Программное обеспечение для компьютерного исследования особенностей элонгации трансляции (на примере одноклеточных организмов рода Mycoplasma)

Соколов В.С., Лихошвай В.А., Матушкин Ю.Г.
Институт цитологии и генетики СО РАН

Введение

Изучение эффективности экспрессии генов, а также ее оптимизация представляют собой актуальную и существенную проблему, как для теоретической, так и практической биологии. Ее исследование важно как с точки зрения фундаментальной науки – получение новых данных по экспрессии генов тех организмов, экспериментальные данные по которым пока еще недоступны, так и в экспериментальном приложении, например, для планирования генно-инженерных экспериментов или максимизации продуктивности продуцентов (генно-модифицированных организмов).

Результаты исследований в данной области показывают, что для одноклеточных организмов и многих многоклеточных существует зависимость уровня экспрессии генов от таких факторов как кодонный состав гена, наличие и распределение вторичных структур в мРНК и «прочность» этих структур [1]. В зависимости от сочетания этих факторов выделяется 5 групп организмов, по-разному оптимизировавших первичную структуру своих генов в процессе эволюции [2].

Программное обеспечение и алгоритм его работы

Для определения принадлежности организма к одной из пяти групп был разработан специальный индекс эффективности элонгации EEI, который имеет смысл усредненного времени элонгации трансляции. Для расчета данного индекса на языке Java была написана специальная программа. На данный момент есть как консольный вариант этой программы, так и ее интернет версия, доступная по следующей ссылке: http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/eei-calculator/index.pl.

Индекс EEI рассчитывается для каждого гена организма по следующей формуле:

$$EEI(i) = u_1 T_a(i) + u_2 T_e(i),$$

где i – номер гена, $u_1=\{0;1\}$; $u_2=\{0;1\}$ – весовые коэффициенты, определяющие учет каждого слагаемого в значении индекса. Всего имеется три нетривиальные комбинации

весовых коэффициентов: u_1=1 и u_2=0 – учитывается только слагаемое $T_a(i)$, u_1=0 и u_2=1 – учитывается только слагаемое $T_e(i)$, u_1=1 и u_2=1 – учитываются оба слагаемых $T_a(i)$ и $T_e(i)$.

Первое слагаемое T_a имеет смысл среднего времени, требуемого для размещения в А-сайте рибосомы изоакцепторной аминоацил-тРНК. Данный параметр рассчитывается по кодонному составу гена, с точностью до коэффициента пропорциональности. Формула для его рассчета:

$$T_a(i)=\sum_{j=1}^{n_i}\beta_{\delta(i,j)}\Big/n_i\ ,\ \beta_{\delta(i,j)}=\frac{\sum_{m=1}^{C}\sqrt{\alpha_m}}{\sqrt{\alpha_{\delta(i,j)}}}\ , \qquad (1)$$

где величина $1/\beta_{\delta(i,j)}$ в простейшем случае интерпретируется как оптимальная относительная концентрация аминоацил-тРНК, комплементарной j-ому учитываемому кодону, а $\alpha_{\delta(i,j)}$ и α_m – частоты использования кодонов $\delta(i,j)$ и m в обучающей подвыборке генов [3].

Второе слагаемое $T_e(i)$ имеет смысл среднего времени, затрачиваемого рибосомой на стадию транслокации, и оценивается по уровню комплементарности i-й мРНК, с точностью до коэффициента пропорциональности:

$$T_e(i) = t_{min}\cdot(1-p(i))+t_{max}\cdot p(i),$$

здесь t_{min} – минимальное условное время транслокации, t_{max} – максимальное условное время транслокации, p(i) – вероятность реализации максимального условного времени транслокации, которая вычисляется по формуле:

$$p(i) = \int_0^{LCI(i)} \frac{k^{n+1}x^n}{G(n+1)}e^{-kx}dx\ ,$$

$k=m/\sigma^2$, $n=(m/\sigma)^2$, где m и σ^2, соответственно, математическое ожидание и дисперсия положительной случайной величины имеющей плотность распределения $\frac{k^{n+1}x^n}{G(n+1)}e^{-kx}$, G(n+1) – Gamma-функция, LCI(i) – индекс локальной комплементарности. Следует отметить, что если в качестве *p(i)* выбирать другие формы S-образной зависимости от аргумента LCI(i), результаты расчетов существенно не изменяются [4].

В программе используются две формы индекса локальной комплементарности:

1) *LCI1* – без учета энергии образования вторичных структур (шпилек). Рассчитывается по формуле (2) на участке длиной m_i нуклеотидов. Данный индекс имеет смысл усредненного количества комплементарных участков в гене (в данной работе m_i бралось равным длине i-го гена плюс 56 нуклеотидов со стороны его 3'- конца):

$$LCI1(i) = \frac{\sum_{m=1}^{m_i - s_{\max} - l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[\sum_{l=l_{\min}}^{l_{\max}} \zeta\left(con(m, m+s-1), \overline{con(m+s+l-1, m+2s+l-2)}\right) \right] \right\}}{m_i - s_{\max} - l_{\max}}, \qquad (2)$$

где *con(i,j)* – контекст гена с i-го по j-й нуклеотиды и $\overline{con(i,j)}$ – комплементарный контекст гена с j-го по i-й нуклеотиды ($i \le j$), ζ(conext1,conext2)=1, если слова conext1 и conext2 идентичны, в противном случае ζ(conext1,conext2)=0. Длина учитываемого инвертированного повтора не меньше s_{min} и не больше s_{max}, расстояние между учитываемыми инвертированными повторами не меньше l_{min} и не больше l_{max} (в работе приняты следующие значения параметров: s_{min}=3, s_{max}=6, l_{min}=3, l_{max}=50).

2) *LCI2* – с учетом энергии образования вторичных структур. Данный индекс имеет смысл усредненной энергии повторов, для которых образование шпильки энергетически возможно. Данный индекс рассчитывается по формуле (3) на участке длиной m_i нуклеотидов (значения параметров следующие: s_{min}=3, s_{max}=6, l_{min}=3, l_{max}=50):

$$LCI2(i) = \frac{\sum_{m=1}^{m_i - s_{\max} - l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[\sum_{l=l_{\min}}^{l_{\max}} \psi\left(con(m, m+s-1), \overline{con(m+s+l-1, m+2s+l-2)}\right) \right] \right\}}{m_i - s_{\max} - l_{\max}}, \qquad (3)$$

где ψ – энергия вторичной структуры, которая подсчитывается стандартным образом [5].

Время, затрачиваемое рибосомой на стадию транспептидации, полагалось равным для всех кодонов и всех генов и поэтому не учитывалось при расчете EEI(i).

После расчетов все гены упорядочиваются по уменьшению индекса EEI(i) и затем среди них выбирается обучающая выборка, состоящая из 150 генов с наименьшими значениями EEI(i). По ней рассчитывается параметр $\beta_{\delta(i,j)}$ из формулы (1), и цикл повторяется, пока состав обучающей выборки не перестанет изменяться. Обычно это происходит за 10-15 итераций.

Данный алгоритм выполняется пять раз, по одному разу для каждого типа индекса:

EEI-1 – учитывается только кодонный состав рамки трансляции. Локальные инвертированные повторы не рассматриваются (u_1=1 и u_2=0);

EEI-2 – учитывается только среднее количество локальных инвертированных повторов (без учета энергии образования шпилек, *LCI1*). Кодонный состав не рассматривается (u_1=0 и u_2=1);

EEI-3 – учитывается усредненная энергия повторов, для которых образование шпильки энергетически возможно (*LCI2*). Кодонный состав не рассматривается (u_1=0 и u_2=1);

EEI-4 – является комбинацией первого и второго типов индекса. Учитываются кодонный состав и среднее количество локальных инвертированных повторов (без учета энергии образования шпилек, *LCI1*) (u_1=1 и u_2=1);

EEI-5 – является комбинацией первого и третьего типов индекса. Учитываются кодонный состав и усредненная энергия повторов, для которых образование шпильки энергетически возможно (*LCI2*) (u_1=1 и u_2=1).

Алгоритм работы программы следующий:

а) программе на вход подается карточка организма в GBK или EMB формате;

б) из карточки экстрагируются нуклеотидные последовательности всех генов;

в) по последовательностям генов рассчитываются частоты кодонов и количество или энергия локальных инвертированных повторов;

г) рассчитываются все пять видов индекса EEI, и идет итерационное упорядочивание генов по уменьшению EEI.

Для определения, какой из видов индекса работает наиболее эффективно в данном организме, используются два параметра M и R. M – среднее положение рибосомных генов в отсортированном списке генов, а R – стандартное отклонение от среднего. Чем выше M и меньше R, т.е. чем ближе к концу списка и плотнее расположены рибосомные гены, тем лучше работает данный вид индекса. Рибосомные гены выбраны в качестве маркеров высоко экспрессирующихся генов по двум основным причинам:

1) известно, что для большинства организмов рибосомные гены являются высоко экспрессирующимися;

2) данные гены легко выделить из общей массы генов организма по ключевым словам в их описании в исходной карточке.

Результаты расчетов индексов для организмов рода *Mycoplasma*

При помощи написанной программы были исследованы геномы 42 видов организмов, принадлежащих к роду *Mycolasma*. Результаты работы представлены в таблице 1.

Как видно из таблицы 1, для большинства организмов лучше всего работает второй вид индекса EEI, т.е. учитывается только фактическая насыщенность рамки считывания локальными инвертированными повторами (без учета энергетики), кодонный состав не учитывается. Но среди общей массы выделяются несколько организмов, у которых наибольшее из всех пяти значений M меньше 30 и значительно отличается от аналогичного значения для остальных организмов. Особенно выделяются *Mycoplasma*

haemocanis str. Illinois, *Mycoplasma haemofelis Ohio2* и *Mycoplasma haemofelis str. Langford 1*. У них наибольшее M находится в районе нуля.

Таблица 1. Результаты работы программы для организмов рода *Mycoplasma*.

Организм	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
Candidatus Mycoplasma haemominutum 'Birmingham 1'	-9	53	23	46	7	52	0	55	-8	52
Mycoplasma agalactiae	19	65	64	41	-24	66	62	50	-25	65
Mycoplasma agalactiae PG2	21	66	64	44	-27	66	68	47	-28	64
Mycoplasma arthritidis 158L3-1	2	61	62	40	-23	62	49	52	-26	58
Mycoplasma bovis Hubei-1	17	61	66	44	-44	57	78	33	-31	64
Mycoplasma bovis PG45	4	67	66	38	-40	62	65	49	-38	61
Mycoplasma capricolum subsp. capricolum ATCC 27343	-59	42	78	20	-26	63	50	54	-45	51
Mycoplasma conjunctivae HRC581	-10	57	67	32	-18	63	59	53	-31	56
Mycoplasma crocodyli MP145	-44	54	76	35	-44	57	58	54	-56	49
Mycoplasma fermentans JER	-34	55	81	22	-42	64	72	41	-61	44
Mycoplasma fermentans M64	-56	45	79	25	-50	61	58	45	-74	26
Mycoplasma gallisepticum str. F	23	59	54	44	1	65	59	48	-3	63
Mycoplasma gallisepticum str. R(high)	11	61	53	44	2	64	45	51	-6	62
Mycoplasma gallisepticum str. R(low)	12	60	53	44	2	64	45	51	-5	61
Mycoplasma genitalium G37	-66	44	59	42	-39	59	-25	59	-63	42
Mycoplasma haemocanis str. Illinois	-19	61	-6	63	-6	70	-22	55	-29	58
Mycoplasma haemofelis Ohio2	-25	49	-21	59	6	67	-39	41	-5	65
Mycoplasma haemofelis str. Langford 1	-25	47	-16	60	3	69	-37	43	-13	62
Mycoplasma hominis ATCC 23114	-10	64	69	30	-28	65	67	47	-39	58
Mycoplasma hyopneumoniae 168	-73	42	68	42	-27	66	-30	69	-63	45
Mycoplasma hyopneumoniae 232	-63	45	77	22	-24	62	-9	68	-56	47
Mycoplasma hyopneumoniae 7448	-67	45	71	33	-27	64	-21	70	-62	44
Mycoplasma hyopneumoniae J	-65	48	71	32	-28	61	-21	70	-61	43
Mycoplasma hyorhinis GDL-1	-24	64	70	38	-23	63	45	65	-39	52
Mycoplasma hyorhinis HUB-1	-16	67	71	40	-18	66	47	60	-34	55
Mycoplasma hyorhinis MCLD	-23	61	80	23	-21	60	52	56	-39	48
Mycoplasma leachii 990146	-56	43	76	23	-29	60	65	42	-48	50
Mycoplasma leachii PG50	-55	45	75	23	-31	61	52	50	-49	50
Mycoplasma mobile 163K	-45	55	70	30	-9	64	50	58	-31	59
Mycoplasma mycoides subsp. capri LC str. 95010	-54	46	78	25	-30	63	44	61	-49	53
Mycoplasma mycoides subsp. capri str. GM12	-50	49	76	25	-24	63	63	45	-39	55
Mycoplasma mycoides subsp. capri str. GM12	-50	49	76	25	-25	63	63	45	-39	56
Mycoplasma mycoides subsp. mycoides SC str. Gladysdale	-39	48	76	29	-24	64	64	46	-38	55
Mycoplasma mycoides subsp. mycoides SC str. PG1	-47	45	78	26	-23	64	61	44	-43	52
Mycoplasma penetrans HF-2	3	66	71	32	-32	65	59	46	-36	63
Mycoplasma pneumoniae FH	-18	58	24	60	-32	57	22	67	-38	52
Mycoplasma pneumoniae M129	-18	55	26	55	-28	56	25	63	-35	48
Mycoplasma pulmonis	-17	58	68	39	3	64	49	62	-18	59
Mycoplasma putrefaciens KS1	-20	63	68	41	-21	63	61	58	-37	59
Mycoplasma suis KI3806	-2	51	26	45	0	50	7	40	-4	44
Mycoplasma suis str. Illinois	-16	45	25	46	-2	49	-4	39	-12	40
Mycoplasma synoviae 53	-28	53	71	33	-4	57	36	58	-30	52

Мы попытались выяснить причину такого небольшого смещения у данных организмов рибосомных генов в сторону низких значений EEI, т.е. более высокой экспрессии. Исследование показало, что рибосомные гены у всех исследованных микоплазм в среднем схожи по количеству локальных инвертированных повторов (Рис. 1), что, в принципе, ожидаемо, т.к. данные гены являются консервативными и не должны были сильно измениться в процессе эволюции у столь близких родственников. Однако

среднее количество повторов по остальным генам сильно отличается у данных организмов (Рис. 1).

На рисунке 1 выделена группа из трех организмов, у которых были наименьшие значения параметра M (вблизи нуля). Группы организмов на краях тоже имели низкие значения M (< 30). Как видно из графика, у данных организмов среднее число повторов у нерибосомных генов снижено, по сравнению с другими организмами, и приближается к среднему значению по рибосомным генам.

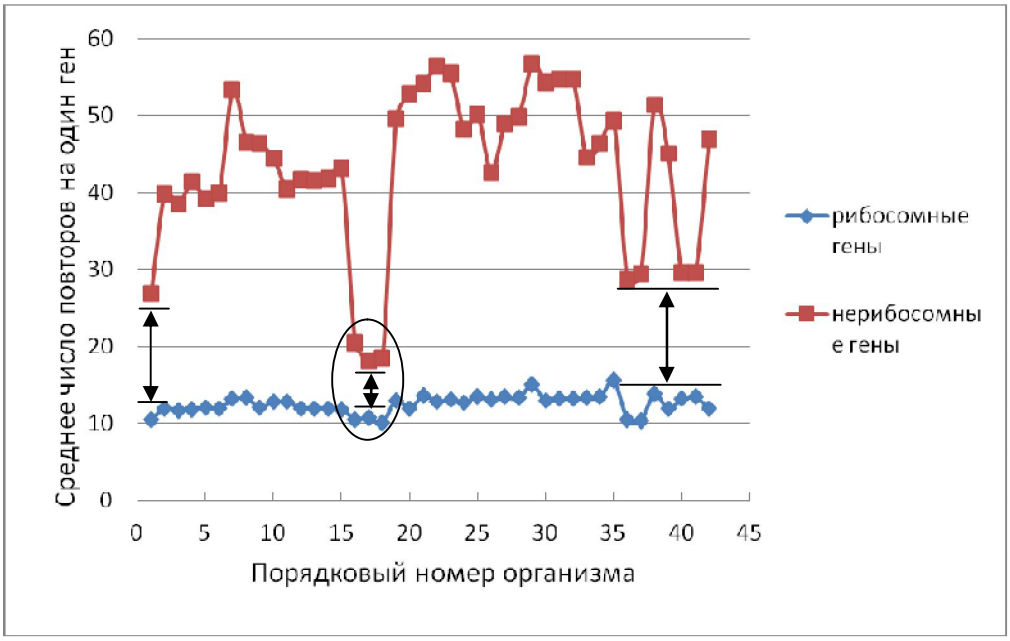

Рис. 1. Среднее количество повторов на один ген у 42 видов организмов рода *Mycoplasma*.

Мы предполагаем, что данные «особые» организмы эволюционировали в сторону общего уменьшения количества повторов, т.е. вторичных структур, в их генах, что привело к меньшему различию между средним количеством повторов в рибосомных и нерибосомных генах. Таким способом они могли повысить уровень экспрессии своих генов. Возможно, это связано с паразитическим образом жизни этих организмов.

Заключение

В работе представлено программное обеспечение для оценки эффективности экспрессии генов организма по их нуклеотидной последовательности. Приведены результаты работы программы для 42 видов организмов принадлежащих к роду *Mycoplasma*. Показано, что у трех видов прошла массовая оптимизация первичной структуры генов по наличию повторов (шпилек в мРНК).

Литература

1. Н.В. Владимиров, В.А. Лихошвай, Ю.Г. Матушкин, Корреляция частот кодонов и потенциальных вторичных структур с эффективностью трансляции мРНК в одноклеточных организмах, Мол. Биол. 2007.

2. Vitali A. Likhoshvai, Yuri G. Matushkin, Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy, FEBS, 2002.

3. Likhoshvai, V.A. (1992) In Modeling and Computer Methods in Molecular Biology and Genetics, Ratner, V.A. and Kolchanov, N.A., Eds. Nova Sc. Publishers: USA, 463-469.

4. Likhoshvai V.A., Matushkin Yu.G. (2000) Molecular Biology (Mosk), 34, 345-350.

5. Turner D.H. and Sugimoto N. (1988) Ann. Rev. Biophys. Chem., 17, 167-192.