Вавиловский журнал генетики и селекции. 2016;20(6):779-786
DOI 10.18699/VJ16.195

Оценка трансляционной значимости характеристик нуклеотидной последовательности мРНК млекопитающих на основе данных рибосомного профилирования

О.А. Волкова[1] ✉, Ю.В. Кондрахин[2, 3], Р.Н. Шарипов[2, 3, 4]

[1] Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
[2] Федеральное государственное бюджетное научное учреждение «Институт вычислительных технологий Сибирского отделения Российской академии наук», Новосибирск, Россия
[3] Общество с ограниченной ответственностью «Институт системной биологии», Новосибирск, Россия
[4] Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

Известно, что характеристики 5′-нетранслируемой последовательности (5′-НТП) мРНК могут оказывать влияние на эффективность и специфичность инициации трансляции. Ранее знания о характеристиках 5′-НТП были получены теоретически и в экспериментах *in vitro* для мРНК отдельных генов, что не давало возможности оценить реальную трансляционную значимость ее параметров. Для выявления трансляционно-значимых характеристик 5′-НТП необходимо проанализировать их связь с трансляционной активностью соответствующих мРНК. Однако до недавнего времени доступные технологии не позволяли получить широкогеномные экспериментальные данные по эффективности трансляции. Благодаря появившейся технологии профилирования рибосом такие данные были получены для мРНК ряда эукариот. Использование их позволяет оценивать и выявлять трансляционно-значимые параметры мРНК, а также предсказывать эффективность трансляции мРНК на основании характеристик ее нуклеотидной последовательности. Цель нашей работы – определение трансляционной значимости отдельных характеристик нуклеотидной последовательности 5′-НТП мРНК на основании соответствующих экспериментальных данных по эффективности трансляции, рибосомному профилированию. Проведен статистический анализ отдельных характеристик нуклеотидных последовательностей мРНК человека и мыши; выявлена их взаимосвязь с соответствующими данными рибосомного профилирования. Были отобраны трансляционно-значимые параметры мРНК, тенденция влияния на эффективность трансляции которых наиболее значима и одинакова для всех трех проанализированных выборок: пурин в –3-позиции стартового кодона, вышележащие стартовые кодоны AUG в 5′-НТП, и комплементарные нуклеотиды G+C в составе 5′-НТП снижают эффективность трансляции; олигонуклеотид CCGCCA в районе 5′-НТП и олигонуклеотиды AAGAAA, AAGAAG, AAGCAG, AAAAAG в составе белок-кодирующей последовательности – усиливают. Разработаны с помощью платформы BioUML набор инструментов, позволяющий анализировать трансляционную значимость отдельных 5′-НТП мРНК, и программа для предсказания эффективности трансляции мРНК на основании ее нуклеотидной последовательности.

Ключевые слова: профилирование рибосом; RiboSeq; 5′-НТП; мРНК; инициация трансляции.

Estimation of translational importance of mammalian mRNA nucleotide sequence characteristics based on ribosome profiling data

O.A. Volkova[1] ✉, Yu.V. Kondrakhin[2, 3], R.N. Sharipov[2, 3, 4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2] Institute of Computational Technologies SB RAS, Novosibirsk, Russia
[3] Institute of Systems Biology, Ltd, Novosibirsk, Russia
[4] Novosibirsk State University, Novosibirsk, Russia

It is known that the 5′ untranslated region (5′ UTR) mRNA characteristics can influence translation initiation efficiency and specificity. Previous knowledge about 5′ UTR characteristics was obtained theoretically and *in vitro* for mRNA of individual genes. It did not allow systematic analysis of mRNA translationally important parameters. To identify the above mentioned 5′ UTR characteristics, it is necessary to analyze their relationships with the translational activity of the corresponding mRNAs. Until recently, there were no experimental data on translation efficiency. Thanks to ribosome profiling technology, genome-wide experimental data of translation efficiency have been obtained for many eukaryotic mRNAs. Now it seems to be possible to reveal translationally important mRNA parameters and predict translation efficiency based on their nucleotide sequences. The aim of this study was to determine the translational significance of individual 5′ UTR characteristics in accordance with experimental ribosome profiling data. A statistical analysis was carried out for revealing relationships between the human and mouse mRNA nucleotide sequence characteristics and ribosome profiling data. Some of the mRNA parameters influencing translation efficiency were most significant, and the same trends for all three samples analyzed were revealed: a purine at start codon context position –3, upstream AUG pre-

ORIGINAL ARTICLE
Received 24.12.2015
Accepted for publication 26.04.2016

✉ e-mail: ov@bionet.nsc.ru

sence and G+C complementary nucleotide concentration reduce translation efficiency; whereas gexonucleotides CCGCCA (5′ UTR) and AAGAAA, AAGAAG, AAGCAG, AAAAAG (CDS) increase translation efficiency. A toolkit that allows analyzing the importance of 5′ UTR characteristics and a program for prediction of translation efficiency were developed on the base of the BioUML platform.

Key words: ribosome profiling; RiboSeq; 5′UTR; mRNA; translation initiation.

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ:

Волкова О.А., Кондрахин Ю.В., Шарипов Р.Н. Оценка трансляционной значимости характеристик нуклеотидной последовательности мРНК млекопитающих на основе данных рибосомного профилирования. Вавиловский журнал генетики и селекции. 2016;20(6):779-786. DOI 10.18699/VJ16.195

HOW TO CITE THIS ARTICLE:

Volkova O.A., Kondrakhin Yu.V., Sharipov R.N. Estimation of translational importance of mammalian mRNA nucleotide sequence characteristics based on ribosome profiling data. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2016;20(6):779-786. DOI 10.18699/VJ16.195

Трансляция оказывает существенное влияние на регуляцию экспрессии генов. Наиболее значима стадия инициации, а элонгация, трансляционные паузы и другие этапы не играют существенной роли (Ingolia et al., 2011). Эффективность трансляции является консервативной в плане эволюции. Это предполагает, что эффективность трансляции обусловлена первичной последовательностью мРНК и ее взаимодействиями с соответствующими трансляционными факторами и 40S субъединицами рибосом.

мРНК эукариотических генов различаются по интенсивности синтеза белка в процессе трансляции. 5′-нетранслируемая последовательность мРНК (5′-НТП, лидерная последовательность) участвует во взаимодействии с факторами аппарата трансляции и 40S субъединицами рибосом (Hinnebusch, Lorsch, 2012; Singh et al., 2012). Показано, что характеристики нуклеотидной последовательности 5′-НТП оказывают существенное влияние на эффективность инициации трансляции (Kozak, 2005; Kochetov, 2008). Были выявлены некоторые из таких функционально значимых характеристик, в частности стабильная вторичная структура (Ding et al., 2012; Li et al., 2012) и потенциальные стартовые кодоны в районе 5′-НТП, которые могут ингибировать трансляцию (Kozak, 2005). Однако имеющаяся информация о структурно-функциональной организации 5′-НТП недостаточна и требует дальнейшего уточнения.

Контекст стартового кодона трансляции в эукариотических мРНК. Считается, что распознавание триплета AUG в качестве стартового кодона зависит от нуклеотидного окружения (контекста). Экспериментально было подтверждено, что аденин или гуанин в позиции –3 (A/G^{-3}) и, в меньшей степени, гуанин в позиции +4 (G^{+4}) относительно кодона AUG способствуют распознаванию его как стартового (оптимальный контекст: $R^{-3}NNAUGG^{+4}$, R = A или G) (Kozak, 1981, 1997). Показано, что у млекопитающих в дискриминации между AUG кодонами в оптимальном и неоптимальном контекстах участвует эукариотический фактор инициации трансляции eIF1 (Pestova, Kolupaeva, 2002; Takacs et al., 2011). На основе экспериментальных данных предполагается, что за узнавание 43S инициаторным комплексом пуринового нуклеотида в позиции –3 ответственно взаимодействие данного нуклеотида с α-субъединицей eIF2, а за узнавание гуанина в позиции +4, возможно, ответственно его взаимодействие с нуклеотидами A1818-A1819 в спирали 44 18S рРНК (Pisarev et al., 2006).

Оценки влияния нуклеотидов белок-кодирующей последовательности (БКП) на распознавание стартового кодона варьируют от существенного (Grunert, Jackson, 1994; Niimura et al., 2003; Nakagawa et al., 2008) и ограниченно значимого (только G в позиции +4; (Kozak, 1997)) до незначительного (Harkins et al., 2005; Xia et al., 2007). Показано, что существуют устойчивые комбинации нуклеотидов в 5′- и 3′-частях контекста стартового кодона, к которым относятся варианты оптимального контекста **A**nnAUGn и **G**nnAUG**G** (Volkova, Kochetov, 2010).

Встречаемость триплетов AUG в составе 5′-НТП. В рамках модели линейного сканирования считается, что в составе 5′-НТП эукариотических мРНК не должны содержаться триплеты AUG (upstream AUG, uAUG), поскольку часть рибосом может распознавать их как стартовые кодоны и это будет снижать эффективность трансляции основной рамки считывания (Kozak, 2005; Kochetov, 2008). Ранее показано, что от 19 до 48 % 5′-НТП мРНК разных видов эукариот содержат uAUG, что значительно реже относительно промоторов, трейлеров и интронов (Rogozin et al., 2001; Volkova, Kochetov, 2010).

Контекстные особенности эукариотических 5′-НТП способствуют формированию менее стабильной вторичной структуры. Считается, что повышенное содержание G и C коррелирует со способностью нуклеотидной последовательности РНК формировать стабильную вторичную структуру, поскольку комплементарное взаимодействие G-C-пар наиболее энергетически выгодно. Вторичная структура в составе 5′-НТП препятствует сканированию мРНК 40S субъединицами рибосом и снижает эффективность инициации трансляции (Kozak, 2005). Так, обнаружено, что 5′-НТП характеризуются даже более высоким содержанием G+C в сравнении с другими функциональными районами мРНК генов. Однако 5′-НТП человека и мыши характеризовались большим дисбалансом в содержании комплементарных нуклеотидов по сравнению с другими функциональными районами генов (Volkova, Kochetov, 2010). Также показано, что вторичная структура, лежащая ниже стартового кодона

в неоптимальном контексте в позициях 13–17, может способствовать его распознаванию 40S субъединицами рибосом (Kochetov et al., 2007).

Ранее было найдено (Kochetov et al., 1998), что трансляционно-значимые характеристики мРНК (контекст старт-кодона, размер 5′-НТП, потенциал формирования вторичной структуры) у выборки высокоэкспрессирующихся генов млекопитающих оптимизированы. мРНК с оптимальным контекстом стартового кодона характеризуются более короткими 5′-НТП, отсутствием (или меньшим числом) uAUG, и они менее склонны к формированию стабильной вторичной структуры в лидерном районе. Однако данные о характеристиках 5′-НТП в основном были получены теоретически и в экспериментах *in vitro* для мРНК отдельных генов, что не позволяло оценить реальную трансляционную значимость ее параметров.

В последние несколько лет благодаря технологии профилирования рибосом (ribosome profiling technique, RiboSeq) (Siwiak, Zielenkiewicz, 2010; Ingolia et al., 2011; Schwanhausser et al., 2011; Weiss, Atkins, 2011; Lee et al., 2012; Michel et al., 2012; и др.) были получены экспериментальные данные по эффективности трансляции мРНК ряда эукариот. Профилирование рибосом – «глобальный снимок всех рибосом, загруженных на мРНК в клетке в одно мгновение», – производится путем количественного преобразования в ДНК защищенных рибосомами от рибонуклеазы участков мРНК и секвенирования всех ДНК.

Процедура рибосомного профилирования осуществляется с помощью соответствующих ингибиторов инициации трансляции (например, харрингтонин) или элонгации (например, циклогексимид) или сначала инициации, а затем элонгации. Обработка харрингтонином ингибирует формирование первой пептидной связи и, таким образом, способствует позиционированию рибосом на стартовых кодонах; циклогексимид ингибирует элонгацию путем подавления перемещения пептидил-тРНК из акцепторного в донорный сайт большой субъединицы рибосомы.

Ранее в мировой практике оценивали уровень трансляции мРНК с помощью анализа микрочипов мРНК, выделенных из полисом (Johannes et al., 1999; Zong et al., 1999; Na, Lee, 2010), и профилирования с использованием аффинной очистки эпитопом-связанных рибосом (Heiman et al., 2008; Sanz et al., 2009). Однако эти методы не дают позиционной и количественной информации в отличие от метода профилирования рибосом. Данная техника важна для аннотирования генома и исследования регуляции экспрессии.

Рибосомное профилирование широко используется для исследования регуляции трансляции мРНК. Так, технология RiboSeq позволила обнаружить новые транслируемые мРНК и альтернативные сайты инициации трансляции (Liu et al., 2011; Michel et al., 2012), в том числе в зависимости от фазы клеточного цикла (Brar et al., 2012), в условиях стресса (Gerashchenko et al., 2012), локализации трансляции в клетке (Reid, Nicchitta, 2012), апоптозе (Zhu et al., 2014). Однако существуют и недостатки данной технологии (Gerashchenko, Gladyshev, 2014).

Выявление трансляционно-значимых характеристик 5′-НТП мРНК важно для предсказания трансляционной активности мРНК (Simon, Ruckenstein, 1966; Heinrich, Rapoport, 1980; Godefroy-Colburn, Thach, 1981; Zouridis, Hatzimanikatis, 2007; Dimelow, Wilkinson, 2009), особенностей контроля экспрессии генов на пост-транскрипционном уровне, а также связи между мутациями в этом районе гена и патологическими состояниями. В этой работе мы используем данные RiboSeq для выявления трансляционно-значимых характеристик мРНК и предсказания эффективности трансляции мРНК по ее нуклеотидной последовательности.

Материалы и методы

Для выявления взаимосвязи между отдельными трансляционно-значимыми характеристиками последовательностей мРНК и соответствующими экспериментально найденными с помощью технологии рибосомного профилирования RiboSeq эффективностями трансляции (количество ридов), известными из литературы и базы данных GEO (http://www.ncbi.nlm.nih.gov/geo/), нами были составлены три следующие выборки мРНК с обработанными данными RiboSeq: GSE30839 (Set 1, *Mus musculus*, эмбриональные стволовые клетки) (Ingolia et al., 2011); GSE31539 (Set 2, *Homo sapiens*, HEK293) (Reid, Nicchitta, 2012); GSE37744 (Set3, *Homo sapiens*, HEK293) (Ingolia et al., 2012).

В качестве трансляционно-значимых характеристик мРНК мы использовали: 1) длину 5′-НТП; 2) наличие оптимального контекста аденина или гуанина в позиции –3 контекста стартового кодона, A/G^{-3}; 3) наличие гуанина в позиции +4 контекста стартового кодона, G^{+4}; 4) содержание кодонов AUG в 5′-НТП, [uAUG]; 5) концентрацию отдельных гексануклеотидов в районах [–100, –1] и [3, 103] относительно стартового кодона; 6) содержание комплементарных нуклеотидов гуанина и цитозина в 5′-НТП, ([G+C]); 7) индексы комплементарности (IC) в позициях [–15, 1] и [1, 20] относительно стартового кодона; 8) длину 3′-НТП; 9) концентрацию комплементарных нуклеотидов C+G в 3′-НТП; 10) длину мРНК; 11) концентрацию комплементарных нуклеотидов C+G в мРНК. Для оценки эффективности трансляции использовали данные RiboSeq: нормированное количество ридов, rpkm (reads per kilobase of transcript per million mapped reads) после обработки харрингтонином.

Были отобраны два специфических набора гексануклеотидов, которые наиболее часто встречались в выборке мРНК, с высокими значениями rpkm и, соответственно, высокой эффективностью инициации трансляции: GCCGCC, CGCCGC, CCGCCA, CCGCCG, CCGCGC, CCCGCC, CTCCGC, CGCGCC и AAGAAG, GAAGAA, CAAGAA, AGAAGC, CCAAGA, AAGCAG, AGAAGA, AAGAAA, AAAAAG, AAGAAC для районов [–100, –1] и [3, 103] относительно стартового кодона соответственно (см. табл. 4).

Индексы комплементарности IC[–15, 1] и IC[1, 20] вычисляли в районах [–15, 1] и [1, 20] соответственно, относительно стартовых кодонов. Комплементарный индекс IC определяли как коэффициент корреляции между частотами комплементарных триплетов. Предполагается, что чем выше индекс IC, тем более стабильная вторичная структура формируется исследуемым районом.

Для оценки и предсказания влияния характеристик нуклеотидной последовательности 5′-НТП мРНК млекопитающих на ее трансляционную эффективность, полученную с помощью рибосомного профилирования, были использованы методы регрессионного, дискриминантного и кластерного анализов.

Для оценки качества предсказания линейной регрессии, построенной методом наименьших квадратов, использовали коэффициент корреляции между наблюдаемыми и предсказанными эффективностями трансляции. Необходимо отметить, что для построения наиболее эффективной регрессионной функции нет необходимости использования всех отобранных трансляционно-значимых характеристистик, например, вследствие наличия мультиколлинеарности (т. е. характеристики мРНК могут коррелировать между собой).

Работа выполнялась на основе платформы BioUML (http://www.biouml.org). Были использованы следующие ее возможности: поддержка работы со многими базами данных; импорт/экспорт данных в разных форматах, работа с таблицами и выборками; поддержка DAS протокола (http://www.biodas.org) для доступа к последовательностям и их аннотациям; геномный браузер, обеспечивающий широкие возможности интерактивной визуализации последовательностей, их аннотаций и данных NGS (поддержка и визуализация данных в форматах SAM/BAM); разнообразные встроенные методы анализа данных; интеграция с R/Bioconductor и Galaxy (https://main.g2.bx.psu.edu/) позволяют использовать внутри BioUML множество других программ, в первую очередь для выравнивания ридов – BWA, Bowtie и др.

Результаты

Предсказание эффективности трансляции на основании трансляционно-значимых характеристик с помощью дискриминантного анализа. Для проведения дискриминантного анализа мы отобрали по 1000 мРНК с высокими и низкими значениями rpkm (с высоким и низким уровнем трансляции соответственно) из выборки Set 1. Использовали следующие характеристики мРНК: [G+C] в 5′-НТП, длина 5′-НТП, A/G^{-3}, IC [–15, 1], IC [1, 20], [uAUG], G^{+4}, описанные в литературе как трансляционно-значимые (табл. 1). Коэффициенты линейной дискриминантной функции определяли стандартным образом.

Среди 2000 проанализированных мРНК из выборки Set 1 с высокими и низкими трансляционными эффективностями при проведении дискриминантного анализа удалось правильно классифицировать 74.8 % (табл. 2).

Регрессионный анализ характеристик 5′-НТП мРНК, связанных с функционированием сигнала инициации трансляции. Проведен регрессионный анализ с целью выявить трансляционно-значимые характеристики 5′-НТП мРНК на основании данных рибосомного профилирования (нормированное количество ридов, rpkm) для выборок Set 1, Set 2 и Set 3 (табл. 3). При построении регрессионной функции были отобраны характеристики мРНК, определенные теоретически или в отдельных экспериментах как трансляционно-значимые. Исходный набор характеристик содержал все (см. Материалы и методы) предсказывающие переменные, регрессоры. На каждом последующем регрессионном шаге удаляли наименее значимую характеристику с максимальным p-value.

Table 1. Fisher linear discriminant function coefficients for translationally important characteristics of mRNAs

mRNA characteristics	Fisher coefficient
A/G^{-3}	-5.8×10^{-5}
G^{+4}	-1.4×10^{-4}
5′-UTR length	1.7×10^{-7}
IC [1, 20]	2.0×10^{-4}
IC [–15, 1]	4.4×10^{-4}
[C,G] in 5′-UTR	0.002
[uAUG]	0.075

Table 2. The accuracy of translation efficiency classification based on translationally important mRNA characteristics

mRNA translation efficiency	Percentage of classified mRNAs	
	correctly	incorrectly
High	0.828	0.172
Low	0.668	0.332
All mRNAs	0.748	0.252

В результате окончательная функция регрессии для выборки Set 1 содержала 14 наиболее значимых характеристик. Коэффициент корреляции между наблюдаемыми и предсказанными значениями достиг максимального значения 0.377. Для предсказания значений из выборки Set 2 было отобрано 19 характеристик мРНК, для Set 3 – 17. Коэффициенты корреляции между наблюдаемыми и предсказанными значениями эффективности трансляции достигли максимальных значений – 0.382 и 0.305 для Set 2 и Set 3 соответственно.

Использование функции линейной регрессии $y = \log(R1)$, где R1 – количество нормированных ридов, rpkm, после обработки харрингтонином, позволяет на основании экспериментальных данных по эффективности трансляции RiboSeq и по теоретически выявленным статистически значимым переменным предсказывать эффективность трансляции мРНК. На базе платформы BioUML мы разработали набор инструментов, позволяющий предсказывать эффективность трансляции по данным нуклеотидной последовательности мРНК (http://micro.biouml.org/bioumlweb/#de=analyses/Methods/Binding%20regions/Ribo-Seq%20and%20mRNA%20features). Доступ к данному набору осуществляется по запросу к авторам статьи.

Анализ частот олигонуклеотидов в районах, прилежащих к стартовому кодону. При формировании набора трансляционно-значимых характеристик мРНК для регрессионного анализа был проведен предварительный

Table 3. Regression coefficients for mRNA characteristics: Set 1, Set 2, Set 3

mRNA characteristics, Set 1	Regression coefficient	Z-score	*p*-value*
	Set 1		
[uAUG]	–18.7745	12.4538	4.8×10^{-15}
5′-UTR length	–0.0061	9.9547	6.4×10^{-15}
mRNA length	7.0×10^{-8}	10.6784	1.5×10^{-13}
AAAAAG [3, 103]	0.1036	6.1414	4.1×10^{-10}
[G + C] in 5′ UTR	–0.6468	5.7385	5.2×10^{-10}
CCGCCA [–100, –1]	0.2425	6.0110	5.9×10^{-10}
IC [–15, 1]	–0.2041	4.2438	1.1×10^{-6}
CCCGCC[–100, –1]	–0.1341	4.1337	1.8×10^{-6}
A/G^{-3}	0.1320	4.0518	3.0×10^{-6}
[G + C] in mRNA	1.2758	5.3203	5.5×10^{-6}
3′-UTR length	4.0×10^{-6}	3.5822	1.7×10^{-5}
AAGCAG [3, 103]	0.0592	3.9310	5.0×10^{-5}
AAGAAA [3, 103]	0.1008	2.6775	0.0037
AAGAAG [3, 103]	0.0321	2.0705	0.0192
	Set 2		
[G + C] in 5′ UTR	1.5874	45.4915	6.7×10^{-18}
[G + C] in mRNA	–2.0886	43.6240	5.4×10^{-18}
mRNA length	6.2×10^{-7}	27.5748	8.8×10^{-17}
3′-UTR length	-5.6×10^{-6}	15.8597	9.1×10^{-17}
CGCCGC [–100, –1]	0.1456	11.1288	8.3×10^{-15}
G^{+4}	0.0491	7.0671	8.0×10^{-13}
CAAGAA [3, 103]	0.1017	6.7531	7.3×10^{-12}
AAAAAG [3, 103]	0.1036	6.1414	4.1×10^{-10}
A/G^{-3}	0.0429	5.0215	2.5×10^{-7}
GAAGAA [3, 103]	0.0659	4.7225	1.1×10^{-6}
AGAAGC [3, 103]	0.0664	4.6729	1.4×10^{-6}
CTCCGC [–100, –1]	0.0724	4.4143	5.0×10^{-6}
CCGCCA [–100, –1]	0.0717	3.9105	6.0×10^{-6}
AAGCAG [3, 103]	0,0592	3.9310	5.0×10^{-5}
5′-UTR length	6.3×10^{-7}	3.1167	9.0×10^{-5}
[uAUG]	–1.0117	–3.2702	0.0005
IC [–15, 1]	–0.0638	–2.9720	0.0015
AAGAAG [3, 103]	0.0491	2.0705	0.0142
AGAAGA [3, 103]	0.0253	1.8736	0.0310
	Set 3		
mRNA length	-9.97×10^{-6}	24.3408	$< 10^{-20}$
[uAUG]	–8.9533	10.3665	$< 10^{-20}$
A/G^{-3}	0.1905	10.1875	$< 10^{-18}$
3′-UTR length	5.1×10^{-6}	8.1743	2.2×10^{-16}
CCGCCA [–100, –1]	0.1569	5.4900	2.0×10^{-9}
IC [–15, 1]	–0.1584	4.6798	1.5×10^{-7}
[G + C] in 5′ UTR	–0.2258	4.3418	7.0×10^{-7}
AAGCAG [3, 103]	0.0903	3.3412	4.2×10^{-5}
[G + C] in 3′ UTR	0.3403	3.2927	4.5×10^{-5}
CTCCGC [–100, –1]	0.0788	3.2086	6.0×10^{-5}

End of Table 3

mRNA characteristics, Set 1	Regression coefficient	Z-score	*p*-value*
CGCCGC [–100, –1]	0.0559	2.9725	0.0014
CAAGAA [3, 103]	0.0829	2.9264	0.0017
AAGAAC [3, 103]	0.0914	2.8492	0.0023
5′-UTR length	-7.0×10^{-8}	2.6223	0.0044
AAGAAG [3, 103]	0.0524	2.3658	0.0090
CCAAGA [3, 103]	0.0527	1.7994	0.0360
[G+C] in mRNA	–0.4714	3.2485	0.5825

*Regressors are arranged in the descending order of statistical significance, *p*-value.

Table 4. Hexanucleotides tending to occur in [–100, –1] and [+3, +103] mRNA regions characterized by high efficiency of translation in samples Set1, Set2, and Set3; $F_{high}/F_{low}>1$

Hexanucleotide	Set1			Set2			Set3		
	F_{low}	F_{high}	F_{high}/F_{low}	F_{low}	F_{high}	F_{high}/F_{low}	F_{low}	F_{high}	F_{high}/F_{low}
[–100, –1] 5′-UTR									
GCCGCC	0.070	0.187	2.671	0.139	0.253	1.820	0.063	0.169	2.68
CGCCGC	0.080	0.161	2.013	0.129	0.240	1.860	0.051	0.148	2.90
CCGCCA	0.046	0.126	2.739	0.040	0.177	4.425	0.037	0.105	2.838
CCGCCG	0.063	0.141	2.238	0.127	0.224	1.764	0.050	0.136	2.720
CCGCGC	0.048	0.110	2.292	0.110	0.148	1.345	0.048	0.110	2.292
CCCGCC	0.083	0.123	1.482	0.117	0.190	1.624	0.067	0.119	1.776
CTCCGC	0.048	0.084	1.750	0.057	0.118	2.070	0.042	0.079	1.881
CGCGCC	0.055	0.094	1.709	0.092	0.150	1.630	0.046	0.098	2.130
[+3, +103] CDS									
AAGAAA	0.031	0.060	1.935	0.046	0.059	1.283	0.027	0.061	2.259
AAAAAG	0.060	0.097	1.617	0.066	0.095	1.439	0.058	0.091	1.569
AAGAAC	0.037	0.054	1.459	0.024	0.071	2.958	0.031	0.050	1.613
AAGAAG	0.071	0.134	1.887	0.090	0.161	1.789	0.073	0.134	1.836
AAGCAG	0.053	0.068	1.283	0.054	0.100	1.852	0.054	0.072	1.333
AGAAGA	0.076	0.105	1.382	0.095	0.127	1.337	0.081	0.115	1.420
AGAAGC	0.056	0.095	1.696	0.082	0.102	1.244	0.056	0.090	1.607
CAAGAA	0.050	0.097	1.940	0.053	0.130	2.453	0.053	0.100	1.887

анализ частот гексануклеотидов в районах, прилежащих к стартовому кодону.

мРНК выборок Set1, Set2 и Set3 были ранжированы по значению нормированного количества ридов (rpkm) и разбиты на группы с высокой и низкой эффективностью трансляции. Был проведен сравнительный анализ частот гексануклеотидов в районах [–100, –1] 5′-НТП и [3, 100] БКП. Для дальнейшего анализа были отобраны гексануклеотиды, частоты которых значимо выше в последовательностях мРНК с высокой эффективностью трансляции (F_{high}) относительно аналогичных частот в последовательностях мРНК, с низкой эффективностью трансляции (F_{low}) (табл. 4). Данные гексануклеотиды, таким образом, являются характеристиками, дискриминирующими мРНК с высокой и низкой эффективностью трансляции. Затем анализ проводился только для гексануклеотидов, чаще встречающихся в последовательностях мРНК, характеризующихся высокой эффективностью трансляции, т. е. $F_{high}/F_{low}>1$.

Обсуждение

Предсказание эффективности трансляции на основании трансляционно-значимых характеристик с помощью дискриминантного анализа. Проведен предварительный дискриминантный анализ с характеристиками мРНК: [C+G] в 5′ НТП, длина 5′ НТП, A/G^{-3}, IC [–15, 1], IC [1, 20], [uAUG], G^{+4}, описанными в литературе как трансляционно-значимые. С использованием функции

Фишера (см. табл. 1) нам удалось с точностью 74.8 % классифицировать 2000 мРНК из выборки Set 1 как характеризующиеся высокой и низкой эффективностью трансляции (см. табл. 2).

Далее с помощью регрессионного анализа мы решили выявить наиболее трансляционно-значимые характеристики мРНК, используя экспериментальные данные рибосомного профилирования как параметры эффективности трансляции.

На основании результатов проведенного регрессионного анализа мы сделали ряд выводов.

1. Полученные с помощью регрессионного анализа результаты не однозначны для трех проанализированных выборок. Тенденции влияния отдельных трансляционно-значимых параметров разнонаправлены. Возможно, это связано с погрешностями эксперимента, различными условиями и особенностями регуляции трансляции для разных типов клеток.

2. Тем не менее можно выделить трансляционно-значимые параметры мРНК, тенденция влияния на эффективность трансляции которых наиболее значима и одинакова для всех трех выборок:

A/G^{-3}, [uAUG], IC [–15, 1] снижают эффективность трансляции;

CCGCCA [–100, –1], AAGAAA [3, 103], AAGAAG [3, 103], AAGCAG [3, 103], AAAAAG [3, 103] усиливают эффективность трансляции.

Обнаруженные закономерности хорошо укладываются в общую схему инициации трансляции в рамках модели «линейного сканирования» (Kozak, 2005) и согласуются с полученными теоретически данными (Volkova, Kochetov, 2010).

3. Регрессионный анализ позволяет выявлять наиболее трансляционно-значимые параметры мРНК и с использованием их предсказывать эффективность трансляции.

4. Данные рибосомного профилирования позволяют выявлять трансляционно-значимые параметры мРНК и использовать полученные данные для предсказания эффективности трансляции мРНК и при клонировании мРНК эукариот методами генной инженерии.

Следует отметить, что гексануклеотиды, отобранные как значимо коррелирующие с эффективностью трансляции на левых флангах стартовых кодонов, существенно отличаются по своему нуклеотидному составу от олигонуклеотидов, отобранных на правых флангах мРНК, характеризующихся высокой эффективностью трансляции, и часто содержат в 5′-НТП GC-богатые гексануклеотиды с гуанином в каждой третьей позиции, в БКП – A-, G- и C-богатые (U избегается) гексануклеотиды. Отметим, что гексануклеотиды, часто расположенные в 5′-НТП мРНК, характеризующихся высокой эффективностью трансляции, имеют сходство с оптимальным контекстом млекопитающих GCCRCC*AUG*G (Kozak, 1986, 1987a, b; De Angioletti et al., 2004). Показано, что олигонуклеотид GCCGCCGCC (с G в каждой третьей позиции) усиливает эффективность трансляции на AUG и неAUG кодонах в отсутствии G в +4 позиции (Kozak, 1987a, 1989). По-видимому, высокая частота встречаемости данных гексануклеотидов в 5′-НТП мРНК, характеризующихся высокой эффективностью трансляции, связана с их способностью усиливать эффективность инициации трансляции как стартовых кодонов.

Отметим, что олигонуклеотиды БКП не позиционированы относительно рамки считывания. Представленные комбинации нуклеотидов приводят к образованию триплетов, кодирующих аминокислоты Ala, Asn, Arg, Ser, Gln, Glu и Pro. Ранее было показано, что данные аминокислоты перепредставлены во 2–4-й позициях белков, и считается, что они способствуют образованию первой пептидной связи (Volkova, Kochetov, 2010). Уридин не встречается в комбинации нуклеотидов в мРНК, характеризующихся высокой эффективностью трансляции. Это позволяет избегать стартового и стоп-кодонов, а также уридина в +4-позиции, который, как показано, уменьшает эффективность трансляции (Kozak, 1997).

Acknowledgments

This work was supported by the Russian Foundation for Basic Research, project 4-04-01284

The authors are grateful to the engineer of the Institute of Computational Technologies SB RAS Ivan Evshin for assistance in data processing.

Conflict of interest

The authors declare no conflict of interest.

References

Brar G.A., Yassour M., Friedman N., Regev A., Ingolia N.T., Weissman J.S. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science. 2012;335:552-557. DOI 10.1126/science.1215110.

De Angioletti M., Lacerra G., Sabato V., Carestia C. Beta+45 G--> C: a novel silent beta-thalassaemia mutation, the first in the Kozak sequence. Br. J. Haematol. 2004;124(2):224-231. DOI 10.1046/j.1365-2141.2003.04754.x.

Dimelow R.J., Wilkinson S.J. Control of translation initiation: a model-based analysis from limited experimental data. J. R. Soc. Interface. 2009;6:51-61. DOI 10.1098/rsif.2008.0221.

Ding Y., Shah P., Plotkin J.B. Weak 5′-mRNA secondary structures in short eukaryotic genes. Genome Biol. Evol. 2012;4(10):1046-1053. DOI 10.1093/gbe/evs082.

Gerashchenko M., Gladyshev V. Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucl. Acids Res. 2014;42(17):e134. DOI 10.1093/nar/gku671.

Gerashchenko M.V., Lobanov A.V., Gladyshev V.N. Genome-wide ribosome profiling reveals complex translational regulatio in response to oxidative stress. Proc. Natl. Acad. Sci. USA. 2012;109:17394-17399. DOI 10.1073/pnas.1120799109.

Godefroy-Colburn T., Thach R.E. The role of mRNA competition in regulating translation. IV. Kinetic model. J. Biol. Chem. 1981;256:11762-11773.

Grünert S., Jackson R.J. The immediate downstream codon strongly influences the efficiency of utilization of eukaryotic translation initiation codons. EMBO J. 1994;13(15):3618-3630.

Harkins S., Cornell C.T., Whitton J.L. Analysis of translational initiation in coxsackievirus B3 suggests an alternative explanation for the high frequency of R+4 in the eukaryotic consensus motif. J. Virol. 2005;79(2):987-996. DOI 10.1128/JVI.79.2.987-996.2005.

Heiman M., Schaefer A., Gong S., Peterson J.D., Day M., Ramsey K.E., Suárez-Fariñas M., Schwarz C., Stephan D.A., Surmeier D.J., Greengard P., Heintz N. A translational profiling approach for the molecular characterization of CNS cell types. Cell. 2008;135(4):738-748. DOI 10.1016/j.cell.2008.10.028.

Heinrich R., Rapoport T.A. Mathematical modelling of translation of mRNA in eukaryotes; steady-states, time dependent processes and

application to reticulocytes. J. Theor. Biol. 1980;86:279-313. DOI 10.5936/csbj.201204002.

Hinnebusch A.G., Lorsch J.R. The Mechanism of eukaryotic translation initiation: new insights and challenges. Cold Spring Harb. Perspect. Biol. 2012;4:a011544. DOI 10.1101/cshperspect.a011544.

Ingolia N.T., Brar G.A., Rouskin S., McGeachy A.M., Weissman J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. 2012;7(8):1534-1550. DOI 10.1038/nprot.2012.086.

Ingolia N.T., Lareau L.F., Weissman J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147(4):789-802. DOI 10.1016/j.cell.2011.10.002.

Johannes G., Carter M.S., Eisen M.B., Brown P.O., Sarnow P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. Proc. Natl. Acad. Sci. USA. 1999;96(23):13118-13123. DOI 10.1073/pnas.96.23.13118.

Kochetov A.V. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. BioEssays. 2008;30(7):683-691. DOI 10.1002/bies.20771.

Kochetov A.V., Palyanov A., Titov I.I., Grigorovich D., Sarai A., Kolchanov N.A. AUG_hairpin: prediction of a downstream secondary structure influencing the recognition of a translation start site. BMC Bioinformatics. 2007;8(1):318. DOI 10.1186/1471-2105-8-318.

Kozak M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res. 1981;9(20):5233-5252. DOI 10.1093/nar/9.20.5233.

Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986;44:283-292. DOI 10.1038/308241a0.

Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 1987a;15(20):8125-8148. DOI 10.1093/nar/15.20.8125.

Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J. Mol. Biol. 1987b;196:947-950. DOI 10.1016/0022-2836(87)90418-9.

Kozak M. Context effects and inefficient initiation at non-AUG codons in eukaryotic cell-free translation systems. Mol. Cell. Biol. 1989;9(11):5073-5080. DOI 10.1128/MCB.9.11.5073.

Kozak M. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and + 6. EMBO J. 1997;16(9):2482-2492. DOI 10.1093/emboj/16.9.2482.

Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 2005;361:13-37. DOI 10.1016/j.gene.2005.06.037.

Lee S., Liu B., Lee S., Huang S.X., Shen B., Qian S.B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. USA. 2012;109(37):E2424-E2432. DOI 10.1073/pnas.1207846109.

Li F., Zheng Q., Vandivier L.E., Willmann M.R., Chen Y., Gregory B.D. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell. 2012;24:4346-4359. DOI 10.1105/tpc.112.104232.

Liu W., Zhao Y., Cui P., Lin Q., Ding F., Xin C., Tan X., Song S., Yu J., Hu S. Thousands of novel transcripts identified in mouse cerebrum, testis, and ES cells based on ribo-minus RNA sequencing. Front. Genet. 2011;2:93. DOI 10.3389/fgene.2011.00093.

Michel A.M., Choudhury K.R., Firth A.E., Ingolia N.T., Atkins J.F., Baranov P.V. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012;22:2219-2229. DOI 10.1101/gr.133249.111.

Na D., Lee S., Lee D. Mathematical modeling of translation initiation for the estimation of its efficiency to computationally design mRNA sequences with desired expression levels in prokaryotes. BMC Syst. Biol. 2010;4:71. DOI 10.1186/1752-0509-4-71.

Nakagawa S., Niimura Y., Gojobori T., Tanaka H., Miura K. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res. 2008;36(3):861-871. DOI 10.1093/nar/gkm1102.

Niimura Y., Terabe M., Gojobori T., Miura K. Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. Nucleic Acids Res. 2003;31(17):5195-5201. DOI 10.1093/nar/gkg701.

Pestova T.V., Kolupaeva V.G. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. Genes Dev. 2002;16:2906-2922. DOI 10.1101/gad.1020902.

Pisarev A.V., Kolupaeva V.G., Pisareva V.P., Merrick W.C., Hellen C.U.T., Pestova T.V. Specific functional interactions of nucleotides at key –3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. Genes Dev. 2006;20:624-636. DOI 10.1101/gad.1397906.

Reid D.W., Nicchitta C.V. Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. J. Biol. Chem. 2012;287:5518-5527. DOI 10.1074/jbc.M111.312280.

Rogozin I.B., Kochetov A.V., Kondrashov F.A., Koonin E.V., Milanesi L. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics. 2001;17(10):890-900. DOI 10.1093/bioinformatics/17.10.890.

Sanz E., Yang L., Su T., Morris D.R., McKnight G.S., Amieux P.S. Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. Proc. Natl. Acad. Sci. USA. 2009;106(33):13939-13944. DOI 10.1073/pnas.0907143106.

Schwanhäusser B., Busse D., Li N., Dittmar G., Schuchhardt J., Wolf J., Chen W., Selbach M. Global quantification of mammalian gene expression control. Nature. 2011;473(7347):337-342. DOI 10.1038/nature10098.

Simon Z., Ruckenstein E. Regulation and synthesis processes in the living cell II. Kinetics of protein synthesis. J. Theor. Biol. 1966;11:299-313. DOI 10.1016/0022-5193(66)90167-6.

Singh C.R., Watanabe R., Chowdhury W., Hiraishi H., Murai M.J., Yamamoto Y., Miles D., Ikeda Y., Asano M., Asano K. Sequential eukaryotic translation initiation factor 5 (eif5) binding to the charged disordered segments of eif4g and eif2β stabilizes the 48S preinitiation complex and promotes its shift to the initiation mode. Mol. Cell Biol. 2012;32(19):3978-3989. DOI 10.1128/MCB.00376-12.

Siwiak M., Zielenkiewicz P.A. A comprehensive, quantitative, and genome-wide model of translation. PLoS Comp. Biol. 2010;6(7):e1000865. DOI 10.1371/journal.pcbi.1000865.

Takacs J.E., Neary T.B., Ingolia N.T., Saini A.K., Martin-Marcos P., Pelletier J., Hinnebusch A.G., Lorsch J.R. Identification of compounds that decrease the fidelity of start codon recognition by the eukaryotic translational machinery. RNA. 2011;17:439-452. DOI 10.1261/rna.2475211.

Volkova O.A., Kochetov A.V. Interrelations between the nucleotide context of human start AUG codon, N-end amino acids of the encoded protein and initiation of translation. J. Biomol. Struct. Dyn. 2010;27(5):611-618. DOI 10.1080/07391102.2010.10508575.

Weiss R.B., Atkins J.F. Translation goes global. Science. 2011;334(6062):1509-1510. DOI 10.1126/science.1216974.

Xia X., Huang H., Carullo M., Betran E., Moriyama E.N. Conflict between translation initiation and elongation in vertebrate mitochondrial genomes. PLoS ONE. 2007;2:e227. DOI 10.1371/journal.pone.0000227.

Zhu L.H., Xu J.X., Zhu S.W., Cai X., Yang S.F., Chen X.L., Guo Q. Gene expression profiling analysis reveals weaning-induced cell cycle arrest and apoptosis in the small intestine of pigs. J. Anim. Sci. 2014;92:996-1006. DOI 10.2527/jas.2013-7551.

Zong Q., Schummer M., Hood L., Morris D.R. Messenger RNA translation state: the second dimension of high-throughput expression screening. Proc. Natl. Acad. Sci. USA. 1999;96(19):10632-10636. DOI 10.1073/pnas.96.19.10632.

Zouridis H., Hatzimanikatis V. A model for protein translation: polyzome self-organization leads to maximum protein synthesis rates. Biophys. J. 2007;92:717-730. DOI 10.1529/biophysj.106.087825.