; Wang *et al*., 1997; Yarnell, Roberts, 1999). Далее в статье такую последовательность будем называть тер-мотивом.

На основании имеющихся данных о структуре ρ-независимого терминатора транскрипции нами был разработан оригинальный компьютерный метод TermPred предсказания тер-мотивов в геномах прокариот. Для этого определяются четыре характеристики фиксированной в окне нуклеотидной последовательности: 1) энергия Гиббса формирования шпильки (G_{score}); 2) насыщенность тимином Т-богатого участка (T_{score}); 3) величина отклонения от оптимальной длины шпильки (C_{score}) и (4) длина интервала между шпилькой и Т-богатым участком (P_{score}). На основе этих характеристик вычисляется индекс потенциала терминации (*TPI*) анализируемой последовательности:

$$TPI = G_{score} + T_{score} + C_{score} + P_{score}. \qquad (1)$$

Превышение значения *TPI* некоторого фиксированного порогового значения служит основанием для идентификации анализируемой последовательности как тер-мотива.

Значение G_{score} (энергия Гиббса шпильки) рассчитывалось программой UNAFold (Markham, Zuker, 2008).

Для расчета величины T_{score} Т-богатого участка длиной 15 нуклеотидов используется формула, предложенная в работах (d'Aubenton Carafa *et al*., 1990; de Hoon *et al*., 2005):

$$T_{score} = T_0 \cdot \sum_{i=0}^{14} \delta_{nuc(i)} \cdot \exp(-\lambda i), \qquad (2)$$

где член $\exp(-\lambda i)$ описывает вклад позиции нуклеотида в значение T_{score}, параметр δ_i – вклад типа нуклеотида, расположенного в i-й позиции по формуле

$$\delta_i = \begin{cases} 1 & nuc\ (i) = t \\ -1 & nuc\ (i) = a \\ -2 & nuc\ (i) = g \text{ или } c \end{cases}$$

Значения параметров δ_i, $\lambda = 0{,}144$ и $T_0 = 2$ подобраны таким образом, чтобы максимизировать и минимизировать предсказание тер-мотивов, истинных и ложных по положению соответственно.

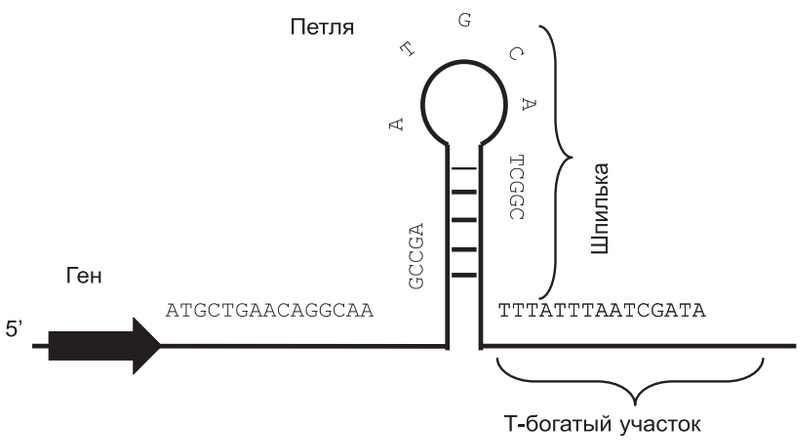

Рис. 4. Структура ρ-независимого терминатора транскрипции: шпилька и Т-богатый участок (Kingsford *et al*., 2007).

Для расчета величины C_{score} используется формула:

$$C_{score} = C_0 \cdot \left(\frac{1 + \left[n/k_{11} \right]^{h_{11}}}{1 + \left[n/k_{12} \right]^{h_{12}}} + \frac{\left[n/k_{12} \right]^{h_{21}}}{1 + \left[n/k_{22} \right]^{h_{22}}} \right), \quad (3)$$

где n – длина исследуемой шпильки, а параметры $C_0 = -5$, $k_{11} = 20{,}5$, $k_{12} = 10$, $k_{21} = 43$, $k_{22} = 43$, $h_{11} = 4$, $h_{12} = 6$, $h_{21} = h_{22} = 10$ описывают значимость отклонения ее длины от некоторой оптимальной для терминации транскрипции. Значения параметров подобраны с учетом минимизации C_{score} для шпилек оптимальной длины.

Для расчета величины P_{score} используется формула:

$$P_{score} = P_0 \cdot (\exp(\eta l) - 1), \quad (4)$$

где l – длина интервала между шпилькой и Т-богатым участком, а $\eta = 2{,}5$ и $P_0 = -1$ – параметры скорости роста P_{score} при увеличении этого разрыва. Значения параметров формулы (4) были подобраны таким образом, чтобы в качестве тер-мотива принимались последовательности, у которых расстояние между шпилькой и Т-повтором не превышает двух нуклеотидов. При этом последовательности с разрывом в два нуклеотида принимаются в качестве тер-мотивов только при очень больших значениях G_{score} и T_{score}.

Анализируемая последовательность принимается в качестве тер-мотива, если энергия Гиббса меньше 1 ккал/моль и величина *TPI* больше нуля.

Предсказание тер-мотивов в геномах бактерий рода *Mycoplasma*

На рис. 5 представлена зависимость плотности тер-мотивов от длины геномов. Отчетливо выявляется более быстрое падение плотности тер-мотивов по сравнению со скоростью уменьшения количества генов в геноме. Корреляция между этими параметрами была достоверной при использовании обеих программ. Для программы TermPred коэффициент корреляции равняется $r = 0{,}46$, $p = 0{,}049$, а для FindTerm – $r = 0{,}53$, $p = 0{,}03$.

Таким образом, полученные расчеты подтверждают высказанную гипотезу об укрупнении структуры транскрипционных единиц генома при уменьшении его длины. Как следует из рис. 5, наблюдаемая тенденция к укрупнению достаточно хорошо выражена: плотность термотивов относительно числа генов падает вдоль тренда примерно на 36 % при расчетах обеими программами.

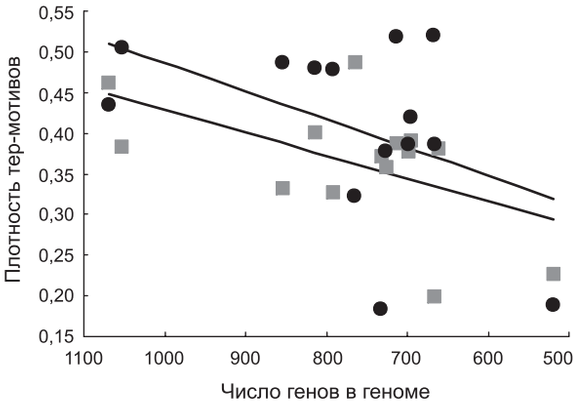

Рис. 5. Зависимость плотности тер-мотивов (число тер-мотивов/число генов) от длины генома (в генах).

Данные, полученные с помощью программы TermPred, обозначены черными кружками, а FindTerm – серыми квадратами.

Результаты качественно не меняются и в случае оценки плотности терм-мотивов в пересчете на длину генома в парах нуклеотидов (данные не приводятся).

Заключение

В работе формулируется гипотеза о существовании эволюционной тенденции к укрупнению оперонных структур в процессе уменьшения размеров генома в ходе дегенеративной эволюции у микоплазм. Для обоснования гипотезы в работе разработан новый метод предсказания сайтов терминации транскрипции (тер-мотивов) и с его помощью проанализированы геномы 14 видов микоплазм, существенно различающихся размерами своих геномов. Альтернативный поиск сайтов терминации транскрипции осуществлялся также программой FindTerm. Первичная оперонная разметка геномов была проведена с использованием двух программ FGenesB и OperonSet. Показано, что с уменьшением размера генома у микоплазм

уменьшается и плотность предсказанных сайтов терминации транскрипции, и плотность предсказанных транскрипционных единиц. Следовательно, среднее количество генов, входящих в предсказанные транскрипционные единицы, возрастает. Эти данные свидетельствуют о том, что в процессе дегенеративной эволюции у микоплазм происходили не только уменьшение размеров генома в связи с потерей целого ряда генов, контролирующих утерянные ими функции и структуры, но также и изменение структуры их геномов в целом, выражающееся в уменьшении числа транскрипционных структур за счет потери части сайтов терминации транскрипции и слияния их в более крупные образования.

Благодарности

Исследования частично поддержаны грантами РФФИ №06-04-49556 и № 08-04-01008, программами РАН «Молекулярная и клеточная биология» (проект № 10.7 «Компьютерное моделирование и экспериментальное конструирование генных сетей») и «Происхождение и эволюция биосферы» (проект №18.13 «Эволюция молекулярно-генетических систем: компьютерный анализ и моделирование», госконтрактом № 10104-37/П-18/110-327/180608/015 «Экосистемно-биоценотические и генетические механизмы биологической эволюции и корреляция биологических событий») и грантом НШ-2447.2008.4. Научная школа Н.А. Колчанова «Биоинформатика и системная компьютерная биология».

Литература

d'Aubenton Carafa Y., Brody E., Thermes C. Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures // J. Mol. Biol. 1990. V. 216. P. 835–858.

de Hoon M.J., Makita Y., Nakai K., Miyano S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species // PLoS Comput. Biol. 2005. V. 1(3). P. e25.

Farnham P.J., Platt T. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro* // Nucl. Acids Res. 1981. V. 9. P. 563–577.

Himmelreich R., Plagens H., Hilbert H. *et al*. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium* // Nucl. Acids Res. 1997. V. 25. P. 701–712.

Kingsford C.L., Ayanbule K., Salzberg S.L. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake // Genome Biol. 2007. V. 8. P. R22.

Lynn S.P., Kasper L.M., Gardner J.F. Contributions of RNA secondary structure and length of the thymidine tract to transcription termination at the *thr* operon attenuator // J. Biol. Chem. 1988. V. 263. P. 472–479.

Markham N.R., Zuker M. UNAFold: software for nucleic acid folding and hybridization // Methods Mol. Biol. 2008. V. 453. P. 3–31.

Matushkin Yu.G., Vishnevskiy O.V., Volod'ko V.B. *et al*. Computer system GenomeMarker for annotation of bacterial genome structure-functional organization // Proc. of the 4th Moscow Intern. Congr. of Biotechnology: State of the Art and Prospects of Development. Moscow, 2007. V. 2. P. 399.

Razin S., Yogev D., Naot Y. Molecular biology and pathogenicity of mycoplasmas // Microbiol. Mol. Biol. Rev. 1998. V. 62. P. 1094–1156.

Wang D., Severinov K., Landick R. Preferential interaction of the his pause RNA hairpin with RNA polymerase beta subunit residues 904-950 correlates with strong transcriptional pausing // Proc. Natl Acad. Sci. USA. 1997. V. 94. P. 8433–8438.

Wilson K.S., von Hippel P.H. Transcription termination at intrinsic terminators: the role of the RNA hairpin // Proc. Natl Acad. Sci. USA. 1995. V. 92. P. 8793–8797.

Yarnell W.S., Roberts J.W. Mechanism of intrinsic transcription termination and antitermination // Science. 1999. V. 28. P. 611–615.

CORRELATION BETWEEN THE OPERON STRUCTURE AND THE GENOME LENGTH IN 14 MYCOPLASM SPECIES

S.A. Lashin [1, 2], Yu.G. Matushkin [1, 2], T.M. Khlebodarova [1], V.A. Likhoshvai [1, 2]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: lashin@bionet.nsc.ru;
[2] Novosibirsk State University, Novosibirsk, Russia

Summary

We have analyzed 14 genomes of various mycoplasm species with quite different size. Both density of predicted transcription termination sites (number of sites scaled with number of genes) and density of predicted transcription units have been shown to decrease in view of genome length reduction. Thereafter an average number of genes in a transcription unit increases. Transcription termination sites (ter-motifs) were predicted with the use of a novel method developed. The hypothesis is formulated and proved that there is an evolutionary tendency of operon structures to increase during degenerative evolution (genome length reduction) of mycoplasms.