Систематическое предсказание функционально связанных генов в геномах бактерий и архей

Функционально связанные гены в геномах бактерий и архей часто организованы в опероны. Тем не менее, состав и архитектура оперонов очень изменчива и часто отличается даже среди тесно связанных генов. Следовательно, чтобы эффективно извлекать надежные функциональные прогнозы для не охарактеризованных генов из сравнительных анализов быстро растущих геномных баз данных, требуются специальные вычислительные подходы. Мы создали протокол для систематического и автоматического определения генов, которые могут быть функционально связаны с геном или локусом «приманкой», основываясь на “relevance metric”. Учитывая набор локусов приманок и базу данных геномов, определяемую пользователем, этот проток сравнивает геномные окрестности приманок, чтобы идентифицировать гены, которые вероятно вязаны с приманкой, вычисляя численность данного гена внутри и вне окрестностей приманки и расстояния до приманки. Мы иллюстрируем эффективность протокола с тремя тестовыми случаями, а именно: гены, связанные с системами CRISPR-Cas, с использованием метрики CRISPRicity, гены, связанные с археальными провирусами и гены, связанные с генами Argonaute в галобактериях. Протокол может быть запущен пользователем с базовыми вычислительными навыками. Вычислительная “стоимость” (время) зависит от размера набора геномных данных и списка референсных локусов и может варьироваться от одного ЦП-часа до сотен часов на суперкомпьютере.

## Введение

Функционально связанные гены бактерий и архей часто образуют опероны, массивы коэкспрессированных и корегулированных генов. Состав и организация гомологичных оперонов даже среди тесно связанных прокариот заметно разнится, причем многие гены встречаются в оперонах только спорадически. Быстрый рост баз данных геномных последовательностей делаем GBA привлекательным и потенциально высокопродуктивным методом для выявления функциональный связей между генами. Основываясь на том факте, что опероны состоят из функционально связанных генов, подход GBA включает определение вероятных функций неохарактеризованных генов, которые периодически обнаруживаются в непосредственной близости от генов с определенными функциями во множестве бактериальных или археальных геномах. Эта методология широко использовалась для прогнозирования функций различных генов и для поиска генов, которые связаны с конкретными функциональными системами. Предсказанные археальные экзосомы — сложного механизма деградации РНК, который в последствии был подтвержден экспериментально, является характерным примером успешного использования GBA. В последнее время таким образом были открыты многочисленные новые CRISPR-Cas системы, вспомогательные гены CRISPR-Cas и гены анти-CRISPR, а во многих случаях подтверждены экспериментально. Однако, учитывая изменчивость организации оперонов среди бактерий и архей, продуктивное использование GBA требует специальной вычислительной процедуры, которая включает оценку релевантности ассоциированных генов. Недавно мы разработали такую вычислительную стратегию для систематического прогнозирования генов, функционально связанных с CRISPR-Cas системами, используя метрику релевантности, которую мы назвали CRISPRicity. Здесь мы представляем протокол Icity, который реализует эту страгетию в общем контексте.

Случайные перестроки генома часто переставляют гены, постоянно разрушая существующие соседние пары генов и создавая новые пары, которых раньше не было, часто объединяя функционально не связанные гены. Следовательно, функционально значимая связь гено может быть надежно обнаружена только в том случае, если сохранение пар генов или более длинных массивов генов находится выше фона, и соответствующие локусы достаточно представлены в геномных базах данных. По крайней мере три фактора усложняю оценку функциональной значимости. Во-первых, те комбинации, которые отражают биологическую функциональность, но редки в природе, соответственно редки в базах данных. Во-вторых, выборка секвенированных геномов чрезвычайно смещена в сторону видов и штаммов, важных для медицины или биотехнологии, что усложняет статистическую оценку наблюдений По обеим этим причинам подход GBA может либо недооценивать, либо переоценивать значимость наблюдаемой связанности и выдавать ложные результаты. Наконец, важным источником проблем является крайняя диверсификация многих функциональных систем, особенно тех, которые участвуют в гонках вооружений между прокариотами и их паразитами (вирусами и другими генетическими подвижными элементами), а также между сосуществующими бактериями и археями). Быстрая эволюция компонент и архитектур функциональных систем, которые учавствуют в таких конфликтах, создает огромное разнообразие вариантов, которые кажутся почти уникальными и их может быть чрезвычайно трудно распознать как представителей гомологичных и функционально аналогичных систем. Отсутствие статистического разрешения часто усложняет выбор сильных кандидатов в гены с использованием GBA.

Протокол Icity был разработан для поиска белок-кодирующих генов, которые связаны с набором приманок (белок-кодирующих генов или иным образом определенных локусов) в больших геномных датасетах с различными таксономическими представлениями. Для выбранного набора приманок геномный контекст анализируется для ранжирования кандидатов в соответствии с их степенью связи с приманкой. Функциональная значимость кандидатов оценивается с использованием следующих переменных: 1) количество вхождений членов кластера генов, кодирующих гомологичные белки вблизи приманок; 2) количество вхождений членов того же кластера во весь геномный датасет; 3) расстояния между членами кластера и ближайшей приманкой. Комбинация этих переменных используется для оценки и ранжирования генов-кандидатов. Дизайн этого подхода основан на нашей предыдущей работе по успешному предсказанию новых систем CRIPR-Cas и спомогательных генов CRISPR-Cas. Было показано, что концептуально подобных подход является продуктивным для идентификации генов, которые достаточно представлены в островах (соседях) генов защиты. Также был разработан ряд других вычислительных методов для идентификации частично косервативных окрестностей генов в бактериальных и археальных геномах и прогнозирования функциональных ассоциаций между генами и белками. Большинство этих процедур нацелены на предсказание оперона, который затем используется для функционального вывода вместе с другими идентифицирующими окрестностями с эволюционно-консервативной локализацией генов или обогащенные гены с определенной функциональной аннотацией. Примечательно, что один из этих методов был успешно применен для описания первых окрестностей cas-гена, как отельной функциональной сети.

Не вдаваясь в подробности этих подходов, отметим, что отличительной способностью Icity является акцент на конкретном наборе приманок, которые были функционально охарактеризованы. Таким образом, протокол Icity Лучше всего подходит для углубленного исследования функциональных сетей, для которых уже имеется некоторая предварительная информация, а не для поиска функциональных ассоциаций de novo. Такой сфокусированный подход делает процедуру очень масштабируемой и позволяет количественно измерить специфичность с достаточной уверенностью предсказать даже сравнительно редкие функциональные ассоциации.

## Протокол Icity

Обоснование и осуществимость

Доступность обширной базы данных разнообразных микробных геномов представляет широкие возможности для расширения наших знаний о функциях генов и эволюции генома у бактерий и архей. Однако, такое разнообразие делает брутфорс анализ непомерно дорогостоящим из-за огромного количества данных геномных последовательностей. Для проведения исчерпывающих поисков в обширных датасетах протокол опирается на permissive clustering parameters (параметры разрешающей кластеризации) для определения семейств белков и, таким образом, минимизации пространства поиска. Этот метод подходит для массового и эффективного распараллеливания, что делает его легко масштабируемым и способным идти в ногу с ростом баз данных прокариотических последовательной.

GBA — это общей подход для нахождения функционально ассоциированных генов, которые в прочих случаях трудно классифицировать и, как таковой, он успешно использовался в течение последних двух десятилетий. Здесь мы объединяем несколько фич, которые ранее использовались в аналогах GBA, а именно: частота появления гена вблизи приманки, относительная частота такого появления по сравниению с общей численностью гена и расстояния от приманки.

Применение и ограничения

Предоставленная процедура может применять для любой геномной многокомпонентной системы, для которой GBA является релевантным, то есть для любых наборов генов, которые образуют опероны, по меньшей мере, в некоторых микробных геномах. Явным ограничением протокола Icity является его зависимость от физической близости функционально связанных генов в микробных геномах. Следовательно, группы генов, которые функционально взаимодействуют, но никогда не кодируются в одном и том же опероне, не могут быть предсказаны. Тем не менее, благодаря эволюционной текучести оперонов, частично перекрывающиеся окрестности генов из разных геномов, образуют обширные связи сети генов, что делает подход Icity применимым для разнообразных микробных функциональных систем, особенно с учетом быстрого роста геномных баз данных. Возможно, некоторые из наиболее многообещающих областей применения Icity — это анализ защиты, трасдукции сигнала, биосинтеза и устойчивости к антибиотикам и сетей вторичного метаболизма, которые характеризуются обширной перетасовкой больших наборов генов среди частично консервативных оперонов. Ниже мы иллюстрируем эту широкую применимость к трем независимым примерам. Эта методология должна быть интересна для широкого круга исследователей в области микробиологии и молекулярной биологии, которые занимаются исследованием микробного генома для поиска новых функций и видов деятельности, в том числе для потенциальных приложений в биотехнологии.

## Дизайн эксперимента

Вся процедура может быть представлена в 7 этапов

1. Определение набора приманок в геномной базе данных (шаги 1-4)
2. Реконструкция окрестностей генов вокруг примаки (шаг 5)
3. Кластеризация белков, кодируемых в окрестностях приманки по сходству последовательности с “permissive cutoff” (разрешающей отсечкой) (шаги 6-8)
4. Анализ распределений семейств белков из окрестностей приманок в геномной базе данных (Шаги 9-11)
5. Определение метрики релевантности и метода ее вычисления (шаг 12)
6. Подсчет вероятных семейств белков-кандидатов в соответствии с их обилием обогащённости в окрестностях приманки и расстоянием до приманки (шаг 13)
7. Ручное курирование выбранных кандидатов

Этот пайплайн берет геномную базу данных (с информацией о структуре хромосомы или контига и список приманок в качестве входных данных (определяются пользователем) и создает список кандидатов белковых семейств, отранжированных по их баллам релевантности. Описание протокола, приведенное в этом разделе подробно описывает этапы 2-5, а для этапов 1,6,7 предполагаются лучшие практики. Общий пайплайн и детальная пошаговая схема протокола показана на рисунках 1 и 2.

Этап 1. Шаги 1-4. Геномная база данных и выбор приманок

Крайне важно, чтобы результаты оставались согласованными между этапами процедуры. Типичная скорость обновления публичных геномных баз данных сопоставима с количеством времени, необходимом для запуска пайплайна, поэтому пользователям следует рассмотреть возможность выполнения всех шагов для замороженной копии геномной базы данных и сопутствующих метаданных, особенно, если база данных поддерживается третьими лицам. База данных должна быть репрезентативной для конкретного анализируемого случая. Например, если приманки представляют в основном один тип, то по крайней мере все геномы из этого типа должны быть включены в базу данных. Приманки для протокола Icity определяются их координатами в секциях генома (хромосомах), скаффолдах или контигах в базе данных генома. Приманки могут быть кодирующими или не кодирующими последовательностями, в том числе белок-кодирующими генами, РНК-генами, целыми операнами или некодирующими фичаси, такми как массив CRISPR. Набор приманок должен быть четко определен и как модно полнее, чтобы обеспечить максимальную разрешающую способность на последующих этапах. Протокол работает с белковыми последовательностями. Следовательно, важно, чтобы проверяемая база данных включала точно аннотированные белки, кодирующие гены. При поиске метагеномных баз данных следует использовать программное обеспечение для прогнозирования генов, такое как GeneMarkS для прогнозирования кодирующих областей последовательности (CDS). Существующая база данных должна быть проверена на плотность кодирования (Мы исключаем >0.6 кодирующих последовательностей на килобазу в качестве порога), а последовательности с низкой плотностью кодирования должны быть заново аннотированы.

Этап 2. Шаг 5. Аннотация белок-кодирующих генов в окрестностях приманки

На этом этапе собирается набор белок-кодирующих генов из окрестностей приманок. Размер области интереса, окружающей приманки, зависит от ожидаемой структуры оперона функциональной системы, которая включает приманку. Например, в тех случаях, когда ожидается, что система будет состоять из двух генов, таких как токсино-антитоксионовые модули bait-flanking области должны быть ограничены одним-двумя генами или 1-2 парой килобаз (Rbp) последовательности генома. Для таких случаев, как CRISPR\_Cas системы с характериными большими оперонами и расширенными некодирующими функциональными элементами (массивами CRISPR) порог расстояния должен быть ослаблен до 10-15 генов или 10-20 килобаз. Размер фланкирующих областей определяет компромисс между сбором всех возможных кандидатов и затратами на вычислительный анализ и ручное курирование. Предварительное знание биологии интересующей системы может быть использовано в качестве дополнительного этапа предварительной фильтрации. Аннотация кодирующих белок генов в окрестностях приманки с использованием баз данных профилей семейства белков позволяет исключить очевидных кандидатов, таких как различные домашние гены, которые не имеют отношения к прогнозируемым функциям анализируемой системы.

Этап 3. Шаги 6-8. Примесный кластерный анализ

Набор белков-кандидатов, сконструированный на стадии 2, кластеризован по сходству последовательностей с помощью разрешающих параметров. Эта стадия превращает набор отдельных генов с их уникальными геномными местоположениями в семейства гомологичных белков, для которых могут быть обнаружены и проанализированы коллективные тенденции. Белковые компоненты многих клеточных систем, в частности те, которые вовлечены в межгеномные конфликты, сильно изменчивы, поэтому порог кластеризации должен быть понижен до самого низкого безопасного уровня, то есть до точки, в которой кластеризация негомологичных последовательностей начинает становиться проблематичной. Выбор конкретного метода кластеризации является гибким и должен быть сделан пользователем. Для нашего выбранного инструмента кластеризации MMseqs2, который, по-видимому, обеспечивает оптимальный компромисс между чувствительностью и скоростью, мы рекомендуем использовать порог идентификации последовательности 0,3 и охват до 0,1 для достижения максимальной чувствительности. Однако, с этим либеральным порогом, необходима проверка качества кластера. Для такой проверки можно выполнить поиск PSIBLAST для всех членов кластера, используя в качестве запроса профиль, полученный из выравнивания кластера, и те белки, которые не обнаружены в этом поиске, с соответствующим e-value (1 × 10-4 был оптимален для наших поисков) порогом отбрасывания. Мы рекомендуем повторно кластеризовать одноэлементные последовательности и повторять проверку качества кластера три-пять раз, а затем кластеризовать оставшиеся синглтоны с более безопасным порогом кластеризации (идентичность последовательности 0,5 и покрытие 0,3 для MMseqs2).

Этап 4. Шаги 9-11. Поиск членов кластера в геномной базе данных

На этом этапе профили, построенные на основе выравниваний семейств белков, собранных на этапе 3, используются для сканирования всей базы геномных данных, чтобы идентифицировать всех членов кластеров и их гомологов. Этот поиск может быть выполнен с использованием PSIBLAST или других инструментов поиска профиля. Выбор инструментов должен основываться на размере базы геномных данных и требуемой чувствительности и / или специфичности. В зависимости от задачи может потребоваться настройка параметров поиска для достижения более высокой чувствительности или, наоборот, для ограничения пространства поиска. Однако при постобработке результатов поиска необходимо решить некоторые проблемы. Сначала удаляются все совпадения, длина выравнивания которых меньше предварительно определенного порога (мы использовали 25% длины профиля кластера). Во-вторых, несколько попаданий в одну и ту же последовательность из разных профилей делятся на набор непересекающихся сегментов с наивысшими баллами. Попадания из профилей-кандидатов в очень отдаленные гомологи (например, общие домены NTPase или бета-пропеллеров) устраняются с помощью следующей процедуры: новые совпадения группируются вместе с исходными последовательностями-кандидатами из соответствующего семейства с использованием тех же разрешающих параметров кластеризации, и кластеры, которые не содержат ни одной из исходных последовательностей, отклоняются. Эта процедура гарантирует, что новые совпадения являются добросовестными членами исходных кластеров с точки зрения сходства последовательностей. Эти этапы постобработки гарантируют, что представление семейств кандидатов в базе данных генома будет оптимизировано с точки зрения, как чувствительности, так и специфичности.