На правах рукописи

Ставровская Елена Дмитриевна

КОМПЬЮТЕРНЫЕ МЕТОДЫ МАССОВОГО АНАЛИЗА РЕГУЛЯЦИИ ТРАНСКРИПЦИИ В БАКТЕРИЯХ

03.00.28 - биоинформатика

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2008

Работа выполнена в Учреждении Российской академии наук Институте проблем передачи информации им. А.А. Харкевича РАН.

Научный руководитель: кандидат физико-математических наук, доктор биологических наук, профессор
Миронов Андрей Александрович

Официальные оппоненты: доктор физико-математических наук, профессор
Туманян Владимир Гаевич
Институт молекулярной биологии имени В. А. Энгельгардта Российской академии наук

кандидат физико-математических наук
Ройтберг Михаил Абрамович
Институт математических проблем биологии Российской академии наук

Ведущая организация: Федеральное государственное унитарное предприятие Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов

Защита диссертации состоится __ _____ 2008 года в ____ часов на заседании диссертационного совета Д.002.007.02 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 __ _____ 2008 года.

Ученый секретарь диссертационного совета Рожкова Г.И.
доктор биологических наук, профессор

Общая характеристика работы

Актуальность темы:

В настоящее время, пожалуй, ни одна из естественных наук не обходится без применения компьютерных методов. Они позволяют моделировать природные процессы и системы, предсказывать их поведение, хранить и обрабатывать большие объемы данных. Биология не является исключением. Более того, на стыке биологии и компьютерных наук появилось новое самостоятельное научное направление — биоинформатика, которая использует компьютерные методы для решения биологических задач.

По мере развития экспериментальных методов секвенирование геномов становится все более быстрым и дешевым процессом. В связи с этим мы получаем все больше геномных последовательностей, которые нуждаются в содержательном описании. Ясно, что экспериментаторам невозможно справиться с таким растущим объемом данных, поскольку эксперимент требует больших временных и денежных затрат. С другой стороны, мы можем с помощью сравнения последовательностей близкородственных геномов предсказывать функции генов и их регуляцию по аналогии с известными геномами. И как раз здесь невозможно обойтись без компьютерных методов. Полученные предсказания можно проверить экспериментально, в этом случае уже понятно, где и что искать, что сильно облегчает работу экспериментаторам. Кроме того, с помощью компьютерных методов можно оценить значимость результатов, полученных в эксперименте.

Одной из важных задач в биоинформатике является поиск сайтов связывания транскрипционных факторов. Этой задачей ученые занимаются на протяжении многих лет, и существует огромное количество алгоритмов для ее решения. Тем не менее, задача является весьма сложной как вычислительно, так и биологически, и на сегодняшний день не существует универсального алгоритма, эффективно решающего задачу за приемлемое время. Основными трудностями при идентификации регуляторных мотивов являются недостаточный либо чрезмерный объем набора исходных последовательностей, слабая консервативность мотива, а также низкая доля последовательностей, содержащих сайт, в исходном наборе. Данная задача может быть сформулирована в терминах оптимизационной задачи и решена известными методами. При этом важно правильно выбрать параметры и оптимизируемый функционал, чтобы решить задачу максимально эффективно.

Алгоритм выделения регуляторных мотивов в наборе областей перед ортологичными генами позволяет найти мотив, но не сам белок-регулятор. Как правило,

один транскрипционный фактор регулирует в геноме сразу несколько генов. Сайты связывания одного белка-регулятора похожи. Поэтому, группируя похожие регуляторные мотивы, мы можем определить потенциальные группы совместно регулируемых генов (регулоны). Более того, если для каких-то генов из регулона известен фактор транскрипции, который их регулирует, то можно предсказать, что он регулирует и остальные гены регулона.

Цели и задачи работы:

Целью данной работы является разработка эффективных методов, алгоритмов и программных приложений для анализа регуляции транскрипции в геномах прокариот.

В ходе работы были поставлены следующие задачи:

1. Исследование возможности применения генетических алгоритмов к решению задачи поиска регуляторных мотивов в наборе областей, взятых перед ортологичными генами в группе близкородственных геномов бактерий.
2. Разработка методики и создание на ее основе программы для оценки статистической значимости экспериментально найденного дополнительного элемента основного промотора в геноме *Thermus aquaticus*.
3. Создание быстрого и эффективного алгоритма для кластеризации регуляторных мотивов и его применение для поиска новых членов известных регулонов, а также новых регулонов.
4. Создание программного конвеера для поиска регуляторных мотивов в рамках функциональных подсистем.

Методика исследования

Создание программных приложений на языке Java в среде программирования Eclipse. Тестирование эффективности алгоритмов на различных искусственных и биологических данных с последующим применением к биологическим задачам выделения регуляторных мотивов.

Научная новизна и практическая ценность

Реализованы генетические алгоритмы с различным способом выбора параметров и целевой функции и проведено их сравнение. Построена новая мера сходства регуляторных мотивов. Алгоритм кластеризации мотивов реализован в виде программного приложения и применялся для поиска новых регулонов, а также новых членов известных регулонов в группах геномов гамма-протеобактерий, фирмикутов и альфа-протеобактерий.

Программное приложение встроено в конвеер выделения регуляторных мотивов в рамках функциональных подсистем.

Апробация работы

Основные положения диссертации были представлены на следующих конференциях:

1. 3rd International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002), 14-20 July 2002, Novosibirsk, Russia.
2. 1st Moscow Conference on Computational Molecular Biology (MCCMB'03), 22-25 July 2003, Moscow, Russia.
3. 4th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004), 25-30 July 2004, Novosibirsk, Russia.
4. International conference Bioinformatics Italian Society (BITS'2005), 2005, Milan, Italy.
5. XII Международной конференции студентов, аспирантов и молодых ученых «Ломоносов», 12-16 апреля 2005, Москва, Россия.
6. 2nd Moscow Conference on Computational Molecular Biology (MCCMB'05), 18-21 July 2005, Moscow, Russia.
7. 5th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2006), 16-22 July 2006, Novosibirsk, Russia.
8. 5th European Conference on Computational Biology (ECCB'2006), 21-24 January 2007, Eilat, Israel.
9. 3rd Moscow Conference on Computational Molecular Biology (MCCMB'07), 27-31 July 2007, Moscow, Russia.
10. Информационные технологии и системы (ИТиС'07), 18-21 сентября 2007, Звенигород, Россия.
11. Научном семинаре Учебно-научного Центра "Биоинформатика" ИППИ РАН, 15 октября 2007, Москва, Россия.

Структура и объем диссертации:

Диссертационная работа содержит 98 страниц, в том числе 19 рисунков, 8 таблиц и список цитируемой литературы из 97 наименований. Текст диссертации включает введение, содержащее постановку задач, обзор литературы, и 4 главы, в которых дано решение поставленных задач и обсуждение результатов, и список цитированной литературы.

Содержание диссертации

В начале диссертационной работы приводится введение.

В главе 1 представлен обзор литературы по теме диссертации.

В главе 2 описываются два генетических алгоритма для решения задачи выделения регуляторных мотивов в наборе последовательностей ДНК. При описании алгоритмов используется ряд абстрактных понятий (их названия совпадают с некоторыми стандартными биологическими терминами, поэтому они в дальнейшем будут выделены курсивом): *геном*, *ген*, *аллель*, *качество генома*, *популяция*, *скрещивание*, *отбор и мутация.* Оба алгоритма осуществляют одну и ту же последовательность шагов, но различаются в определении понятия *геном* и его составляющих (*ген*, *аллель*, *качество генома*). Вначале дадим общее определение используемых понятий и опишем основной алгоритм.

Геном представляет собой набор *генов*. Конкретное значение, принимаемое *геном*, называется *аллелем*. Каждый *геном* характеризуется своим *качеством*.

Популяцией назовем набор *геномов* с конкретными фиксированными значениями *генов*. На каждой итерации алгоритм изменяет состав популяции с помощью следующих действий:

Скрещивание: случайным образом выбирается пара *геномов* и порождается новый, у которого одна часть берется из первого генома, а другая из второго. Позиция разреза *i* выбирается случайным образом.

Мутация: выбираем случайный *ген* в случайном *геноме* и меняем его *аллель* на другой случайным образом.

Отбор: исключается один *геном* с самым плохим *качеством*.

Алгоритм создает популяцию и последовательно осуществляет итерации, каждая из которых включает в себя операцию *скрещивания*, *мутации* и *отбора*. Критерием остановки алгоритма служит близкое к 0 значение среднеквадратичного отклонения функции *качества* по всей *популяции*, либо достижение предельного значения количества итераций.

Определение** генома **в Алгоритме 1: Каждый *ген* соответствует фрагменту ДНК из исходного набора, а каждый *аллель* – позиции в этом фрагменте (стартовой позиции сайта). Конкретный *геном* есть набор *аллелей*, то есть *геном* соответствует набору сайтов, по одному из каждого исходного фрагмента ДНК. Специфическое значение *аллели* (NAN)

означает, что данный фрагмент не содержит сайта. *Качество генома* определяется как информационное содержание соответствующего набора сайтов:

$$I = \sum_{k=1}^{l} \sum_{i=A,C,G,T} f(i,k) \log[\ f(i,k) / 0.25\ ] \quad (1)$$

$$f(i,k) = (n(i,k) + 0.35\sqrt{N}) / (N + 4 * 0.35\sqrt{N}) \quad (2)$$

где *f(i, k)* – частота встречаемости нуклеотида *i* в позиции *k*, $0.35\sqrt{N}$ и $4 * 0.35\sqrt{N}$ – псевдоотсчеты. Значение коэффициента 0.35 подобрано эмпирическим путем.

Определение** генома **в Алгоритме 2: Здесь под *геномом* будем понимать слово, представляющее собой потенциальный мотив. Каждая позиция в этом слове является *геном*, а нуклеотид, стоящий в этой позиции, *аллелем*. *Качество генома* определяется следующим образом:

- Для каждого фрагмента ДНК исходного набора ищется подпоследовательность, которая более всего похожа на слово-*геном* (имеющая наименьшее количество отличающихся нуклеотидов). Таким образом, мы получаем набор сайтов, по одному из каждой последовательности.
- В полученном множестве сайтов выбираем подмножество, обеспечивающее максимальное информационное содержание (1). *Качество генома* определяется как информационное содержание этого подмножества сайтов.

На графиках изображено поведение математического ожидания и среднего квадратичного отклонения *качества* с ростом количества итераций при различном числе *геномов* в *популяции* для алгоритмов 1 (Рисунок 1) и 2 (Рисунок 2).

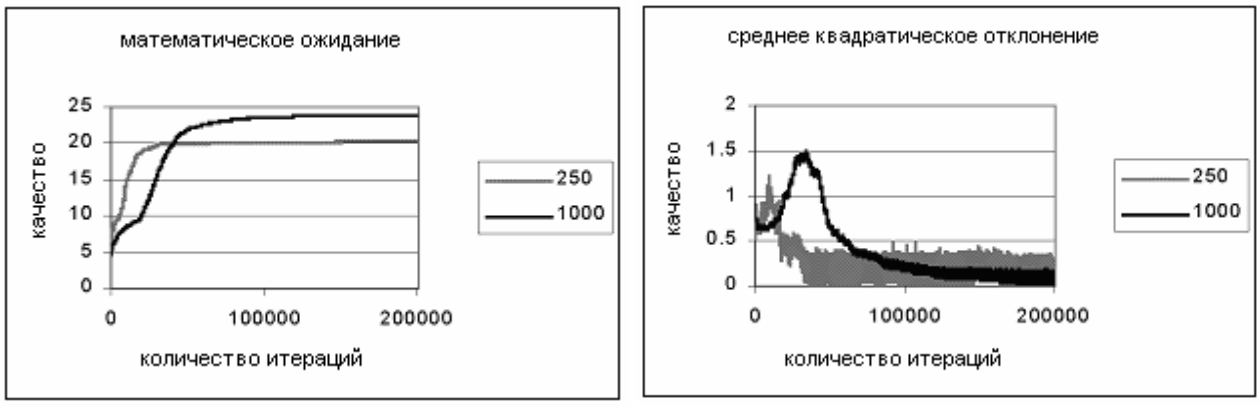


Рисунок 1. Графики поведения математического ожидания и среднего квадратичного отклонения функции *качества* для Алгоритма 1 на искусственной выборке (для *популяций* размером 250 и 1000 геномов).

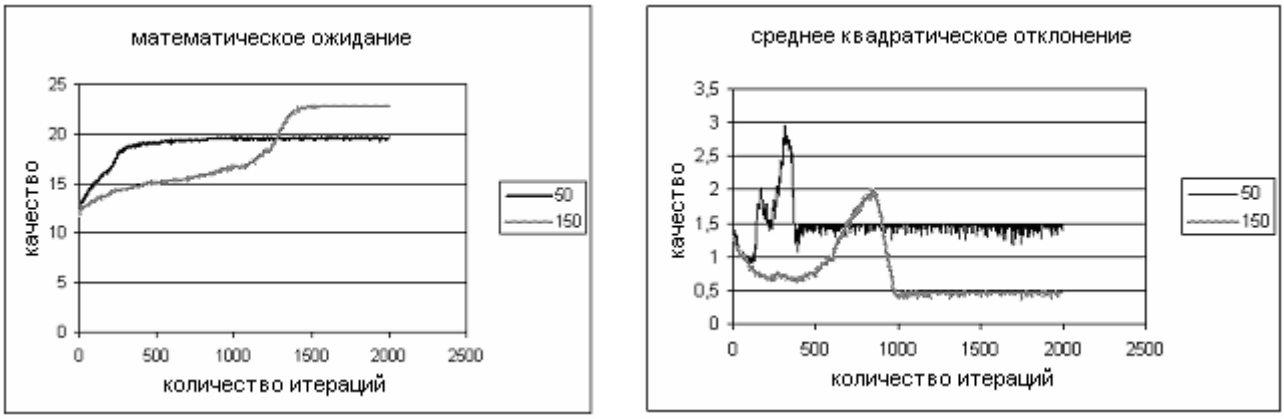


Рисунок 2. Графики поведения математического ожидания и среднего квадратичного отклонения функции *качества* для Алгоритма 2 на искусственной выборке (для *популяций* размером 50 и 150 геномов).

Нетрудно заметить, что для большего размера *популяции* выше максимальное значение, к которому сходится математическое ожидание, и меньше значение среднего квадратичного отклонения, хотя стационарные значения достигаются позже. Таким образом, чем больше число *геномов* в *популяции*, тем точнее находится мотив. Критерием остановки процесса служит прекращение роста математического ожидания функции *качества* (выход на горизонталь) и близкое к нулю значение среднего квадратичного отклонения. Алгоритм 2 требует меньшего размера *популяции*, чем Алгоритм 1. Количество итераций до достижения стационарного режима у Алгоритма 2 также меньше, но каждая итерация занимает большее время.

Оба алгоритма были протестированы на искусственной выборке. Тестирование показало, что Алгоритм 2 более эффективен, чем Алгоритм 1. Алгоритм 2 был также протестирован на двух реальных выборках из генома *E.coli* (кишечной палочки). Для каждой выборки были известны положения сайтов в исходных последовательностях. При составлении тестов мы вырезали один за другим сайты из последовательностей выборки. Таким образом, все меньшее число последовательностей в тесте содержало искомый мотив. Каждый раз мы убирали самый сильный сайт, то есть сайт, имеющий наибольший вес относительно матрицы позиционных весов мотива. Результаты представлены на рисунках 3 и 4. Для сравнения, мы применили к тем же выборкам алгоритм *SeSiMCMC*[1]. Для оценки полученных результатов был использован следующий критерий:

[1] *Favorov A. V., Gelfand M. S., Gerasimova A. V., Ravcheev D. A., Mironov A. A., Makeev V. J.* A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length // Bioinformatics – 2005 – Vol. 21(10) – Pp. 2240-2245.

Опр. Истинный сайт (real) называется найденным (found), если существует предсказанный, который перекрывает его не менее, чем на половину.

Опр. Предсказанный сайт (predicted) называется хитом (hit), если существует истинный, перекрывающий его не менее, чем на половину.

Далее вычисляются две функции S_f = found/real и S_h = hit/predicted.

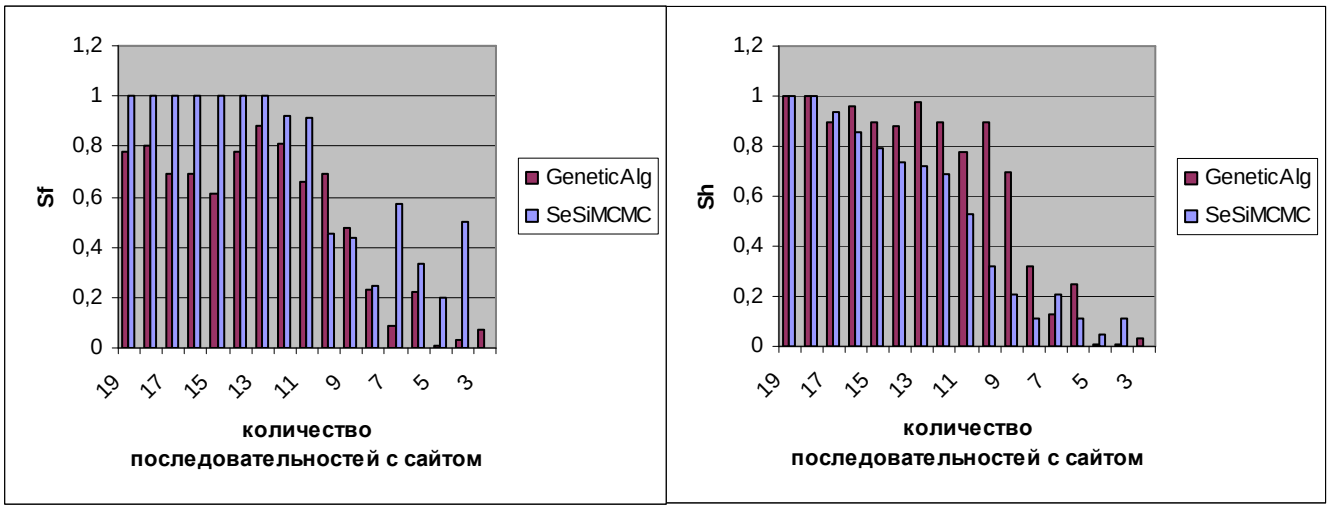


Рисунок 3. Значения функций Sf и Sh для пуринового теста для Алгоритма 2 и SeSiMCMC

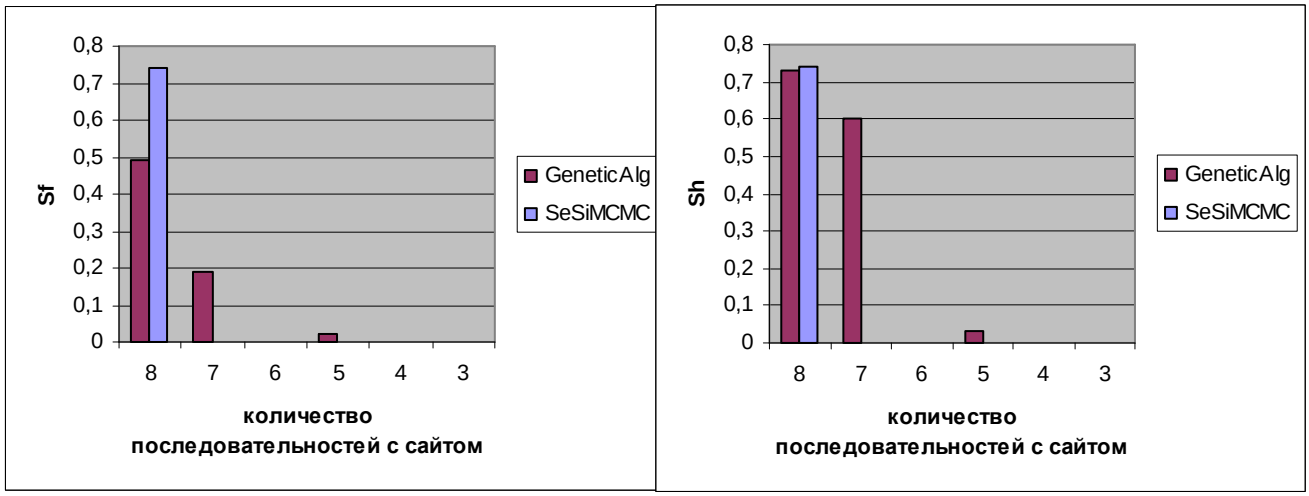


Рисунок 4. Значения функци функций Sf и Sh для аргининового теста для Алгоритма 2 и SeSiMCMC

В главе 3 рассказывается об оценке статистической значимости эксперимента, в результате которого был найден дополнительный элемент базального промотора в геноме *T.thermophilus*. Вначале дается краткое описание эксперимента. Целью эксперимента было установить участки ДНК *Thermus thermophilus*, которые связываются с базальным сигма-

фактором и, тем самым, важны для инициации транскрипции. Применялся метод рандомизации двух цепей исходной ДНК с последующей выборкой тех участков, которые хорошо связываются с белком (SELEX, systematic evolution of ligands by exponential evolution). Были выделены одноцепочечные аптамеры ДНК, которые связываются с сигма-субъединицей с высокой аффинностью. Эти аптамеры содержали мотив, подобный боксу «-10» промотора, за которым следовал новый дополнительный мотив из 4-х нуклеотидов. Далее было показано, что мотив дополнительного аптамера работает как элемент базального промотора и допускает узнавание промотора РНК-полимеразой *T.aquaticus* в отсутствие бокса «-35» промотора.

Из литературы и открытых баз данных известно мало последовательностей промоторов бактерий рода *Thermus*, большая часть этих последовательностей принадлежит геному *T.thermophilus*. На рисунке 13 представлены все известные -35/-10 последовательности промоторов из *T.thermophilus*.

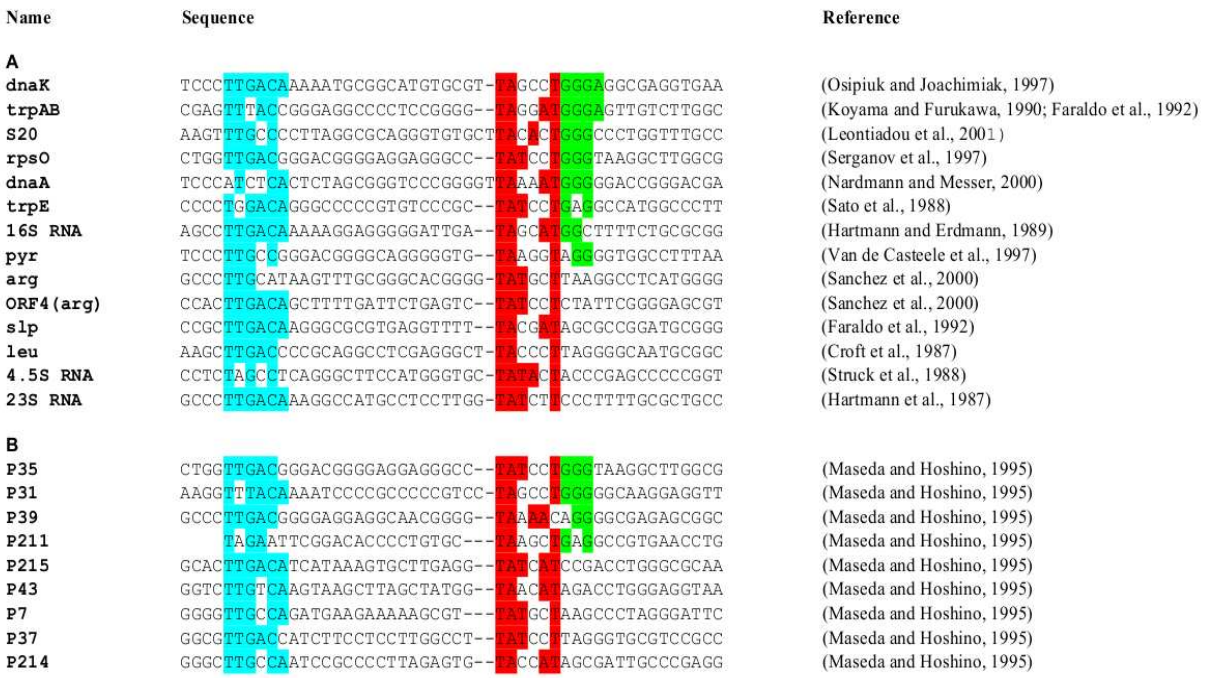


Рисунок 5. Последовательности промоторов *T.thermophilus*

Из рисунка видно, что два промотора содержат участок GGGA, найденый в аптамерах базального σ-фактора. Существенная часть промоторов (7 из 22) содержат последовательность GGG перед боксом «-10» промотора. Для контроля, только один промотор содержит последовательность CCC. Ограниченное количество анализируемых промоторов не позволяет сделать статистически значимых выводов, но все же эти данные демонстрируют, что GGGA-подобные мотивы играют важную роль при распознавании

естественных промоторов *T.thermophilus.* Для дальнейшего анализа встречаемости GGGA-подобных мотивов в промоторах *T.thermophilus* был проведен биоинформатический анализ промоторов в данном геноме.

Для предсказания последовательностей промоторов были построены позиционные весовые матрицы (профили) для консервативных блоков -35-го и -10-го боксов промотора, основываясь на 22 промоторах из *T.aquaticus* и *T.thermophilus*, известных из литературы. Полученными профилями просматривались две выборки межгенных участков всего генома *T.thermophilus.* Первая выборка состояла из областей между генами, которые транскрибируются дивергентно (дивергоны), а вторая – из генов, которые тринскрибируются конвергентно (конвергоны). Дивергоны должны содержать промоторы, которые используются при транскрипции в обоих направлениях, а конвергоны не должны содержать промоторов и, таким образом, служат контрольной выборкой. В итоге, была составлена выборка из 308 дивергонов длиной в среднем 135 нуклеотидов и 404 конвергона длиной в среднем 133 нуклеотида.

В каждой последовательности из двух выборок с помощью профиля определялся потенциальный промотор как подпоследовательность, получившая наибольший вес по данному профилю. Затем были удалены потенциальные промоторы, вес которых был меньше худшего веса промотора из исходного набора (по которому строился профиль). После этого осталось 115 потенциальных промоторов в дивергонах (37% от общего количества дивергонов) и 63 в конвергонах (16% от общего числа конвергонов).

Большая доля потенциальных промоторов в дивергонах по сравнению с конвергонами показывает, что построенные профили имеют некоторую, хотя и не абсолютную, предсказательную силу, сравнимую с другими инструментами предсказания промоторов[2].

Затем мы проверили, присутствует ли в предсказанных промоторах сразу после -10-го бокса мотив GGG (мы не искали полный мотив GGGA, поскольку ранее было показано, что последний нуклеотид A менее важен для промотора, чем первые GGG). Мы также подсчитали количество потенциальных промоторов, у которых по крайней мере два из трех нуклеотидов сразу после -10-го бокса были G. В качестве контроля, мы посчитали количество потенциальных промоторов, имеющих по крайней мере два C в этих позициях. Результат показал, что 33% промоторов в дивергонах содержат две или более G сразу после -10-го бокса, в то время как лишь 19% содержат два или более C. Для сравнения, в

[2] *Robison K., McGuire A.M., Church G.M.* A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome // J Mol Biol. – 1998 – Vol. 284(2) – Pp. 241-254.

контрольной выборке промоторов в конвергонах 24% содержат два или более G и 22% два или более C. Таким образом, мы показали, что предсказанные промоторы в дивергонах *T.thermophilus* насыщены GGG-подобными мотивами сразу после -10-го бокса по сравнению с контрольной выборкой. Аналогичный анализ 197 промоторов *E.coli* из базы данных *DPInteract*[3] показал, что только 7% из них содержат GGG-подобный мотив и 11% CCC-подобный мотив, что подтверждает специфичность GGG-подобных мотивов после -10 бокса для генома *T.thermophilus.*

В главе 4 описывается алгоритм кластеризации регуляторных мотивов ClusterTree-RS. Алгоритм состоит из двух частей – построение бинарного дерева и его обход. На стадии обхода алгоритм выделяет узлы дерева, которые соответствуют кластерам.

Каждому узлу дерева соответствует множество выравненных сайтов, а каждому листу – один мотив из исходного набора. Построение бинарного дерева осуществляется при помощи алгоритма простого связывания, который является итеративным процессом. На каждой итерации имеется набор поддеревьев (сначала – набор листьев, т.е. полный набор исходных мотивов). Алгоритм выбирает два наиболее близких поддерева и объединяет их в одно. В качестве меры расстояния между поддеревьями используется *функция сходства* соответствующих матриц выравниваний, которая будет рассмотрена чуть ниже. Корень полученного поддерева содержит все сайты своих правого и левого дочерних узлов. За одну итерацию количество поддеревьев уменьшается на 1, и в итоге мы получаем одно бинарное дерево, в корне которого лежат все мотивы исходной выборки.

Функция сходства между поддеревьями (т.е. между соответствующими им мотивами) вычисляется по формуле

$$D = \sum_{k=1}^{l} I_k \frac{\sum_{i=A,C,G,T} (f_1(i,k) - \bar{f}_1(k))(f_2(i,k) - \bar{f}_2(k))}{F_1 F_2}, \qquad (3)$$

$$F_j = \sqrt{\sum_{i=A,C,G,T} (f_j(i,k) - \bar{f}_j(k))^2},$$

где I_k – информационное содержание в позиции k для общего набора сайтов, $f_j(i, k)$ – частота встречаемости нуклеотида i в позиции k для набора сайтов поддерева j, $\bar{f}_j(k)$ – среднее значение частоты в столбце k, $0.25\alpha\sqrt{N_j}$ и $\alpha\sqrt{N_j}$ – псевдоотсчеты. Чем больше D, тем ближе поддеревья.

На этапе обхода алгоритм просматривает все узлы дерева и выделяет те из них, которые соответствуют кластерам. Каждому узлу дерева соответствует набор сайтов,

[3] *Robison K., McGuire A.M., Church G.M.* A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome // J Mol Biol. – 1998 – Vol. 284(2) – Pp. 241-254.

полученный слиянием двух наборов сайтов дочерних узлов. Если группы сайтов дочерних узлов похожи, то есть, схожи соответствующие им матрицы счетчиков нуклеотидов, то они могут принадлежать одному кластеру (т.е. нести один и тот же мотив). Если же матрицы сильно различаются, то это означает, что в данном месте дерева сливаются две различные группы сайтов, соответствующие двум различным мотивам. Для того, чтобы определить, являются матрицы счетчиков нуклеотидов схожими или различными, вычисляется логарифм отношения правдоподобия:

$$R = \sum_{k=1}^{l} \log \left[\frac{(N_1 + L - 1)!(N_2 + L - 1)! \prod_{i=1}^{L} (n_1(i,k) + n_2(i,k))!}{(L-1)!(N_1 + N_2 + L - 1)! \prod_{i=1}^{L} (n_1(i,k))! \prod_{i=1}^{L} (n_2(i,k))!} \right], \qquad (4)$$

где L – размер алфавита (L = 4), $n_j(i,k)$ количество нуклеотидов типа i в позиции k в дочернем узле, N_j – количество сайтов в дочернем узле. Узел соответствует кластеру, если логарифм отношения правдоподобия является положительным для этого узла и отрицательным для его родительского узла.

При применении базового алгоритма к реальным данным возникает ряд сложностей. Во-первых, непонятно, как вычислять функцию сходства D между мотивами разной длины? Во-вторых, программы поиска мотивов часто находят мотив неточно, то есть с некоторым смещением или частично. Поэтому при кластеризации мотивы нужно выравнивать друг относительно друга.

Алгоритм *ClusterTree* накладывает сравниваемые мотивы друг на друга со всеми допустимыми смещениями и выбирает такое наложение, при котором функция D сходства между мотивами принимает максимальное значение (т.е. мотивы наиболее скоррелированы). При выравнивании мотивов алгоритм *ClusterTree-RS* дополняет все сайты до длины объединения, подгружая необходимые участки из генома. Таким образом, мы восполним недостающую информацию в случае, если мотив был найден не точно. К тому же, теперь мы вычисляем сходство между мотивами одинаковой длины.

Бывает так, что мотив, найденный с помощью программы поиска, содержит ложно предсказанные сайты. Даже если доля таких сайтов в мотиве невелика, они искажают мотив. Отсеять такие сайты на начальном этапе довольно трудно, поскольку исходные мотивы могут быть идентифицированы по небольшому количеству последовательностей, и у нас нет никакой априорной информации о том, каким должен быть правильный мотив. Однако когда схожие мотивы собраны в один кластер, в нашем распоряжении имеется мотив, состоящий из достаточно большого количества сайтов. Для улучшения качества полученных кластеров алгоритм осуществляет процедуру удаления ложных сайтов.

Для того, чтобы сравнить алгоритм ClusterTree-RS с другими существующими методами кластеризации, он был протестирован на искусственной выборке, описанной в (Qin *et. al.* 2003)[4]. Для построения тестовой выборки были выбраны 43 транскрипционных фактора *E.coli* из базы данных *DPInteract* и соответствующие им регуляторные сайты (общим числом 356). Частотная матрица для каждого фактора обрезалась справа до длины 15 и дополнялась справа и слева 3 случайными позициями. Для каждой матрицы порождались искусственные регуляторные мотивы с таким же распределением частот нуклеотидов, количество таких мотивов совпадало с числом известных сайтов для этого регулятора. Таким образом, каждый тестовый файл содержал 356 (по числу сайтов) искусственных регуляторных мотивов. Число сайтов в мотиве выбиралась случайным образом в трех диапазонах: 2-4, 2-10 и 5-10 сайтов. Для каждого диапазона было создано по 100 тестовых файлов. Результаты тестирования представлены в Таблице 1.

Таблица 1. Сравнение средних значений ошибки для различных существующих методов кластеризации.

Размер мотива	**Ошибка[a]**	**KM[b]**	**HC-1[c]**	**HC-2[d]**	**BMC-1[e]**	**BMC-2[f]**	**ClusterTree-RS[g]**
2-4 сайтов	**Частота ошибок, % (E, σ)**	**32,9 (5,2)**	**26,3 (6,1)**	**63,7 (1,7)**	**9,0 (1,7)**	**5,8 (0,2)**	**10,7 (2,0)**
	Кол-во кластеров, (E, σ)	43	43	250 (4,2)	34,4 (1,5)	38,0 (1,3)	49,8 (2,2)
2-10 сайтов	**Частота ошибок, % (E, σ)**	**33,4 (3,9)**	**14,1 (3,1)**	**25,7 (2,4)**	**3,3 (1,0)**	**2,5 (0,1)**	**5,7 (1.0)**
	Кол-во кластеров, (E, σ)	43	43	118,9 (6,7)	40,6 (1,2)	41,6 (0.6)	46,8 (1,7)
5-10 сайтов	**Частота ошибок, % (E, σ)**	**31,6 (4,6)**	**3,9 (1,1)**	**11,0 (1,5)**	**2,6 (0,4)**	**2,2 (0,0)**	**3,6 (0,7)**
	Кол-во кластеров, (E, σ)	43	43	66,0 (3,9)	41,4 (0,7)	42,0 (0,1)	43.2 (1.0)

[a]Величина ошибки определяется как среднее число неправильно определенных в кластер, усредненное по 100 тестам и данное в процентах.

[b]Метод K-средних выполнен с помощью функции *kmeans* из пакета программ Splus. Для подсчета матрицы расстояний использовалось расстояние Кульбака-Лейбера. Количество кластеров k=43 полагалось известным.

[c]Иерархическая кластеризация с помощью программ CompareACE и Tree[5]. Значение порога выбрано таким образом, чтобы получилось правильное число кластеров (43).

[4] *Qin Z.S., McCue L.A., Thompson W., Mayerhofer L., Lawrence C.E., Liu J.S.* Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites // Nat Biotechnol. – 2003 – Vol. 21(4) – Pp. 435-439.

[5] *Hughes J.D., Estep P.W., Tavazoie S., Church G.M.* Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae // J Mol Biol. – 2000 – Vol. 296(5) – Pp. 1205-1214.

[d]Иерархическая кластеризация теми же методами, но пороговое значение фиксировано и равно рекомендованному значению 0.7.
[e]Сокращенная версия алгоритма BMC[6] (без выбора ширины мотива).
[f]Полная версия алгоритма BMC.
[g]ClusterTree-RS.

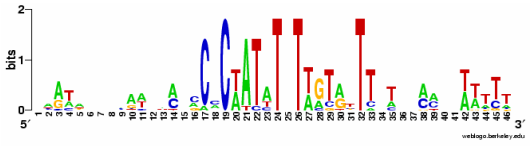
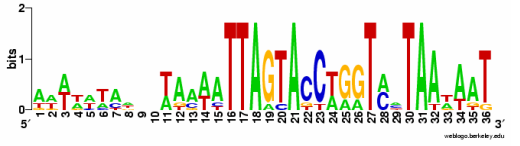
Из таблицы видно, что алгоритм дает неплохие результаты, уступая только алгоритму BMC, который определяет мотивы в кластеры в соответствии с апостериорным распределением вероятностей. Однако, наш алгоритм работает существенно быстрее алгоритма BMC (выборку из 215 мотивов программа BMC обрабатывает за время порядка 2,5 часов, программа ClusterTree-RS за время порядка 5 минут), поскольку не переопределяет многократно мотивы в кластеры, а строит единую иерархическую структуру. Когда мы пытаемся осуществить кластеризацию всех потенциальных регуляторных мотивов для одного генома (группы родственных геномов), мы имеем выборку из нескольких тысяч мотивов. Наш алгоритм является кубическим по времени и обработает такую выборку за несколько часов, в то время как алгоритм BMC экспоненциален по времени и будет работать несколько недель.

Алгоритм был применен к выборкам потенциальных регуляторных мотивов из гамма-протеобактерий (EC выборка) и фирмикутов (BS выборка). Потенциальные регуляторные мотивы были получены путем применения программы поиска регуляторных мотивов *SeSiMCMC* к наборам областей, выделенных перед ортологичными генами[7]. Были найдены кластеры, соответствующие известным мотивам, которые содержали новые сайты (то есть были найдены новые гены для известных регулонов). Были также обнаружены новые мотивы (то есть новые потенциальные регулоны), которые перечислены в Таблице 2:

[6] *Qin Z.S., McCue L.A., Thompson W., Mayerhofer L., Lawrence C.E., Liu J.S.* Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites // Nat Biotechnol. – 2003 – Vol. 21(4) – Pp. 435-439.

[7] *Danilova L.V., Lyubetsky V.A., Gelfand M.S.* An algorithm for identification of regulatory signals in unaligned DNA sequences, its testing and parallel implementation // In Silico Biol. – 2003 – Vol. 3(1-2) – Pp.33-47.

Таблица 2. Потенциальные новые регулоны, найденные при применении алгоритма ClusterTree-RS к выборкам ES и BS.

выборка	N	Гены	функция	logo
EC	1	*nrdD*	фермент; метаболизм 2'-деоксирибонуклеотидов	
		nrdA	фермент; метаболизм 2'-деоксирибонуклеотидов	
		ubiE	фермент; биосинтез кофакторов, переносщики: менахинон, убихинон	
		proS	фермент; аминоацил-тРНК-синтетаза, модификация тРНК	
BS	2	*pyrR*	аттенюация (антитерминация) пиримидинового опрерона (pyrPBCADFE) в присутствии UMP (биосинтез пиримидина)	
		pyrP	биосинтез пирримидинов	
		*pyrF**	биосинтез пирримидинов биосинтез пиримидинов	
BS	3	*ylpC (fapR)*	Неизвестная функция	
		yhfB (yhfC)	Неизвестная фенкция	

* Указано имя гена *B.subtilis*, соответствующее мотиву, однако сайт непосредственно перед геном *B.subtilis* не найден или удален из кластера как ложный. Кластер содержит сайты, найденные перед ортологичными генами в родственных организмах.

Был произведен анализ геномов альфа-протеобактерий. Ортологические ряды были получены из базы данных Phogs[8]. Потенциальные регуляторные сигналы были выделены с помощью программы SignalX[9]. Было получено 9 потенциальных новых регулонов: центральный метаболизм, транспортеры азотсодержащих кислот, кислород-индуцируемая регуляция фиксации азота, биосинтез тиамина, биосинтез рибосомальных белков, метаболизм ДНК и РНК, метаболизм цинка, защита от кислорода и Метаболизм марганца.

В главе 5 описывается конвеер для автоматического поиска регуляторных мотивов перед генами в рамках функциональных подсистем. Конвеер позволяющий осуществлять поиск потенциальных регуляторных мотивов в рамках одной функциональной подсистемы для заданной группы геномов.

[8] *Merkeev I.V., Novichkov P.S., Mironov A.A.* PHOG: a database of supergenomes built from proteome complements // BMC Evol Biol – 2006 – Vol. 22 – Pp. 6-52

[9] *Mironov A.A., Vinokurova N.P.*, Gel'falnd M.S. Software for analyzing bacterial genomes // Mol Biol (Mosk). – 2000 – Vol. 34(2) – Pp. 253-262.

Функциональные подсистемы были взяты из базы данных SEED[10]. Подсистема состоит из функциональных ролей, которые составляют определенный биологический процесс или структурный комплекс. Функциональная роль – это функция, которую выполняет белок. Подсистему можно рассматривать как обобщенный метаболический путь и представить в виде обобщенной таблицы. Каждая колонка таблицы соответствует одной функциональной роли подсистемы, а строка – одному геному. Ячейка таблицы содержит гены, которые кодируют белки с выбранной функцией в рассматриваемом геноме.

Конвеер включает в себя процедуру, которая убирает слишком близкие геномы из исходной группы. Для определения расстояний между геномами мы использовали филогенетическое дерево организмов, которое построено по множественному выравниваю высококонсервативных белков из 74 кластеров ортологичных генов из базы данных COGs[11]. Нами было выбрано пороговое значение 0,01. Геномы, расстояние между которыми меньше порога, считаются слишком близкими. Пороговое значение подобрано так, что близкими геномами считаются штаммы и некоторые близкие виды (например, *Neisseria gonorrhoeae* и *Neisseria meningitidis*). Поскольку функциональная подсистема может содержать гены не из всех запрашиваемых геномов, отсев близких геномов производится после проецирования группы геномов на исследуемую подсистему.

Конвеер реализует три стратегии поиска мотивов. Первая – искать мотивы отдельно для каждого генома (перед генами из каждой строки таблицы), а затем кластеризовать полученные мотивы. Вторая – искать мотивы в рамках каждой функциональной роли (перед генами из каждого столбца таблицы) и кластеризовать результаты. Третья – искать мотивы для всех генов подсистемы (перед всеми генами таблицы).

Для поиска мотива в наборе областей перед генами мы использовали алгоритм *SignalX*.

Для кластеризации полученных мотивов мы использовали алгоритм ClusterTree-RS. Конвеер позволяет искать палиндромные безделеционные мотивы фиксированной длины. На вход подается описание функциональной подсистемы и группа геномов, в рамках которых осуществляется поиск, а также длина искомого мотива.

Каждый найденный мотив характеризуется функциями оценки качества мотива – информационным содержанием и p-value. Об истинности мотива можно судить по

[10] *Overbeek R. et al.* The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes // Nucleic Acids Research – 2005 – Vol. 33(17) – Pp. 5691-5702.

[11] *Tatusov R.L., Koonin E.V., Lipman D.J.* A genomic perspective on protein families // Science. – 1997 – Vol. 278(5338) – Pp. 631-637.

значениям этих функций, а также по структуре самого сигнала. Как правило, это определяет эксперт. Для того, чтобы облегчить ему задачу, мы попытались классифицировать ложнопредсказанные мотивы и выработать способы их устранения.

В областях перед генами встречаются так называемые участки низкой сложности, состоящие из нуклеотидов двух (реже одного) типов (например, G и C), повторяющихся с некоторой периодичностью. Такие участки могут иметь функциональное значение, однако они редко являются сайтами связывания факторов транскрипции. Если эти области встречаются перед несколькими генами выборки, то при поиске эти участки в силу искаженного нуклеотидного состава определяются как консервативный потенциальный мотив. Причем, если область низкой сложности длиннее искомого мотива, то в силу периодичности будет считаться, что она содержит несколько перекрывающихся сайтов. Например, в последовательности GCGCGCGCGCGCGCGCGCGC (20 нуклеотидов) содержится 3 консервативных сайта длины 16 со сдвигом 2. Такие ложные мотивы составляют большую часть перепредсказаний.

Для борьбы с перепредсказаниями такого типа мы применяем процедуру фильтрации областей низкой сложности. Перед применением программы поиска все последовательности сканируются окном заданной длины. Каждая следующая подпоследовательность получена из исходной последовательности путем сдвига окна на один из шагов (от одного до пяти нуклеотидов). Если соседние подпоследовательности отличаются не более чем на заданное количество замен, то участок, покрываемый этими подпоследовательностями, объявляется областью низкой сложности. Каждый нуклеотид участка низкой сложности заменяется в последовательности на X, и в дальнейшем эти участки игнорируются.

При поиске палиндромных мотивов находятся мотивы, у которых одно плечо более консервативно, чем другое. При кластеризации такая ассиметрия может накапливаться, поскольку мотивы группируются за счет консервативности одного плеча. При этом в другом плече группируемых мотивов нуклеотидный состав может несколько отличаться, поэтому после кластеризации распределение нуклеотидов в этом плече становится близким к случайному. Ясно, что такие мотивы уже не являются палиндромными.

Для борьбы с такими перепредсказаниями перед кластеризацией во всех мотивах число сайтов удваивается путем добавления к каждому сайту комплементарного. Таким образом, мотивы уравновешиваются и становятся строгими палиндромами.

На рисунке 14 изображены диаграммы, соответствующие трем моделям. Точки на диаграммах соответствуют полученным мотивам. Черными квадратами изображены известные мотивы (охарактеризованные экспертом как правильно найденные). По осям

диаграмм — характеристики качества мотивов: log(p_value) и информационное содержание (inf). Из сопоставления диаграмм а) и б) видно, что при добавлении комплементарных сайтов области черных и серых точек перекрываются меньше. При добавлении фильтра областей низкой сложности количество перепредсказаных мотивов сильно сократилось. На диаграмме можно выделить область, которая содержит все подтвержденные мотивы (log(p_value)<-100 и inf>10), и область, которая практически не содержит неподтвержденных мотивов (log(p_value)<-200 и inf>12.5).

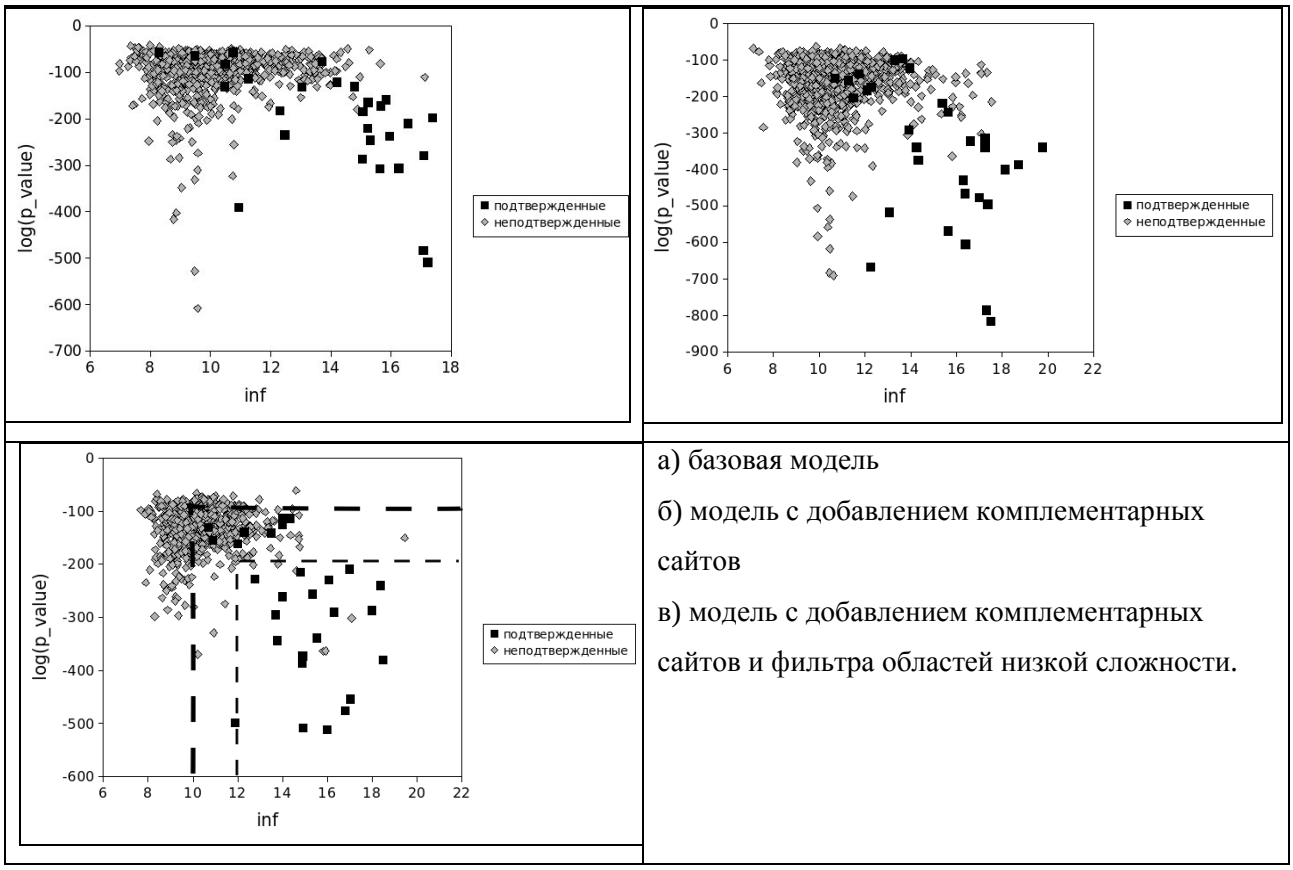


Рисунок 14. Сопоставление характеристик качества (log(p_value) и информационного содержания) найденных мотивов.

Выводы.

1. Разработана и реализована программно методика применения генетитических алгоритмов к задаче поиска регуляторных мотивов в наборе областей перед ортологичными генами в геномах прокариот. Показано, что с помощью этих алгоритмов можно эффективно искать сайты связывания транскрипционных факторов белок-ДНКового взаимодействия.
2. Разработана и применена методика для проверки статистической значимости результатов эксперимента, показавшего новый дополнительный элемент основного промотора в геноме *Thermus aquaticus.*

3. Создано и тестировано эффективное программное средство – алгоритм и реализующая его компьютерная программа для кластеризации регуляторных мотивов из геномов прокариот. На ее основе предсказаны новые потенциальные члены известных регулонов, а также потенциальные новые регулоны в геномах гамма-протеобактерий, фирмикутов и альфа-протеобактерий.
4. Создан и тестирован программный конвеер для поиска регуляторных мотивов в рамках функциональных подсистем, в который встроена процедура кластеризации.

Список публикаций по теме диссертации.

1. *Stavrovskaya E.D., Mironov A.A.* Two genetic algorithms for identification of regulatory signals // In Silico Biol. 2003. Vol. 3(1-2). P. 49-56.
2. *Stavrovskaya E.D., Mironov A.A.* Clustering regulatory signals by binary trees // Biophysics (Moscow). 2003. Vol. 48 Suppl. 1. P. S17-S20.
3. *Ставровская Е.Д., Макеев В.Ю., Миронов А.А.* ClusterTree-RS: алгоритм кластеризации регуляторных сигналов с помощью бинарного дерева // Молекулярная биология. 2006. Т. 40. №3. С. 465-473.
4. *Feklistov A., Barinova N., Sevostyanova A., Heyduk E., Bass I., Vvedenskaya I., Kuznedelov K., Merkiene E., Stavrovskaya E., Klimasauskas S., Nikiforov V., Heyduk T., Severinov K., Kulbachinskiy A.* A basal promoter element recognized by free RNA polymerase sigma subunit determines promoter recognition by RNA polymerase holoenzyme // Mol Cell. 2006. Vol. 23. № 1. P. 97-107.
5. *Миронов А.А., Ставровская Е.Д., Макеев В.Ю.* Способ исследования совместной регуляции генов бактерий и прогнозирования содержания новых регулонов и функций генов. 2006. регистрационный номер 2006127264.
6. *Stavrovskaya E.D., Mironov A.A.* A genetic algorithm for identification of regulatory signals. // Proc. 3d International Conference on Bioinformatics of Genome Regulation and Structure (BGRS’2002). 2002. Vol. 1. P. 26-27.
7. *Stavrovskaya E.D., Mironov A.A.* Binary tree for clusterization of regulatory signals // Proc. Moscow Conference on Computational Molecular Biology (MCCMB'03). 2003. P. 218-219.
8. *Stavrovskaya E.D., Mironov A.A.* Binary tree for clustering of regulatory signals // Proc. 4th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS’2004). 2004. Vol. 1. P.195-199.
9. *Stavrovskaya E.D., Mironov A.A.* Binary tree for clustering of regulatory signals. // Proc. International conference BITS’2005. 2005. P. 85.

10. *Ставровская Е.Д., Миронов А.А.* ClusterTree: программа кластеризации регуляторных сигналов с помощью бинарного дерева // Материалы XII Международной конференции студентов, аспирантов и молодых ученых «Ломоносов». 2005. С. 36-37.
11. *Stavrovskaya E.D., Makeev V.J., Mironov A.A.* ClusterTree-RS: The binary tree algorithm for identification of co-regulated genes by clustering regulatory signals // Proc. Moscow Conference on Computational Molecular Biology (MCCMB'05). 2005. P. 385.
12. *Stavrovskaya E.D., Makeev V.J., Merkeev I.V., Mironov A.A.* Tool for automatic aetection of co-regulated genes. // Proc. 5'th European Conference on Computational Biology (ECCB'2006). 2007.
13. *Stavrovskaya E.D., Makeev V.J., Merkeev I.V., Mironov A.A.* Tool for automatics detection of co-regulated genes // Proc. 5th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2006). 2006. Vol.1. P.172-175.
14. *Stavrovskaya E.D., Cipriano M., Dubchak I.L., Mironov A.A., Gelfand M.S.* Automated search for regulatory motifs in upstream regions of genes from the functional subsystems // Proc. Moscow Conference on Computational Molecular Biology (MCCMB'07). 2007. P. 283.
15. *Ставровская Е.Д., Сиприано М., Дубчак И.Л., Миронов А.А., Гельфанд М.С.* Автоматический поиск регуляторных сигналов перед генами в рамках функциональных подсистем. // Труды конференции Информационные технологии и системы (ИТиС'07). 2007. С. 330-331.

Благодарности: Автор выражает искреннюю благодарность своим научным руководителям, Андрею Александровичу Миронову и Михаилу Сергеевичу Гельфанду, за руководство, помощь и поддержку при выполнении диссертации, а также Роману Сутормину, Всеволоду Юрьевичу Макееву, Ольге Калининой и Дмитрию Виноградову, за участие, ценные советы и продуктивное обсуждение.

- ДЛЯ ЗАМЕТОК –

Ставровская Елена Дмитриевна

КОМПЬЮТЕРНЫЕ МЕТОДЫ МАССОВОГО АНАЛИЗА РЕГУЛЯЦИИ ТРАНСКРИПЦИИ В БАКТЕРИЯХ

Работа посвящена разработке эффективных методов, алгоритмов и программных приложений для анализа регуляции транскрипции в геномах прокариот. Исследовалась возможность применения генетических алгоритмов к задачи поиска регуляторных мотивов в наборе областей, взятых перед ортологичными генами в группе близкородственных геномов. Разработаны генетические алгоритмы с различным способом выбора параметров и целевой функции и произведено их сравнения. Показано, что генетический алгоритм, использующий алгебраический способ описания мотива, является более эффективным.
Разработана и применена методика для оценки статистической значимости экспериментально найденного дополнительного элемента основного промотора в геноме *Thermus aquaticus.*
Построена новая мера сходства регуляторных мотивов. На ее основе построен быстрый и эффективный алгоритм кластеризации регуляторных мотивов. С помощью алгоритма предсказаны новые члены известных регулонов, а также потенциальные новые регулоны в геномах гамма-протеобактерий, фирмикутов и альфа-протеобактерий.
Создан и тестирован программный конвеер для поиска регуляторных мотивов в рамках функциональных подсистем, в который встроена процедура кластеризации.

Stavrovskaya Elena Dmitrievna

COMPUTATIONAL METHODS FOR THE LARGE-SCALE ANALYSIS OF TRANSCRIPTIONAL REGULATION IN BACTERIAL GENOMES

Efficient methods, algorithms and applications for the analysis of transcription regulation in procariotic genomes were developed. Genetic algorithms were applied to identification of the regulatory motifs in a set of orthologous upstream regions from closely related genomes. We developed and compared genetic algorithms with different selection of parameters and criterion function and demonstrated that the genetic algorithm with an algebraical motif interpritation is more efficient.

We developed and applied technique to estimate the statistical significance of an additional basal promoter element experimentally found in the *Thermus aquaticus* genome.

We devepoled a new motif similarity measure. Based on this measure we developed a fast and efficient algorithm for clustering regulatory motifs. Using this algorithm we predicted new potential members of known regulons and new potential regulons in the genomes of the gamma-proteobacteria, firmicutes and alpha-proteobacteria.

We developed and tested a pipeline searching for regulatory motif upstream of genes from functional subsystems. This pipeline utilizes the developed clustering procedure.