УДК 577.217.53:577.322.52:004.738

ELOE – ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ОЦЕНКИ ЭФФЕКТИВНОСТИ ЭЛОНГАЦИИ ТРАНСЛЯЦИИ ГЕНОВ

© 2014 г. **В.С. Соколов[1], Б.С. Зураев[1, 2], С.А. Лашин[1, 2], Ю.Г. Матушкин[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия,
e-mail: sokovlad1@bionet.nsc.ru;
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

Поступила в редакцию 18 сентября 2014 г. Принята к публикации 8 октября 2014 г.

Многие современные исследования изучают важную характеристику гена – эффективность его экспрессии. Как известно, она определяется на уровнях транскрипции, трансляции, посттрансляционной модификации и др. В работе представлена программа EloE (Elongation Efficiency), сортирующая гены организма в порядке уменьшения их предполагаемой скорости элонгации трансляции на основе анализа их нуклеотидных последовательностей. Полученные теоретические данные достоверно коррелируют с доступными экспериментальными данными по экспрессии генов различных организмов, например *S. cerevisiae* и *H. pylori*. Также программа выявляет предпочтительные кодоны в геноме организма и строит распределение стабильности потенциальных вторичных структур в районах 5′- и 3′-концов мРНК. Программа может быть использована для предварительной оценки уровня экспрессии генов исследуемого организма, экспериментальные данные для которого еще не доступны. Результаты работы EloE могут быть переданы в сторонние программные инструменты, которые моделируют искусственные генетические конструкции для генно-инженерных экспериментов.

Ключевые слова: индекс эффективности элонгации, веб-приложение, эффективность трансляции, вторичные структуры.

ВВЕДЕНИЕ

Эффективность экспрессии генов организма определяется на многих уровнях, важнейшими из которых являются транскрипция, трансляция, посттрансляционная модификация. Изучение факторов, регулирующих трансляцию, – актуальная задача современной биологии. Результаты ее решения могут быть использованы, например, при создании генно-инженерных конструкций и принесут большую пользу в таких областях, как медицина и сельское хозяйство.

В работах (Thanaraj, Argos, 1996; Lopinski *et al.*, 2000; Takyar *et al.*, 2005; Eck, Stephan, 2008) показано, что у большинства прокариотических, а также многих эукариотических видов уровень экспрессии генов зависит от их кодонного состава и от наличия и стабильности вторичных структур в мРНК. Эти факторы влияют на скорость движения рибосомного комплекса по мРНК в процессе трансляции и тем самым – на скорость синтеза белка. На данный момент разработано множество различных индексов для выявления предпочтительных кодонов (Ikemura, 1981; Bennetzen, Hall, 1982; McLachlan *et al.*, 1984). Индекс адаптации кодонов (CAI) – один из первых в этом ряду (Sharp, Li, 1986). Также существует большое количество программ для оценки насыщенности мРНК вторичными структурами (Zuker *et al.*, 1999; Hofacker, 2003; Zuker, 2003).

В статье Н.В. Владимирова с соавт. (Vladimirov *et al.*, 2007) описано пять типов эволюционной оптимизации первичной структуры генов, основанных на факторах, влияющих на эффективность экспрессии генов на уровне трансляции. К этим факторам относятся: час-

тоты кодонов в гене, наличие и распределение вторичных структур в мРНК и стабильность этих структур. Разные комбинации названных параметров формируют пять типов эволюционной оптимизации, которые учитывают:

1) только кодонный состав мРНК;
2) только количество локальных инвертированных повторов в мРНК;
3) только энергетическую стабильность потенциальных шпилек в мРНК;
4) кодонный состав и количество локальных инвертированных повторов в мРНК;
5) кодонный состав и энергетическую стабильность потенциальных шпилек в мРНК.

Индекс эффективности элонгации (ИЭЭ, EEI – Elongation Efficiency Index), предложенный в статье В.А. Лихошвая и Ю.Г. Матушкина (2000), позволяет классифицировать большинство прокариот и некоторых эукариот (в основном одноклеточных, например дрожжей) по этим пяти типам оптимизации первичной структуры генов. Данный индекс оценивает предполагаемую эффективность прохождения стадии элонгации трансляции для каждого гена организма. Поскольку элонгация является одной из наиболее энерго- и времязатратных стадий трансляции, на основе индекса ИЭЭ можно сделать предположения об эффективности трансляции в целом (Там же).

Наша работа посвящена исследованию связи контекстных характеристик генов с их эффективностью трансляции. В представленной программе для исследования эффективности элонгации трансляции был выбран именно ИЭЭ (EEI) (Лихошвай, Матушкин, 2000; Likhoshvai, Matushkin, 2002), поскольку он позволяет учитывать в расчетах как кодонный состав гена, так и его насыщенность локальными инвертированными повторами (потенциальными шпильками в составе вторичной структуры мРНК).

Данный индекс позволяет ранжировать по эффективности элонгации трансляции гены даже тех организмов, для которых другие индексы, основанные на учете частот использования кодонов, не работают (Лихошвай, Матушкин, 2000). Также ранее было показано, что ИЭЭ коррелирует с другими параметрами, оценивающими эффективность экспрессии генов, в частности с плотностью нуклеосомной упаковки в промоторном районе генов дрожжей (Vladimirov *et al.*, 2007; Матушкин и др., 2013). Для массового анализа геномов различных организмов была необходимость в создании общего программного интерфейса с возможностью изменения параметров расчетов, доступного в сети интернет и позволяющего производить анализ сразу нескольких (до нескольких тысяч) геномов за один запуск. Такая задача была решена в форме специального веб-приложения.

РЕЗУЛЬТАТЫ

Для классификации видов по пяти типам эволюционной оптимизации первичной структуры их генов и оценки их эффективности элонгации трансляции создано web-приложение EloE, доступное по адресу http://www-bionet.sscc.ru:7780/EloE. Вид интерфейса представлен на рис. 1.

Исходные данные составляют файлы с аннотированной нуклеотидной последовательностью полного генома в формате gbk (данные могут быть взяты в базе GenBank ftp://ftp.ncbi.nih.gov/genbank/genomes). Для произведения расчетов требуется создание zip-архива с аннотированными геномами (gbk) исследуемых организмов. Геном каждого организма должен располагаться в архиве в отдельной папке. Архив загружается в программу с помощью кнопки Upload. Все результаты, в том числе список генов организма, отсортированных по индексу ИЭЭ, сохраняются в файлы и могут быть загружены после окончания расчетов (кнопка Download results).

Основные файлы с результатами располагаются для каждого организма в отдельной папке Organism_name:

1) organism_name_all.txt – файл со всеми пятью типами индекса ИЭЭ, рассчитанными для всех генов организма, учитываемых в расчетах;
2) organism_name_eeiN.txt (N = {1, 2, 3, 4, 5}) – файл только с тем типом индекса ИЭЭ, который работает в данном организме;
3) organism_name_genes_and_flanks.txt – файл с подробной информацией по каждому гену и его нуклеотидной последовательностью с флангами;
4) organism_name_number_eei.txt – файл с указанием номера гена, его позиции в опероне (только для прокариот) и значения ИЭЭ;

5) organism_name_ribpos.txt – файл с расположением генов рибосомных белков в списках генов организма, отсортированных по каждому из пяти типов ИЭЭ в порядке увеличения (рис. 2 и 3).

Общие результаты по всем анализируемым геномам собраны в одном файле organism_index.txt. Данные во всех файлах разделены знаком табуляции.

Интерфейс программы позволяет изменять параметры расчетов: размеры фланкирующих районов генов, длины инвертированных повторов в мРНК и расстояние между мономерами этих повторов. В начале/конце первого/последнего кодирующего экзона обычно расположены специфические кодоны, характерные именно для сайтов начала/конца трансляции. Поэтому их учет может негативно повлиять на расчеты ИЭЭ. В программе можно указать количество кодонов, которые не будут учтены в расчетах, или поставить галочку Use auto calculation of flanks' length. Тогда программа сама определит оптимальное количество неучитываемых кодонов. Также можно заказать дополнительные выходные файлы.

Одной из возможностей программы является генерация файла organism_name_lciij_profile_out.xlsx (при установленной галочке Calculate Local Complementarity Index for individual nucleotides). В нем хранятся значения для построения профилей индексов локальной комплементарности (ИЛК, LCI – Local Complementarity Index), которые строятся в web-приложении. Индекс локальной комплементарности имеет смысл среднего количества локальных совершенных инвертированных повторов определенной длины в мРНК. Такие повторы потенциально могут образовывать шпильки в составе вторичной структуры мРНК и замедлять движение рибосомного комплекса в процессе элонгации трансляции (Lopinski *et al.*, 2000; Takyar *et al.*, 2005). Индекс локальной комплементарности индивидуального нуклеотида показывает среднюю стабильность шпилек, в образовании которых может принимать участие данный нуклеотид. Индекс рассчитывается в районах старт-кодона (± 600 нуклеотидов) и стоп-кодона (± 600 нуклеотидов) трансляции. Для этого вместе с последовательностью гена из файла gbk экстрагируются фланкирующие районы длиной 600 нуклеотидов. Изменение длины экстрагируемых флангов не влияет на расчеты индексов ИЭЭ.

Для каждого организма EloE строит график с позициями генов рибосомных белков для каждого типа ИЭЭ (рис. 2 и 3). Гены на графике отсор-

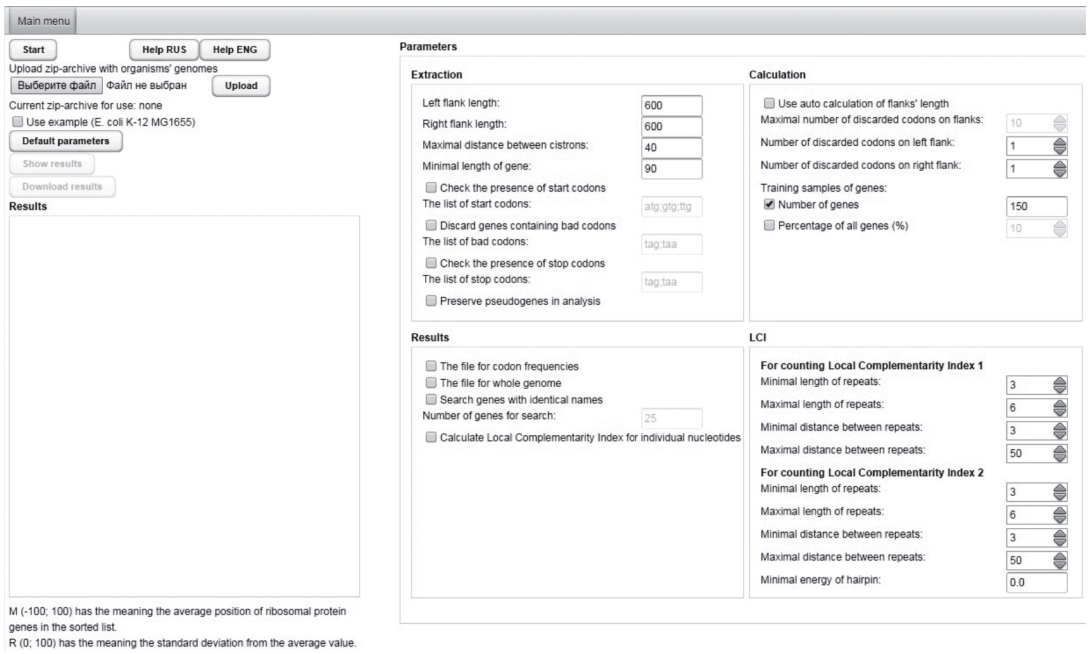

Рис. 1. Интерфейс web-приложения EloE.

тированы в порядке увеличения ИЭЭ. Таким образом, наилучший тип ИЭЭ для организма – это такой тип, для которого гены рибосомных белков расположены правее и плотнее (рис. 2 и 3). Как видно из рис. 2, в *E. coli* лучше всего работает первый тип ИЭЭ, т. е. эффективность элонгации в большей степени зависит от частот кодонов в гене. У *Mycoplasma fermentans* JER (рис. 3) лучше работает второй тип ИЭЭ, т. е. эффективность элонгации в основном определяется количеством инвертированных повторов в гене.

К особенностям программы EloE относятся:

1) возможность обработки более одного генома (до нескольких тысяч) за один запуск;

2) расчет дополнительных параметров и их визуализация, например ИЛК индивидуальных нуклеотидов (LCI) (рис. 4).

Приведенный на рис. 4 профиль ИЛК индивидуальных нуклеотидов отображает среднюю по всем генам организма стабильность потенциальных вторичных структур в районе старт- и стоп-кодона трансляции. Спад профиля в районе старт-кодона (рис. 4, *а*) говорит о потенциально меньшей стабильности шпильки в этом районе мРНК. С другой стороны, пик профиля в районе стоп-кодона (рис. 4, *б*) указывает на повышенную стабильность шпильки.

Выходные файлы содержат такие параметры генов, как: индекс ИЭЭ, индекс ИЛК, GC-состав, длина, позиция в опероне (для прокариот) и др. Дополнительной функцией программы является построение усредненных профилей стабильности вторичных структур в районах 5′- и 3′-концов мРНК всех генов организма. Для прокариот можно выбирать, по каким генам строить профиль: по всем или только по первым, средним, последним или единственным цистронам в операх.

При помощи программы EloE было проведено исследование геномов 62 штаммов *Mycoplasma* (Sokolov *et al.*, 2014). У группы микоплазм (*C.M. haemolamae*, *C.M. haemominutum*,

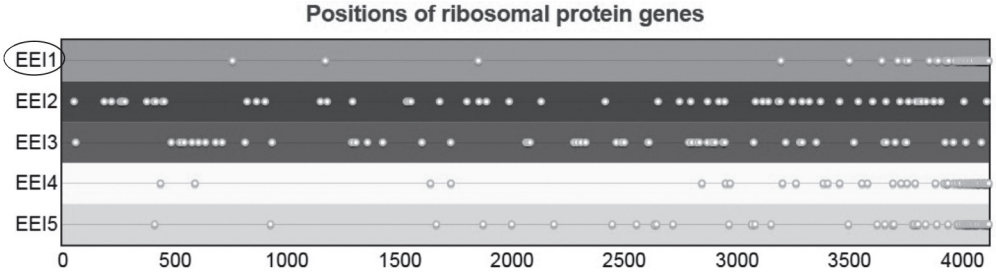

Рис. 2. Распределение генов рибосомных белков (точки) в списке генов *E. coli* K-12 MG1655, расположенных слева направо в порядке увеличения EEI1-5. Наилучший тип индекса (EEI1) выделен кружком – гены рибосомных белков расположены правее и плотнее.

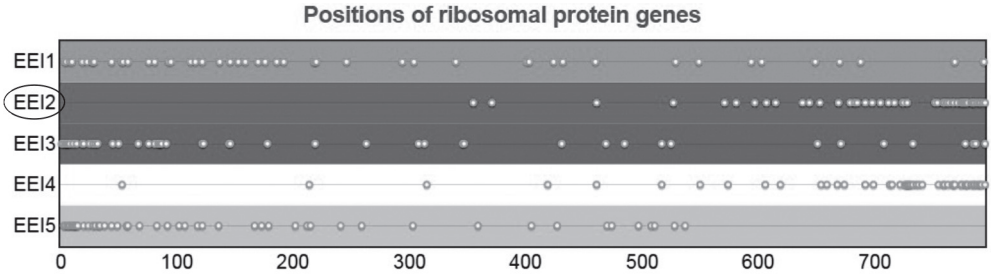

Рис. 3. Распределение генов рибосомных белков (точки) в списке генов *Mycoplasma fermentans* JER, расположенных слева направо в порядке увеличения EEI1-5. Наилучший тип индекса (EEI2) выделен кружком – гены рибосомных белков расположены правее и плотнее.

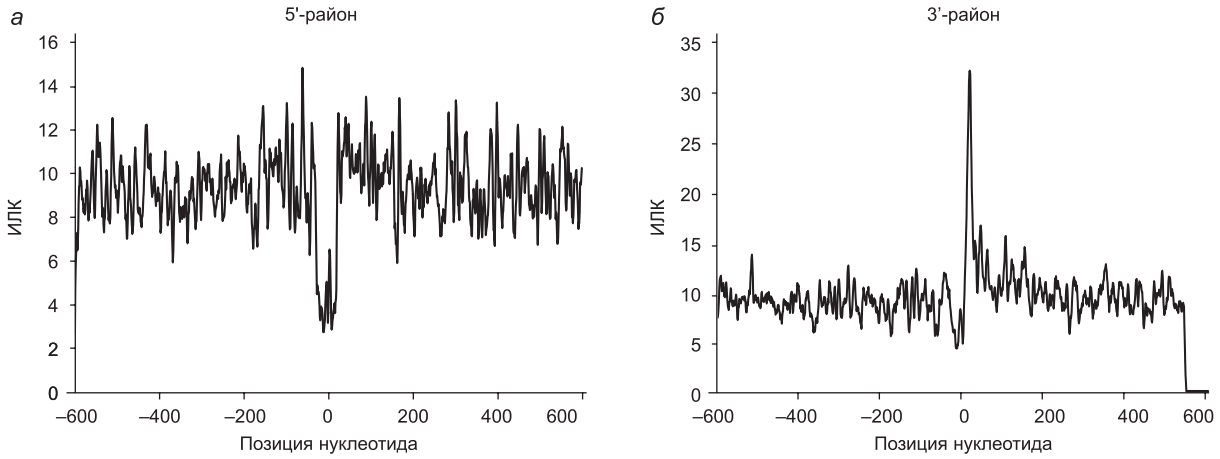

Рис. 4. Визуализация профиля среднего значения ИЛК индивидуальных нуклеотидов по всем генам *Mycoplasma fermentans* JER. Ноль на оси абсцисс на графике *а* – старт-кодон трансляции, на графике *б* – стоп-кодон.

M. haemocanis, *M. haemofelis*, *M. pneumoniae*, *M. suis*) впервые выявлено пониженное содержание в генах локальных инвертированных повторов по сравнению с другими микоплазмами. Также при построении профилей распределения локальных инвертированных повторов в районах старт- и стоп-кодонов трансляции у *M. haemofelis* обнаружен не характерный для остальных микоплазм пик в районе старт-кодона.

ЗАКЛЮЧЕНИЕ

Эффективность экспрессии гена – одна из его главнейших характеристик. Программа EloE позволяет ранжировать гены организма по вычисляемой эффективности одной из важнейших стадий трансляции – элонгации. Учет одновременно кодонного состава и локальных инвертированных повторов в мРНК позволяет программе EloE анализировать более широкий класс организмов, для которых учета только данных по частотам использования кодонов недостаточно. Дополнительные результаты, такие как график распределения стабильности вторичных структур вблизи флангов генов, позволяют более детально исследовать особенности нуклеотидных последовательностей.

Эти данные могут быть полезны во многих областях современных исследований. Особенно это важно при изучении организмов, для которых еще не получены экспериментальные данные по экспрессии их генов. Программа находится в открытом доступе по адресу http://www-bionet.sscc.ru:7780/EloE.

БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке гранта РФФИ № 13-04-00620, государственного контракта № 1/223-114 и бюджетного проекта № VI.61.1.2.

ЛИТЕРАТУРА

Лихошвай В.А., Матушкин Ю.Г. Предсказание эффективности экспрессии генов по их нуклеотидному составу // Молекулярная биология. 2000. Т. 34. № 3. С. 406–412.

Матушкин Ю.Г. и др. Эффективность элонгации генов дрожжей кореллирует с плотностью нуклеосомной упаковки в 5′-нетранслируемом районе // Математическая биология и биоинформатика. 2013. Т. 8. № 1. С. 248–257.

Bennetzen J.L., Hall B.D. Codon selection in Yeast // J. Biol. Chem. 1982. V. 257. No. 6. P. 3026–3031.

Eck S., Stephan W. Determining the relationship of gene expression and global mRNA stability in *Drosophila melanogaster* and *Escherichia coli* using linear models // Gene. 2008. V. 424. No. 1. P. 102–107.

Hofacker I.L. Vienna RNA secondary structure server // Nucleic acids research. 2003. V. 31. No. 13. P. 3429–3431.

Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli system // J. Molecular Biology. 1981. V. 151. No. 3. P. 389–409.

Likhoshvai V.A., Matushkin Y.G. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy // FEBS letters. 2002. V. 516. No. 1. P. 87–92.

Lopinski J.D., Dinman J.D., Bruenn J.A. Kinetics of ribosomal pausing during programmed–1 translational frameshifting // Mol. Cell. Biol. 2000. V. 20. No. 4. P. 1095–1103.

McLachlan A.D., Staden R., Boswell D.R. A method for measuring the non-random bias of a codon usage table // Nucleic acids research. 1984. V. 12. No. 24. P. 9567–9575.

Sharp P.M., Li W.H. An evolutionary perspective on synonymous codon usage in unicellular organisms // Journal molecular evolution. 1986. V. 24. No. 1-2. P. 28–38.

Sokolov V.S., Likhoshvai V.A., Matushkin Y.G. Gene expression and secondary mRNA structures in different Mycoplasma species // Russian Journal Genetics: Applied Research. 2014. V. 4. No. 3. P. 208–217.

Takyar S., Hickerson R.P., Noller H.F. mRNA helicase activity of the ribosome // Cell. 2005. V. 120. No. 1. P. 49–58.

Thanaraj T.A., Argos P. Ribosome-mediated translational pause and protein domain organization // Protein Science. 1996. V. 5. No. 8. P. 1594–1612.

Vladimirov N.V., Likhoshvai V.A., Matushkin Y.G. Correlation of codon biases and potential secondary structures with mRNA translation efficiency in unicellular organisms // Molecular Biology. 2007. V. 41. No. 5. P. 843–850.

Zuker M. Mfold web server for nucleic acid folding and hybridization prediction // Nucleic acids research. 2003. V. 31. No. 13. P. 3406–3415.

Zuker M., Mathews D.H., Turner D.H. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide // RNA biochemistry and biotechnology. Springer Netherlands, 1999. P. 11–43.

ELOE: A WEB APPLICATION FOR ESTIMATION OF GENE TRANSLATION ELONGATION EFFICIENCY

V.S. Sokolov[1], B.S. Zuraev[1, 2], S.A. Lashin[1, 2], Yu.G. Matushkin[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: sokovlad1@bionet.nsc.ru;
[2] Novosibirsk National Research State University, Novosibirsk, Russia

Summary

Expression efficiency is one of major characteristics of genes considered in a number of modern investigations. It is known that gene expression efficiency in an organism is regulated at many stages: transcription, translation, posttranslational protein modification, and others. In this study, a special EloE (Elongation Efficiency) web application is described. It sorts genes in an organism in the order of decreasing theoretical rate of the elongation stage of translation deduced from their nucleotide sequences. The predictions done in this way show a significant correlation with available experimental data on gene expression in various organisms, for instance, *S. cerevisiae* and *H. pylori*. In addition, the program identifies preferential codons in a genome and defines the distribution of stability of potential secondary structures in 5′ and 3′ regions of mRNA. EloE can be useful in preliminary estimation of translation elongation efficiency of genes in organisms for which experimental data are not available yet. Some results can be used, for instance, in other programs modeling artificial genetic constructs in gene engineering experiments.

Key words: elongation efficiency index; web application; translation efficiency; secondary structures.