Функциональная аннотация и сравнительный анализ геномов *Streptomyces* spp.

Марина Борисова
МГУ им.М.В.Ломоносова, ФББ
ИППИ им.А.А.Харкевича РАН, УНЦ «Биоинформатика»
biomidgyy@gmail.com

Дмитрий Малько
Институт общей генетики им.Н.И.Вавилова РАН
dmitry.malko@gmail.com

Аннотация

Актинобактерии — это большая группа семейств, родов и видов, распространённых во многих средах обитания: встречаются преимущественно в почве, а также и в морских осадках, выступают в качестве симбионтов и паразитов. Виды рода Streptomyces — Грам-положительные нитчатые бактерии, которые производят множество вторичных метаболитов, в том числе антибиотиков. Эти бактерии продуцируют более двух третей всех клинически важных антибиотиков и потому вызывают высокий фармакологический и индустриальный интерес.

Филогенетический анализ двух штаммов из рода Streptomyces, выделенных из разных губок, показал их близость (99,9% идентичности генов 16S рРНК). Секвенирование, аннотация и анализ геномов этих штаммов Streptomyces показали, что у них был общий почвенный предок с Streptomyces albus. Несмотря на высокое сходство генов «домашнего хозяйства» в изучаемых штаммах и S.albus, некоторые гены биосинтеза вторичного метаболизма могут сохраняться в одном и отсутствовать в другом штамме, подтверждая гипотезу о различных путях эволюции этих Streptomyces.

Наши исследования проясняют эволюцию изучаемых стрептомицет в том числе при переходе от почвенного образа жизни к морскому.

1. Введение

1.1. Общие сведения

На сегодняшний день в таксономической базе данных NCBI род *Streptomyces* насчитывает 806 видов (на декабрь 2011 [1]). Из такого огромного количества видов только для восьми доступны полные последовательности геномов.

Среды обитания *Streptomyces*, почва и осадочные слои морской воды, разнообразны и различаются наличием питательных веществ. Их предпочтения в питании не уникальны и у них есть конкуренты. Секвенирование геномов ряда стрептомицет показало, что все они содержат гены биосинтеза нескольких вторичных метаболитов [2–4]. Гены, кодирующие ферменты для производства отдельных вторичных метаболитов, ко-локализованы и часто объединены с одним или двумя генами, регулирующими их транскрипцию [5].

1.2. Адаптация к морской среде обитания

Способность бактерий приспосабливаться к значительным внешним изменениям осмотического давления является основой выживания [6]. Осмоадаптация бактерии часто включает в себя накопление растворимых веществ, таких как глицин-бетаина. Эти молекулы организм может накапливать, как производя их сам, так и получая из окружающей среды. Помимо этого бактерия может пережить осмотический стресс, сопрягая транспорт веществ с транспортом воды [6]. Ещё один способ регулировать транспорт веществ — пропускать его через механочувствительный канал большой проводимости. Этот связанный с мембраной канал, регулируемый растяжением, встречается во многих бактериях и считается экстренным способом сбрасывания тургорного давления после внезапного осмотического шока [7]. Подобные структуры, возможно, помогают хозяевам преодолевать трудности при переходе от морской среды в пресную.

Есть и ещё один способ адаптации: накачивать внутрь клетки Na^+, то есть производить дополнительные транспортёры Na^+ [8].

1.3. Сложности геномного анализа рода *Streptomyces*

Современные методы секвенирования порождают серьёзные вычислительные проблемы, связанные с короткими длинами секвенированных фрагментов, огромным объёмом данных и метод-специфичными ошибками. Такие ошибки часто встречаются в областях, содержащих большое количество ГЦ-пар. Последовательности нуклеотидов Г-Ц образуют стабильную вторичную структуру, которая плохо разрушается и мешает отжигу праймеров. ПЦР таких структур часто не даёт нужных продуктов [9]. Помимо этого, к метод-специфичным ошибкам относится сдвиг рамки считывания в гомополимерных регионах. Восстановить области с потерями и ошибками нельзя. Следовательно, такие данные влияют на сравнительно-геномный анализ с использованием филогенетических методов [10].

Перечисленные проблемы особенно влияют на функциональную аннотацию геномов рода *Streptomyces,* так как последние содержат много белков с повторами, мультиблочные структуры, например, поликетид синтазы, синтазы нерибосомных пептидов (NRPS) и серин-треониновые киназы. Поэтому чем более фрагментированным получается геном, тем больше частота ложно-негативных предсказаний по гомологии [10].

Ещё одна серьёзная проблема аннотации плохо секвенированных, ГЦ-богатых геномов — ошибки определения открытых рамок (orf) считывания, возникающие вследствие малого числа стоп-кодонных триплетов [3].

Программы отбирают наиболее вероятные кодирующие области. Одним из параметров отбора перекрывающихся orf является длина: чем длиннее, тем лучше. В ГЦ-богатых геномах стоп-кодоны встречаются редко, и часто возникает ситуация, когда длинный *orf1* лежит комплементарно к относительно короткому, но — при дальнейшем исследовании — осмысленному *orf3* (Рисунок 1). При функциональной аннотации отбор *orf1* усиливается ещё и в силу большого количества очень близких последовательностей в базе данных NCBI (не все геномы проверяются после автоматического аннотирования). В результате аннотирования генома получаем записанный в геном длинный «гипотетический белок» и теряем короткий осмысленный, со значимой функцией.

Данную проблему можно решить, если на первом шаге использовать курируемые базы данных (например, SwissProt); на втором — использовать базу данных NCBI без повторов; а на третьем — рассматривать при решении конфликтных ситуаций функции белков.

1.4. Вторичный метаболизм

Благодаря своим антибактериальным, противораковым противогрибковым, противоглистным и иммуносупрессивным веществам [9], [11] стрептомицеты являются объектом активного изучения. Естественно, что главный интерес исследователей вызывает устройство вторичного метаболизма.

Химия вторичного метаболизма структурно разнообразна и основана на ряде молекулярных скелетов, таких как поликетиды, β-лактамы, пептиды и пирролы [6], [7]. Считают, что основная природная функция вторичных метаболитов — это антибактериальная [12], торможение роста конкурирующих организмов. Однако это не единственная функция. Вторичные метаболиты также могут быть сигнальными молекулами, активируя мутационные процессы и подвижность, ингибируя ощущения кворума [13].

В настоящей работе были исследованы актинобактерии рода *Streptomyces*, выделенные из двух разных видов морских губок, собранных на дне фьорда Тронхеймс (Норвегия). Филогенетический анализ генов 16S рРНК из

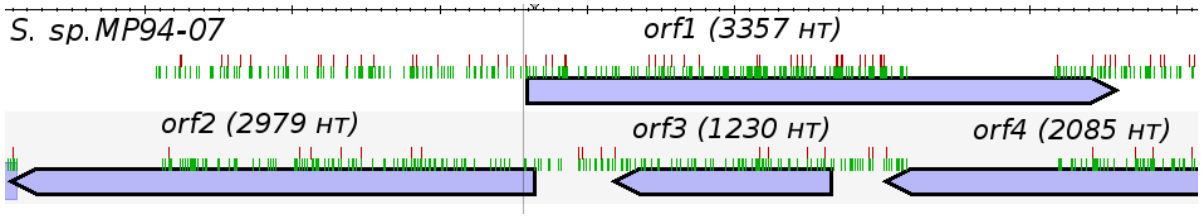

Рисунок **1.** Пример решения отбора перекрывающихся **orf** в пользу неверно аннотированной**.** orf1 — гипотетический белок, orf2 — FscD (один из белков G-рецептора), orf3 — модульная поликетидсинтаза, orf4 — FscD.

секвенированных последовательностей показал, что выделенные штамы близки друг к другу и *S.albus*.

Была поставлена задача изучить два штамма, относящихся к роду *Streptomyces*: проаннотировать геномы и исследовать гены вторичного метаболизма.

2. Материалы и методы

2.1. Геномы штаммов

Два секвенированных штамма *Streptomyces* sp.MP94-07 и *Streptomyces* sp.MP94-10 были предоставлены группой Сергея Зотчева (Норвежский университет наук и технологий, Трондхейм, Норвегия).

Для сравнительно-геномного анализа также были использованы геномы видов из рода *Streptomyces* с полными последовательностями из GenBank (на сентябрь 2011) [1]: *Streptomyces avermitilis* MA-4680, *Streptomyces bingchenggensis* BCW-1, *Streptomyces coelicolor* A3(2), *Streptomyces flavogriseus* ATCC 33331, *Streptomyces griseus subsp. griseus* NBRC 13350, *Streptomyces scabiei* 87.22, *Streptomyces sp.* SirexAA-E и *Streptomyces violaceusniger* Tu 4113. Все перечисленные организмы являются почвенными.

Помимо этого были использованны геномы *Streptomyces albus* J1074 (наземный), а также отдельные локусы: энтероцин (морской бактерии *Streptomyces maritimus*) и последовательности фрагментов из GenBank (AJ561198.1, U82965.2, Y16952.3 и AY116644.1), выделенных из почвенных бактерий.

Всего в сравнительно-геномном исследовании было использовано 9 полных геномов и 2 новых генома рода *Streptomyces,* выделенных из гомогената губок *Geodia barretti* и *Phakellia ventilabrum.*

2.2. Таксономический и филогенетический анализ

Для построения таксономического дерева были использованы последовательности 16S рРНК из полных геномов и полученные при секвенировании 16S рРНК новых штаммов. С помощью MUSCLE [14] было построено множественное выравнивание 11 последовательностей, с помощью алгоритма объединения соседей из пакета PHYLIP [15] было расчитано дерево. Все расчёты проводились с параметрами по умолчанию. Для визуализации дерева была использована программа Dendroscope [16].

2.3. Функциональная аннотация. Этап первый

Геномы *S.* sp.MP93-07 и *S.* sp.MP94-10 были аннотированы с последовательным использованием Mauve [17], BLAST [18] и GLIMMER [4].

Были построены ортологические кластеры в пакете OrthoMCL [19]. Каждый из 13045 кластеров ортологичных генов в четырёх геномах (*S.* sp.MP93-07, *S.* sp.MP94-10, *S.albus J1074* и объединённом) был описан с помощью четверного паттерна. Если кластер содержал несколько генов из одного генома, избыточные гены считались паралогами. Гены, не попавшие в кластеры, являются видоспецифичными и часть из них, возможно, являются ошибками GLIMMER и других программ распознавая генов.

2.4. Функциональная аннотация. Этап второй

Результаты построения ортологических кластеров были использованы для общего функционального анализа геномов. Каждой группе ортологов была приписана общая фукнция. Если присутствовали разные аннотации, то выбиралась наиболее специализированная. Далее все функции классифицировались по основным функциональным категориям COG [20].

Из всех групп функций наибольший интерес представляет группа вторичного метаболизма, а именно биосинтез антибиотиков. Для этого функция каждого ортологического кластера была изучена с точки зрения участия в биосинтезе антибиотиков. Наиболее вероятные функции были выписаны и дальше рассматривалась их геномная локализация. Для этого рассматривалось генное окружение каждого выписанного ортологического кластера. Если ортологи гена (ген 1) рядом с исходным (ген 2) располагались в двух и более геномах одинаково относительно гена 2, то ген 1 считался участником локуса гена 2. Из большого числа полученных локусов были выбраны те, у которых кодируемые белки вероятнее всего участвуют в биосинтезе антибиотиков.

2.5. Поиск Na^+ транспортёров

Для определения потенциальных Na^+ транспортёров из базы данных NCBI были извлечены все белки, проаннотированные как «Na+ symporter» или «Na+ antiporter». Кроме того, белки, проаннотированные как «Na+ transporter» были извлечены из другой базы

данных, TDCB [21]. Извлечённые белки были использованы для определения с помощью BLAST ортологических кластеров, содержащих Na^+ транспортёры. Параметры поиска совпадали с параметрами, используемыми при аннотации геномов.

3. Результаты и обсуждения

3.1. Филогенетический анализ

Предварительный анализ генов 16S рРНК из предоставленных геномов *S.* sp.MP94-07 из *P.ventilabrum* и *S.* sp.MP94-10 из *G.barrette* показал близость видов (совпадение=99,9%). С точки зрения эволюции было интересно изучить эти два генома на наличие общего предка до пространственного разделения.

Геном *S.* sp.MP94-07 был собран из 413 неперекрывающихся последовательностей общим размером около 7.07Mb. Геном *S.* sp.MP94-10 меньше, 6,89Mb, был составлен из 1133 подпоследовательностей. После предсказания открытых рамок считывания (orf) и генов с использованием гомологии и статистических методов было получено 6986 и 6091 генов для *S.* sp.MP94-07 и *S.* sp.MP94-10, соответственно.

Предварительный анализ генов «домашнего хозяйства» подтвердил близкое родство двух изучаемых геномов: 99-100% совпадения на нуклеотидном и белковом уровнях. Помимо этого, при выравнивании 16S рРНК генов остальных геномов оказалось, что ближайшим предком новых штаммов является *S.albus*.

3.2. Кластеры ортологичных генов

Были построены кластеры ортологичных генов для анализа геномного содержания трёх штаммов (*S.* sp.MP94-07, *S.* sp.MP94-10 и *S.albus*) (Рисунок 2).

На основании диаграммы проведён анализ потеряных генов. Для этого анализа было сделано допущение, что если ген присутствует в одном из трёх штаммов и хотя бы в одном штамме из объединённого генома, то такой ген присутствует и в геноме общего предка трёх штаммов, рассматриваемых отдельно. Всего получилось, что потери генов составляют: 13% в линии *S.albus*, 11% общих для *S.* sp.MP94-07 и *S.* sp.MP94-10, 5% в линии *S.* sp.MP94-07 и 19% — в *S.* sp.MP94-10.

Предположения о приобретённых генах сделать сложно, так как в геномах имеется много генов, особенно геном-специфичных, которые могут повлиять на перепредсказание.

3.3. Функциональная классификация

Было изучено функциональное разделение генов, общих для разных совокупностей геномов, на основании диаграммы Венна (Рисунок 2).

В качестве меры отклонения от случайности была использована величина, равная квадрату разности между ожидаемым (O) и наблюдаемым (H) числом в ячейках таблицы (Таблица 1), делённому на ожидаемое.

$$\frac{(O-H)^2}{O} > 13 \qquad (1).$$

Это отношение должно быть больше 13. Только при этом условии результат остаётся значимым при сокращении данных до таблицы сопряжённости 2x2 и применения теста χ^2 с p=0,05, с учётом поправки Бонферони для множественного тестирования. Категории, представленные во всех трёх исследуемых штаммах и хотя бы в одном другом, т. е.

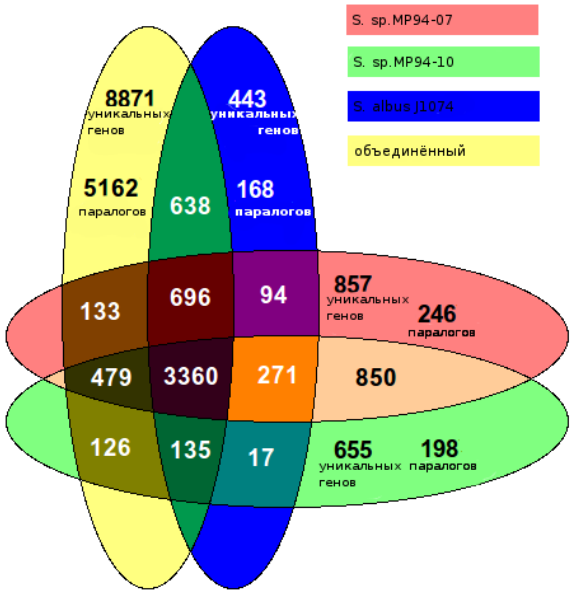

Рисунок **2.** Кластеры ортологичных генов (в четырёх геномах: S. sp.MP94-07, S. sp.MP94-10, S.albus и объединённый).

перепредставленные во всех *Streptomyces*, были «регуляция» и, как и ожидалось, «плохо характеризованные гены», имеющие только общую аннотированную функцию (Таблица 1). С другой стороны, «вторичный метаболизм» перепредставлен среди общих генов *S.* sp.MP94-07 и *S.* sp.MP94-10, но не среди их общих генов с *S.albus*. Вероятно, это определяет штамм-специфичную, спорадическую потерю некоторых генов, несмотря на близость этих штаммов.

3.4. Приспособленность к морской воде

При выделении и культивировании предоставленных видов было отмечено, что штаммы *S.* sp.MP94-07 и *S.* sp.MP94-10 заметно лучше пролиферируют и размножаются на среде с добавлением морской воды.

Для объяснения этого приспособления к морской воде были исследованы кандидаты на Na^+-транспортёры, закодированные в новых геномах, но не в *S.albus*. В результате были найдены два гена из семейства KefB.

3.5. Гены вторичного метаболизма

При сравнительно-геномном анализе ко-локализации генов (см. Материалы и Методы) с функцией биосинтеза антибиотиков были найдены следующие локусы: локусы лантибиотика, DpgC и аминокумариновый, тиопептидный, трипептидный, валанимициновый, грамицидиновый, кандициновый, энтероциновый локусы и ещё несколько менее специфичных. Всего описано 29 локусов.

Ещё один подход к поиску биосинтетических локусов — прямой поиск интересующих путей. Для этого были использованы уже известные локусы из GenBank. После поиска ортологов в исследуемых геномах, определялись новые локусы. Так были определены локусы биосинтеза аминокумарина, трипептида (нерибосомного) и энтероцина.

4. Выводы

1. Филогенетический анализ показал, что у новых штаммов *S.* sp.MP94-07 и *S.* sp.MP94-10 из губок *Phakellia ventilabrum* и *Geodia barretti*, соответственно, есть общий ближайший родственник, *S.albus* J1074.
2. Потери генов в изучаемых геномах по сравнению с общим предком составляют: 11% общих для *S.* sp.MP94-07 и *S.* sp.MP94-10, 5% в линии *S.* sp.MP94-07 и 19% — в *S.* sp.MP94-10.
3. Классы функций «регуляция» и «плохо характеризованные гены» перепредставлены в *Streptomyces* spp., использованных в данной работе, а «вторичный метаболизм» перепредставлен общими генами *S.* sp.MP94-07 и *S.* sp.MP94-10, но не общими с другими геномами.
4. Приспособленность к морской воде можно объяснить наличием двух генов, кодирующих Na^+-транспортёры семейства KefB, ортологов которых нет в почвенном родственнике выделенных штаммов, *S.albus* J1074.
5. В исследованных штаммах описано 29 локусов синтеза антибиотиков: локусы лантибиотика, DpgC и аминокумариновый, тиопептидный, трипептидный, валанимициновый, грамицидиновый, кандициновый, энтероциновый локусы и ещё несколько менее специфичных.

5. Благодарности

Работа выполнена под руководством д.б.н., проф. М.С.Гельфанда.

Ссылки

[1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D32–D37, Jan. 2011.

[2] M. Nett, H. Ikeda, and B. S. Moore, “Genomic basis for natural product biosynthetic diversity in the actinomycetes,” *Nat Prod Rep*, vol. 26, no. 11, pp. 1362–1384, Nov. 2009.

[3] K. J. Hoff, “The effect of sequencing errors on metagenomic gene prediction,” *Bmc Genomics*, vol. 10, no. 1, p. 520, 2009.

[4] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, “Improved microbial gene identification with GLIMMER.,” *Nucleic Acids Res*, vol. 27, no. 23, pp. 4636–4641, Dec. 1999.

[5] G. P. van Wezel and K. J. McDowall, “The regulation of the secondary metabolism of Streptomyces: new links and experimental advances,” *Natural Product Reports*, vol. 28, no. 7, p. 1311, 2011.

[6] R. D. Sleator and C. Hill, “Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence,” *FEMS Microbiol. Rev.*, vol. 26, no. 1, pp. 49–71, Mar. 2002.

[7] S. I. Sukharev, P. Blount, B. Martinac, and C. Kung, “Mechanosensitive channels of Escherichia coli: the MscL gene, protein, and activities,” *Annu. Rev. Physiol.*, vol. 59, pp. 633–657, 1997.

[8] B. Palenik, B. Brahamsha, F. W. Larimer, M. Land, L. Hauser, P. Chain, J. Lamerdin, W. Regala, E. E. Allen, J. McCarren, I. Paulsen, A. Dufresne, F. Partensky, E. A. Webb, and J. Waterbury, "The genome of a motile marine Synechococcus," *Nature*, vol. 424, no. 6952, pp. 1037–1042, Aug. 2003.

[9] F. Hube, P. Reverdiau, S. Iochmann, and Y. Gruel, "Improved PCR method for amplification of GC-rich DNA sequences," *Molecular biotechnology*, vol. 31, no. 1, pp. 81–84, 2005.

[10] J. L. Klassen and C. R. Currie, "Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation," *BMC genomics*, vol. 13, no. 1, p. 14, 2012.

[11] H. Gross, "Genomic mining--a concept for the discovery of new bioactive natural products," *Curr Opin Drug Discov Devel*, vol. 12, no. 2, pp. 207–219, Mar. 2009.

[12] D. A. Hopwood, "Streptomyces genes: from Waksman to Sanger," *J. Ind. Microbiol. Biotechnol.*, vol. 30, no. 8, pp. 468–471, Jul. 2003.

[13] G. Yim, H. Huimi Wang, and J. Davies FRS, "Antibiotics as signalling molecules," *Philos Trans R Soc Lond B Biol Sci*, vol. 362, no. 1483, pp. 1195–1200, Jul. 2007.

Таблица **1.** Распределение функций генов.

albus — ост	94-10 — ост	94-10 — albus	94-10 — albus — ост	94-07 — ост	94-07 — albus	94-07 — albus — ост	94-07 — 94-10	94-07 — 94-10 — ост	94-07 — 94-10 — albus	все	Функция
Хранение информации и процессинг											
9	1	0	4	1	1	12	8	6	0	118	Трансляция, рибосомные структуры и биогенез
0	0	0	1	0	0	0	1	0	0	5	РНК процессинг и модификации
6	1		3	3	1	7	4	4	0	47	Транскрипция
22	15	1	7	15	3	27	22	9	2	82	Репликация, рекомбинация и репарация
0	0	0	0	0	0	0	0	0	0	2	Структуры хроматина и динамика
Клеточные процессы, передача сигналов											
1	0	0	3	0	0	7	1	1	1	22	Контроль клеточного цикла, клеточное деление, деление хромосом
41	6	0	10	11	4	22	49	58	10	104	Защитные механизмы, вторичный метаболизм
34	10	1	12	9	6	43	24	24	8	56	Механизмы передачи сигналов, регуляция
20	5	0	8	7	4	15	24	21	9	149	Клеточная стенка, мембрана, секреция, везикулярный транспорт и подвижность
19	1	1	5	4	4	17	16	6	8	89	Посттрансляционные модификации, шапероны, пептидазы
Метаболизм											
23	1	0	4	2	0	15	6	13	1	108	Получение и преобразование энергии
22	0	1	5	5	1	27	10	21	6	125	Транспорт и метаболизм углеводов
29	0	1	9	5	2	23	32	24	7	190	Транспорт и метаболизм аминокислот
8	0	0	0	3	1	5	2	2	1	62	Транспорт и метаболизм нуклеиновых кислот
18	1	0	2	0	0	25	19	7	3	82	Транспорт и метаболизм коферментов
7	1	1	1	3	2	18	12	10	3	74	Транспорт и метаболизм липидов
21	1	1	2	2	0	15	10	13	5	100	Транспорт и метаболизм неорганических ионов
Плохо характеризовано											
56	13	2	21	12	11	70	47	48	38	187	Предсказана только основная функция

Серый фон ячейки — перепредставленные функции. Ост — группа, включающая в себя 8 геномов.

[14] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.

[15] J. Felsenstein, “PHYLIP-Phylogeny Inference Package (Version 3.2),” *Cladistics*, vol. 5, pp. 164–166, 1989.

[16] D. H. Huson, D. C. Richter, C. Rausch, T. Dezulian, M. Franz, and R. Rupp, “Dendroscope: An interactive viewer for large phylogenetic trees,” *BMC Bioinformatics*, vol. 8, p. 460, 2007.

[17] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, “Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements,” *Genome Res*, vol. 14, no. 7, pp. 1394–1403, Jul. 2004.

[18] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, “NCBI BLAST: a better web interface,” *Nucleic Acids Res*, vol. 36, no. Web Server issue, pp. W5–W9, Jul. 2008.

[19] L. Li, C. J. Stoeckert, and D. S. Roos, “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes,” *Genome Res*, vol. 13, no. 9, pp. 2178–2189, Sep. 2003.

[20] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale, “The COG database: an updated version includes eukaryotes,” *BMC Bioinformatics*, vol. 4, p. 41, Sep. 2003.

[21] M. H. Saier, M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan, “The Transporter Classification Database: recent advances,” *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D274–D278, Jan. 2009.