Обзор генома бактерии *Burkholderia cepacia JBK9*

Быкова Дарья[1]

[1]Факультет биоинженерии и биоинформатики Московского государственного университета

РЕЗЮМЕ

В данной работе был проведён анализ генома и протеома бактерии *Burkholderia cepacia* штамм JBK9. В ходе работы была опровергнута гипотеза о случайном распределении генов по цепям ДНК, найдено несколько фаговых вставок и потенциальных оперонов.

1 ВВЕДЕНИЕ

Burkholderia cepacia – аэробная грамотрицательная бактерия, потенциально опасная для людей с ослабленным иммунитетом. В частности, вызывает пневмонию у больных кистозным фиброзом. [1,2] Также является патогеном некоторых растений, таких как лук, где была впервые обнаружена, и табак[2]. Геном *B. cepacia* имеет большой размер относительно других бактерий: состоит из трёх хромосом, суммарно содержащих 8,42181 млн. п.н.

2 МАТЕРИАЛЫ И МЕТОДЫ

В данной работе использовалась программа Microsoft Office Excel 2013. Также были написаны вспомогательные программы на Python. Для получения необходимых данных использовались базы данных NCBI (идентификаторы хромосом - CP013730.1, CP013731.1, CP013732.1) и Uniprot.

3 РЕЗУЛЬТАТЫ

3.1 Распределение генов белков и генов РНК по категориям

В таблице 1 представлено распределение генов по категориям: гены рибосомальных белков, транспортных белков, гипотетических, всех остальных, а также гены тРНК, рРНК и других РНК. В данном случае в категорию «другие РНК» попадает всего один ген, кодирующий РНКазу Р.

Категория	Кол-во генов
рибосомальные белки	68
транспортные белки	817
гипотетические белки	1644
другие белки	4815
тРНК	69
рРНК	18
другие РНК	1

Таблица 1. Количественное аспределение генов

Псевдогены не учитывались. Классификация генов белков проводилась при помощи текстового фильтра («ribosom» и не «non-ribosomal» для рибосомальных белков и «transport» - для транспортных). Примерное число генов на 1 млн. п.н. составило 876.

3.2 Распределение длин белков в протеоме бактерии

Был проведён анализ протеома бактерии, построена гистограмма, отражающая распределение длин белков (рис. 1). Как видно из приведённой гистограммы, наибольшее количество белков лежит в диапазоне длины 150 – 525 аминокислот. Средняя длина составила 352 а.к., медиана – 312 а.к., стандартное отклонение составило примерно 228 а.к., что говорит о высокой вариативности внутри рассматриваемого множества белков. Гипотетические белки при составлении данной гистограммы не учитывались. Самые длинные белки – пептидогликан-связывающий белок (4565 а.к.) и синтетаза нерибосомных пептидов (4467 а.к.). Самый короткий - УДФ-3-О-(3-гидроксимиристоил) глюкозамин N-ацилтрансфераза (28 а.к.). Большие размеры белков можно объяснить тем, что это необходимо для успешного связывания других крупных молекул: нерибосомные пептиды часто имеют разветвлённые и/или циклические структуры [3], пептидогликан или муреин имеет полимерное строение. С другой стороны, малые размеры полезны для транспортных белков, а также для некоторых белков-ферментов.

3.3 Распределение генов по цепочкам ДНК

Результаты анализа распределения генов белков и РНК, а также псевдогенов по прямой и комплементарной цепочкам ДНК представлены в таблице 2.

Гены	+	-
белок-кодирующие	3514	3830
псевдогены	41	49
рРНК	9	9
тРНК	34	35

Таблица 2. Распределение генов по цепям ДНК

Из таблицы следует, что гены тРНК и рРНК распределяются по цепочкам ДНК равномерно, т.е., по-видимому, случайным образом. Впрочем, их количество недостаточно, чтобы судить об этом. В работе была проверена гипотеза о том, что белок-кодирующие гены

*To whom correspondence should be addressed.

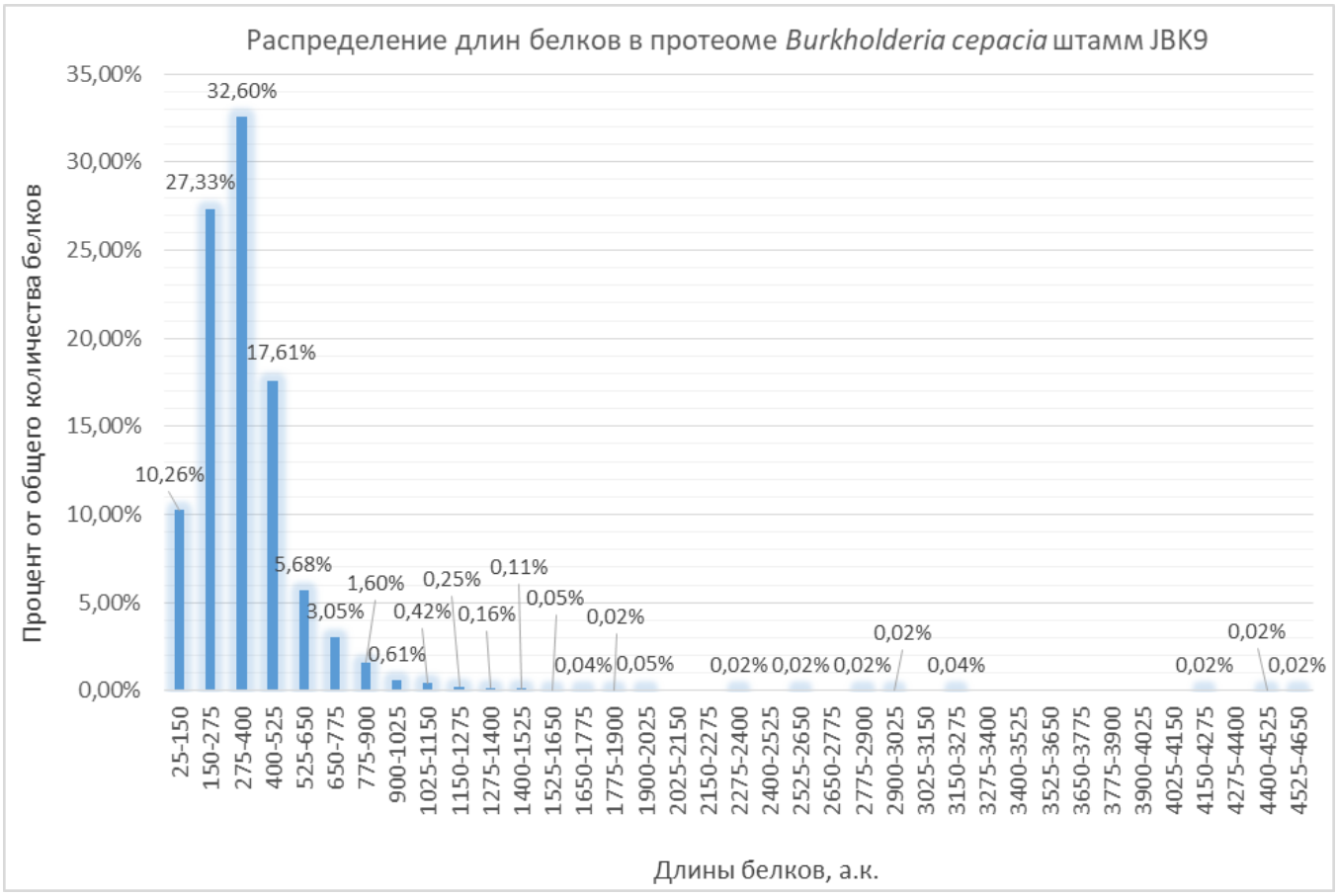

Рис.1. Распределение длин белков в протеоме бактерии

также распределены по цепям случайно. Была написана вспомогательная программа, считающая вероятность полученного отклонения от ожидаемого результата (половина от общего количества генов) при условии, что вероятность обоих событий - «ген на прямой цепи» и «ген на комплементарной цепи», - равна 50%. Было выявлено, что вероятность такого отклонения (158) не превышает 0,1%, что опровергает первоначальную гипотезу о случайном распределении. Причины данного явления не ясны, так как причисление цепи к «+» или «-» не несёт биологического смысла. Существуют два пути объяснения: человеческий фактор (ошибки в методе либо технические особенности аннотирования генов) либо реальное природное явление. Автор, исходя из предположения, что гены данной бактерии действительно распределены по цепям ДНК неслучайно, выдвинул идею о том, что по суммарной длине гены прямой и комплементарной цепи всё же совпадают. Однако эта гипотеза не подтвердилась: суммарная длина белок-кодирующих генов на комплементарной цепи оказалась много больше, чем на обратной (3,450822 млн. п.н. против 3,786861 млн. п.н.). По количеству квазиоперонов комплементарная цепь также обгоняет прямую. Возможно, объяснение можно построить в другую сторону: оперон представляет собой группу генов, вовлечённых в один и тот же процесс и лежащих на одной цепи. Опероны формировались в процессе эволюции, и, возможно, «главный» ген «вытягивал» за собой остальные, т.е. возникала перегруппировка генов в цепях, в ходе которой гены одного оперона выстраивались друг за другом на одной из них.

3.4 Поиск квазиоперонов

В данной работе последовательность генов считалась квазиопероном, если:

1) Гены расположены друг за другом на одной и той же цепи ДНК
2) Расстояние между ними не больше 100 п.н.

Квазиоперон мог состоять и из одного гена. Результаты приведены в таблице 3 и на рис.2. Также была построена гистограмма, отражающая распределение количества генов в квазиоперонах.

порог	количество	среднее	максимальное	медиана

	квазиоперонов			
100	4286	1,71	26	1

Таблица 3. Квазиопероны

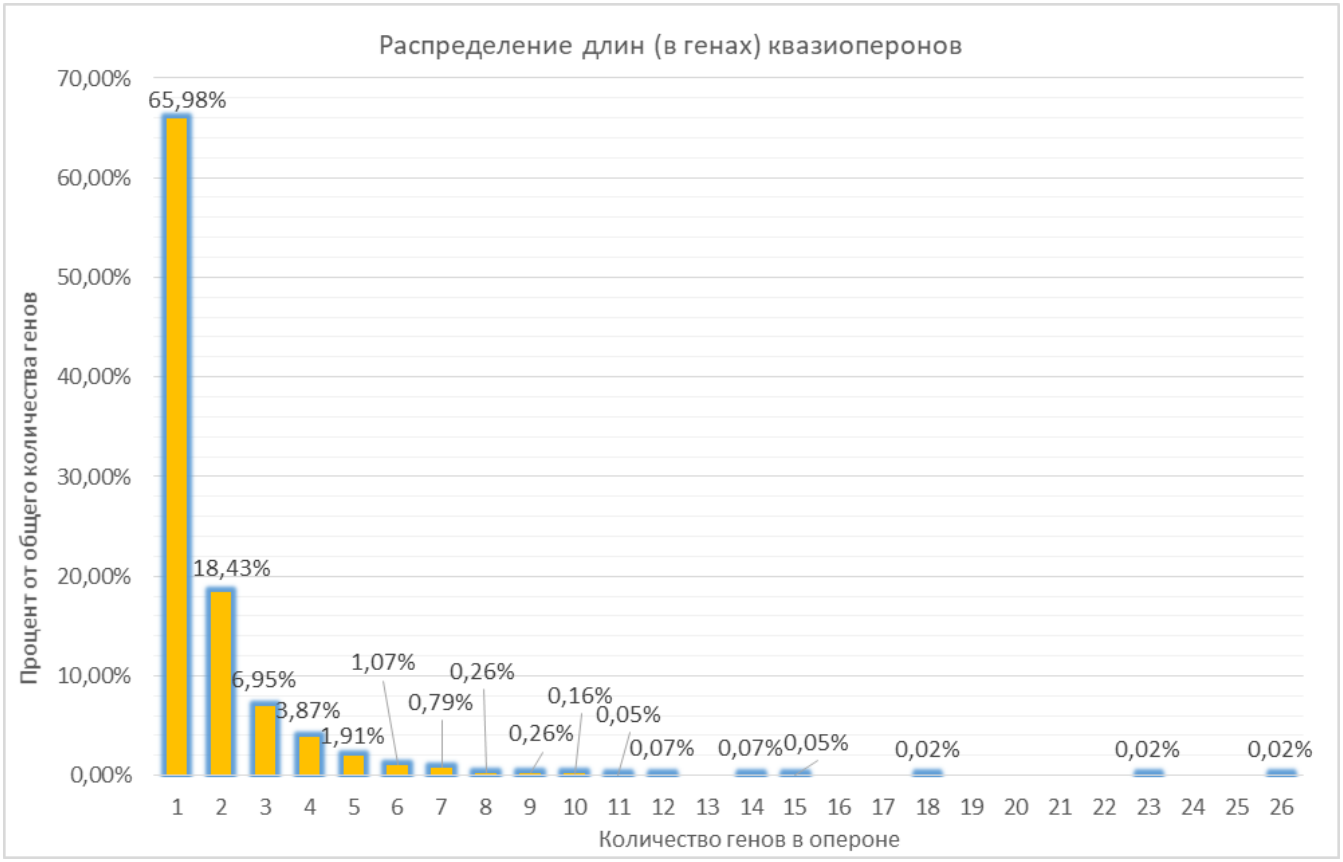

Рис. 2. Распределение квазиоперонов

Как нетрудно заметить из приведённых данных больше половины генов расположены относительно обособлен

но от других. При увеличении длины квазиоперонов их встречаемость резко снижается. Оперонов, состоящих из двух генов, уже примерно в 3,5 раза меньше, чем единичных генов (18,43% против 65,98%). Резкое снижение встречаемости происходит вплоть до длины в пять генов, более длинные опероны встречаются уже более равномерно. Подавляющее большинство квазиоперонов укладывается в рамки длины 1-6 генов. Полученные результаты, конечно, предварительны и носят характер довольно грубой оценки. Для получения более полной информации следовало бы рассмотреть все опероны, найти в базах данных информацию о взаимосвязи между близко расположенными на цепи белками. В данной работе автор ограничился рассмотрением только самых длинных квазиоперонов. Самый длинный квазиоперон (26 генов, координаты 1678455, 1699772 на хромосоме 1), по-видимому, оказался фаговой вставкой: среди 26 белков встречается 8, аннотированных как «phage tail protein» (белок хвоста фага), а также «portal protein», «major capsid protein E», «phage baseplate protein», - тоже белки фагов.

В некоторых других длинных квазиоперонах замечено многократное повторение одних и тех же генов. Так в квазиопероне длиной в 18 генов (координаты 3128189, 3156944 на хромосоме 2) встречается 7 генов, кодирующих белки, относящихся к семейству «type VI secretion protein», а в квазиоперон с координатами 70084, 76755 на хромосоме 1 полностью состоит из рибосомальных белков, а также в его состав входит фактор инициации трансляции. Интересно, что, помимо уже описанной, была обнаружена ещё одна фаговая вставка с координатами 241426, 252753 на хромосоме 1. Была проведена попытка поиска некоторых из полученных квазиоперонов в базах данных, но

оказалось, что для рассматриваемого штамма таких данных не приводится. Впрочем, есть сходные данные для других штаммов того же вида. Автор попытался использовать эту информацию, предварительно проведя выравнивание и заметив высокую степень сходства между некоторыми штаммами. Однако в дальнейшем оказалось, что у разных бактерий даже похожие участки генома аннотированы по-разному. Возможно, это связано с тем, что для рассматриваемой бактерии приведено ещё недостаточно экспериментальных данных. Таким образом, вопрос об операх остаётся открытым и требует дальнейшего исследования.

3.5 Статистические данные о пересечениях генов

Было рассчитано количество генов, имеющих пересечения. В таблице 4 приведены результаты расчёта. В подавляющем большинстве случаев пересечение происходит со сдвигом рамки считывания. Чаще пересекаются гены, лежащие на одной цепи, чем на разных. Однако в пяти случаях из шести, когда длина пересечения кратна трём, пересекающиеся гены располагаются на разных цепях. В шестом случае мы имеем дело с псевдогеном.

количество пересечений	% пересечений	Максимальная длина пересечения	Сдвиги рамки считывания	на одной цепи	на разных цепях	Средняя длина пересечения	Медиана
780	10,37%	83	774	742	38	7	4

Таблица 4. Пересечения генов

3.6 Статистика белков по категориям достоверности их существования

Для получения информации о существовании белков использовалась база данных Uniprot. Как оказалось, экспериментальных данных о существовании белков рассматриваемого штамма приведено не было. Большинство белков было предсказано без опоры на экспериментальные данные об их экспрессии или гомологию (452), остальные – предсказаны по гомологии (251).

4 БЛАГОДАРНОСТИ

Автор выражает благодарность преподавателям кафедры биоинформатики Факультета биоинженерии и биоинформатики Московского государственного университета за пояснения, облегчившие работу, а также студентам того же факультета Белоусовой Евгении и Перевощиковой Кристине за содержательные обсуждения результатов.

5 ССЫЛКИ

[1]https://emedicine.medscape.com/article/237122-overview#showall
[2]https://en.wikipedia.org/wiki/Burkholderia_cepacia_complex
[3] https://en.wikipedia.org/wiki/Nonribosomal_peptide
[4] Complete genome sequences for 59 burkholderia isolates, both pathogenic and near neighbor. Johnson SL, et al. Genome Announc 2015 Apr 30
[5]https://www.ncbi.nlm.nih.gov/genome/10703?genome_assembly_id=262259
[6] http://www.uniprot.org/help/protein_existence

6 СОПРОВОДИТЕЛЬНЫЕ МАТЕРИАЛЫ

http://kodomo.fbb.msu.ru/~darya_by/DBykova_suplimentary.xlsx