Краткий обзор генома бактерии Alicycliphilus denitrificans с использованием вычислительных методов

Бурлака А.А.

РЕЗЮМЕ

Были взяты из базы данных файлы с последовательностью генома бактерии и с информацией о ее генах. На их примере выявлены основные характеристики генома бактерии: размер, среднее количество нуклеотидов в геноме, распределение нуклеотидов в геноме, количество и распределение генов по двум цепям, место GC-перекоса, частоты к-меров и кодонов.

ВВЕДЕНИЕ

На примере Alicycliphilus denitrificans выявлены закономерности в геноме с помощью биоинформатических методов. Исследовать геном и научиться применять для этого вычислительные методы послужило целью для этой работы.

Рассматриваемая бактерия является первой выделенной в культуру из своего рода. Это грамотрицательная подвижная неспорообразующая бактерия, является факультативным анаэробом.[1] [2]

Классификация[3]:

Таблица 0. Классификация.

Таксон	Название
Домен	Bacteria
Филум	Proteobacteria
Класс	Betaproteobacteria
Отряд	Burkholderiales
Семейство	Comamonadaceae
Род	Alicycliphilus

МАТЕРИАЛЫ И МЕТОДЫ

Исходная база данных[4]

Из нее взяты файлы с информацией о генах, последовательностях и размерах генома Alicycliphilus denitrificans.

Создана электронная таблица с информацией о генах.

С помощью команд пакета EMBOSS на сервере kodomo.msu.ru были разделены изначально записанные в одном файле последовательности хромосомы и плазмиды.

с помощью команды EMBOSS: wordcount -wordsize 1 получено количество каждого нуклеотида.

С помощь команды fuzznuc –complement из того же пакета был получен столбец координат GC пар. Метод неэкономичный. На основе этого построен график количества GC пар на протяжении генома.

С помощью команды cusp того же пакета была получена таблица с информацией о частоте использования кодонов для кодирования каждой аминокислоты.

Написан сценарий, читающий входной файл и создающий файл, в котором три отделенных знаками табуляции столбца содержат информацию о количестве букв G и C на интервалах в 500 букв.

Написан сценарий, считающий количество всех возможных к-меров заданной длины, затем записывающий их в файл, готовый для экспорта в ЭТ.

Методы ЭТ MS Excel: Создана электронная таблица с информацией о генах, с помощью вспомогательного листа gene2, сортировка по значениям столбца, МАКС, ВПР, СЧЁТЕСЛИ, СЧЁТЕСЛИМН, БИНОМРАСП, ЕСЛИ, специальная вставка, графики, гистограммы, линии трендов.

Сайт для сравнения своих результатов по поиску ориджина.[5]

РЕЗУЛЬТАТЫ

ОБЩИЕ СВЕДЕНИЯ. На основе базы данных получены размеры генома, хромосомы и плазмиды (Приложение с диаграммами, Табл. 1). На основе таблицы с генами получены данные о количестве генов каждого типа (Приложение с диаграммами, Табл. 2).

РАСПРЕДЕЛЕНИЕ НУКЛЕОТИДОВ. С помощью команды посчитано количество нуклеотидов в последовательностях хромосомы, плазмиды и всего генома (Приложение с диаграммами, Табл. 3). Число букв A примерно равно числу букв T, а число букв G приблизительно равно числу букв C в последовательности одной цепочки геномной ДНК, т.е. вероятность нахождения каждого из нуклеотидов комплементарной пары на одной из цепей близка к случайной, что подтверждается биномиальным распределением. Однако существует стабильное отклонение от случайной вероятности в количестве каждой из пар. Так, процент GC пар составил 68% по хромосоме и 64% по плазмиде, а отношение AT/GC – 0.48 по хромосоме и 0.56 по плазмиде.

Полученные с помощью сценария данные о содержании GC пар на интервалах в 500 п.н. на протяжении генома позволили построить графики AT/GC и GC/AT (Приложение с диаграммами, Рис. 1). Линии тренда этих графиков говорят о стабильности этих величин, однако можно наблюдать пики этих отношений. То же самое демонстрируют графики количества GC пар на интервалах (Приложение с диаграммами, Рис. 2).

РАСПРЕДЕЛЕНИЕ ГЕНОВ. С помощью методов ЭТ Excel получены данные о распределении типов генов по цепям (Приложение с диаграммами, Табл. 4). Для каждого типа распределение генов по цепям близко к случайному, или количество таких генов слишком мало, чтобы достоверно это утверждать. Так же был проведен анализ распределения генов на протяжении генома (Приложение с диаграммами, Рис. 3). Заметной закономерности не выявлено, в среднем гены распределены равномерно. Можно выделить скопление генов транспортных РНК около 1574111 координаты и отдельные пики, означающие большое количество генов закодированных на одной цепи близко друг к другу.

GC-ПЕРЕКОС. Полученный с помощью описанного в материалах и методах сценария файл импортировался в ЭТ, после чего считались значения $\frac{G-C}{G+C}$, а затем строился график «первообразной» для этой функции.

Поскольку при усреднении исходной функции (Приложение с диаграммами, Рис. 4) получается:

$$f(x) = \begin{cases} C, x < x_0 \\ -C, x > x_0 \\ 0, x = x_0 \end{cases}$$

,где функция берется от номера интервала, а C – константа, первообразная имеет вид «угла» вершиной вверх (Приложение с диаграммами, Рис. 5). В итоге GC-перекос происходит в том месте, где исходная функция принимает значение 0, или первообразная достигает максимума.

Это все позволяет определить точку перекоса с точностью до 500 п.н.: 2379500, что довольно близко к значению, полученному на сервисе (Приложение с диаграммами, Рис. 6).

К-МЕРЫ. С помощью сценария для поиска к-меров были созданы и экспортированы в ЭТ файлы с количеством к-меров в хромосоме. После приведения данных к виду O/E (O – наблюдаемое, E – ожидаемое, $\left(\frac{1}{4}\right)^k$) и сортировки по убыванию этого отношения, получены соответствующие гистограммы (Приложение с диаграммами, Рис. 7). Заметно сразу, что GC-богатые к-меры представлены значительно больше ожидаемого, AT-богатые – меньше ожидаемого.

КОДОНЫ. После обработки таблицы с частотами кодонов в кодирующих последовательностях были получены соответствующие гистограммы (Приложение с диаграммами, Рис. 8). Самые представленные аминокислоты – аланин, лейцин, глицин, валин – алифатические. Для каждой аминокислоты наиболее представленный кодон отличается от наименее представленного большим содержанием GC пар. Аналогично к-мерам во всем геноме больше всего GC-богатых кодонов.

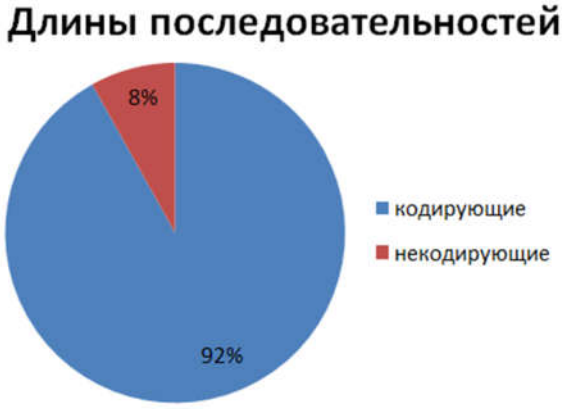

Рисунок 9.

Доля кодирующей части генома существенно выше не кодирующей (Рис. 9).

ОБСУЖДЕНИЕ

Исследование показало, что распределение нуклеотидов из комплементарной пары по каждой из цепей – случайно, как и распределение генов по цепям ДНК. А соотношение комплементарных пар не такое однородное: наблюдаются пики отношения их количеств, но процентное содержание каждой из пар постоянно на протяжении каждой из последовательностей. Вероятно, это индивидуальное свойство генома, и тогда различное соотношение нуклеотидов в хромосоме и плазмиде означает, что плазмида не принадлежала геному бактерии раньше.

GC-перекос показывает явное различие между разнонаправленными цепями ДНК. Например, репликация идет несимметрично относительно цепей, поэтому, скорее всего, точка пика графика (Приложение с диаграммами, Рис. 5) обозначает ориджин репликации. Вероятно, различие состава лидирующей и отстающей цепей указывает направление, в котором идет репликация, и в том месте, где она разнонаправленна, оказывается точкой начала репликации.

Распределение к-меров и кодонов удивительно напоминает распределение частоты слов в языке[6]. Наиболее важные и везде используемые слова представлены гораздо шире, чем специальные и необходимые в определенных ситуациях. К-меры и кодоны, имеющие отношение O/E равным единице, наверное, не имеют важного значения в геноме. Редкие – важные индикаторы и команды. Часто используемые выполняют «повседневную» функцию. Например, стоп-кодоны могут встречаться один раз в конце гена, потому что они имеют ключевое значение для производства белка: терминируют трансляцию. Предпочтение GC-богатых последовательностей, видимо, связано с поддержанием особого уровня GC пар в геноме. Почему в геноме больше всего закодировано алифатических аминокислот, мне не известно.

ЗАКЛЮЧЕНИЕ

Можно сказать, что наряду со случайными факторами (распределение нуклеотидов и генов по цепям), строение генома определяется и неслучайными: содержанием каждой из пар, предпочтением определенных к-меров и кодонов, GC-перекосом. Следовательно, можно сделать вывод, что геном обладает неслучайным смыслом, и, возможно, его сохранение обеспечивается давлением отбора.

СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

Последовательность генома Alicycliphilus denitrificans
https://kodomo.fbb.msu.ru/~burlaka.a/term1/BURLAKA_genome.fasta
Приложение с диаграммами
https://kodomo.fbb.msu.ru/~burlaka.a/term1/diagrams.pdf
ЭТ с сопроводительными материалами к обзору
https://kodomo.fbb.msu.ru/~burlaka.a/term1/BURLAKA_suppl_fin.xlsx
Дополнительные файлы к обзору
https://kodomo.fbb.msu.ru/~burlaka.a/term1/suppl/

СПИСОК ЛИТЕРАТУРЫ

1) «Isolation and characterization of Alicycliphilus denitrificans strain BC, which grows on benzene with chlorate as the electron acceptor» Sander A. B. Weelink et al. American Society for Microbiology Journals, опубликовано онлайн 27.10.2008. https://pubmed.ncbi.nlm.nih.gov/18791031/

2) «Characterization of the Polyurethanolytic Activity of Two Alicycliphilus sp. Strains Able To Degrade Polyurethane and N-Methylpyrrolidone» Alejandro Oceguera-Cervantes et al. American Society for Microbiology Journals, опубликовано онлайн 26.09.2007. https://aem.asm.org/content/73/19/6214.full

3) Классификация бактерии Alicycliphilus denitrificans на сайте:
https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=179636&lvl=3&lin=f&keep=1&srchmode=1&unlock

4) Файлы с данными о бактерии, использованные в обзоре:
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/012/530/715/GCF_012530715.1_ASM1253071v1

5) Сервис для расчета GC-перекоса:
http://genskew.csb.univie.ac.at/

6) Закон Ципфа:
https://ru.wikipedia.org/wiki/Закон_Ципфа