

На правах рукописи



Сорокин Валерий Андреевич

СРАВНИТЕЛЬНО-ГЕНОМНЫЙ АНАЛИЗ
СИСТЕМ БАКТЕРИАЛЬНОГО ИММУНИТЕТА

Специальность 03.00.28 - биоинформатика

Автореферат
диссертации на соискание ученой степени
кандидата биологических наук

Москва - 2009

Работа выполнена на Факультете биоинженерии и биоинформатики Московского государственного университета имени М.В. Ломоносова и в Учебно-научном центре «Биоинформатика» учреждения Российской академии наук Института проблем передачи информации им. А.А. Харкевича РАН.

Научный руководитель:

доктор биологических наук, профессор,
Гельфанд Михаил Сергеевич

Официальные оппоненты:

доктор биологических наук,
Карягина Анна Станиславовна,
ГУ НИИ эпидемиологии и микробиологии им. Н.Ф.Гамалеи

кандидат физико-математических наук,
Макеев Всеволод Юрьевич,
ФГУП ГНЦ РФ "ГосНИИГенетика"

Ведущая организация:

Учреждение Российской академии наук
Институт молекулярной биологии им.
В.А.Энгельгардта РАН

Защита диссертации состоится 10 декабря 2009 года в 14⁰⁰ часов на заседании диссертационного совета Д 002.077.02 при учреждении Российской академии наук Институте проблем передачи информации им. А.А. Харкевича РАН по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, д.19, стр.1.

С диссертацией можно ознакомиться в библиотеке учреждения Российской академии наук института проблем передачи информации им. А.А. Харкевича РАН.

Автореферат разослан 7 ноября 2009 года

Ученый секретарь диссертационного совета
доктор биологических наук, профессор



Рожкова Г.И.

Общая характеристика работы

Актуальность темы

Данная работа посвящена исследованию двух систем бактериального иммунитета. Актуальность темы следует из двух основных соображений.

Знания, полученные в результате исследований таких систем, могут найти практическое применение в промышленности. Так, в пищевой промышленности бактериальные культуры используются для получения сыров, йогуртов, кефиrow и других молочных продуктов, а также различных приправ. Заражение культуры фаговой инфекцией чревато остановкой производства, дезинфекцией всех производственных мощностей, наработкой культуры с нуля и повторным запуском производства. Издержки от этого могут быть очень велики, особенно если принять во внимание масштабность производства. Понимание принципов работы систем бактериального иммунитета – прямой путь к управляемому повышению устойчивости промышленных культур к фаговым инфекциям.

С другой стороны, поскольку взаимодействие фаг-бактерия отличается особой динамичностью, системы бактериального иммунитета представляют собой удобную модель для изучения молекулярной и геномной эволюции. Наряду со сравнительно хорошо изученными системами рестрикции-модификации, построенными на базе систем типа токсин-антитоксин, в настоящей работе рассмотрены и недавно открытые CRISPR-системы, которые, как полагают, участвуют в анти-фаговой защите клетки, используя механизм, схожий с механизмом РНК-интерференции в эукариотах. Как будет показано в дальнейшем на примере таких систем, в геномах можно обнаружить следы взаимодействия фаг-бактерия, а именно, ответа бактериальной клетки на внедрение фага.

Системы рестрикции-модификации (РМ-системы), как правило, состоят из двух ферментов, один из которых способен узнавать определенные участки (короткие последовательности) ДНК и химически их модифицировать (метилировать), а другой, узнавая те же самые участки ДНК, способен вносить двуцепочечный разрыв в этот участок ДНК в случае, если участок не подвергся модификации. Система устроена таким образом, что ДНК клетки-хозяина оказывается полностью метилированной, в отличие от последовательностей ДНК внедряющихся фагов, которые эффективно деградируются рестриктазой РМ-системы. Система противофаговой защиты должна обладать сложной системой регуляции, т.к. в противном случае клетке-носителю может быть нанесен непоправимый вред. Одним из компонентов системы регуляции РМ-систем является С-белок, типичный представитель НТН-семейства (факторы транскрипции с ДНК-связывающим доменом спираль-поворот-спираль). С-белки, наряду с их авторегуляторными сайтами, являлись одним из объектов данного исследования.

Типичная CRISPR-система представляет собой кассету, предположительно содержащую информацию о геномах тех фагов, которые уже атаковали клетку, и набор генов, продукты которых позволяют использовать эту информацию для противодействия повторно атакующим клетку фагам. Кассета представляет собой последовательность ДНК, состоящую из коротких (25-45 п.о.) уникальных участков (так называемых спейсеров), которые разделены точными прямыми повторами примерно такой же длины. Показано, что спейсерные последовательности похожи на участки геномов некоторых фагов, а последовательности белков, закодированные в наборе генов, обслуживающих кассету, содержат мотивы, сходные с мотивами ферментов, проявляющих нуклеазную активность. По всей видимости, за счет комплементарного узнавания ДНК внедряющегося фага взаимодействует со спейсерами кассеты, что приводит к её деградации в результате действия белков, закодированных генами CRISPR-системы. Этому процессу предшествует процесс обучения – встраивание элементов генома внедряющегося фага в кассету в виде новых спейсерных последовательностей.

Цель исследования

Целью исследования являлось изучение С-белков и их авторегуляторных сайтов, а также исследование CRISPR-систем в метагеноме методами сравнительной геномики.

Задачи исследования

1. Полномасштабное предсказание новых представителей семейства С-белков и предсказание авторегуляторных сайтов для предсказанных генов С-белков.
2. Классификация сайтов связывания С-белков и выявление их структурных особенностей.
3. Описание генетической структуры локусов предсказанных генов С-белков.
4. Решение проблемы большого числа ложных предсказаний, наблюдаемых при применении известных алгоритмов к предсказанию CRISPR-подобных структур в метагеноме и полномасштабное предсказание CRISPR-подобных структур в метагеномных последовательностях *Sorcerer II*.
5. Проверка гипотезы о том, что в спейсерных последовательностях CRISPR-кассет популяции бактерий преимущественно содержатся элементы фагов из этой же популяции.
6. Анализ эволюционных событий, наблюдаемых в CRISPR-кассетах.
7. Формирование удобных баз данных, доступных через Интернет и содержащих результаты исследования С-белков и CRISPR-систем. Базы данных должны обладать удобной системой запросов как по идентификаторам, параметрам и ключевым словам, так и по сходству с фрагментами, заданными пользователем.

Научная новизна и практическое значение:

1. В работе идентифицировано 169 новых представителей семейства С-белков. До начала исследования было известно лишь 46 представителей этого семейства.
2. Предсказан 201 авторегуляторный сайт генов С-белков (169 сайтов для ранее неизвестных генов С-белков, 32 для известных генов С-белков, содержащихся в Rebase). До начала исследования число известных сайтов не превышало 30, при этом лишь 8 сайтов сохранилось в Rebase.
3. Предсказанные авторегуляторные сайты разбиты на 10 отличающихся структурных групп (мотивов), из которых лишь 3 были описаны ранее.
4. Описаны локусы, содержащие гены С-белков, и показано, что в них часто встречаются гены фаговой активности. Впервые описаны локусы, содержащие пары С-белков из различных подсемейств. Для ряда мотивов описаны дополнительные сайты связывания.
5. Разработана процедура фильтрации ложных предсказаний CRISPR-подобных структур.
6. Показано, что в спейсерных элементах CRISPR-кассет бактерий из фиксированного местообитания значимо чаще встречаются элементы из фагов из того же местообитания.
7. Описаны элементарные эволюционные события, наблюдаемые в выборке предсказанных CRISPR-кассет, и реконструирована эволюционная история родственных CRISPR-кассет.
8. Разработаны удобные базы данных, позволяющие независимым исследователям получить доступ ко всем результатам работы.

Апробация работы

Материалы исследований по теме диссертации были представлены на российских и международных конференциях: XV Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов-2008», Москва (диплом за лучший доклад); Первый международный форум по нанотехнологиям «Роснанофорум-2008», Москва (диплом за стендовый доклад); Российско-германский симпозиум по системной биологии, 2008, Москва; Berlin Summer Meeting: Computational & Experimental Molecular Biology, 2008, Берлин, Германия; 7th European Conference on Computational Biology (ECCB'08) & 5th BITS Meeting, 2008, Кальяри, Сардиния, Италия; Конференция "Информационные технологии и системы ИТиС'08", 2008, Геленджик; 16th Albany Conversation, 2009, Олбани, США; 17th Annu. Int. Conf. on Intelligent Systems for Molecular Biology and 8th European Conf. On Computational Biology ISMB/ECCB'09, 2009, ISMB/ECCB'09, 2009, Стокгольм, Швеция.

Объем и структура диссертации

Диссертационная работа изложена на 85 страницах машинописного текста и состоит из введения и четырех разделов. Первый раздел включает обзор литературы по теме диссертации. Второй раздел содержит описание использованных в работе программ и алгоритмов, а также подробное описание разработанных в работе методов. Третий раздел содержит описание проведенных в работе исследований и состоит из двух частей, каждая из которых включает изложение полученных в работе результатов и их обсуждение: первая часть посвящена предсказанию и анализу С-белков и их авторегуляторных сайтов; вторая часть посвящена предсказанию и анализу CRISPR-систем в метагеноме Sorcerer II. В четвертом разделе содержится описание построенных баз данных.

Список цитируемой литературы, приведенный в конце диссертации, содержит 87 наименований. Работа содержит 14 рисунков и 5 таблиц.

Содержание работы

Первый раздел: Обзор литературы

Первый раздел содержит обзор основных литературных источников, затрагивающих системы рестрикции-модификации и CRISPR-системы. В этом разделе дается классификация РМ-систем, объясняется принцип их работы, а также то, чем обусловлена потребность в сложной системе регуляции транскрипции генов РМ-систем. Кратко дается обзор основных механизмов, реализующих регуляцию транскрипции, один из которых использует отдельный белок-регулятор – С-белок. Далее приводится обзор основных экспериментальных фактов, которые были установлены для С-белков, а также описание работ, посвященных биоинформатическому анализу С-белков.

Далее, в первом разделе приводятся основные известные сведения о CRISPR-системах. В частности, описывается функциональное назначение каждой из составных частей CRISPR-системы: лидерной последовательности, кассеты, состоящей из спейсеров и повторов, а также группы *cas*-генов. Излагаются результаты работ, в которых выдвигается гипотеза об участии CRISPR-систем в противофаговой обороне. Приводятся результаты исследований «динамики» CRISPR-систем, в частности, результаты изучения систем в метагеномных образцах. В завершение, приводится описание крупнейшего на сегодняшний момент метагенома – Sorcerer II, который был объектом исследования.

Второй раздел: Материалы и методы

Этот раздел содержит описание основных источников информации, методов, алгоритмов и программ, использованных в работе.

В качестве источников информации используются база данных Rebase и последовательности метагенома Sorcerer II. Далее следует описание подхода к предсказанию С-белков и их авторегуляторных сайтов, разработанного в данной работе. В контексте подхода описывается применение стандартных биоинформатических инструментов: программ поиска сходных последовательностей BLAST и HMMER, программы множественного выравнивания последовательностей MUSCLE, пакета программ для филогенетического анализа PHYLIP и т.д.

Излагаются базовые принципы работы трех известных алгоритмов, которые были использованы для предсказания CRISPR-кассет в метагеноме Sorcerer II, и описывается подход, разработанный для повышения надежности предсказания. Приводится методика тестирования основной гипотезы: в спейсерных последовательностях бактерий, обитающих в определенной экологической зоне, содержатся элементы геномов сосуществующих с ними фагов.

Третий раздел: Результаты и обсуждение

1. Исследование С-белков

1.1. Предсказание новых С-белков

В качестве отправной точки анализа были выбраны 46 белков из Rebase. С помощью процедуры *tblastn* (из пакета программ BLAST) для каждого из 46 С-белков независимо был произведен поиск сходных последовательностей в базе GenBank. Сравнительно строгий по-

рог по е-значению был выбран, чтобы гарантировать отсутствие сходств с другими представителями НТН-семейства (например, фаговыми репрессорами). После ручного анализа результатов было отобрано 245 генов, каждый из которых потенциально мог кодировать С-белок. Для этих генов было проведено уточнение точки начала трансляции с учетом того, что средняя длина С-белка не превосходит 70 аминокислот и различий в паттерне эволюции белок-кодирующих и некодирующих областей.

1.2. Предсказание авторегуляторных сайтов

Поскольку выборка идентифицированных потенциальных С-белков содержала довольно далекие группы С-белков, можно было ожидать, что выборка авторегуляторных сайтов также будет довольно разнородна (действительно, как оказалось в дальнейшем, авторегуляторные сайты предсказанных С-белков образуют 10 мотивов, каждый из которых имеет свои структурные особенности).

Поэтому было решено начать анализ со сравнительно небольших подгрупп наиболее близких С-белков. Было построено филогенетическое дерево (пакет PHYLIP, процедура *proml*) всех 291 членов из выборки (245 предсказанных и 46 С-белков из Rebase). Незначительно отличающийся вариант этого дерева, на котором представлены лишь С-белки с предсказанными сайтами связывания, приведен на рис. 1. Анализ дерева выявил несколько обособленных групп белков, которые были проанализированы отдельно.

Для каждого члена группы извлекался 5'-участок, предшествующий предсказанному гену С-белка. Длина участка была фиксирована и составляла 100 пар оснований. Для наиболее близких (в метрике филогенетического дерева) представителей группы строилось выравнивание извлеченных участков. Далее это выравнивание итеративно расширялось за счет 5'-участков других, все более и более далеких членов группы. На каждой итерации качество выравнивания контролировалось вручную. Как только качество признавалось неудовлетворительным, итерации останавливались, и оставшиеся на выравнивании консервативные участки объявлялись предсказанными авторегуляторными сайтами связывания соответствующих гипотетических С-белков.

Как следует из описания процедуры, она чувствительна к правильности предсказания старт-кодона гена гипотетического С-белка. Поэтому, с использованием матрицы весов, построенной на основании уже предсказанных авторегуляторных сайтов, были уточнены имеющиеся предсказания, идентифицирован ряд новых авторегуляторных сайтов, и в ряде случаев уточнена позиция старт-кодона предсказанного гена С-белка.

Всего сайты связывания были предсказаны для 201 (из 291) потенциальных С-белков.

1.3. Сравнение предсказанных сайтов с экспериментально подтвержденными сайтами

Для восьми генов С-белков, содержащихся в Rebase, сайты связывания были определены экспериментально. Согласно литературным данным, еще для 24 белков сайты были определены биоинформатическими методами. Все предсказания, сделанные в результате анализа, совпадали с уже известными сайтами. Однако, поскольку полученная выборка авторегуляторных сайтов была на порядок больше, это позволило обнаружить существенно больше элементов структуры сайтов, чем было охарактеризовано до настоящего анализа.

1.4. Анализ структуры предсказанных сайтов связывания

Известные к моменту начала работы сайты можно было разделить на три большие структурные группы: группа сайтов, подобных сайту С-белка С.PvuII, сайт С-белка С.EcoRV и сайт С-белка С.EcoO109I.

Структура сайтов первой группы была частично описана в литературе (Mruck I. et al., 2007): типичный мотив представляет собой два палиндрома, разделенных консервативной четырехнуклеотидной спейсерной последовательностью. Два палиндрома представляют собой непосредственно места связывания димеров С-белка, при этом с одним плечом палиндрома (С-боксом) связывается один (из двух белков димера) С-белок. Парам консервативных динуклеотидов, расположенным по краям сайта, не было уделено внимания, хотя данная

структурная особенность вызывает интерес. Два других сайта содержали один палиндром, при этом палиндромы отличались как между собой, так и от палиндромов группы сайтов, похожих на сайт белка C.PvuII.

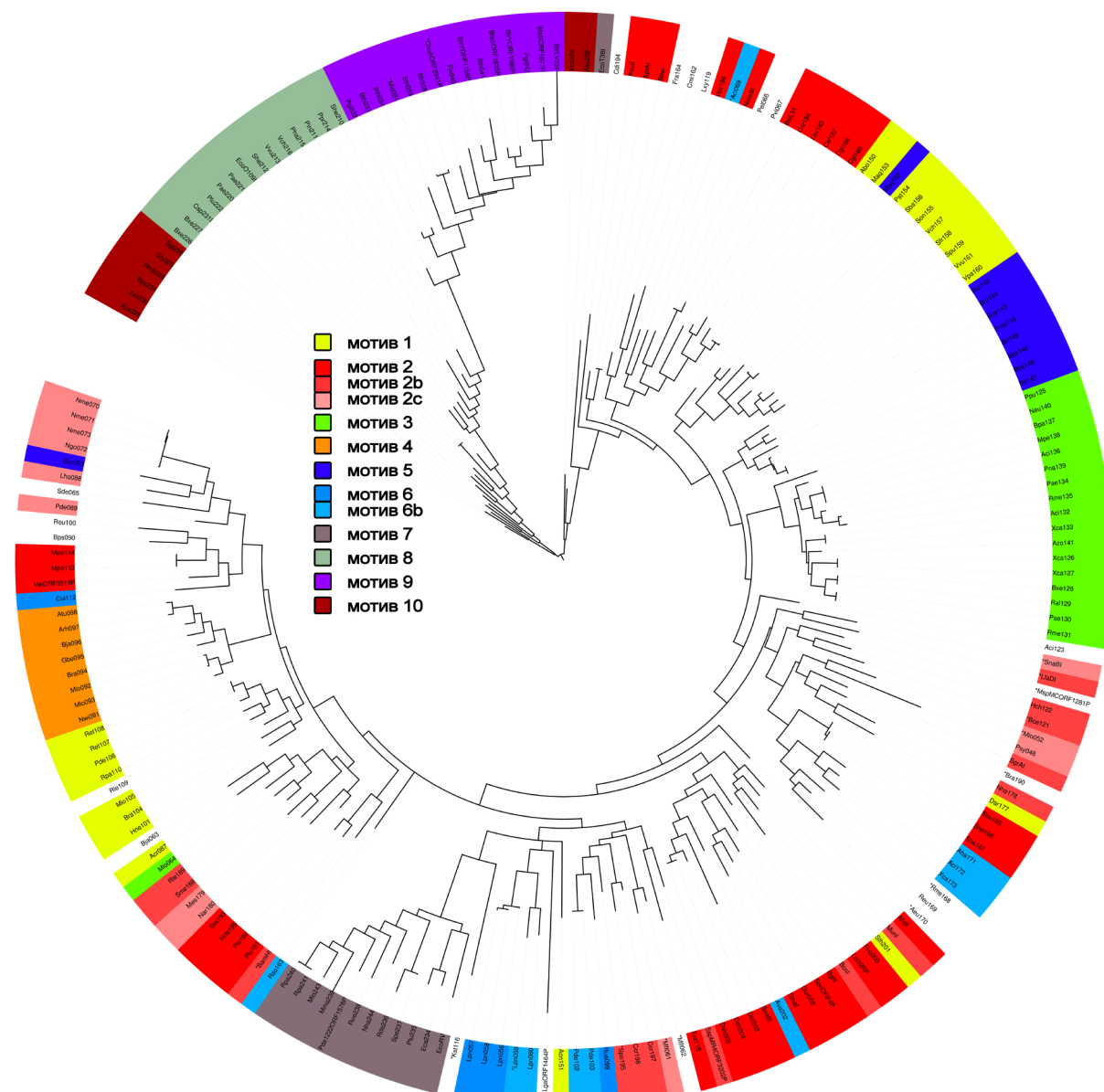


Рис.1 Филогенетическое дерево предсказанных С-белков, для которых был предсказан авторегуляторный сайт. Принадлежность авторегуляторного сайта к одному из 10 мотивов указана цветом.

Набор предсказанных в диссертации сайтов был дополнительно разделен на подгруппы с помощью процедуры ClusterTree-RS (предоставлена Е. Ставровской). Процедура выявила 10 явно различающихся кластеров (в дальнейшем именуемых мотивами), в которые входил 181 сайт (из 201 предсказанного сайта). Изображение мотивов приведено на рис. 2.

Мотивы 7 и 8 соответствуют сайтам связывания, подобным сайтам связывания С-белков C.EcoRV и C.EcoO109I. Однако, мотивы содержат соответственно 13 и 14 позиций, что соответствует дополнительным консервативным позициям, составляющим сайт: две позиции было обнаружено для мотива 7, и одна позиция была обнаружена для мотива 8. Кроме этого, в мотивах 7 и 8 разделитель между плечами палиндрома составляет всего одну позицию, а в ранее предсказанных сайтах разделитель был существенно длиннее.

Мотивы 1-6 обладают рядом общих структурных элементов. Архитектура «усредненно-го» мотива описывается следующей формулой:

$$Z-X-N-X^*-[GT\text{-ядро}]-x-n-x^*-Z^*,$$

где Z – краевые консервативные комплементарные нуклеотиды, X-N-X* и x-n-x* – копии палиндрома, звездочка * означает комплементарность элемента структуры, а нижний регистр букв означает тот факт, что 3' копия гораздо больше похожа на общий консенсус, нежели 5' копия.

Мотивы 1, 2, 4 точно подходят под описание, задаваемое этой формулой. Мотив 3 не содержит тройных краевых нуклеотидов (Z), мотивы 5 и 6, напротив, не содержат копий палиндромов (X-N-X*), но содержат внешние консервативные комплементарные нуклеотиды. Все известные на момент начала исследования С-белки с сайтами, похожими на сайты С-белка C.PvuII, соответствовали мотиву 2.

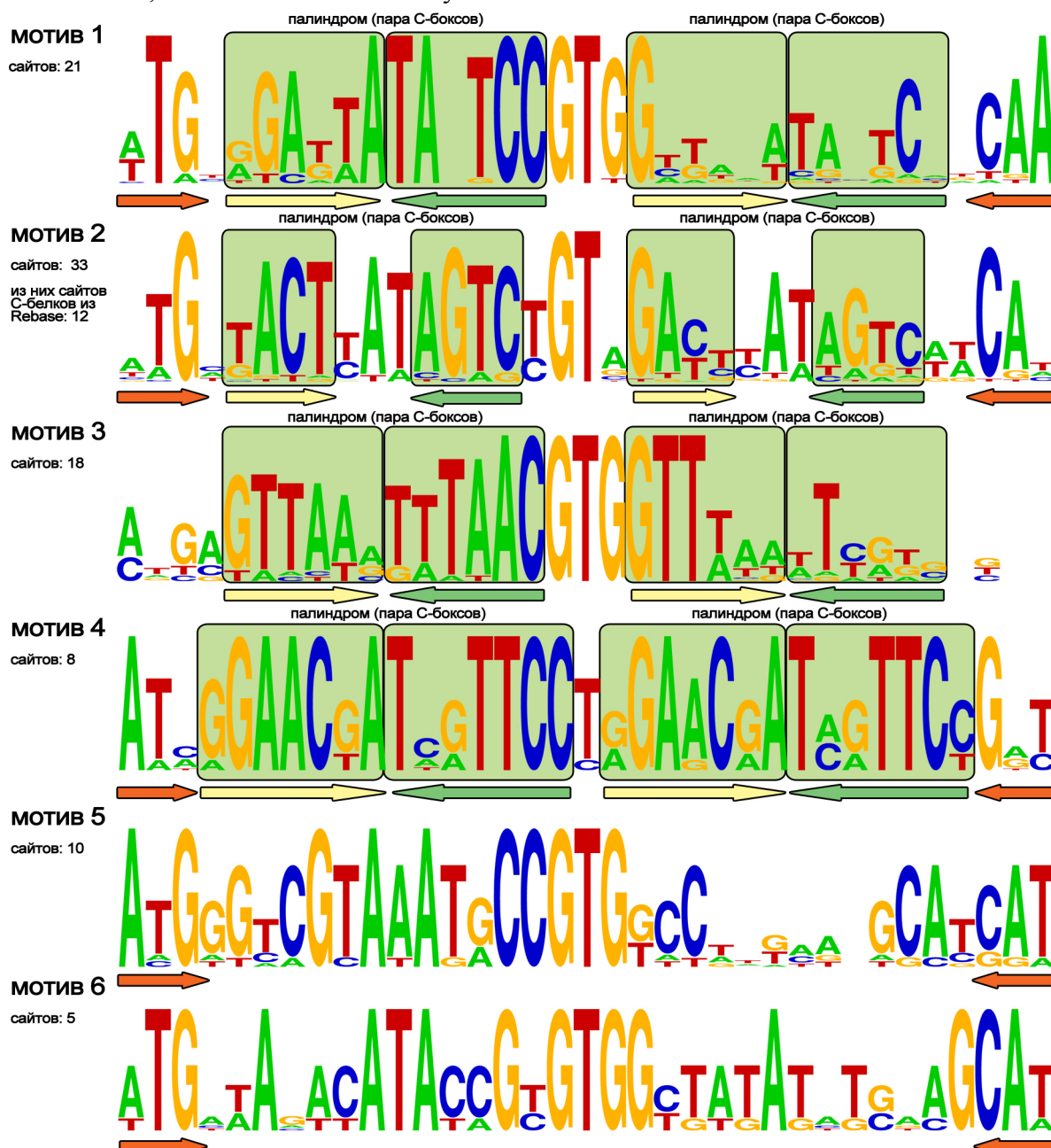


Рис.2а Профили мотивов 1-6 предсказанных авторегуляторных сайтов. Окрашенными прямоугольниками обозначены С-боксы. Палиндромные элементы структуры обозначены стрелками.

Мотив 9 представляет собой короткую консервативную последовательность. Её длина (10 п.о.), а также отсутствие какой-либо симметрии, характерной для сайтов связывания, не позволяет предполагать в ней полноценный авторегуляторный сайт связывания и, возможно, указывает на то, что эти сайты являются ошибочно предсказанными.

Мотив 10 содержит один палиндром, отличающийся от палиндромов всех остальных предсказанных сайтов С-белков.

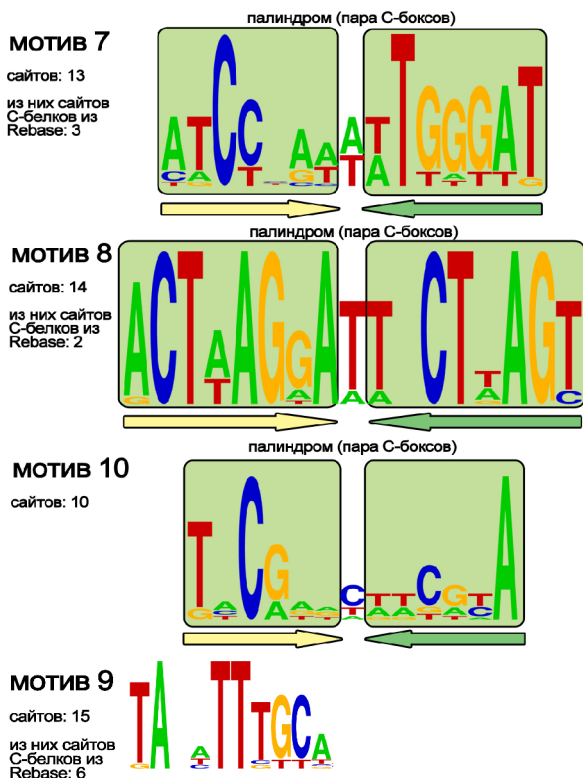


Рис.26 Профили мотивов 7-10 предсказанных авторегуляторных сайтов. Окрашенными прямоугольниками обозначены С-боксы. Палиндромные элементы структуры обозначены стрелками.

ние с использованием позиционных весовых матриц, построенных по уже идентифицированным сайтам, которое позволило обнаружить для каждого авторегуляторного сайта из мотивов 7, 8 и 10 дополнительные проксимальные сайты связывания. Однако, в отличие от мотивов 1-4, где проксимальные и дистальные сайты располагались на фиксированном расстоянии друг от друга, расстояние между соответствующими сайтами мотивов 7, 8 и 10 не было фиксированным, что является принципиальным отличием мотивов 7, 8 и 10 от мотивов 1-4.

1.6. Дополнительные исследования предсказанных генов С-белков и их авторегуляторных сайтов

Был проведен ряд дополнительных исследований, связанных с предсказанными генами С-белков и их авторегуляторными сайтами. Во-первых, исследовалась возможность дополнительной регуляции активности РМ-систем через трансляцию CR-оперонов с безлидерных матриц. Наличие этого механизма было продемонстрировано в литературе на примере РМ-системы PvuII, для которой расстояние между предсказанным сайтом (принадлежащим к мотиву 2) и старт-кодоном гена С-белка составляет 18 п.о. В результате исследования распределения расстояний между предсказанными авторегуляторными сайтами, принадлежащими к С.PvuII-подобным мотивам 1-6, и старт-кодонами предсказанных генов С-белков, был обнаружен пик, соответствующий расстоянию в 17-18 п.о. Дополнительный анализ таких случаев не выявил в разделяющих 17-18 нуклеотидах присутствия последовательности Шайна-Дальгарно (AGGAG), что может означать, что трансляция CR-оперонов идет с безлидерных РНК-матриц.

Во-вторых, исследовалось геномное окружение предсказанных генов С-белков на предмет наличия в нем генов, сходных с генами РМ-активности (метилазы, рестриктазы), а также генов плазмидно-фаговой активности (гены межбактериального транспорта, гены мобильных элементов и т.д.) В окрестности 3000 п.о. относительно старта трансляции предсказанного гена С-белка были вручную предсказаны все открытые рамки считывания длиной не менее 100 аминокислот. Размер окна (3000 п.о. в обе стороны) выбирался таким образом,

1.5. Предсказание дополнительных (проксимальных) сайтов связывания

Кристаллографический анализ С-белков C.AhdI и C.BclI показал, что С-белки существуют в форме димеров (Sawaya M.R. et al., 2005, McGeehan J.E. et al., 2005). Кроме этого, все экспериментально изученные С-белки имели пары сайтов связывания (проксимальный и дистальный) непосредственно перед CR-опероном, что обеспечивало сначала активацию, а затем репрессию транскрипции этого оперона.

Сайты, принадлежащие мотивам 1-4, уже содержат в себе два участка связывания С-белка. В соответствии с теорией, левый (дистальный) участок связывания обладает большим сродством к общему консенсусу, нежели правый (проксимальный) участок сайта.

Мотивы 7, 8 и 10 состоят из одного четко выраженного палиндрома, что является недостаточным для реализации активационно-репрессивного механизма. Было проведено дополнительное исследование

чтобы примерно соответствовать размеру типичных известных РМ-систем: BamHI~2.3 тыс. п.о.; PvuII~1.8 тыс. п.о.; AhdI~3.5 тыс. п.о.

Потенциальные белки, кодируемые открытыми рамками считывания, были исследованы на предмет сходства с белками из базы данных GenBank (процедура blastp), а также с белковыми семействами из базы Pfam (процедура hmmsearch). В диссертации приводится подробное описание результатов. Кратко они представлены в табл. 1.

В-третьих, в ходе исследования было обнаружено семь случаев, когда в одном локусе (на расстоянии менее 3000 п.о.) содержалось два предсказанных С-белка. В то время как пять пар могут объясняться обычной дупликацией одиночного гена С-белка (на филогенетическом дереве, приведенном на рис. 1, участники пяти пар находятся по соседству друг с другом), для двух оставшихся пар картина значительно более сложная. Во-первых, авторегуляторные сайты для участников двух пар принадлежат структурно разным мотивам, и, во-вторых, гены соответствующих С-белков на филогенетическом дереве не находятся по соседству друг с другом.

сколько локусов содержит...	хотя бы один ген РМ активности	хотя бы один ген фаговой активности	хотя бы по одному гену РМ и фаговой активности	всего локусов
предсказанные С-белки	39 (23%)	115 (68%)	27 (16%)	169 (100%)
С-белки из Rebase	32 (100%)	26 (78%)	26 (81%)	32 (100%)

Табл.1 Статистика анализа геномного окружения предсказанных С-белков и С-белков из Rebase.

Вероятнее всего, эти две пары являются следствием недавних «перетасовок» генетического материала, когда мобильные элементы (гены, похожие на гены мобильных элементов, были обнаружены неподалеку от генов С-белков обеих пар), захватывая части РМ-систем и перемещаясь по геному, смешивают захваченные куски геномов.

2. Исследование CRISPR-систем

2.1. Создание первоначального набора CRISPR-кассет

Основной идеей исследования был анализ CRISPR-систем, предсказанных в метагеноме Sorcerer II. В диссертации приводится подробное обоснование, согласно которому исследование CRISPR-систем представляет особый интерес в метагеномных последовательностях.

Метагеном Sorcerer II, расположенный в свободном доступе на сайте CAMERA, состоял из 3081849 последовательностей ДНК (контигами). Совокупная длина метагенома составляла около 4.5 млрд. п.о.

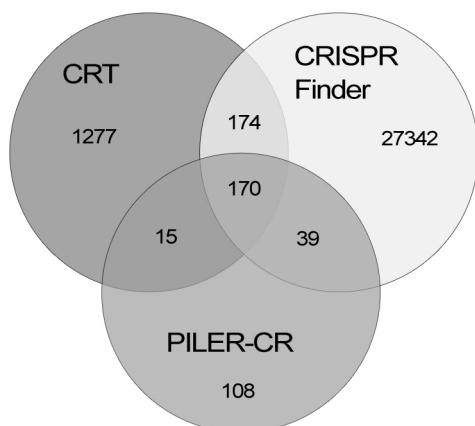


Рис.3 Диаграмма, демонстрирующая степень перекрытия результатов применения трех программ CRT, PILER-CR, CRISPRFinder к метагеному Sorcerer II. Числа означают количество предсказаний.

Для предсказания CRISPR-кассет в метагеноме использовались три алгоритма: CRT, PILER-CR и CRISPRFinder. Несмотря на то, что все три программы разработаны для предсказания CRISPR-кассет, предсказанные наборы кассет перекрывались лишь незначительно (см. рис. 3).

Размеры соответствующих наборов отличались как количественно, так и качественно. Так, алгоритм CRISPRFinder трактовал некоторые артефакты сборки контигов как CRISPR-подобные структуры.

Был разработан специальный алгоритм, который позволил из полного набора кассет выявить 190 CRISPR-кассет, являющихся наиболее вероятными предсказаниями. Формирование набора началось со 170 CRISPR-кассет, обнаруженных тремя программами одновременно. После этого к ним

были добавлены 10 CRISPR-кассет, которые были обнаружены в контигах, содержащих последовательности, похожие на *cas*-гены. Наконец, все обнаруженные CRISPR-кассеты были сгруппированы (кластеризованы) по последовательности повторов и 10 кассет, которые попали в кластеры, содержащие ранее отобранные CRISPR-кассеты, также были добавлены в набор предсказаний повышенной надежности. В результате, окончательная выборка состояла из 190 кассет.

2.2. Исследование построенного набора CRISPR-кассет

Кроме описательного анализа полученного набора, CRISPR-кассеты исследовались на предмет проверки следующей гипотезы: в спейсерных последовательностях обнаруженных CRISPR-кассет содержатся элементы геномов фагов из той же географической области.

Эту гипотезу удалось подтвердить с помощью специально разработанной процедуры. Процедура использует тот факт, что каждому контигу приписана метка, обозначающая определенную географическую область, в которой был собран данный образец. Если спейсерная последовательность, найденная в одном контиге, похожа на участок какого-то другого контига (который может являться элементом генома некультивируемого в лабораторных условиях фага, поэтому неизвестного на сегодняшний день), то с большой вероятностью метки, приписанные обоим контигам, должны совпадать. Для контроля проводились процедуры перемешивания меток с сохранением всех суммарных распределений (число меток на контиг, чисто меток на область и т.д.). В результате было показано, что статистическая значимость сделанного наблюдения составляет по крайней мере $1e-04$.

Наряду с этим, в диссертации уделяется внимание таксономическому составу контигов, в которых были предсказаны CRISPR-кассеты. Предсказание таксономической принадлежности контига осуществлялось, в основном, через анализ сходных с участками контига последовательностей из GenBank, для которых аннотация известна. С помощью этой методики удалось осуществить предсказание таксономической группы для 46% контигов, содержащих предсказанные CRISPR-кассеты. В двух кластерах (26 и 40) таксономические группы существенно различались, что может указывать на горизонтальный перенос CRISPR-систем из одного вида в другой. Так, в кластере 26 кассета c1105 была приписана к Betaproteobacteria, а кассета c0309 – к Cyanobacteria. В кластере 40 кассета c0368 была приписана к Gammaproteobacteria, а c1199 – к Chlamydia.

2.3. Исследование эволюции CRISPR-кассет

Кластеры CRISPR-кассет дают возможность изучить эволюцию CRISPR-кассет. В самом деле, если две кассеты обладают похожими повторами, то вероятно, что кассеты произошли от одной общей предковой CRISPR-кассеты. В этом случае различия между спейсерными последовательностями CRISPR-кассет дают возможность изучить набор элементарных событий, которые меняют спейсерный состав CRISPR-кассет.

В результате ручного анализа было выявлено шесть классов элементарных событий: 1) простая вставка/делеция, 2) вставка/делеция двух и более смежных спейсеров, 3) смежная дупликация одного спейсера, 4) несмежная дупликация одного спейсера, 5) смежная дупликация нескольких последовательных спейсеров, 6) несмежная дупликация нескольких последовательных спейсеров. Наряду с этими классами было выявлено четыре сложных случая.

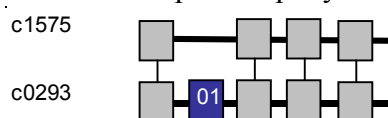


Рис. 4 Схематичное изображение простой вставки/делеции. Прямоугольники обозначают спейсеры, идентичные спейсеры соединены линиями.

Иллюстрация самого простого из элементарных событий – вставка/делеция одного спейсера, приведена на рис. 4.

Наиболее интересным представляется четвертый сложный случай, который проиллюстрирован парой кассет c1575-c1340 на рис. 5.

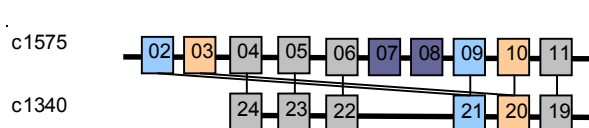


Рис. 5 Схематическое изображение четвертого сложного случая. Прямоугольники обозначают спейсеры, идентичные спейсеры соединены линиями.

спейсеры 25 и 26, которые идентичны спейсерам 02 и 03 кассеты c1575. Основанием для такого сильного предположения может служить тот факт, что последовательность из трех спейсеров (04-06 в c1575 и 22-24 в c1340) идентична в обеих кассетах.

Четвертый раздел: Описание созданных баз данных

Четвертый раздел работы содержит краткое описание двух построенных в работе баз данных, содержащих все основные результаты исследований и доступных через Интернет. Одна база данных содержит результаты исследования С-белков: последовательности С-белков, координаты и ориентацию предсказанных генов в GenBank последовательности, координаты и ориентацию предсказанных авторегуляторных сайтов, последовательности, координаты и ориентацию предсказанных близлежащих генов, их функциональную аннотацию, предсказанную исходя из результатов поиска сходных последовательностей в базах GenBank и Pfam. В базе реализовано графическое представление локуса каждого предсказанного гена С-белка, на котором графически маркируется информация о предсказанной функциональной аннотации.

Вторая база данных содержит результаты исследования CRISPR-систем в метагеноме Sorcerer II. Интерфейс базы позволяет сортировать по широкому набору параметров и визуально отображать предсказанные в данной работе кассеты. На базе алгоритма BLAST реализован механизм поиска последовательностей, сходных со спейсерами и повторами предсказанных кассет, а также сходных с участками контигов, в которых были предсказаны CRISPR-кассеты. Наконец, на базе алгоритма HMMER реализован поиск последовательностей, сходных с кластерами повторов, построенных в данной работе.

Выводы

1. Предсказано 169 потенциальных новых членов семейства С-белков и соответствующих им авторегуляторных сайтов. Для 26 известных С-белков предсказаны неизвестные ранее авторегуляторные сайты. Для 8 С-белков независимое предсказание авторегуляторного сайта совпало с сайтом, определенным экспериментально.
2. Предложены 10 четко отличающихся по последовательности мотивов авторегуляторных сайтов, которые содержат предсказанные в данной работе авторегуляторные сайты. Все известные на момент начала исследования сайты соответствуют усеченным версиям трех мотивов из десяти, т.е. остальные семь мотивов авторегуляторных сайтов были впервые описаны в данной работе.
3. Показано, что в распределении расстояния между сайтами мотивов 1-6 и началом гена потенциального С-белка наблюдается ярко выраженный пик, соответствующий 17-18 п.о. В участках, разделяющих сайты и старт-кодона, не выявлено последовательностей Шайна-Дальгарно. Это указывает на возможность использования РМ-системами безлидерных транскриптов для отсрочки трансляции рестриктазы.
4. Впервые описано семь локусов, в которых находятся два потенциальных С-белка. При этом филогенетический анализ показывает, что в пяти локусах пара С-белков возникла из-за дупликации, а в остальных двух локусах — в результате горизонтального переноса.
5. Было исследовано геномное окружение потенциальных С-белков. В 23% случаев по соседству с генами С-белков были обнаружены гены РМ-активности, а в 68% случаев по соседству были обнаружены гены фаговой активности.
6. Разработан метод идентификации надежно предсказанных кассет в метагеномных данных большого объема.

7. Подтверждена гипотеза о том, что в спейсерных последовательностях кассет из данной экологической ниши находятся элементы геномов сосуществующих в этой экологической нише фагов.
8. Проанализированы эволюционные события, наблюдаемые в популяции предсказанных CRISPR-кассет, которые образуют в шесть классов элементарных событий.
9. Все результаты оформлены в виде баз данных, доступных через сеть Интернет. В базе реализован поиск сходных последовательностей для фрагмента, указанного пользователем, а также гибкая система запросов по широкому набору параметров.

Список работ, опубликованных по теме диссертации

- [1] Artamonova I.I., Gelfand M.S., Sorokin V.A. A recently discovered type of the prokaryotic immunity, the CRISPR system, in metagenomes. Российско-германский симпозиум по системной биологии. 2008. Москва.
- [2] Сорокин В., Гельфанд М.С. Предсказание и анализ тонкой структуры сайтов С-белков бактериальных систем рестрикции-модификации. XV Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов-2008», Москва (диплом за лучший доклад).
- [3] Сорокин В. Метагеномный анализ CRISPR-систем прокариотического иммунитета. Первый международный форум по нанотехнологиям «Роснанофорум-2008». Москва (диплом за стендовый доклад).
- [4] Artamonova I.I., Gelfand M.S., Sorokin V.A. Metagenomic evidence of the CRISPR systems, the recently discovered type of the prokaryotic immunity. Berlin Summer Meeting «Computational & Experimental Molecular Biology», 2008, Berlin, Germany
- [5] Artamonova I.I., Gelfand M.S., Sorokin V.A. Prokaryotic immunity systems of the CRISPR type in metagenomes. 7th European Conference on Computational Biology (ECCB'08) & 5th BITS Meeting, 2008, Cagliari, Sardinia, Italy.
- [6] Артамонова И.И., Гельфанд М.С., Сорокин В.А. CRISPR-системы в метагеномах. Конференция «Информационные технологии и системы ИТиС'07», 2008. Геленджик, Россия.
- [7] Sorokin V., Severinov K., Gelfand M.S. Systematic prediction of control proteins and their DNA binding sites // Nucleic Acids Research. 2009. V.37. N. 2. P. 441-451
- [8] Artamonova I.I., Sorokin V.A., Gelfand M.S. Evolutionary dynamics of CRISPR-cassettes in the metagenome Sorcerer II // Journal of Biomolecular Structure and Dynamics. 2009. C. 26. N. 6 (Proc. 16th Albany Conversation). P. 883.
- [9] Artamonova I.I., Sorokin V.A., Gelfand M.S. Browsing CRISPR-cassettes in the Sorcerer II metagenome. 17th Annu. Int. Conf. on Intelligent Systems for Molecular Biology and 8th European Conf. On Computational Biology ISMB/ECCB'09, 2009, Stockholm, Sweden.

- ДЛЯ ЗАМЕТОК -