На правах рукописи

СОКОЛОВ ВЛАДИМИР СЕРГЕЕВИЧ

КОМПЬЮТЕРНОЕ ИССЛЕДОВАНИЕ КОНТЕКСТНЫХ ХАРАКТЕРИСТИК ОТКРЫТЫХ РАМОК СЧИТЫВАНИЯ, СВЯЗАННЫХ С ЭФФЕКТИВНОСТЬЮ ЭЛОНГАЦИИ ТРАНСЛЯЦИИ, У ОДНОКЛЕТОЧНЫХ ОРГАНИЗМОВ

03.01.09. Математическая биология, биоинформатика

Автореферат
диссертации на соискание ученой степени
кандидата биологических наук

Новосибирск

2015

Работа выполнена в лаборатории молекулярно-генетических систем Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики СО РАН», г. Новосибирск, Россия.

Научный руководитель: кандидат биологических наук, старший научный сотрудник **Матушкин Юрий Георгиевич**

Официальные оппоненты **Бажан Сергей Иванович**, доктор биологических наук, доцент, заведующий теоретическим отделом, ФБУН Государственный научный центр вирусологии и биотехнологии «Вектор», р. п. Кольцово, Новосибирская область.

Щербаков Дмитрий Юрьевич, доктор биологических наук, профессор, заведующий лабораторией геносистематики, ФГБУН Лимнологический институт СО РАН, г. Иркутск.

Ведущее учреждение ФГБУН Институт общей генетики им. Н.И. Вавилова РАН, г. Москва.

Защита диссертации состоится «___» _____ 201_ г. на утреннем заседании диссертационного совета Д 003.011.01 по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук в ИЦиГ СО РАН в конференц-зале Института по адресу:

630090, г. Новосибирск, проспект ак. Лаврентьева, 10
тел./факс: (383) 363-49-06; e-mail: dissov@bionet.nsc.ru.
факс: (383) 333-12-78

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института www.bionet.nsc.ru.

Автореферат разослан «___» _____ 2015 г.

Ученый секретарь
диссертационного совета,
доктор биологических наук Т.М. Хлебодарова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Трансляция – это процесс синтеза белка из аминокислот на матрице информационной (матричной) РНК (мРНК), осуществляемый рибосомой. Это очень сложный, многостадийный процесс, в котором принимает участие огромное количество разнообразных молекул. Выделяют три основных стадии трансляции: инициацию, элонгацию и терминацию. Стадия инициация считается лимитирующим звеном трансляции [Kaczanowska and Rydén-Aulin, 2007]. Однако после инициации, элонгация является самой время- и энергозатратной. Время прохождения каждой из стадий вносит свой вклад в суммарное время трансляции. Соответственно, скорость синтеза белка – эффективность трансляции, зависит от эффективности каждой стадии.

На эффективность стадий трансляции оказывают влияние различные факторы. Например, для прокариот показана связь эффективности инициации с наличием в районе старт кодона трансляции определенной последовательности нуклеотидов, последовательности Шайна-Дальгарно (ШД) [Kaczanowska and Rydén-Aulin; 2007]. Другие исследования показали, что эффективность элонгации зависит от кодонного состава открытых рамок считывания (ОРС) [Varenne et al., 1984; Sorensen et al., 1989] и от вторичной структуры мРНК [Takyar et al., 2005; Tuller et al., 2011]. Однако, несмотря на огромное количество исследований, эта область остается недостаточно изученной, и предсказание эффективности трансляции мРНК у многих организмов является актуальной проблемой.

Важность оценки эффективности трансляции связана с таким понятием, как гетерологичная экспрессия [Welch et al., 2009 a, b]. Гетерологичной называется экспрессия чужеродного гена или искусственной генетической конструкции в целевом организме. В настоящее время известны структуры геномов большого числа одноклеточных организмов. Многие из них рассматриваются в качестве кандидатов для использования в биотехнологических процессах и экспериментах. Это часто требует экспрессии различных генетических конструкций в целевых организмах. Для максимизации эффективности гетерологичной экспрессии требуются знания о механизмах и факторах, ее определяющих, в том числе и знания об эффективности трансляции. Поэтому одной из актуальных задач современной биоинформатики является изучение различных характеристик мРНК, влияющих на эффективность трансляции. Кроме этого, сравнительный анализ трансляционно значимых параметров мРНК генов различных организмов ценен сам по себе, так как является источником информации об эволюционных аспектах формирования этих признаков, имеющих как универсальные для всех, так и видоспецифичные особенности.

Исследование контекстных характеристик ОРС, связанных с эффективностью трансляции, актуально как для одноклеточных, так и для многоклеточных организмов. Однако наличие у многоклеточных организмов тканеспецифичной экспрессии [Dittmar et al., 2006] не позволяет с достаточной точностью выявлять у них эти особенности.

В Институте цитологии и генетики был разработан математический индекс эффективности элонгации трансляции EEI (elongatioin efficiency index), позволяющий оценивать эффективность элонгации трансляции генов организма

на основании их нуклеотидного состава [Лихошвай и Матушкин, 2000]. Данный индекс имеет смысл средней скорости движения рибосомы по мРНК в процессе элонгации трансляции. EEI учитывает кодонный состав ОРС и локальные совершенные инвертированные повторы (потенциальные вторичные структуры в мРНК). В зависимости от того, какие из этих факторов являются определяющими при оценке эффективности элонгации трансляции, у исследуемого организма определяется тип эволюционной оптимизации его генома для увеличения эффективности процесса элонгации трансляции генов.

Целью данной работы является: исследование контекстных характеристик открытых рамок считывания, связанных с эффективностью элонгации трансляции, у одноклеточных организмов.

В соответствии с поставленной целью были сформулированы следующие **задачи**:

1) Разработать доступную через Интернет программную реализацию самообучающегося алгоритма расчета индекса эффективности элонгации трансляции EEI;
2) Классифицировать секвенированные геномы одноклеточных организмов по типам эволюционной оптимизации процесса элонгации трансляции;
3) Исследовать связанные с процессом трансляции особенности структурно-функциональной организации открытых рамок считывания у различных одноклеточных организмов;
4) Изучить взаимосвязь между эффективностью инициации транскрипции и эффективностью элонгации трансляции у *S. cerevisiae* и *S. pombe*.

Научная новизна и практическая значимость работы

Разработанное веб-приложение EloE (http://www-bionet.sscc.ru:7780/EloE) позволило впервые провести анализ полных геномов 2771 одноклеточного организма.

В результате анализа организмов, принадлежащих к роду *Mycoplasma*, у группы видов обнаружено сниженное количество локальных инвертированных повторов в генах по сравнению с другими микоплазмами. Филогенетическое исследование *Mycoplasma* позволило установить возможную связь эволюционной оптимизации первичной структуры генов данных организмов с их средой обитания. Также было установлено наличие достоверной отрицательной корреляции между GC-составом генома и степенью эволюционной оптимизации первичной структуры генов для повышения эффективности элонгации трансляции. Показано, что *M. haemofelis*, возможно, обладает отличным от других микоплазм механизмом регуляции процесса инициации трансляции.

Анализ нуклеотидных последовательностей генов и их предковых форм у архей позволил установить, что наиболее сильные изменения в первичной структуре генов, связанные с оптимизацией элонгации трансляции, происходили при радикальной смене среды обитания данных организмов. Также для архей было показано, что температура среды обитания данных организмов не коррелирует с влиянием потенциальных вторичных структур в мРНК на эффективность элонгации трансляции.

При анализе генов дрожжей выявлено наличие корреляции между потенциалом формирования нуклеосом и индексом эффективности элонгации трансляции, что подтверждает предположение о согласованной оптимизации

процессов транскрипции и трансляции. Обнаружено различие между *S. cerevisiae* и *S. pombe* по форме корреляции между потенциалом формирования нуклеосом и индексом эффективности элонгации трансляции для высоко- и низкоэкспрессирующихся генов. Проведен сравнительный анализ геномов этих организмов для выявления причин данного различия.

Результаты данной работы могут быть использованы в генно-инженерных экспериментах для создания искусственных генетических конструкций. Оптимизация первичной структуры нуклеотидных последовательностей позволит увеличить эффективность их трансляции и тем самым повысить уровень их экспрессии в целевых организмах.

Также результаты могут быть полезны при работе с малоизученными организмами, для которых не доступны экспериментальные данные по экспрессии генов. Предсказанные уровни эффективности элонгации трансляции в первом приближении позволяют оценить эффективность экспрессии исследуемых генов.

В теоретическом плане данная работа содержит новую информацию по связанным с эффективностью элонгации трансляции особенностям геномов разнообразных организмов (архей, микоплазм, дрожжей). Эти знания могут послужить основой для проведения новых экспериментов или объяснения особенностей процесса трансляции.

Положения, выносимые на защиту

1) У семи видов *Mycoplasma* (*C. M. haemolamae*, *M. haemocanis*, *M. wenyonii*, *M. haemofelis*, *M. pneumonia*, *C. M. haemominutum*, *M. suis*), в процессе эволюции прошла массовая минимизация количества локальных совершенных инвертированных повторов (потенциальных шпилек) в мРНК.
2) *M. haemofelis* радикально отличается от остальных проанализированных видов микоплазм наличием более стабильных потенциальных вторичных структур в мРНК в районе старт-кодона трансляции, что может быть связано с альтернативным механизмом регуляции инициации трансляции у данного вида.
3) Индекс эффективности элонгации трансляции генов *S. cerevisiae* значимо коррелирует с экспериментально определенной плотностью нуклеосомной упаковки во фланкирующем 5'-районе ДНК выше старта трансляции мРНК.

Апробация результатов

Данная работа была представлена на следующих конференциях: XIII всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям, Новосибирск, 2012; Moscow conference on computational molecular biology, MCCMB'13, Москва, 2013; 5th international young scientists school «Systems biology and bioinformatics», SBB’2013, Новосибирск, 2013; VI съезд Вавиловского общества генетиков и селекционеров (ВОГиС) и ассоциированные генетические симпозиумы, Ростов-на-Дону, 2014 (диплом 3-ей степени); The 9th international conference on bioinformatics of genome regulation and structure\System biology, BGRS\SB'2014, Новосибирск, 2014; 6th international young scientists school «Systems biology and bioinformatics», SBB’2014, Новосибирск, 2014.

Публикации

По материалам диссертации опубликовано 3 работы в журналах, входящих в список ВАК.

Структура и объем работы

Диссертационная работа состоит из списка сокращений, введения, обзора литературы, методов и алгоритмов, результатов и обсуждений, заключения, выводов, списка литературы и приложения. Работа изложена на 163 страницах, содержит 74 рисунка и 9 таблиц. Библиографический указатель литературы включает 196 источников, из них 2 отечественных и 194 зарубежных.

Личный вклад автора

Основные результаты работы получены автором самостоятельно. Разработка веб-приложения EloE проводилась совместно с Б. С. Зураевым (создание веб-интерфейса, подключение программы UNAFold), к.б.н. С. А. Лашиным (консультации по коду программы), д.б.н. В. А. Лихошваем (консультации по алгоритмам программы) и к.б.н. Ю. Г. Матушкиным (консультации по алгоритмам программы). Исследование оптимизации первичной структуры генов архей в процессе эволюции проводилось совместно с к.б.н. К. В. Гунбиным (реконструкция предковых форм генов архей). Исследование взаимосвязи между эффективностью элонгации трансляции генов дрожжей и плотностью их нуклеосомной упаковки в 5'-НТР проводилось совместно с к.б.н. В. Г. Левицким (расчет потенциала формирования нуклеосом), д.б.н. Ю. Л. Орловым (экспериментальные данные по плотности нуклеосомной упаковки), д.б.н. В. А. Лихошваем и к.б.н. Ю. Г. Матушкиным.

МЕТОДЫ И АЛГОРИТМЫ

Геномные последовательности

Целевыми объектами исследования являются одноклеточные организмы. Данный выбор был сделан по нескольким причинам. Во-первых, тканеспецифичная экспрессия, присущая многоклеточным, может скрывать или сглаживать влияние изменений нуклеотидного состава генов на эффективность экспрессии. Во-вторых, одноклеточные характеризуются более высокими значениями эффективной численности популяции и скорости размножения. И поскольку все изменения, происходящие в геноме одноклеточных, напрямую отражаются на приспособленности всего организма, эволюция нуклеотидных последовательностей у них идет быстрее, чем у многоклеточных.

В качестве исходных данных использовались геномные последовательности прокариот и одноклеточных эукариот в gbk формате, скачанные из базы данных NCBI GenBank (ftp://ftp.ncbi.nih.gov/genomes/). В исследование брались только те организмы, у которых геном был полностью секвенирован и аннотирован. Для проведения расчетов была использована база данных GenBank по состоянию на 13 июня 2013 г.

Индекс эффективности элонгации трансляции EEI

Для оценки эффективности элонгации трансляции генов исследуемых организмов в данной работе использовался индекс эффективности элонгации трансляции EEI (Elongation Efficiency Index). EEI был разработан сотрудниками ИЦиГ СО РАН В. А. Лихошваем и Ю. Г. Матушкиным [Лихошвай, Матушкин, 2000]. Данный индекс рассчитывается для каждого гена организма и имеет смысл средней скорости прохождения стадии элонгации трансляции. Выбор в пользу данного индекса был сделан на основании следующих его особенностей.

Во-первых, для расчета EEI кроме аннотированного генома организма не требуются никакие дополнительные данные, например, выборки наиболее

высокоэкспрессируемых генов. В качестве данной выборки используются гены рибосомных белков (ГРБ), которые, как известно, являются одними из высокоэкспрессирующихся в большинстве одноклеточных организмов [Sharp and Li, 1986; Владимиров, 2007].

Во-вторых, данный индекс позволяет учитывать не только кодонный состав гена, как большинство других индексов, но и его насыщенность локальными инвертированными повторами (потенциальными вторичными структурами в мРНК). Это дает возможность применять EEI при исследовании организмов, для которых показано отсутствие корреляции между кодонным составом генов и эффективностью их экспрессии или отсутствие неравномерности по использованию синонимичных кодонов.

В-третьих, данный индекс позволяет показать, какая из характеристик (кодонный состав или потенциальные вторичные структуры) оказывает основное влияние на эффективность трансляции.

Индекс EEI рассчитывается по следующей формуле:

$$EEI(i) = K / (w_1 T_a(i) + w_2 T_e(i)), \qquad (1)$$

i – номер гена, K – нормирующий множитель, обеспечивающий границы индекса от 0 до 10, $w_1 = \{0, 1\}$ и $w_2 = \{0, 1\}$ – индикаторные коэффициенты, определяющие учет слагаемых в значении индекса. Слагаемое $T_a(i)$ имеет смысл среднего времени размещения в А-сайте рибосомы изоакцепторной аминоацил-тРНК. Слагаемое $T_e(i)$ имеет смысл среднего времени, затрачиваемого рибосомой на стадию транслокации.

В зависимости от значений, принимаемых индикаторными коэффициентами, получаются пять типов индекса EEI:

1) $EEI1 = K/T_a$ – учитывается только кодонный состав гена;
2) $EEI2 = K/T_e(LCI1)$ – учитывается только количество вторичных структур в мРНК;
3) $EEI3 = K/T_e(LCI2)$ – учитывается только стабильность вторичных структур в мРНК;
4) $EEI4 = K/(T_a+T_e(LCI1))$ – учитываются и кодонный состав, и количество вторичных структур в мРНК;
5) $EEI5 = K/(T_a+T_e(LCI2))$ – учитываются и кодонный состав, и стабильность вторичных структур в мРНК.

Таким образом, для каждого гена исследуемого организма рассчитывается пять типов индекса EEI. Для определения, какой из типов индекса лучше всего оценивает эффективность элонгации трансляции в конкретном организме, в списках генов, отсортированных по каждому из пяти типов индекса EEI, выделяются гены рибосомных белков и рассчитываются их ранг (M) и стандартное отклонение (R) по формулам:

$$M = \frac{1}{N_{rib}} \sum_{i=1}^{N_{rib}} x_i, \qquad (2)$$

$$R = \sqrt{\frac{1}{N_{rib}} \sum_{i=1}^{N_{rib}} (M - x_i)^2}, \qquad (3)$$

где N_{rib} – количество ГРБ, x_i – позиция ГРБ в отсортированном по увеличению индекса EEI списке генов. Для каждого организма считается основным тот тип индекса EEI, для которого параметр M принимает наибольшее значение, а параметр R – наименьшее.

Программа EloE

Для проведения исследований в рамках данной работы было создано специальное программное обеспечение EloE (Elongation Efficiency), реализующее описанный выше алгоритм расчета индекса EEI. Данная программа в виде веб-приложения доступна по ссылке http://www-bionet.sscc.ru:7780/EloE/. Также на программу EloE было получено авторское свидетельство №2014662021, зарегистрированное 19.11.2014, «Программа для автоматической оценки эффективности элонгации трансляции генов различных организмов (EloE)», авторы: Соколов В. С., Зураев Б. С., Генаев М. А.

Статистический анализ

При получении описанных ниже результатов статистическая обработка данных и их визуализация были проведены с помощью языка программирования R и программ RStudio и Microsoft Excel. Для расчета корреляций в работе использовались коэффициенты корреляции Пирсона и Спирмена. Также при анализе для оценки различий между исследумыми выборками использовались t-критерий Стьюдента, критерий Манна-Уитни, φ-критерий Фишера. На нормальность выборки проверялись при помощи теста Шапиро-Вилка.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Исследование геномов одноклеточных организмов при помощи программы EloE

При помощи программы «EloE» был проведен анализ 2582 геномов бактерий. На Рисунке 1 представлен график распределения организмов по типам индекса EEI. Как видно из графика, большинство бактерий относится к первому типу индекса EEI1 – 45,35%. У данных организмов эффективность элонгации трансляции в большей степени зависит от частот кодонов в генах, чем от потенциальных вторичных структур в мРНК. Типичный представитель – *Escherichia coli* K-12.

Было проанализировано 165 геномов архей. График распределения организмов по типам индекса EEI представлен на Рисунке 2. Большинство организмов (58,79%) принадлежит к 4-му типу индекса EEI. Это говорит о существенной роли кодонного состава и количества потенциальных вторичных структур в определении эффективности элонгации трансляции. Типичный представитель – *Methanococcus voltae* A3. На основании данного распределения высказано предположение, что археи, возможно, лишены энергозависимого механизма расплетания вторичных структур в мРНК.

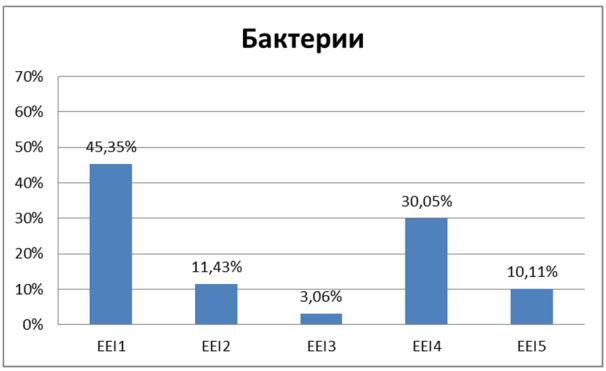

Рис. 1. Распределение 2582 организмов бактерий по пяти типам индекса EEI.

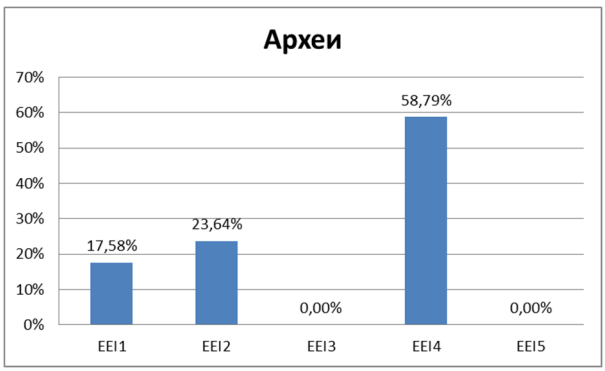

Рис. 2. Распределение 165 организмов архей по пяти типам индекса EEI.

Также было проанализировано 24 генома одноклеточных эукариот. График распределения организмов по типам индекса EEI представлен на Рисунке 3. Большинство организмов (62,50%) относятся к первому типу EEI1. Типичный представитель – *Saccharomyces serevisiae* S288C.

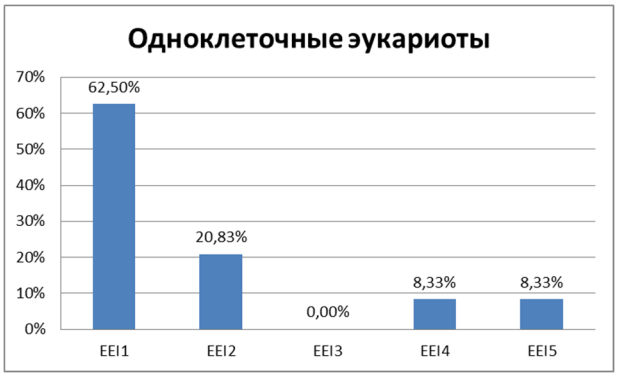

Рис. 3. Распределение 24 одноклеточных эукариот по пяти типам индекса EEI.

Подробное исследование организмов, принадлежащих к роду *Mycoplasma*

Суммарное распределение 62 проанализированных штаммов, принадлежащих к 27 видам *Mycoplasma*, по типам индекса представлено на Рисунке 4. Как можно видеть, в большинстве штаммов работает второй тип индекса EEI2 – эффективность элонгации трансляции зависит только от количества потенциальных вторичных структур в мРНК и не зависит от кодонного состава генов.

Почти у всех проанализированных видов *Mycoplasma* ГРБ хорошо определяются как высокоэкспрессируемые, значения ранга M для них высоки. Но есть виды со значительно более низкими значениями параметра M: *C. M. haemominutum*, *M. suis*, *M. pneumoniae* и, особенно, *M. haemocanis* и *M. haemofelis*. У данных видов практически отсутствует смещение ГРБ в сторону высокоэкспрессирующихся.

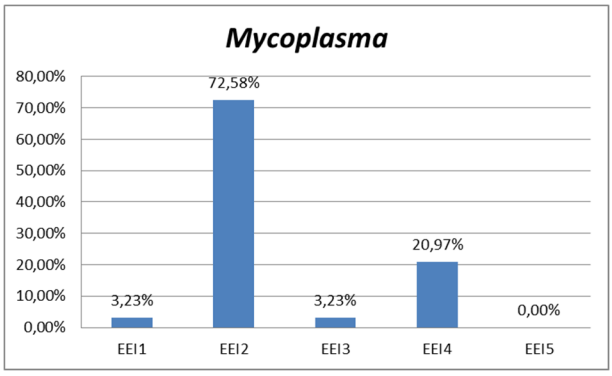

Рис. 4. Распределение 62 штаммов микоплазм по типам индекса EEI.

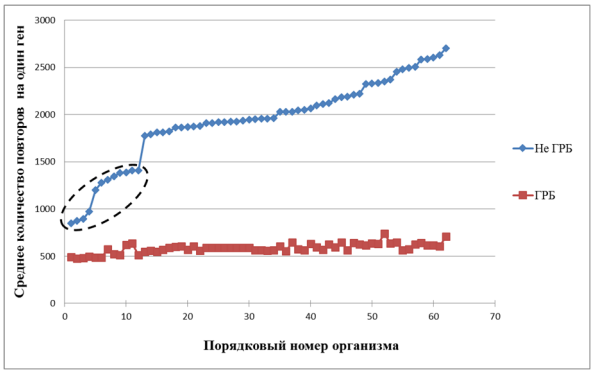

Рис. 5. Среднее число совершенных локальных инвертированных повторов на один ген для каждого штамма. Пунктиром выделены штаммы с параметром M ≤ 30.

Для исследуемых микоплазм были рассчитаны средние значения количества совершенных локальных инвертированных повторов на один ген. Полученные результаты были отсортированы по увеличению среднего количества локальных инвертированных повторов на один нерибосомный ген

(ген, не являющийся ГРБ) и отображены на графике, представленном на Рисунке 5. Установленный факт неравномерности по среднему числу локальных инвертированных повторов в генах разных штаммов *Mycoplasma* является новым и ранее неизученным.

Филогенетический анализ исследуемых *Mycoplasma*

Чтобы понять, почему же именно у данных штаммов (ограничены пунктиром на Рисунке 5) наблюдается такое низкое значение среднего количества локальных инвертированных повторов на один ген, был проведен филогенетический анализ исследуемых *Mycoplasma*.

Филогенетическое дерево *Mycoplasma*, построенное на основе анализа последовательностей 16S рРНК, было взято из статьи [Peters et al., 2008] и представлено на Рисунке 26. На дереве зеленым цветом отмечены виды со значением параметра M > 30, а оранжевым – с M ≤ 30. Видно, что почти все виды, отмеченные оранжевым, (кроме *M. pneumoniae*) попадают в группу гемоплазм. Достоверность преобладания особых видов микоплазм с параметром M ≤ 30 в группе гемоплазм была рассчитана при помощи φ-критерия Фишера. Было выделено две группы организмов: «гемоплазмы» и «не гемоплазмы». После этого проверялась гипотеза о различии частот встречаемости видов с параметром M ≤ 30 в каждой из групп. Показано преобладание особых микоплазм в группе гемоплазм с достоверностью более 99% (P < 0,01).

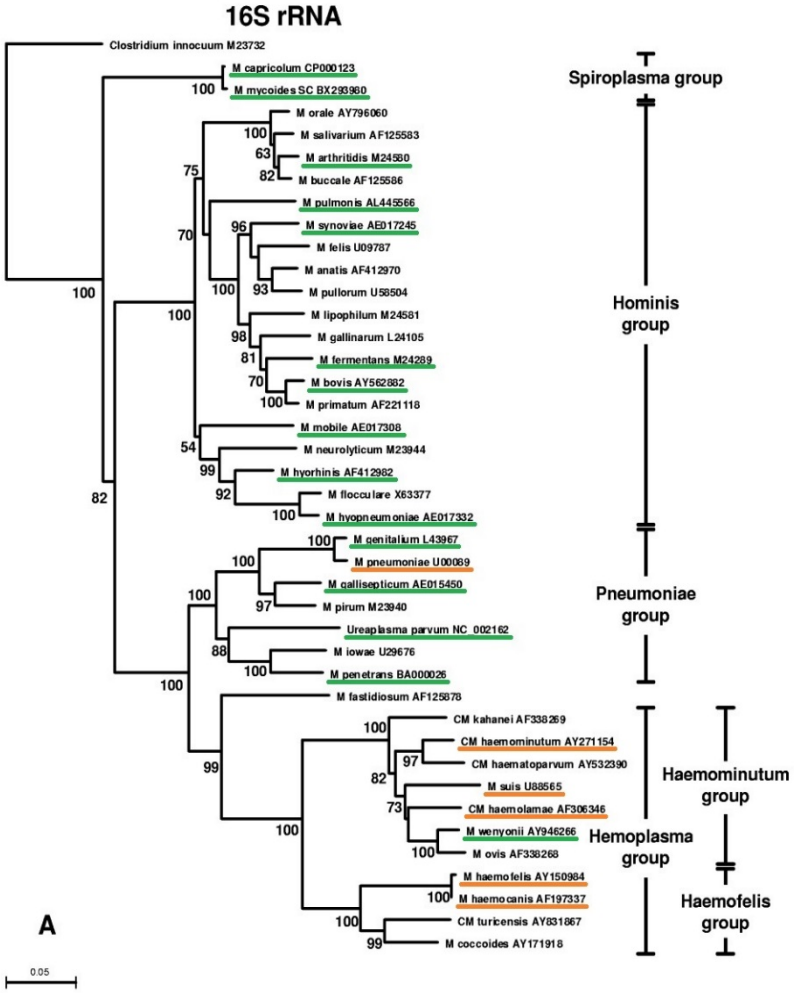

Рис. 6. Филогенетическое дерево *Mycoplasma*, построенное на основе анализа последовательностей 16S рРНК. Данные взяты из статьи [Peters et al., 2008]. Линией зеленого цвета подчеркнуты виды с параметром M > 30, а оранжевого – с M ≤ 30.

Возможно, в особых условиях обитания на поверхности или внутри эритроцитов данные виды эволюционировали в сторону уменьшения количества

совершенных локальных инвертированных повторов (потенциальных вторичных структур) в генах (выделены пунктирным овалом на Рисунке 5). В качестве гипотезы можно предположить, что таким образом они уменьшили энергетические затраты на процесс трансляции, чтобы повысить эффективность их экспрессии.

Анализ профилей LCI индексов индивидуальных нуклеотидов

Для более подробного изучения распределения вторичных структур в генах организмов рода *Mycoplasma* были рассчитаны специальные индексы локальной комплементарности для каждого нуклеотида в гене и на его флангах (LCI(*i*, *j*), где *i* – номер гена, *j* – номер нуклеотида в гене). Данный индекс показывает среднюю энергию потенциальных вторичных структур, в образовании которых принимает участие конкретный нуклеотид. После расчетов все гены одного организма выравнивались по старт (стоп) кодону трансляции и рассчитывались средние значения индексов LCI(*i*, *j*). Часть полученных результатов представлена на Рисунках 7-10.

Рис. 7. Профиль средних значений LCI(*i*, *j*) по всем генам *M. fermentans* JER (0 – старт-кодон).

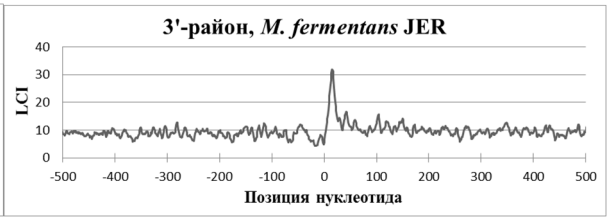

Рис. 8. Профиль средних значений LCI(*i*, *j*) по всем генам *M. fermentans* JER (0 – стоп-кодон).

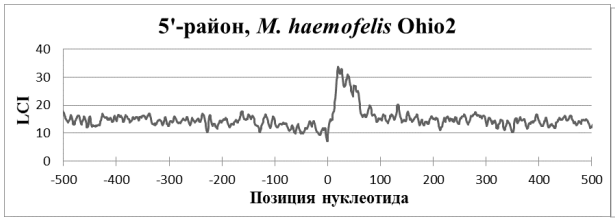

Рис. 9. Профиль средних значений LCI(*i*, *j*) по всем генам *M. haemofelis* Ohio2 (0 – старт-кодон).

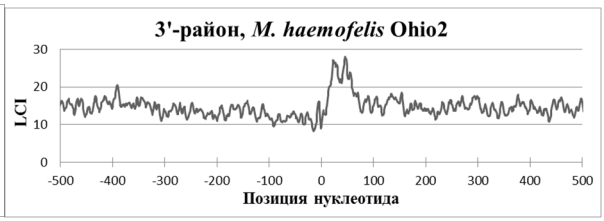

Рис. 10. Профиль средних значений LCI(*i*, *j*) по всем генам *M. haemofelis* Ohio2 (0 – стоп-кодон).

У большинства *Mycoplasma* профиль средних значений LCI(*i*, *j*) для 5′- и 3′-районов гена имеет вид, как у *M. fermentans* JER (Рисунки 7-8). В районе старт кодона трансляции наблюдается характерный спад профиля, а в районе стоп кодона – пик. Спад профиля в 5′-районе, т.е. сниженная стабильность потенциальных шпилек, вероятнее всего, способствует сборке рибосомного комплекса и началу трансляции. Наличие пика в 3′-районе говорит о повышенной стабильности потенциальных шпилек в данной области, которые могут отвечать, наоборот, за терминацию трансляции или, возможно, транскрипции.

Интересной особенностью обладают профили для 5′-районов у *M. haemofelis*. У данного вида вместо спада профиля наоборот наблюдается его повышение (Рисунок 9), что говорит о повышенной стабильности потенциальных шпилек в данной области. Пока сложно сказать что-либо о причинах данного явления и о том, почему именно *M. haemofelis* обладает

данной особенностью. Но, как уже было сказано выше, данный организм обитает в особых условиях (поверхность или внутриклеточное пространство эритроцитов), которые, возможно, способствовали эволюции первичной структуры его генов в сторону уменьшения количества локальных инвертированных повторов. Поэтому возможно, что шпильки в 5′-районе могут отвечать именно за регуляцию инициации трансляции у данного вида. Вероятно, у *M. haemofelis* есть особенный механизм регуляции инициации трансляции, отличный от механизмов в других *Mycoplasma*.

Связь между GC-составом и эволюционной оптимизацией первичной структуры генов *Mycoplasma* для повышения эффективности элонгации трансляции

Как известно из литературы, эффективность трансляции может зависеть от GC-состава генов. Для проверки данного факта для 62 исследуемых штаммов *Mycoplasma* был построен график зависимости параметра M, показывающего эффективность определения ГРБ как высокоэкспрессирующихся, от GC-состава генома данных организмов (Рисунок 11).

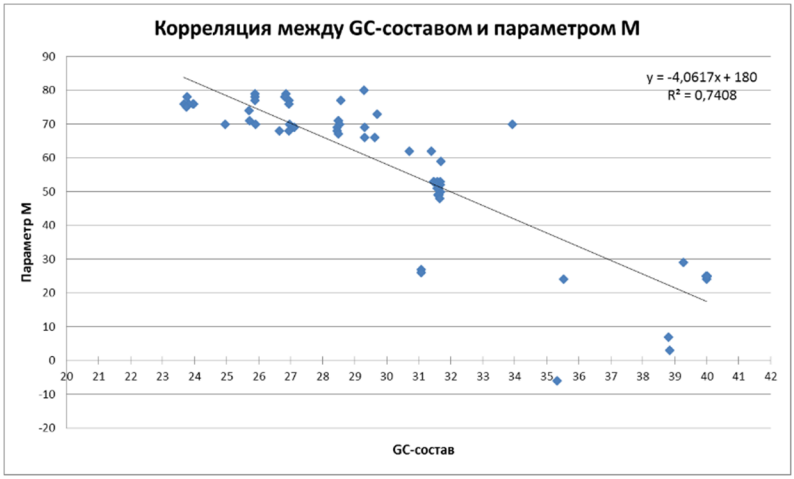

Рис. 11. Зависимость параметра M от GC-состава генома для 62 штаммов *Mycoplasma* (r = -0,85; P $<1{,}97\times10^{-18}$). Большинство штаммов с параметром M ≤ 30 характеризуются наибольшими значениями GC-состава генома.

Как видно из Рисунка 11, имеет место достоверная отрицательная корреляция r = -0,85 ($P<1{,}97\times10^{-18}$) между GC-составом генома и степенью эволюционной оптимизации первичной структуры его генов для повышения эффективности элонгации трансляции. Отсюда следует, что большинство штаммов с параметром M ≤ 30 характеризуются наибольшими значениями GC-состава генома. Этот факт хорошо согласуется с предположением о том, что у данных особенных микоплазм локальные инвертированные повторы редко встречаются в геноме из-за повышенной стабильности шпилек, которые они могут сформировать. Так как GC-богатые шпильки более стабильны, чем AT-богатые.

Исследование оптимизации первичной структуры генов архей в процессе эволюции

В рамках работы по изучению эволюции белков представителей царства Archaea был проведен анализ нуклеотидных последовательностей генов, кодирующих исследуемые белки, на предмет значительных изменений в уровне эффективности/скорости элонгации трансляции в процессе эволюции.

Для анализа брались нуклеотидные последовательности генов ныне существующих организмов, принадлежащих к царству Archaea, а также нуклеотидные последовательности генов их предков. Предковые последовательности генов были реконструированы К. В. Гунбиным на основе

трех альтернативных предковых последовательностей белков (реконструированных с использованием WAG, LG и JTT матриц при помощи пакета RAxML v.7.4.2 и пакета prank v.121218 для реконструкции паттерна делеций/инсерций), а также с использованием информации о кодирующих современные белки последовательностях ДНК. Реконструкция производилась на основе принципа парсимонии – минимального количества нуклеотидных замен, требующихся для объяснения замен аминокислот.

После того, как последовательности генов были получены, для каждой из них рассчитывался индекс EEI2. Был выбран именно этот индекс, т.к. реконструированные последовательности имеют вероятностный характер, а индекс EEI2 меньше других изменяется при небольших изменениях в первичной последовательности гена. Результаты записывались в соответствующие им позиции на построенном филогенетическом дереве. Далее проводился анализ данного дерева для обнаружения наибольших изменений в значениях индекса EEI2 в соседних узлах дерева.

Использовались два способа оценки эволюционных изменений значений индекса EEI2. Первый способ основан на учете значений индекса только самих последовательностей. Для этого для каждого соседства рассчитывалось значение Z_{score} (изменение величины EEI2 в единицах стандартного отклонения). Второй способ учитывал кроме значений индексов самих последовательностей еще и значения их ближайших предков. Для этого вместо исходного индекса EEI2 использовалась величина равная модулю разности $|EEI2_{предок} - EEI2_{потомок}|$. Далее для данной величины аналогично первому способу вычислялся Z_{score}. Итоговые результаты в виде основных соседств филогенетического дерева архей с наибольшими значениями Z_{score} представлены в Таблице 1. Также в таблице приведены условия обитания организмов, принадлежащих к данным группам.

Таблица 1. Основные соседства, в которых происходят наиболее существенные преобразования в контроле экспрессионного статуса генов, измеренного с помощью EEI2.

Соседство	**Значение Z_{score}**	**Изменение условий обитания в соседстве**
Учет значения индекса только самой последовательности		
Halobacteriales-Methanomicrobia	**2,13**	**Halobacteriales:** оптимум t=31-50°C; питание органическим субстратом; аэробы; галофилы
		Methanomicrobia: оптимум t=0-70°C; питание неорганическим субстратом; анаэробы; галофобы
Desulfurococcales1-Thermoproteales	**1,79**	**Desulfurococcales1:** оптимум t=85-106°C; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности и любых глубинах
		Thermoproteales: оптимум t=75-104°C; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности

Desulfurococcales1-Sulfolobales	**1,55**	**Desulfurococcales1:** оптимум t=85-106°C; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности и любых глубинах
		Sulfolobales: оптимум pH~2; оптимум t=40-85°C; питание органическим и неорганическим субстратом; факультативные аэробы
Учет значения индекса как самой последовательности, так и ближайшего предка		
Desulfurococcales1-Thermoproteales	**2,16**	**Desulfurococcales1:** оптимум t=85-106°C; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности и любых глубинах
		Thermoproteales: оптимум t=75-104°C; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности
Archaeoglobales-Methanomicrobia	**1,91**	**Archaeoglobales:** оптимум t=55-90°C; питание неорганическим субстратом; анаэробы; галофобы
		Methanomicrobia: оптимум t=0-70°C; питание неорганическим субстратом; анаэробы; галофобы
Halobacteriales-Methanomicrobia	**1,72**	**Halobacteriales:** оптимум t=31-50°C; питание органическим субстратом; аэробы; галофилы
		Methanomicrobia: оптимум t=0-70°C; питание неорганическим субстратом; анаэробы; галофобы

Наибольшие значения Z_{score} соответствуют наибольшим изменениям индексов EEI2. Например, для первого соседства Halobacteriales–Methanomicrobia: Halobacteriales обитают в аэробных условиях и являются галофилами, а Methanomicrobia обитают в анаэробных и являются галофобами. Таким образом, из Таблицы 1 видно, что при изменении условий обитания организмов, наблюдаются наиболее значительные перестройки в первичной структуре их генов, что свидетельствует о приспосабливании к новым условиям.

Исследование зависимости между влиянием потенциальных вторичных структур в мРНК у архей и температурой их среды обитания

Как было показано, в большинстве архей работают 2 и 4 тип индекса EEI. Другими словами, потенциальные вторичные структуры в мРНК являются одними из основных параметров в определении эффективности элонгации трансляции генов. Поскольку данные организмы являются пойкилотермными, низкая температура обитания потенциально может увеличить влияние стабильности вторичных структур на эффективность трансляции, а высокая –

уменьшить. Для ответа на вопрос, влияет ли температура среды обитания организма на учет потенциальных вторичных структур в мРНК при определении эффективности элонгации трансляции, был рассчитан коэффициент корреляции между данной температурой и наибольшим параметром M организма.

Достоверной корреляции между температурой обитания организма и влиянием потенциальных вторичных структур в мРНК на эффективность элонгации трансляции обнаружено не было. Коэффициент корреляции равен r = -0,15 (P < 0,11). Это может быть связано с тем, что археи, по-видимому, лишены энергозависимого механизма расплетания вторичных структур в мРНК, как было отмечено выше.

Связь между GC-составом и эволюционной оптимизацией первичной структуры генов архей для повышения эффективности элонгации трансляции

Для проверки зависимости эффективности элонгации трансляции от GC-состава генов у архей был построен график зависимости параметра M, показывающего эффективность определения ГРБ как высокоэкспрессируемых, от GC-состава генома данных организмов (Рисунок 12). Имеет место достоверная отрицательная корреляция между GC-составом генома и степенью эволюционной оптимизации первичной структуры его генов для повышения эффективности элонгации трансляции ($r = -0{,}40$; $P < 8{,}65\times10^{-8}$).

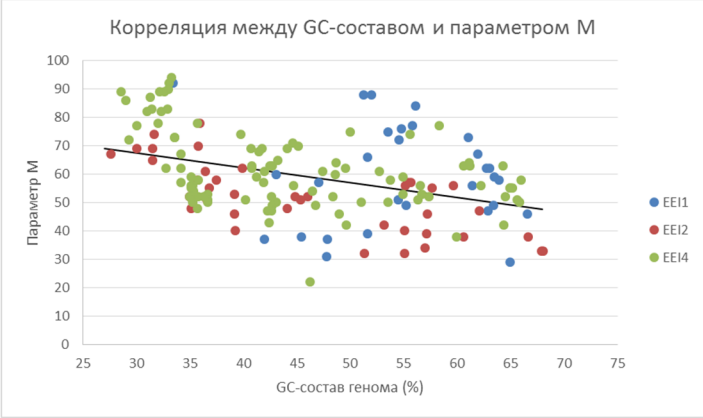

Рис. 12. Зависимость параметра M от GC-состава генома для 135 видов архей ($r = -0{,}40$; $P < 8{,}65\times10^{-8}$). Синим цветом выделены организмы, в которых работает EEI1, красным – EEI2, зеленым – EEI4. Черная линия – линия регрессии.

Исследование взаимосвязи между эффективностью элонгации трансляции генов дрожжей и плотностью их нуклеосомной упаковки в 5' фланкирующем районе

Проверяемая гипотеза заключалась в следующем: для эффективной экспрессии генов необходимы согласованно оптимизированные процессы транскрипции и трансляции, в частности – инициации транскрипции и элонгации трансляции. Такая корреляция была найдена между потенциалом формирования нуклеосом (ПФН) в 5´ фланкирующих областях генов дрожжей видов *S. cerevisiae* и *S. pombe* со значением EEI соответствующих генов. ПФН – функция, которая характеризует вероятность расположения нуклеосомы в заданном сайте последовательности. Для расчета ПФН использовалась программа RECON [Matushkun et al., 2010].

У *S. pombe* для 15% генов с наибольшими значениями эффективности трансляции имеется достоверная отрицательная корреляция между EEI и ПФН в интервале [-330, -130] относительно старта транскрипции. Т.е., чем слабее нуклеосомная упаковка в 5´-области, тем выше эффективность элонгации трансляции мРНК. В интервале [0; +400] также наблюдается достоверная

отрицательная корреляция для высоко транслируемых последовательностей. Видимо, это связано с особенностями нуклеотидного и динуклеотидного состава кодирующих частей генов.

У *S. cerevisiae* наоборот обнаружена достоверная положительная корреляция между EEI и ПФН для 15% генов с низкими значениями EEI (медленно транслируемых) в интервалах [-120; -40], [-230; -170], [-600; -370] относительно старт кодона трансляции. Т.е., чем крепче нуклеосомная упаковка в 5′-области, тем выше эффективность элонгации трансляции мРНК. Высоко достоверная положительная корреляция между EEI и ПФН для медленно транслируемых генов в начале кодирующей части [0; +400], видимо, связана с нуклеотидным и динуклеотидным составом этих последовательностей.

Для подтверждения теоретических результатов были рассчитаны коэффициенты корреляции между EEI генов *S. cerevisiae* и экспериментальными данными по расположению нуклеосом в геномной ДНК (Рисунок 13). Для медленно транслируемых последовательностей в интервале [-110, 0] корреляция ($r \leq 0{,}25$) достоверна и положительна. Этот результат совпадает с теоретическими данными, что подтверждает их достоверность. Для высоко транслируемых последовательностей корреляция ($r \geq -0{,}15$) в этом же интервале достоверна и отрицательна.

Дополнительный анализ кодирующих частей генов *S. cerevisiae и S. pombe* показал, что различия в значениях коэффициентов корреляции для кодирующих районов генов связано именно с различием в их первичной структуре. Отбор кодонов в генах данных организмов идет по-разному. Это согласуется с данными из литературы о том, что у разных видов дрожжей ортологичные гены по-разному адаптируют свой кодонный состав к похожим пулам тРНК в зависимости от функциональной значимости этих генов для жизни каждого конкретного организма.

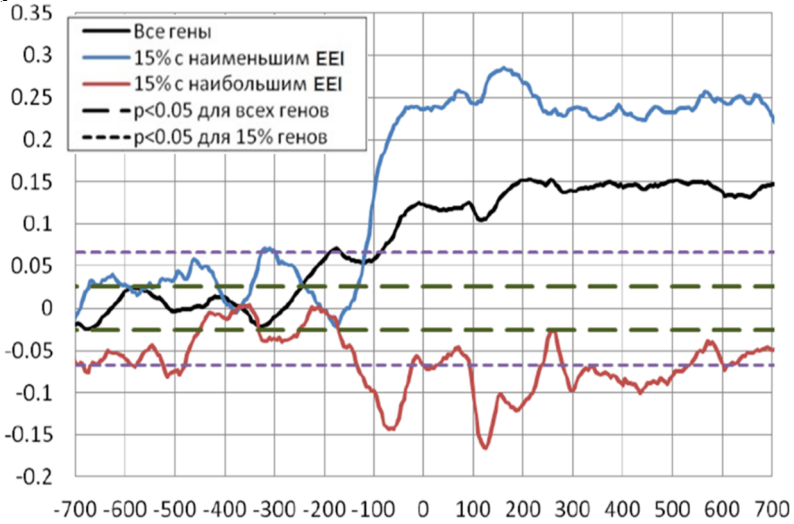

Рис. 13. Коэффициенты корреляции (ось ординат) между EEI и экспериментальными данными по плотности нуклеосомной упаковки, полученной прямым секвенированием нуклеосомной ДНК для *S. cerevisiae*: для всех последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует положению нуклеотида A в старт-кодоне трансляции ATG. Горизонтальные линии – зоны достоверности ($P < 0{,}05$) по критерию Фишера. Каждая точка графика – усреднение по 10 точкам расчета.

ВЫВОДЫ

1) Создано веб-приложение EloE, реализующее самообучающийся алгоритм расчета индекса эффективности элонгации EEI, которое обеспечивает достоверный высокопроизводительный анализ эффективности элонгации трансляции генов одноклеточных организмов и позволяет обрабатывать до нескольких тысяч геномов за один запуск.
2) С помощью программы EloE проведена классификация 2771 генома различных одноклеточных организмов (бактерии, археи, эукариоты) по пяти типам эволюционной оптимизации процесса элонгации трансляции. Показано различие между бактериями и археями по предпочтительному типу оптимизации: у большинства проанализированных бактерий (45,35% от 2582) основную роль в определении эффективности элонгации трансляции играет только кодонный состав генов; у большинства архей (58,79% от 165) – кодонный состав генов и количество локальных инвертированных повторов в мРНК.
3) Показано наличие достоверной отрицательной корреляции ($r = -0{,}85$; $P < 1{,}97\times10^{-18}$) между GC-составом генома и степенью эволюционной оптимизации первичной структуры генов *Mycoplasma* для повышения эффективности элонгации трансляции.
4) Показано, что у семи видов *Mycoplasma* (*C. M. haemolamae*, *M. haemocanis*, *M. wenyonii*, *M. haemofelis*, *M. pneumonia*, *C. M. haemominutum*, *M. suis*), в отличие от 20 других, в процессе эволюции прошла массовая минимизация количества локальных совершенных инвертированных повторов (потенциальных шпилек) в мРНК.
5) Анализ профилей стабильности потенциальных вторичных структур в мРНК у *Mycoplasma* показал, что *M. haemofelis* радикально отличается от остальных видов микоплазм наличием более стабильных потенциальных вторичных структур в мРНК в районе старт-кодона трансляции.
6) Показано наличие достоверной отрицательной корреляции ($r = -0{,}40$; $P < 8{,}65\times10^{-8}$) между GC-составом генома и степенью эволюционной оптимизации первичной структуры генов архей для повышения эффективности элонгации трансляции.
7) Показано наличие значимой корреляции между потенциалом формирования нуклеосом и индексом эффективности элонгации трансляции генов *S. cerevisiae* и *S. pombe*.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи:

1. Матушкин Ю. Г., Левицкий В. Г., **Соколов В. С.**, Лихошвай В. А., Орлов Ю. Л. Эффективность элонгации генов дрожжей коррелирует с плотностью нуклеосомной упаковки в 5′-нетранслируемом районе. Математическая биология и биоинформатика. 2013. 8(1):248–257.

2. **Sokolov V. S.**, Likhoshvai V. A., Matushkin Yu. G. Gene expression and secondary mRNA structures in different *Mycoplasma* species. Russian Journal of Genetics: Applied Research, 2014, 4(3):208–217.

3. **Sokolov V. S.**, Zuraev B. S., Lashin S. A., Matushkin Yu. G. EloE: web application for estimation of gene translation elongation efficiency. Russian Journal of Genetics: Applied Research. 2015. 5(4):335–339.

4. **Sokolov V. S.**, Zuraev B. S., Lashin S. A., Matushkin Yu. G. A web application for automatic prediction of gene translation elongation efficiency. Journal of Integrative Bioinformatics. 2015. 12(1):257–264.

Авторское свидетельство

Соколов В. С., Зураев Б. С., Генаев М. А. «Программа для автоматической оценки эффективности элонгации трансляции генов различных организмов (EloE)», № 2014662021 от 19.11.2014.

Тезисы конференций:

1. **Соколов В. С.**, Лихошвай В. А., Матушкин Ю. Г. Программное обеспечение для компьютерного исследования особенностей элонгации трансляции (на примере одноклеточных организмов рода *Mycoplasma*). XIII всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям, 2012. http://conf.ict.nsc.ru/ym2012/ru/reportview/138699.

2. **Соколов В. С.**, Матушкин Ю. Г. Компьютерное исследование особенностей элонгации трансляции у *Mycoplasma*. VI съезд Вавиловского общества генетиков и селекционеров (ВОГиС) и ассоциированные генетические симпозиумы, 2014. С. 64–65. Диплом 3-ей степени.

3. **Sokolov V. S.**, Likhoshvai V. A., Matushkin Yu. G. Gene expression and mRNA secondary structures in *Mycoplasma* strains. 5th international young scientists school «Systems biology and bioinformatics», SBB'2013. http://conf.ict.nsc.ru/SBB2013/reportview/158983.

4. **Sokolov V. S.**, Likhoshvai V. A., Matushkin Yu. G. Computational study of translation elongation features in *Mycoplasma*. Moscow conference on computational molecular biology, MCCMB'13. http://mccmb.belozersky.msu.ru/2013/abstracts/abstracts/156.pdf.

5. **Sokolov V. S.**, Gunbin K. V., Matushkin Yu. G. Variation of elongation efficiency index of Archaea genes during evolution. The 9th international conference on bioinformatics of genome regulation and structure\System biology, BGRS\SB'2014. P. 153.

6. **Sokolov V. S.**, Zuraev B. S., Lashin S. A., Matushkin Yu. G. EloE – web application for estimation of translation elongation efficiency of genes in various organisms. The 9th international conference on bioinformatics of genome regulation and structure\System biology, BGRS\SB'2014. P. 152.

7. **Sokolov V. S.**, Matushkin Yu. G. Analysis of Bacteria and Archaea genomes available in GenBank database by “EloE” program. 6th international young scientists school «Systems biology and bioinformatics», SBB'2014. P. 33.