Федеральное государственное автономное образовательное учреждение
высшего образования
«**СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**»
Институт фундаментальной биологии и биотехнологий
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ / Кратасюк В.А.
«___» _____2020 г.

**БАКАЛАВРСКАЯ РАБОТА**

06.03.01. Биология

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТИПОВ ВНУТРЕННЕЙ СТРУКТУРИРОВАННОСТИ ГЕНОМОВ БАКТЕРИЙ ДЛЯ РАЗЛИЧНЫХ ДЛИН ФРАГМЕНТОВ С УЧЁТОМ ЗНАЧЕНИЙ GC-СОСТАВА

Руководитель _____ д.ф.-м.н. М. Г. Садовский

Выпускник _____ В.С. Густов

Красноярск 2020

# РЕФЕРАТ

Выпускная квалификационная работа по теме «Сравнительный анализ типов внутренней структурированности геномов бактерий для различных длин фрагментов с учётом значений GC-состава» содержит 46 страниц текстового документа, 24 использованных источника, 31 иллюстрацию, 6 таблиц.

ГЕНОМ БАКТЕРИЙ, ЧАСТОТНЫЕ СЛОВАРИ, ВНУТРЕННЯЯ СТРУКТУРА ГЕНОМА, GC-СОСТАВ

Цель работы – выявить различия в наблюдаемых структурах геномов бактерий для разных параметров их фрагментации и установить связь этих различий с таксономией носителей геномов.

Изучение биологических макромолекул, таких как ДНК, а также поиск новых способов их изучения являются актуальными задачами в связи с их фундаментальным значением в жизни любого организма. В настоящее время обработка информации, записанной в геноме всё ещё не является простым процессом, в связи с чем актуален поиск новых способов определения функций генома.

В работе был проведён анализ геномов методом частотных словарей. Объектом исследования стали геномы бактерий различных видов.

Выявлены различные структуры при разных длинах окна считывания. Проведён анализ распределения GC-состава в структурах и его влияние на тип структуры. Проведён анализ связи структур и таксономии растения.

**СОДЕРЖАНИЕ**

# ВВЕДЕНИЕ

Современная биология имеет очень множество смеженных наук, изучающих не только сами организмы, но и их составные части. Такими частями могут являться ткани, клетки, органеллы или даже макромолекулы. Одной из самых изучаемых макромолекул является ДНК.

Объектом в данном исследовании стали бактерии. Они являются очень удобным организмом для изучения ДНК, так как их геном сравнительно мал.

Развитие технологий позволило изучать ДНК разнообразными методами. Один из таких методов – определение структуры генома при помощи частотных словарей [1]. Данный метод уже применялся как к бактериям в общем [2], так и к отдельным их группам [3, 4]. Такими же методами исследуются и геномы органелл: митохондрий [5] и хлоропластов [6, 7, 8, 9]. Метод частотных словарей позволяет строить пространственные структуры, которые могут нести в себе информацию о функциях и таксономии носителя генома, что делает изучение данного метода актуальным вопросом.

Цель данной работы: выявить различия в наблюдаемых структурах геномов бактерий для разных параметров их фрагментации и установить связь этих различий с таксономией носителей геномов.

Для достижения цели были поставлены следующие задачи:

- Выделить структуры при различных длинах окон считывания;
- Проследить путь изменения структур при изменении длины окна считывания;
- Определить однородность распределения GC-состава в структурах и изменение данного распределения при перемене длины окна считывания;
- Определить, наблюдается ли соответствие между таксоном растения и свойственной этому таксону структурой.

Основные результаты работы были доложены на конференции «Нейроинформатика, её приложения и анализ данных» (Красноярск, 27-29

сентября 2019 г.)

Основные результаты работы опубликованы в сборнике «Нейроинформатика, её приложения и анализ данных: Материалы XXVII Всероссийского семинара»

# 1 Обзор литературы

## 1.1. Бактерии

Бактерии — домен микроорганизмов, генетический материал которых не обособлен от цитоплазмы ядерной мембраной. Самая древняя группа организмов, с момента их появления на Земле насчитывается около 3,5 млрд. лет. Распространены практически повсеместно: населяют почву, глубинные земные слои, горячие источники, пресные и морские водоёмы, другие организмы, радиоактивные отходы.

Форма и размер бактериальных клеток в значительной степени варьируют: от нанобактерий размером 0,5 мкм до бактерий родов *Macromonas* и *Achromatium*, достигающих в длину до 100 мкм. Наиболее распространённые формы бактериальных клеток – палочковидная, сферическая и извитая, однако также существуют организмы, обладающие звёздчатыми, треугольными и квадратными клетками. Бактерии могут существовать одиночно, образовывать пары, формировать колонии.

Трудно переоценить значимость бактерий для хозяйственного значения. Бактериальные организмы нашли свое применение не только в традиционных областях, таких как изготовление пищевых продуктов в процессах брожения, но также и в процессах переработки отходов, биоремедиации, в промышленности и в сельском хозяйстве, в медицине и исследованиях. Бактерии применяют для устранения последствий разлива нефти, ввиду их способности к разрушать нефтяные углеводороды [10]. Метанообразующие бактерии способствуют окислению водорода углекислотой, продуктом восстановления которой впоследствии становится метан [11]. В химической промышленности бактериальные организмы являются важным элементом процесса производства чистых энантиомеров химических веществ, применяемых в агрохимии и фармацевтической отрасли [12]. При помощи информации о бактериальном метаболизме и экспрессии генов бактерий стало возможным промышленное производство терапевтически важных соединений, таких как факторы роста, антитела,

инсулин [13].

Довольно часто дифференциация видов бактерий в пределах одного рода является затруднительной процедурой. Классические методы микробиологии не всегда позволяют добиться успеха.

### 1.2. Генетическая система бактерий

Генетический материал бактерий организован в виде компактно уложенной молекулы ДНК, она локализована в определенных участках цитоплазмы, и, в сравнении с эукариотическими организмами, не отделена от нее ядерной мембраной. Структура дезоксирибонуклеиновой кислоты представляет собой правозакрученную антипараллельную двойную спираль Геномная ДНК и организованные с ней белки в совокупности образуют зону, именуемую нуклеоидом [14]. Типичный бактериальный нуклеоид содержит одну двунитевую кольцевую молекулу ДНК [15]. В нуклеоиде бактерий насчитывается около 4 тыс. генов.

Существенное различие в структурной организации между прокариотической и эукариотической ДНК состоит в том, что нуклеоид не содержит гистоновых белков, они содержат лишь гистоноподобные белки, принимающие участие в процессе компактизации ДНК, а именно это белки H-NS, IHF и HU. Среди бактериальных организмов только гены архей содержат интроны, в целом же геном организован компактно, количество некодирующих последовательностей минимально. Понятие «ген» подразумевает дискретный участок молекулы ДНК, содержащий специфическую последовательность, в которой заключена информацию о структуре и свойствах кодируемых белков. Геном имеет повышенный кодирующий потенциал, но при этом невелик по размеру, ввиду того, что несколько рамок считывания одной и той же генетической последовательности используется для кодирования белков [16].

Длинна ДНК бактерий составляет около 5 млн. п.н., содержание в ней пар оснований АТ и ГЦ является важным диагностическим признаком, так

как оно постоянно для определенного вида [17].

Геномы прокариотических организмов содержат не только хромосомные гены, которые кодируют РНК и пептиды, но также и иные генетические структуры. К числу таких структур относятся плазмиды – генетические внехромосомные элементы, способные к автономной репликации. Утрата плазмид не приводит к гибели клетки, поэтому содержащаяся в них генетическая информация не является жизненно важной. Тем не менее, наличие плазмид в бактериальной клетке обеспечивает эффективную адаптацию к различным условиям среды и поддерживает стабильность протекания эволюционного процесса [18]. Плазмиды представлены в виде кольцевых ДНК, длина которых может достигать до нескольких сотен тысяч пар оснований. Признаки, приобретаемые совместно с плазмидами, нередко определяют наименование этих плазмид, например, плазмида R, наделяющая организм резистентностью к воздействию антибиотиков [19].

Автономность плазмидной репликации обеспечивается сайтом инициации репликации (origin of replication) в совокупности с набором генов, содержащих информацию о пептидах. Плазмиды не обладают полным набором ферментов для собственной репликации, поэтому для полноценного функционирования репликационного аппарата требуются компоненты клетки-хозяина. Для разных плазмид, обладающих одинаковыми ori сайтами, сосуществование в одной клетке невозможно, поэтому несовместимость плазмид определяется именно специфичность сайтов ori [20]. Ряд плазмид имеет свойство интегрироваться в хромосому организма-хозяина и производить репликацию в ее составе в форме эписомы. Количество плазмидных копий в клетках широко варьируется: от одной до сотен копий на одну хромосому. Тем не менее, количество копий плазмид строго регулируется, так как неограниченная репликация плазмидных молекул ДНК чревата гибелью клетки [21].

Вариабельность геномов внутри видов сподвигла условно поделить

наборы генов на базовый и условно вспомогательный. Базовые консервативные гены обеспечивают функционирование информационных систем репликации, трансляцию, транскрипцию, формирование клеточных структур, определяющих видовую принадлежность, основные пути метаболизма. К вспомогательным относят гены, ответственные за приспособленность к определенной экологической нише, а также гены, осуществляющие контроль метаболизма и морфофизиологических признаков. Многие подобные гены в основном локализованы на геномных элементах, плазмидах, геномных островках, необязательно имеющихся у всех штаммов одного вида организмов [22].

### 1.3 GC-состав

Данный термин означает долю цитозина и гуанина среди всех нуклеотидов, встречающихся в геноме. Процентное содержание этой пары нуклеотидов может влиять на структуру генома. Кроме того, GC-состав может сильно варьировать у различных таксонов. Так, у хлоропластов мхов очень низкий процент GC-состава, в среднем он не превышает 0,3. Хлоропласты папоротников имеют GC-состав в районе 0,41. В геноме хлоропластов древних споровых растений GC-пары занимают примерно половину генома [23]. Влияние GC-состава на структуру генома, зачастую, имеет решающую роль и напрямую влияет на получаемую структуру [1, 2].

## 2 Материалы и методы

### 2.1 Базы данных

Для исследования было отобрано 46 геномов бактерий. Данные были получены из открытой базы данных EMBL-банка (релиз от 5 мая 2015 года). Данная база собирает геномы из различных источников и имеет большое разнообразие организмов для выбора.

Соотнесение систематики, представленной в информации сопровождающей геномы с актуальной на сегодняшний день происходило при помощи базы данных «Integrated Taxonomic Information System». Данная база данных является одной из крупнейших в своём роде и содержит информацию о таксономии большинства известных науке видов живых существ.

### 2.2 Построение частотных словарей

Так как весь геном любого живого организма состоит из нуклеотидов, его можно представить, как последовательность символов. Длина такой последовательности обозначается $\aleph$. Вся последовательность состоит из нуклеотидов и так как данная работа рассматривает последовательности ДНК, то данными нуклеотидами являются: аденин, гуанин, цитозин и тимин. То есть, последовательность может состоять только из 4 символов алфавита $\aleph = \{A, C, G, T\}$. Для каждой из таких последовательностей строится частотный словарь. Его толщина равна 3. Частотный словарь триплетов — это список всех троек идущих подряд нуклеотидов, в котором указана частота встречаемости этих троек. Не сложно посчитать, что всего возможно существование 64 вариаций триплетов. Частотой триплета является отношение копий данного триплета, к общему числу триплетов:

$$f_\omega = \frac{n_\omega}{N} \tag{1}$$

Для построения частотных словарей триплетов вся последовательность

обрабатывалась окнами считывания, имеющими длину $\Delta$ и шаг $t$. Задачи данной работы предполагают изменение этих значений. Для каждого положения $i$ окна считывания вычислялся частотный словарь $W_3^{(i)}$. Он соответствовал одной из точек в 64-мерном пространстве. После, строилось распределение словарей в пространстве частот, а анализировался вид данных распределений в пространстве первых трёх главных компонент, вычисленных для 63-мерного пространства. Триплет с минимальным стандартным отклонением исключался поскольку сумма всех частот не должна быть равна единице. Выбор данного триплета обоснован тем, что он вносит наименьший вклад в различимость фрагментов генома [24].

Точки могли быть окрашены в зависимости от поставленной задачи. Так, одним из вариантов окраски является окраска по количеству GC-состава. В данном варианте, точки, содержащие в себе наибольше количество GC-состава, окрашиваются в красный цвет, а содержащие наименьшее количество – в синий. Зелёным цветом обозначено среднее значение GC-состава.

Другим вариантом было окрашивание в соответствии с относительной фазой. Расположение рассматриваемого участка в кодирующей или же некодирующей последовательности генома определяет относительную фазу. Участок может относиться к кодирующим, только если он целиком попал в кодирующую область исследуемой последовательности. В том случае, если участок является кодирующим, для него возможны 6 вариантов маркировки, а именно: $B_0$, $B_1$, $B_2$, $F_0$, $F_1$, $F_2$. Буква в данной маркировке обозначает прямое и обратное направление аннотирования. Для прямого используется символ $F$, а для обратного $B$. Цифра в свою очередь обозначает остаток от деления на 3 разности номеров центрального символа участка и перового символа кодирующей области, к которой этот участок относится.

Точки, соответствующие относительным фазам $F_0$ и $B_0$, помечались

пурпурным и сиреневым цветом со, точки, соответствующие относительным фазам $F_1$ и $B_1$, были окрашены в зеленый и голубой цвет, а точки, соответствующие относительным фазам $F_2$ и $B_2$, были обозначены желтым и оранжевым цветом соответственно. Для точек, соответствующих некодирующим областям был выбран коричневый цвет.

### 2.3 Использованное программное обеспечение

Построение распределения словарей в пространстве частот производилось при помощи программного обеспечения *VidaExpert*. Данная программа открыта для использования и несёт в себе множество функций для работы с данными различного вида.

Для расчётов и построения графиков и таблиц было использовано программное обеспечение *Microsoft Excel*.

## 3 Результаты работы

Изъято в связи с авторскими правами со страницы 13 по страницу 41.

# ЗАКЛЮЧЕНИЕ

Геномы бактерий могут образовывать различные структуры в зависимости от длины окна считывания. При длине окна считывания, равной 603 нуклеотидам, было выделено 4 структуры, одна из которых имеет 2 различные конформации. Данные структуры были названы «Шестилучевая», «Совпадающие треугольники», «Параллельные треугольники», «Перпендикулярные треугольники с жёлто-оранжевым хвостом» и «Перпендикулярные треугольники с голубо-зелёным хвостом».

Распределение GC-состава внутри данных структур показало, что структуры «Шестилучевая» и «Совпадающие треугольники» имеют градиентное распределение GC-состава, в то время как остальные — хаотичное. В результате исследования среднего содержания GC-состава в различных типах структур выявлено, что каждая из структур имеет свой характерный диапазон процентного содержания GC-состава.

При увеличении окна считывания до 60000 нуклеотидов образуются совершенно иные структуры. Их основное отличие заключается в нитчатом составе. Так, структуры, выделяемые при чётном шаге, состоят из 3 раздельных нитей, а при нечётном из одной длинной непрерывной нити. Всего было выделено 3 такие структуры. Им были присвоены названия «Клубок», «Снитч» и «Две плоскости». GC-состав внутри данных структур всегда распределяется градиентно, однако средние его значения в структуре могут сильно варьировать, что говорит о независимости типа структуры от содержания в ней GC-состава.

Показано, что постепенное увеличение длины фрагментов, по которым определяется структура, приводит к её плавной трансформации. Полученные данные дают основание полагать, что любая структура, выделяемая при длине окна считывания в 603 нуклеотида, может преобразовываться в любую структуру, выделяемую при длине окна считывания в 60000 нуклеотидов.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Gorban A. N., Zinovyev A. Y., Popova T. G. Seven clusters in genomic triplet distributions //In silico biology. – 2003. – Т. 3. – №. 4. – С. 471-482.
2. Gorban A., Popova T., Zinovyev A. Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences //Physica A: Statistical Mechanics and its Applications. – 2005. – Т. 353. – С. 365-387.
3. Сенашова М. Ю., Садовский М. Г. Пространственная структура геномов цианобактерий //Международный журнал прикладных и фундаментальных исследований. – 2017. – №. 11-2. – С. 255-259.
4. Горбань И. К., Густов В. С., Сенашова М. Ю, Садовский М. Г. Структура геномов цианобактерий для разной толщины словарей //Нейроинформатика, её приложения и анализ данных. – 2019. – С. 31-36.
5. Косарев Р. Е., Сенашова М. Ю., Садовский М. Г. Пространственная структура митохондриальных геномов растений и животных //Нейроинформатика, её приложения и анализ данных. – 2019. – С. 58-66.
6. Sadovsky M. G. et al. Seven-cluster structure of larch chloroplast genome. – 2015.
7. McFadden G. I. Chloroplast origin and integration //Plant Physiology. – 2001. – Т. 125. – №. 1. – С. 50-53.
8. Сенашова М. Ю., Садовский М. Г. Структура геномов хлоропластов водорослей //Международный журнал прикладных и фундаментальных исследований. – 2018. – №. 1. – С. 121-125.
9. Горбань И. К., Густов В. С., Сенашова М. Ю, Садовский М. Г. Некодирующие области геномов хлоропластов отражают таксономию их носителей //Нейроинформатика, её приложения и анализ данных. – 2018. – С. 35-53.
10. Cohen Y. Bioremediation of oil by marine microbial mats //International Microbiology. – 2002. – Т. 5. – №. 4. – С. 189-193.
11. Каллистова А. Ю. и др. Образование и окисление метана прокариотами //Микробиология. – 2017. – Т. 86. – №. 6. – С. 661-683.
12. Liese A., Villela Filho M. Production of fine chemicals using biocatalysis

//Current opinion in biotechnology. – 1999. – Т. 10. – №. 6. – С. 595-603.

13. Graumann K., Premstaller A. Manufacturing of recombinant therapeutic proteins in microbial systems //Biotechnology Journal: Healthcare Nutrition Technology. – 2006. – Т. 1. – №. 2. – С. 164-186.

14. Лысак, В. В. Микробиология : учебное пособие / В. В. Лысак.- Минск: БГУ, 2007. – 430 с.

15. Гусев, М. В. Микробиология / М. В. Гусев, Л. А. Минеева. – Москва: МГУ, 2004. – 448 с.

16. Шестаков, С. В. Как происходит и чем лимитируется горизонтальный перенос генов у бактерий / С. В. Шестаков // Экологическая генетика. –2007.– Т. 5, № 2. – С. 12–24.

17. Boto, L. Horizontal gene transfer in evolution: facts and challenges / L. Boto // Proc. Roy. Soc., 2010. - V. 277. - P. 819–827.

18. Shintani, M.Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy / M. Shintani, Z. Sanchez, K. Kimbara //Frontiers In Microbiology. – 2015. - V.6. – P. 242-243.

19. Равин, Н.В.Геном Прокариот / Н.В. Равин, С.В. Шестаков // Вавиловский журнал генетики и селекции. – 2013. – Т.17, №4. – С. 972-982.; Квитко, К.В. Генетика микроорганизмов: учеб.пособие / К.В. Квитко, И.А Захаров; под ред. А.В. Пиневича. 2-е изд. СПб.: Изд.дом СПб.ун-та, 2012.- 268 с.

20. Гигани, О. Б. Плазмиды: монография / О. Б.,Гигани, О.О. Гигани. — М.: РУСАЙНС, 2017. — 154 с.

21. Canchaya, C. Phage as agents of lateral gene transfer / C. Canchaya [et al.] // Current Opin. Microbiol. — 2003. — V. 6. — P. 417–424.

22. Смирнов, Г.Б. Механизмы приобретения и потери генетической информации бактериальными геномами / Г.Б., Смирнов // Усп.соврем. биологии. – 2008. - Т. 28, № 1. - С. 52–76.

23. Садовский М. Г., Сенашова М. Ю., Малышев А. В. Восьмикластерная

структура геномов хлоропластов наземных растений //Журнал общей биологии. – 2018. – Т. 79. – №. 2. – С. 124-134.

24. Горбань А. Е., Зиновьев А. Ю., Питенко А. А. ViDaExpert: программа для нелинейной визуализации и анализа многомерных данных. – 2014.

Федеральное государственное автономное образовательное учреждение
высшего образования
**«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
/ Кратасюк В.А.
«22» июня 2020 г.

**БАКАЛАВРСКАЯ РАБОТА**

06.03.01. Биология

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТИПОВ ВНУТРЕННЕЙ СТРУКТУРИРОВАННОСТИ ГЕНОМОВ БАКТЕРИЙ ДЛЯ РАЗЛИЧНЫХ ДЛИН ФРАГМЕНТОВ С УЧЁТОМ ЗНАЧЕНИЙ GC-СОСТАВА

Руководитель 24.06.2020 д.ф.-м.н. М. Г. Садовский

Выпускник 24.06.2020 В.С. Густов

Красноярск 2020

Федеральное государственное автономное образовательное учреждение
высшего образования
**«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»**
Институт фундаментальной биологии и биотехнологии
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
/ Кратасюк В.А.
«22» июня 2020 г.

# БАКАЛАВРСКАЯ РАБОТА

06.03.01. Биология

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ ТИПОВ ВНУТРЕННЕЙ СТРУКТУРИРОВАННОСТИ ГЕНОМОВ БАКТЕРИЙ ДЛЯ РАЗЛИЧНЫХ ДЛИН ФРАГМЕНТОВ С УЧЁТОМ ЗНАЧЕНИЙ GC-СОСТАВА

Руководитель 24.06.2020 д.ф.-м.н. М. Г. Садовский

Выпускник 24.06.2020. В.С. Густов

Красноярск 2020