Journal of Bioinformatics and Computational Biology

МИНИ-ОБЗОР ГЕНОМА И ПРОТЕОМА БАКТЕРИИ ROSEATELES DEPOLYMERANS

Пушкарев С.В.

Факультет биоинженерии и биоинформатики, Московский государственный университет им. М.В. Ломоносова
Ленинские горы МГУ 1 стр. 73, г. Москва, 119234, Российская Федерация
spush@kodomo.fbb.msu.ru

Roseateles depolymerans – относительно недавно выделенный из речных вод вид β-протеобактерий. В этой статье я произвел анализ генома и протеома штамма КСТС 42856, а также продемонстрировал случайное распределение генов между прямой и комплементарной цепями ДНК с вероятностью 0.5. Были получены данные о перекрывании генов, требующие дальнейшего изучения.

Ключевые слова: Roseateles depolymerans; Анализ генома; Перекрывание генов.

1. Введение

1.1. *Описание объекта*

Roseateles depolymerans – подвижные облигатно аэробные грамотрицательные палочки из класса β-протеобактерий. Из этого класса R.depolymerans – первый вид, для которого было показано наличие бактериохлорофилла a, позволяющего им питаться фототрофно[1]. Геном R.depolymerans состоит из одной кольцевой хромосомы длиной 5681722 пар оснований[2]. Данный вид продемонстрировал способность разлагать некоторые виды полиэфиров[3], что позволяет говорить о его возможном практическом применении в сфере переработки полимеров.

1.2. *Гипотеза случайного распределения генов*

Была сформулирована следующая гипотеза: «Гены распределены между прямой и комплементарной цепями случайно с вероятностью 0.5». В дальнейшем я произвел ее проверку, используя геном R.depolymerans.

2. Материалы и методы

Для проверки гипотезы о случайном распределении генов между прямой и комплементарной цепями был написан скрипт[4] на языке python. Остальная работа была выполнена в программе MS Office Excel версии 2013 года. Данные для анализа были взяты из feature-table с ftp-сервера NCBI[5]. Использованные колонки из

feature-table: «#feature», «class», «start», «end», «strand», «product_accession», «name», «locus_tag», «product_length». Из них была составлена плоская таблица. Её и последующие расчеты можно найти в сопроводительных материалах[4] к статье. Данные о функциях белков и степени достоверности их существования были взяты в базе данных Uniprot.

2.1. *Количество генов белков и РНК по категориям*

2.1.1. *Гены белков*

С помощью фильтра «содержит ribosomal и не содержит RNA» в столбце «name» был получен первичный список белков (60 позиций). Из него были вручную удалены белки, не несущие структурной функции (4 белка, модифицирующие белки рибосомы: locus-tag RD2015_639, RD2015_1917, RD2015_3366, RD2015_4154 и 1 RHF-подобный белок[6]: locus-tag RD2015_4562. Применяя фильтр «содержит hypothetical» к столбцу «name» был получен список гипотетических белков. Посторонних результатов в нем не было обнаружено. Аналогично было проделано для транспортных белков (фильтр: «содержит transport»). Посторонних результатов в нем не было найдено за исключением «transport-associated protein»: locus-tag RD2015_53, назначение которого не удалось достоверно установить[7]. В связи с этим он не был посчитан в категории «транспортные» и был включен в «остальные». Количество всех белков было получено, считая число элементов в отфильтрованном по «содержит CDS» столбце «#feature». Вычитая ранее рассмотренные категории, я получил число белков для категории «остальные».

2.1.2. *Гены РНК*

Применяя фильтр «содержит tRNA» к столбцу «#feature» было установлено число генов, кодирующих тРНК. Аналогичным образом, заменяя «tRNA» на «rRNA» в тексте фильтра, получено количество генов, кодирующих рРНК.
Скрыв с помощью фильтра в столбце «#feature» «tRNA» и «rRNA» а также все другие строки, не имеющие отношения к генам РНК, получил число «остальной» РНК.

2.1.3. *Псевдогены и среднее число генов на 1 млн. п.о.*

Скрыв в столбце «class» все, кроме «pseudogene», получил число псевдогенов.
Скрыв фильтром все поля кроме «gene» в столбце «#feature», нашел количество всех генов, включая неработающие псевдогены. Найдя отношение полученного числа и размера генома (в млн. п.н.), установил среднее число генов на 1 млн. п.н.

2.2. *Статистика длин белков*

По столбцу «product_length» с помощью функций excel нашел наименьший и наибольший белок, вычислил среднюю длину и среднее квадратическое отклонение (Использовал «=СТАНДОТКЛОН.Г»). Размер кармана был подобран вручную для получения наиболее информативной гистограммы. Гистограмма была построена в excel. Дополнительно, отсортировав длины белков, было взято 5 самых длинных и 5 самых коротких полипептидов. С помощью функции «=ВПР» было установлено соответствие между длиной и названием белка (столбец «name» исходной таблицы).

2.3. *Распределение генов по цепям*

2.3.1. *Таблица распределения генов*

С помощью функции «=СЧЕТЕСЛИМН» были найдены значения в каждой ячейке. Чтобы найти все РНК, одним из аргументов функции было использовано выражение «*RNA».

2.3.2. *Проверка гипотезы случайного распределения генов по цепям*

Гипотеза была проверена пятью запусками скрипта[4] (аргументы: 2359 2519 1000). 2359 и 2519 – суммарные количества генов на цепях в геноме.

2.4. *Статистика по квазиоперонам*

Для значения максимального расстояния, на котором гены считаются квазиопероном, было выбрано 100 п.н. («порог»). В дальнейшем анализе были использованы колонки «start», «end» и «strand». С помощью фильтра эти три колонки были разделены на шесть в зависимости от цепи. Используя формулы excel, все квазиопероны при пороге 100 п.н. были занумерованы, для каждого было посчитано количество входящих генов. Для гистограммы было посчитано количество оперонов с числом белков от 1 до максимально найденного. Также было посчитано суммарное число всех квазиоперонов. Стоит еще отметить, что последний ген не объединяется с первым, так как расстояние между ними слишком велико[2].

2.5. *Анализ пересечения генов*

2.5.1. *Число пересечений, данные по цепям*

Используя те же исходные данные, что и в предыдущем пункте, пользуясь формулами excel, все последовательно идущие в геноме гены были протестированы на перекрывание. Стоит отметить, что последний ген не пересекается с первым, так как расстояние между ними слишком велико[2]. Далее было установлено сколько пересечений происходит между генами, расположенными на одной цепи, и сколько – между генами на разных цепях (см. таблицу 2).

2.5.2. *Пять наибольших пересечений*

Была установлена длина пересечений в случае каждой пары генов. Для пяти пар с наибольшим пересечением была построена таблица (см. Результаты и обсуждение, таблица 3). Названия генов были найдены в feature-table с помощью «=ВПР».

2.5.3. *Расчет гистограммы распределения числа пересечений генов по размеру их пересечения*

С помощью формул excel для каждого уникального размера пересечения генов было рассчитано число пар генов с таким пересечением. Была построена диаграмма (см. рис.4)

2.6. *Анализ данных uniprot*

Загрузив в uniprot product_accession всех белков, получили данные о них из базы данных. По этим данным была построена сводная таблица типа подтверждения существования белков (см. таблицу 4).

3. Результаты и обсуждение

3.1. *Количество генов белков и РНК по категориям*

Таблица 1. Количество генов белков и РНК по категориям.

	Категория	Количество генов
Гены белков	Рибосомальные	55
	Транспортные	310
	Гипотетические	1325
	Остальные	3083
Гены РНК	Транспортные	57
	Рибосомальные	12
	Остальные	2

Учитывая псевдогены (их 34), в геноме R.depolymerans 4878 гена. Количество генов на 1 млн. п.н. составило 858 гена.

3.2. *Статистика длин белков*

Длины белков варьируют в диапазоне от 29 до 4828 а.о. Средняя длина белка – 347 а.о. Среднее квадратическое отклонение – 258 а.о. Самый длинный белок (4828 а.о.) – это Hemagglutinin-related protein. Остальные отдельно рассмотренные белки (4 самых длинных и 5 самых коротких) – hypothetical.

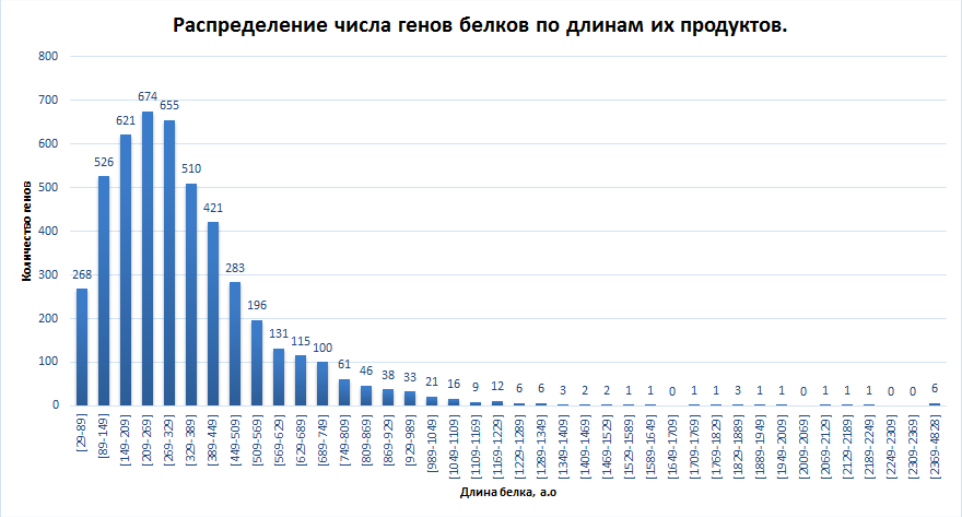

Рис.1. Гистограмма распределения количества генов белков по их длинам.

Как видно из диаграммы, белки длинной более тысячи аминокислот довольно редки. Больше всего белков в диапазоне от 209 до 269 а.о.

3.3. *Распределение генов по цепям*

3.3.1. *Таблица распределения генов*

Таблица 2. Распределение генов по цепям.

Распределение генов по цепям ДНК			
Цепь	**Гены белков**	**Гены РНК**	**Псевдогены**
+	2299	41	19
-	2474	30	15

3.3.2. *Проверка гипотезы*

По итогам пяти последовательных запусков скрипта гипотеза оказалась верной. В свете того, что принципиально две цепи ничем не отличаются с химической точки зрения, наша гипотеза подтверждает равнозначность цепей ДНК.

3.4. *Статистика по квазиоперонам*

Заметим, что при увеличении «порога» квазиоперона, количество квазиоперонов уменьшается: порог = 200 -> 2023, и наоборот: порог = 50 -> 3107. Для порога = 100 п.н. количество квазиоперонов составило 2634.

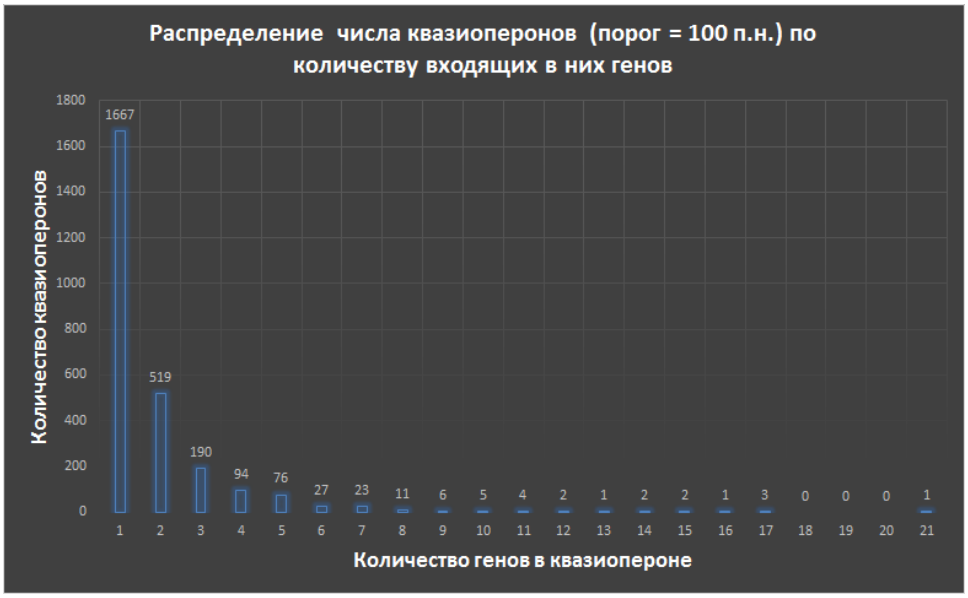

Рис.2. Гистограмма распределения числа квазиоперонов по числу входящих генов.

На гистограмме видно, что большая часть квазиоперонов при пороге 100 п.н. представляет собой отдельные гены (1667). Квазиопероны из двух генов встречаются уже примерно в три раза меньше (519). Примечательно наличие больших квазиоперонов, включающих более 10 генов.

3.5. *Анализ пересечения генов*

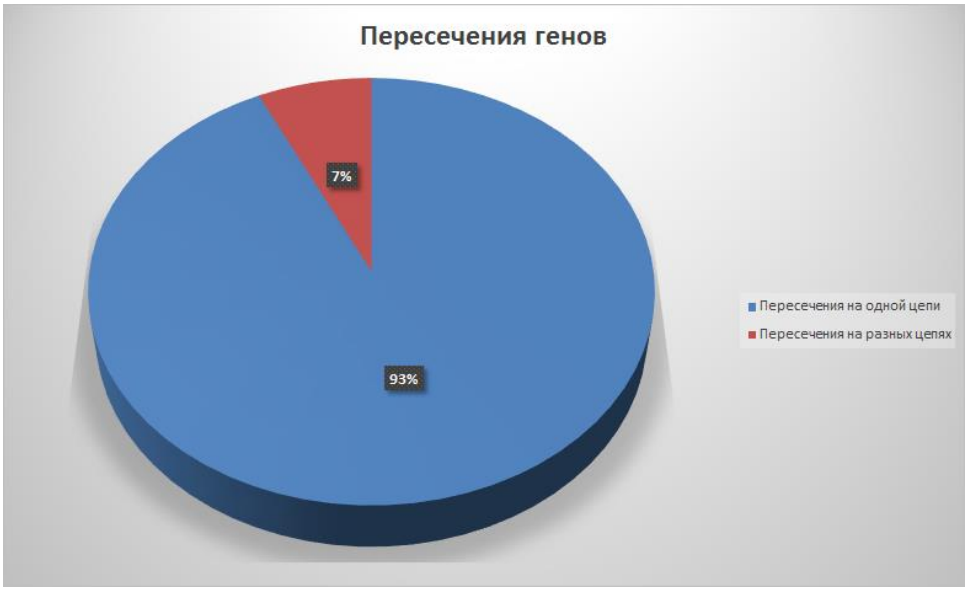

Рис.3. Процентное соотношение пересечений генов с одной и с разных цепей.

Согласно диаграмме, большинство (741 из) пересекающихся генов находятся на одной цепи. Случаи, когда они находятся на разных цепях (57), редки, однако гены из пяти пар наиболее полно перекрывающихся генов расположены именно на разных цепях. Это можно видеть из таблицы 3.

Таблица 3. Пять пар наиболее перекрывающихся генов.

Координаты пары*	**Перекрывание (п.н.)**	**Взаимная ориентация**	**Первый ген**	**Второй ген**
2399445-2401428	161	На разных цепях	DNA topoisomerase IB (ALV06523.1)	D,D-heptose 1,7-bisphosphate phosphatase (ALV06524.1)
2159519-2161455	160	На разных цепях	hypothetical protein (ALV06313.1)	pseudogene (-)
1296555-1298067	89	На разных цепях	membrane protein (ALV05636.1)	N-linked glycosylation glycosyltransferase PglG (ALV05637.1)
208575-211662	86	На разных цепях	Chloride channel core (ALV04689.1)	Histidine kinase(ALV04690.1)

2865077-2867345	74	На разных цепях	LysR family transcriptional regulator (ALV06916.1)	Mercuric reductase (ALV06917.1)

*Приведены координаты начала первого гена и конца второго.

В целом, для всего генома было показано, что чаще всего гены перекрываются на 4 нуклеотида.

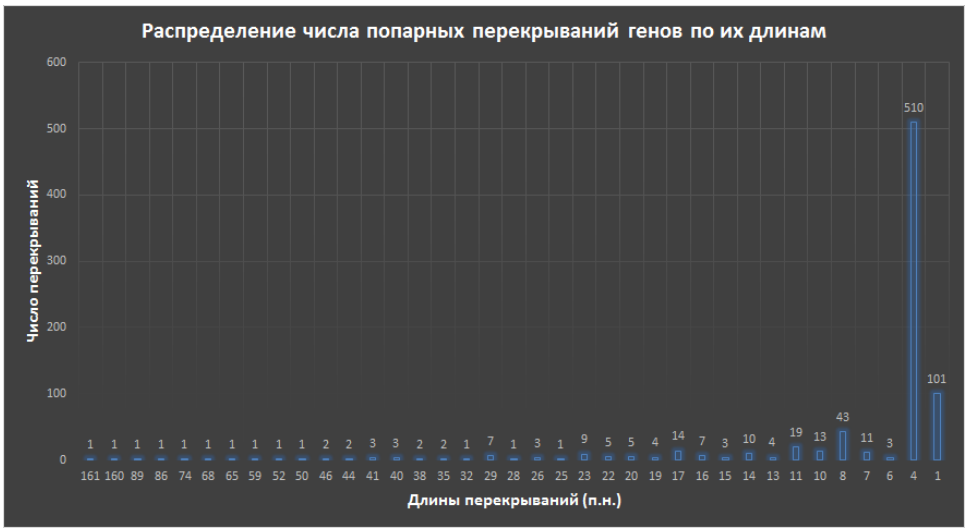

Рис.4. Описывает встречаемость пар генов в геноме в зависимости от величины их перекрывания.

3.6. *Анализ данных uniprot*

Таблица 4. Подтверждение существования белка из базы uniprot.

Тип подтверждения	Установлено из гомологии	Белок предсказан*	Свидетельство на уровне транскрипта
Количество генов	1292	3474	2
Процентное соотношение	27,10%	72,86%	0,04%

*Подробнее о типах подтверждения можно узнать на сайте uniprot: http://www.uniprot.org.

Поскольку взятый геном не являлся референтным, было вполне ожидаемо увидеть большое количество предсказанных белков. Неожиданностью было, что 4773 идентификаторам белков было сопоставлено 4768 белков.

4. Заключение

В этой статье был произведен краткий анализ генома и протеома бактерии R.depolymerans. Дополнительно, были получены данные, пробуждающие любопытство и требующие дальнейшего теоретического обоснования. Сформулирую полученные необъясненные результаты в виде вопросов:

1) Почему перекрываний генов, расположенных на одной цепи, намного больше перекрываний генов с разных цепей?
2) Почему для длин перекрываний, в остатке при делении на три дающих 2, пар генов больше, чем для перекрываний дающих 1 при рассмотрении рис.4 для длины перекрываний от 7 до 20? Почему совсем нет перекрываний длины 5?
3) Является ли простым совпадением тот факт, что самые длинные перекрывания получены для генов, расположенных на разных цепях?

5. Сопроводительные материалы

Сопроводительные материалы могут быть найдены по адресу: http://kodomo.fbb.msu.ru/~spush/term1/pr13-14.html.

6. Список литературы

1. Suyama T. et al., Roseateles depolymerans gen. nov., sp. nov., a new bacteriochlorophyll a-containing obligate aerobe belonging to the beta-subclass of the Proteobacteria. Int J Syst Bacteriol. 1999 Apr; 49 Pt 2:449-57.
2. NCBI: Roseateles depolymerans strain KCTC 42856, complete genome; GenBank: CP013729.1.
3. Aamer Ali Shah et al., Purification and properties of novel aliphatic-aromatic co-polyesters degrading enzymes from newly isolated Roseateles depolymerans strain TB-87. Polymer Degradation and Stability Volume 98, Issue 2, February 2013,p. 609-618.
4. Сопроводительные материалы могут быть найдены по адресу http://kodomo.fbb.msu.ru/~spush/term1/pr13-14.html.
5. Assembly: GCA_001483865.1.
6. ID в UniprotKB – A0A0U3NA26 (A0A0U3NA26_9BURK).
7. ID в UniprotKB – A0A0U3L916 (A0A0U3L916_9BURK).