На правах рукописи

Кулаковский Иван Владимирович

Построение оптимальных моделей ДНК-сайтов связывания факторов транскрипции высших эукариот на основе данных различных экспериментальных методов

03.00.28 Биоинформатика

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва — 2009

Работа выполнена в лаборатории биоинформатики и системной биологии Учреждения Российской академии наук Института молекулярной биологии им. В.А. Энгельгардта РАН

Научный руководитель:
д.ф.-м.н., проф. Владимир Гайевич Туманян

Официальные оппоненты:
д.б.н. Мария Георгиевна Самсонова,
Санкт-Петербургский государственный политехнический университет
к.ф.-м.н. Андрей Владимирович Алексеевский,
Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского, МГУ им. М.В. Ломоносова

Ведущая организация:
Учреждение Российской академии наук
Институт математических проблем биологии РАН

Защита состоится 25 ноября 2009 года в 14 часов на заседании Диссертационного совета Д002.077.02 при Учреждении Российской академии наук Институте проблем передачи информации им. А.А. Харкевича РАН по адресу:
127994, г. Москва, ГСП-4, Большой Каретный переулок, 19, стр. 1.

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

Автореферат разослан « » октября 2009 года.

Ученый секретарь Диссертационного совета
д.б.н., профессор Рожкова Галина Ивановна

Общая характеристика работы

Актуальность темы

В настоящее время интенсивное развитие экспериментальных методов молекулярной биологии позволило получить практически полностью расшифрованные последовательности геномов множества организмов, в том числе высших эукариот, включая человека. С появлением современных высокопроизводительных методов секвенирования можно ожидать экспоненциального роста количества расшифрованных геномов.

Наряду с областями, кодирующими белки, рибосомные и транспортные РНК, значительную часть генома занимают некодирующие области, в том числе и имеющие регуляторное значение. Особый интерес представляют сегменты ДНК, содержащие участки связывания белков-факторов регуляции транскрипции (ТФ). Взаимодействие транскрипционных факторов с ДНК является одним из важнейших механизмов регуляции экспрессии генов. Задача идентификации участков, непосредственно взаимодействующих с регуляторными белками, или сайтов связывания ТФ (ССТФ) в геномах эукариотических организмов осложняется малой длиной сайтов и объединением их в регуляторные модули, представляющие собой сложно организованные кластеры ССТФ в пределах сравнительно коротких сегментов ДНК.

Для правильного понимания функционирования регуляторных каскадов необходимо четко идентифицировать сайты связывания ТФ для каждого белка и установить их локализацию в геноме. Появление высокопроизводительных экспериментальных методов анализа связывания ТФ с ДНК на основе иммунопреципитации хроматина вызывает потребность в новых методах и инструментах *in silico*, ориентированных на обработку большого объема данных. Одновременно в компьютерный анализ необходимо вовлечь и результаты, полученные традиционными методами идентификации ССТФ *in vitro*. Таким образом, возникает необходимость в новых биоинформатических методах

построения оптимальных моделей ССТФ на основе различных типов экспериментальных данных.

Для многих ТФ достаточно хорошо описана специфичность связывания; созданы публично и коммерчески доступные базы данных, содержащие информацию о характерных закономерностях в последовательностях олигонуклеотидов, отличающихся высокой аффинностью к ТФ (так называемые мотивы связывания). Однако, различные экспериментальные методы и различные алгоритмы обработки данных приводят к тому, что в открытых источниках присутствуют различные профили связывания для одного и того же ТФ. Таким образом, наряду с определением специфичности связывания и оценки корректности моделей необходимы подходы для сравнения моделей, построенных различными способами на основе данных, полученных с использованием различных экспериментальных методов.

Цели и задачи исследования

Целью работы является разработка и программная реализация биоинформатических методов построения оптимальных моделей ДНК-сайтов связывания транскрипционных факторов с использованием различных типов экспериментальных данных.

Были поставлены следующие задачи:

- Разработать методику построения оптимальной модели ССТФ на базе результатов традиционных (догеномных) экспериментальных методов.
- Разработать методику построения оптимальной модели ССТФ путем интеграции результатов экспериментов по иммунопреципитации хроматина и результатов традиционных методов.
- Реализовать алгоритмы, соответствующие разработанным методам, в виде программных инструментов.
- Верифицировать построенные модели с использованием экспериментальных данных для различных транскрипционных факторов.

Научная новизна

Научная новизна данного исследования характеризуется разработкой новых алгоритмов, позволяющих более точно и эффективно использовать существующие экспериментальные данные по локализации ССТФ в природных и синтетических нуклеотидных последовательностях. Для ряда факторов, связывание которых с ДНК было исследовано несколькими экспериментальными методами, по данным, полученным с помощью каждого из этих методов, построены модели последовательностей, распознаваемых фактором (мотивы) и проведено сравнение достоверности этих моделей с помощью разработанных программных инструментов. Впервые построена коллекция моделей ССТФ транскрипционных факторов *Drosophila melanogaster* путем систематической интеграции данных, полученных с помощью различных экспериментальных методов.

Практическое значение

Разработанные программные средства могут быть применены для эффективного построения оптимальных моделей ССТФ при анализе новых экспериментальных данных, которые получены или будут получены в будущем для различных биологических видов. Программные инструменты могут быть использованы как для непосредственного поиска ССТФ в нуклеотидных последовательностях, так и для верификации альтернативных подходов при моделировании ССТФ. Программные инструменты и созданная коллекция моделей ССТФ предоставлены в открытый доступ.

Апробация работы

Материалы исследований по теме диссертации докладывались и обсуждались на международных научных конференциях BGRS (International Conference on Bioinformatics of Genome Regulation and Structure, Новосибирск 2008) и MCCMB (Moscow Conference on Computational Molecular Biology, Москва 2007, 2009); на конференции молодых ученых ИТиС (Информационные технологии и системы, Звенигород 2007, Геленджик 2008); на IV съезде Российского общества биохимиков

и молекулярных биологов (Новосибирск 2008); на симпозиуме Helmholtz Russian-German Workshop on Systems Biology (Москва 2008).

По материалам диссертации опубликовано 10 печатных работ, включая две статьи в реферируемых журналах, а также тезисы докладов научных конференций.

Объем и структура диссертации

Диссертационная работа изложена на 100 страницах машинописного текста и включает в себя введение, обзор литературы, четыре главы, содержащие результаты и обсуждение, выводы и список литературы из 168 наименований. Работа содержит 18 рисунков, 3 таблицы и 5 приложений.

Содержание работы

Введение и обзор литературы

Раздел содержит аналитический обзор современной литературы, посвященный сравнению современных экспериментальных методов локализации ССТФ в последовательностях нуклеиновых кислот, и описание подходов к построению биоинформатических моделей. Кроме того, в разделе излагается мотивировка постановки решаемых в работе задач и описание их значимости. Идея большинства алгоритмов поиска ССТФ основана на предположении, что различные сайты связывания одного и того же белка схожи между собой, т.е. их выравнивания обладают консервативными позициями. Это позволяет свести задачу поиска набора сайтов, узнаваемых одним и тем же транскрипционным фактором, к задаче нахождения множества сходных фрагментов (мотива) в выборке экспериментально полученных последовательностей ДНК, для которых показано связывание с ТФ. Таким образом, последовательности, включающие в себя ССТФ, должны допускать безделеционное множественное локальное выравнивание (БЛВ, Waterman, 1986) с высоким сходством выравненных сегментов ДНК. БЛВ может быть легко преобразована в классическую модель мотива, а именно матрицу позиционных весов (МПВ, Schneider, 1986). Под мотивом мы везде понимаем БЛВ и соответствующую ему МПВ.

Глава 1. Моделирование ССТФ в виде МПВ

В данном разделе обсуждаются особенности подхода, используемого в рамках работы для моделирования нуклеотидных последовательностей ССТФ в виде МПВ.

1.1 Представление мотива

БЛВ может быть естественным образом представлено в виде матрицы позиционных подсчетов (МПП) в которой каждый элемент $x_{\alpha j}$, $j=1...m$ представляет собой число встреч нуклеотида α в j-й колонке выравнивания. При классическом подходе предполагается, что элементы МПП — целые числа. Мы допускаем, что БЛВ может содержать нескольких слов (длины m, m-меров) из одной и той же последовательности. Кроме того, предполагается, что выравниваемые последовательности могут вносить разный вклад в итоговый мотив, что формализуется путем присвоения последовательностям весов w_i. Пусть последовательность s_i, принадлежащая исходному набору из N последовательностей, имеет вес w_i и содержит k_i слов из БЛВ. В этом случае вклад каждого слова в МПП будет равен w_i/k_i (и $x_{\alpha j}$ могут быть вещественными числами). Таким образом, принятый в работе подход отличается от традиционного, в котором каждому слову, принадлежащему выравниванию, соответствует вклад в некоторые элементы МПП, равный единице.

1.2 Выбор оптимального БЛВ

Для выбора оптимального БЛВ (ОБЛВ) было предложено использовать дискретное информационное содержание (ДИС):

$$ДИС = \sum_{j=1}^{m} I_j = \sum_{j=1}^{m} \frac{1}{N} \left(\sum_{\alpha \in \{A,C,G,T\}} \ln x_{\alpha,j}! - \ln N! \right) \qquad (1)$$

где m – длина выравнивания (мотива), I_j – ДИС одной колонки. Для подсчета lnx! выбрано приближение Стилтьеса (т. е. для вещественных x используется значение гамма-функции). ДИС является аналогом традиционного информационного содержания (Schneider, 1986) для больших выборок и дает более корректную оценку в случае малых выборок (что типично при анализе данных, полученных с помощью SELEX и футпринтинга с ДНКазой I). ОБЛВ соответствует максимальному ДИС для

заданного набора последовательностей (ДИС отрицательно и имеет потенциальный максимум в нуле, соответствующий 100% консервативным колонкам).

1.3 Построение МПВ

Для построения МПВ по БЛВ используется формула:

$$S_{\alpha,j} = \ln\left(\frac{x_{\alpha,j} + aq_\alpha}{(N+a)\,q_\alpha}\right) \qquad (2)$$

Здесь q_α соответствует фоновому распределению нуклеотидов (например, средняя встречаемость нуклеотидов в геноме), a – псевдоотсчет (Lifanov, 2003). Для каждого m-мера МПВ позволяет вычислить качество этого слова s, которое можно использовать как характеристику правдоподобия того, что данное слово является сайтом связывания $s(w) = \sum_{j=1}^{m} S_{w[j],j}$, где w_j - это j-ая буква слова w.

1.4 Критерий качества слова

Для оценки качества слова с помощью МПВ необходимо получить сумму элементов МПВ, соответствующих буквам слова. Превышение некоторого порогового значения качества слова является критерием того, что указанное слово является сайтом связывания исследуемого фактора. Для выбора порогового значения мы строим распределение качества всех возможных m-меров (для мотивов длины меньше или равной 10) либо случайно выбранных 4^{10} m-меров. Считая, что распределение значений качества m-меров на МПВ может быть приближено нормальным, мы выбирали пороговые значения как среднее качество плюс одно, два или три стандартных отклонения (t_1, t_2, t_3).

1.5 Сравнительная оценка предсказательной способности мотивов

С помощью МПВ и критерия качества слова, можно осуществить предсказание ССТФ в исходных экспериментальных данных. Для сравнения эффективности различных мотивов на заданном наборе последовательностей мы использовали модифицированные ROC-кривые (Receiver Operating Characteristic), которые выражают зависимость между чувствительностью (долей истинных предсказаний) и избирательностью (долей ложных предсказаний). Множество истинных

предсказаний (true positive) составляет набор лучших вхождений мотива в последовательности тестового набора, имеющие вхождения с оценкой МПВ выше заданного порогового значения. Множество ложных предсказаний (false positive) составляют все слова с оценкой МПВ выше того же порога, которые могут присутствовать в случайной последовательности. Таким образом, ROC-кривая будет представлена графиком, на котором значения по оси Y соответствуют доле экспериментально определенных последовательностей, имеющих лучшее вхождение выше порога; по оси X – *P-value* (вероятность встретить слово с оценкой МПВ выше порога в случайной последовательности). Под случайной последовательностью мы понимаем последовательность независимых случайных испытаний над ДНК-алфавитом с фоновыми вероятностями нуклеотидов, оцененными из полного генома. При сравнении мотивов длиной меньше или равной m, длина случайной последовательности выбирается как $3m$-2. Это позволяет учесть возможные эффекты самопериодичности мотива. Для подсчета точных значений *P-value* использован алгоритм AhoPro (Boeva, 2007).

Для суммарной оценки качества мотива подсчитывали число точек ROC кривых (каждая точка соответствует одной последовательности в наборе), в которых чувствительность конкретного мотива оказывалась выше всех остальных, участвовавших в сравнении (т. е. ROC-кривая этого мотива проходит слева от остальных, см. рисунок 2). Фактически, в этом случае мы оцениваем, как часто мотив показывает лучшую чувствительность при фиксированной доле ложных предсказаний. Зная, что в наборе могут содержаться ошибочные последовательности, не содержащие реальных ССТФ и, следовательно, вхождений мотива, из рассмотрения разумно исключать ряд точек в правой области ROC-кривой (например, ограничиваясь фиксированным процентом рассматриваемых последовательностей, либо минимально допустимыми порогами для мотивов, либо ограничивая диапазон значимых *P-value*). При анализе малых выборок в главе 2 мы используем диапазон 10-90% последовательностей. В главе 3 для больших выборок

мы ограничиваемся диапазоном (0, 0.1) для *P-value* и нижней границей на пороги каждого мотива заданной как t_3.

Глава 2. Построение моделей мотивов нуклеотидных последовательностей, распознаваемых белками-регуляторами транскрипции, на основе данных традиционных экспериментальных методов

Исследование ССТФ получило существенное развитие с появлением новых высокопроизводительных экспериментальных методов анализа на базе иммунопреципитации хроматина (ChIP) с последующей гибридизацией выделенной ДНК на микрочипах (Horak and Snyder, 2002), либо непосредственным секвенированием (Mardis, 2007). Эти методы дают возможность определения ССТФ в масштабах полного генома. Недостатком метода ChIP оказалась сравнительно большая длина выделяемых фрагментов ДНК, что во многих случаях затрудняет однозначное определение ССТФ. Однако, привлечение дополнительных данных, полученных традиционными методами футпринтинга (Galas and Schmitz, 1978) и SELEX (systematic evolution of ligands by exponential enrichment, Tuerk and Gold, 1990), часто дает возможность корректной интерпретации и верификации данных ChIP. В частности, на основе ССТФ-содержащих искусственных олигонуклеотидов, получаемых с использованием SELEX-подобных методов, можно непосредственно строить ОБЛВ. Для ряда высших эукариот накоплен большой экспериментальный материал, систематизированный в базах данных (Kolchanov, 2002; Sandelin, 2004). Нами был разработан набор инструментов для работы с коммерчески доступной базой данных TRANSFAC (Matys, 2003), осуществляющий картирование на геном выбранного организма данных для заданного ТФ, извлечение сегментов ДНК с соответствующими фланкирующими областями и корректное построение ОБЛВ с помощью опубликованного ранее алгоритма SeSiMCMC (Favorov, 2005). Эта методика была успешно применена для определения мотива связывания белка Sp1 у *Homo sapiens*.

Длина прилегающих фланкирующих областей футпринта (области, защищенной молекулой ТФ от действия ДНКазы) оказывает чрезвычайно сильное

влияние на идентифицируемый мотив ССТФ при рассмотрении выборки, состоящей из малого числа последовательностей, либо сложно идентифицируемого мотива. Возникает необходимость в более точных методах использования информации, содержащейся в указанных сегментах, прилегающих к исходным-содержащим ССТФ. Особенно важным это становится при анализе данных, полученных футпринтингом ДНКазой I. Природные олигонуклеотиды, полученные с помощью этого метода, трудно использовать непосредственно, поскольку их длина и позиционирование ССТФ существенно зависят от многих факторов, связанных с условиями эксперимента и дальнейшей процедурой сбора результатов различных экспериментов в единую базу данных. Не менее важное значение имеют особенности работы ДНКазы I, а именно неравномерное внесение разрывов в последовательности, различающиеся по GC-составу, и возможный сдвиг футпринта относительно реального сайта, защищенного от действия ДНКазы наличием связанного с ним изучаемого ТФ.

Для выработки универсальной методики получения ОБЛВ на базе результатов футпринтинга нами был разработан новый алгоритм Bigfoot. Были использованы данные футпринтов для различных ТФ *D. melanogaster*, картированные на геном и систематизированные в курированной БД Drosophila DNase I Footprint Database (Bergman, 2005).

2.1 Алгоритм Bigfoot

Алгоритм выполняет поиск оптимального мотива путем построения ОБЛВ в заданном диапазоне длин (перебираемом в направлении от наименьшей к наибольшей) на базе футпринтов, взятых с соответствующими фланкирующими областями. Bigfoot предполагает, что каждая последовательность содержит одно значимое вхождение мотива. Проблема отсечения фрагментов, являющихся ошибками эксперимента, решается для уже построенного ОБЛВ, как описано в разделе 2.2. Для исследуемого ТФ длина фланков выбирается как разность длин самого короткого футпринта и стартовой длины мотива (если мотив длиннее, чем самый короткий футпринт). Длина фланков увеличивается при переходе от длины *m*

к *m+1*. Построив ОБЛВ с максимальным ДИС, алгоритм определяет оптимальную длину мотива. Эта процедура основана на оценке значимости колонок путем применении критерия χ^2 к колонкам выравнивания и идентификации наиболее стабильного ядра (наиболее консервативной части) мотива для разных длин ОБЛВ. Помимо определения оптимальной длины мотива эта процедура решает проблему переобученности модели, которая часто возникает на малых выборках (что типично для классических экспериментальных методов исследования ССТФ).

Для поиска ОБЛВ заданной длины Bigfoot использует модификацию классического подхода максимизации правдоподобия (Bailey and Elkan, 1994). МП-процедура для заданной стартовой МПВ заключается в выборке слов с максимальными оценками из каждой последовательности и итеративной перестройке МПВ на базе каждого нового набора слов. В случае наличия в последовательности нескольких слов с одинаковой наилучшей для данной последовательности оценкой, используются все слова, причем им присваиваются веса в соответствии с методикой, описанной ранее. В качестве стартовых МПВ Bigfoot использует МПВ построенные из всех возможных пар слов, содержащихся в последовательностях, используемых для построения ОБЛВ.

2.2 Коллекция DMMPMM

В результате работы был создан автоматизированный программный конвейер DMMPMM (Drosophila Melanogaster Major Position Matrix Motifs, http://line.imb.ac.ru/DMMPMM/) для построения коллекции основных мотивов ССТФ *D. melanogaster*, заданных в форме МПВ. Данные БД футпринтов были загружены в

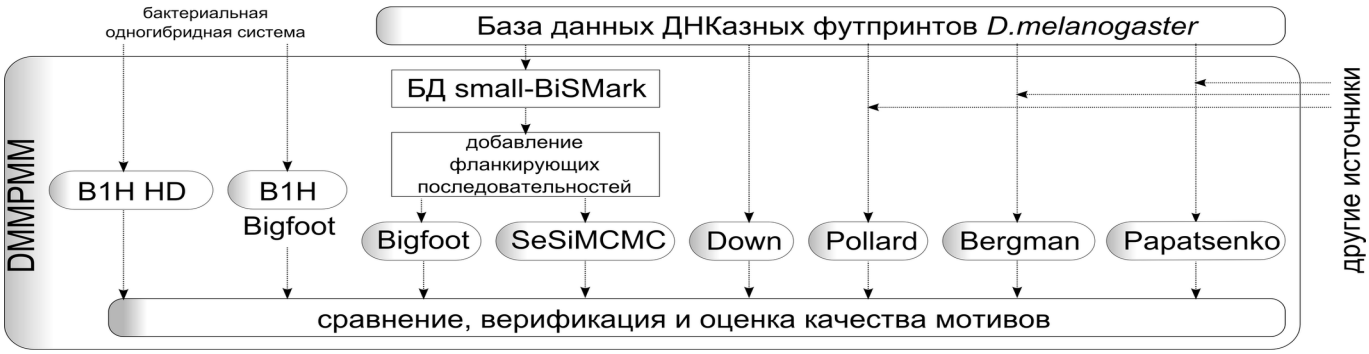

Рисунок 1. Схема построения коллекции DMMPMM.

специализированную БД (см. Главу 4). Для построения мотивов использовали опубликованный ранее алгоритм SeSiMCMC и новый алгоритм Bigfoot.

В качестве данных использовались результаты ДНК футпринтига для 41 ТФ, т.е. для всех факторов *D. melanogaster*, для которых доступны футпринты числом более восьми. Для сравнительного анализа были использованы существующие коллекции мотивов (Papatsenko, 2007; Pollard 2006; Down, 2007), частично основанные на БД футпринтов, а также коллекция мотивов (Bergman, 2005), построенная на базе SELEX. Кроме того, были использованы данные бактериальной одногибридной системы (B1H, Noyes, 2008). В этом случае для определения мотивов ТФ (за исключением независимой коллекции мотивов гомеодоменных (HD) белков) использовали алгоритм Bigfoot. Для построения МПВ по ОБЛВ псевдоотсчет был принят равным 1; в рамках МП-процедуры Bigfoot использовал псевдоотсчеты, равные числу последовательностей в ОБЛВ.

Для оценки количества футпринтов, не содержащих ССТФ, мы выбрали пороговые значения на качество лучшего слова в последовательности (см. Главу 1): лучше, чем t_3 (мотив присутствует); хуже, чем t_1 (мотив отсутствует). Дополнительно для подсчета числа случаев, когда рассмотрение фланкирующих областей дало возможность восстановить вхождения мотивов, мы ввели порог t_2: успешно восстановленным сайтом считали ситуацию, когда добавление фланков к последовательности, в которой изначально максимальное качество МПВ не превышало t_1, позволяло найти сайт лучше t_2.

В результате удалось установить, что только для 25% ТФ все доступные футпринты для всех коллекций мотивов содержали хорошее вхождение МПВ. Для других 25% все коллекции мотивов детектировали наличие футпринтов, не содержащих сайта. В оставшихся 50% случаев «ошибочные» футпринты детектировались мотивами только из некоторых коллекций. Добавление фланкирующих последовательностей длиной 2 п.н. позволяло восстановить более 50% утерянных сайтов; добавление фланков длины равной длине мотива доводило это число до 80%.

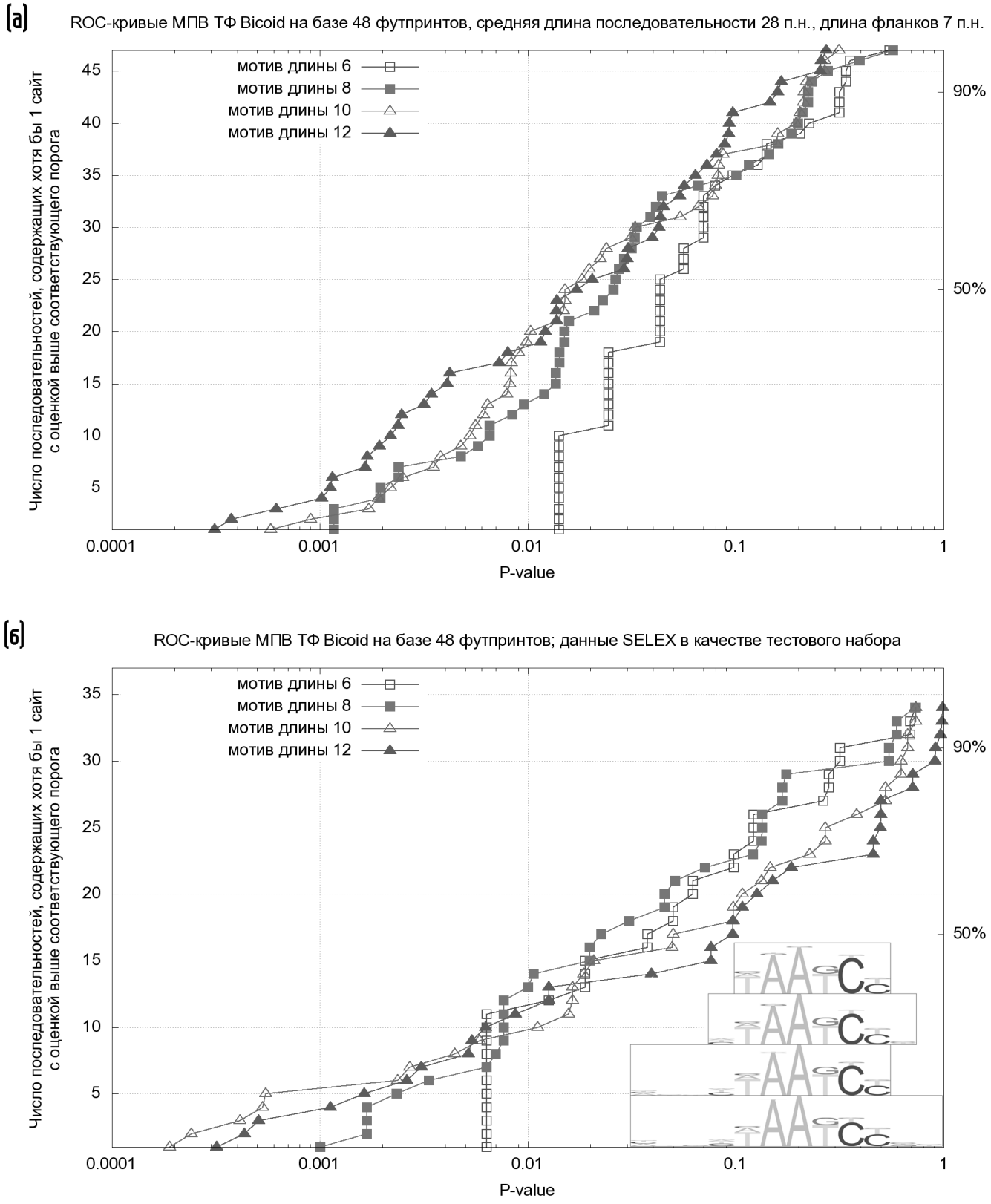

Рисунок 2. Проверка переобученности мотивов на базе двух источников данных. (а) ROC-кривые для мотивов разных длин, построенных Bigfoot на наборе футпринтов для фактора Bicoid. Качество мотива растет с ростом длины (соотв. увеличению числа свободных параметров), что демонстрирует переобученность МПВ. (б) ROC-кривые с использованием данных SELEX в качестве тестового набора. Оптимальная длина мотива здесь составляет 8 п.н.

Кроме того, рассмотрение фланкирующих областей позволило построить более точные МПВ. Даже с учетом жесткой процедуры определения длины, мотивы, построенные с помощью Bigfoot, в 10% случаев оказались лучшими среди всех коллекций. В случае отключенной процедуры определения длины Bigfoot показал лучшую чувствительность в 60% случаев. Кроме того, удалось установить, что курированные коллекции мотивов и мотивы, созданные на основе данных SELEX и B1H-системы, обладают плохой предсказательной способностью на наборах футпринтов. Важным результатом также является демонстрация существенной переобученности моделей (рисунок 2), построенных всего на одном экспериментальном источнике данных, что дает дополнительный стимул к интеграции различных источников.

Глава 3. Интеграция данных, полученных различными экспериментальными методами для определения мотивов ССТФ в последовательностях ДНК

Одной из важных проблем в определении ССТФ является идентификация длины участка ДНК, несущего сигнал специфического связывания с регуляторным белком. Практика показывает, что эта проблема не может быть однозначно решена при использовании данных ChIP, SELEX или футпринтига, рассмотренных по отдельности. Интеграция данных, полученных различными экспериментальными методами, позволяет не только эффективно решить эту проблему, но и получить в общем случае мотивы, обладающие лучшей предсказательной способностью.

Первый подход, который мы применили при исследовании ТФ Sp1 человека, заключался в том, чтобы совместно использовать неточные данные ChIP-chip и SELEX для построения совместной модели мотива. В этом случае данные SELEX использовались для расстановки якорей внутри последовательностей, полученных ChIP-chip. Под якорями мы понимаем слова, отличающиеся от найденных с помощью SELEX не более чем на фиксированное число замен. Затем, с помощью SeSiMCMC мы строили «заякоренное» БЛВ, т. е. такое, в котором каждое слово выравнивания пересекалось с якорным словом не менее чем в одной позиции. Совместное использование ChIP-chip и SELEX в этом случае позволило выявить

мотив, обладающий значительным сходством с извлеченным из футпринтов, по сравнению с исходным, построенным исключительно на базе SELEX.

Для *D. melanogaster* мы разработали более гибкий подход, позволивший интегрировать все доступные на сегодня источники данных, сравнение которых проводилось в главе 2. Дополнительно мы использовали данные ChIP-chip (MacArthur, 2009). Были выбраны сегменты длиной 500 п.н., несущие сигнал ChIP-chip с интенсивностью, входившей в сотню лучших значений по геному (в тех случаях, когда были доступны данные нескольких независимых экспериментов, соответствующие геномные сегменты составляли единую выборку). С учетом этих данных был создан набор из 39 ТФ, для которых были доступны последовательности из двух и более экспериментальных источников, использованных для коллекции DMMPMM. Принималось предположение, что каждая последовательность содержит единственное вхождение мотива неизвестной длины.

3.1 Алгоритм Chipmunk

Алгоритм Chipmunk разработан для определения мотива из данных ChIP с возможностью использования данных, полученных традиционными методами, как дополнительных источников информации. В качестве входных данных Chipmunk использует n независимых наборов последовательностей; различные наборы соответствуют различным источникам данных. Последовательностям присваиваются веса, равные в пределах одного набора. Пусть N – полное число последовательностей во всех n наборах, l_k – число последовательностей в k-м наборе; тогда вес каждой последовательности в k-м наборе принимается равным $\frac{N}{(nl_k)}$. ОБЛВ строится на объединенном наборе взвешенных последовательностей с суммарным весом N равным их полному числу. Вес определяет вклад конкретной последовательности в мотив; подобная схема присваивания весов позволяет сбалансировать вклады разных экспериментальных источников в том случае, если соответствующие наборы последовательностей

содержат существенно различное число последовательностей (как правило, это сотни сегментов для ChIP-chip и десятки для других методов). Значение псевдоотсчета принимали равным натуральному логарифму числа последовательностей.

Для поиска ОБЛВ заданной длины Chipmunk использует модификацию алгоритма максимизации ожидания с бутстраппингом. Основой алгоритма является процедура двухэтапной оптимизации (ДО) МПВ. На первом этапе предварительной оптимизации выбирается случайное подмножество последовательностей (возможно с дубликатами) общим весом порядка $\sqrt{N}$. Выполняется фиксированное число итераций МП-оптимизации на выбранном подмножестве последовательностей. Полученная в результате матрица подвергается МП-оптимизации до сходимости на полном наборе последовательностей. МП позволяет существенно (до 50% в зависимости от набора последовательностей и МПВ) уменьшить число итераций по сравнению с прямой оптимизацией МПВ на полном наборе последовательностей.

ДО, будучи примененной к любой МПВ, приводит ее в некий локальный максимум ДИС на полном наборе. В случае применения ДО к МПВ, которая уже представляет некий локальный максимум, раунд оптимизации на подмножестве служит для выхода из этого локального максимума, а полученная в результате промежуточная матрица может быть оптимизирована до сходимости на полном наборе к другому (или прежнему) локальному максимуму за малое число итераций.

Алгоритм состоит из двух вложенных циклов. Внешний цикл инициирует тривиальную МПВ на базе случайного слова. Внутренний цикл запускает ДО для текущей МПВ которая является либо тривиальной (для первого прохода), либо результатом предыдущего прохода цикла (если на нем была получена матрица с ДИС большим, чем у лучшей МПВ, пришедшей из предыдущих циклов), либо (в противном случае) лучшей МПВ из предыдущих циклов. На каждом проходе внутреннего цикла при необходимости обновляется лучшая текущая МПВ, которая может быть сохранена при каждом окончании внутреннего цикла.

В большинстве реализацией МП-алгоритмов рассматривается большое число

стартовых тривиальных МПВ; при этом каждая из них может давать результат как более близкий к оптимальному мотиву, так и довольно далекий. Chipmunk предлагает использование сравнительно небольшого количества стартовых матриц и итеративное применение ДО к текущем локальному максимуму. На обработку принципиально ошибочных предположений (соответствующих неудачным случайным стартовым матрицам) тратится минимум вычислительного времени, поскольку неудачная проба (имеющая низкое ДИС по окончании одного прохода) заменяется лучшей МПВ на текущий момент работы алгоритма.

Для выбора оптимальной длины ОБЛВ в заданном диапазоне длин алгоритм перебирает их, начиная с наибольшей, и останавливается тогда, когда находит так называемый *сильный* мотив. Выберем два пороговых значения на ДИС: *T* и *t*. *T* соответствует ДИС колонки, в которой представлены только 3 нуклеотида из возможных 4x; *t* соответствует колонке, в которой два из четырех нуклеотидов представлены в два раза более часто, чем два оставшихся. Будем считать, что мотив состоит из нескольких доменов в том случае, если одна или несколько внутренних колонок выравнивания имеют ДИС хуже *t*. Сильным называется мотив, для которого в целом и для каждого входящего в него домена выполняется следующее условие: крайние колонки имеют ДИС не хуже *t*, присутствует не менее одной колонки с ДИС лучше *T*. Использование такого критерия возможно благодаря сравнительно большим выборкам, получающимся при объединении различных экспериментальных источников.

3.2 Коллекция iDMMPMM

Созданная коллекция iDMMPMM (impoved Drosophila Melanogaster Major Position Matrix Motifs, коллекция улучшенных основных МПВ мотивов *D. melanogaster*, http://line.imb.ac.ru/iDMMPMM) содержит 39 интегрированных мотивов, построенных с помощью Chipmunk как ОБЛВ объединенной взвешенной выборки.

Для ROC-анализа мы создали независимые наборы последовательностей. Для этого выбрали 300 сегментов длиной 500 п.н., соответствующих лучшим сигналам

ChIP-chip для каждого фактора и исключили из этой выборки лучшие 100, использованные при построении мотивов (при наличии независимых экспериментов выборки последовательностей объединялись). От 50 до 90% точек ROC-кривых интегрированных мотивов соответствуют наилучшей чувствительности для 90% факторов (пример для фактора Knirps приведен на рисунке 3).

Анализ мотивов из различных независимых источников позволяет утверждать, что в более чем 80% случаев интеграция данных позволила построить мотив с лучшими показателями чувствительности и избирательности в широком диапазоне порогов МПВ. Случаи, когда интеграция источников приводила к ухудшению мотива, соответствуют либо недостаточному объему данных, либо реальному существованию различных функциональных мотивов для одного ТФ.

Глава 4. Программная реализация

В качестве эффективного стандартного формата передачи данных в рамках работы был разработан язык разметки small-BiSMark(small Biological Sequence Markup Language) на основе XML и набор инструментов для поддержки базы данных (работающей под управлением MySQL 5), адаптированной для импорта и экспорта наборов сегментов ДНК и моделей ССТФ на их основе. Кроме того, разработана многофункциональная библиотека для решения различных вычислительных задач, интегрирующая новые алгоритмы Bigfoot и Chipmunk, а также существующие инструменты AhoPro и SeSiMCMC.

Программные конвейеры для сборки коллекций мотивов, автоматизированной верификации и сравнения реализованы на базе Ruby 1.8. Вычислительно-нагруженные этапы алгоритма Bigfoot и алгоритм Chipmunk реализованы на базе Java 1.6. Созданный набор программных средств базируется исключительно на свободном программном обеспечении, является кросс-платформенным и был успешно протестирован на платформах Windows (XP32/64) и Linux (Ubuntu 8.04).

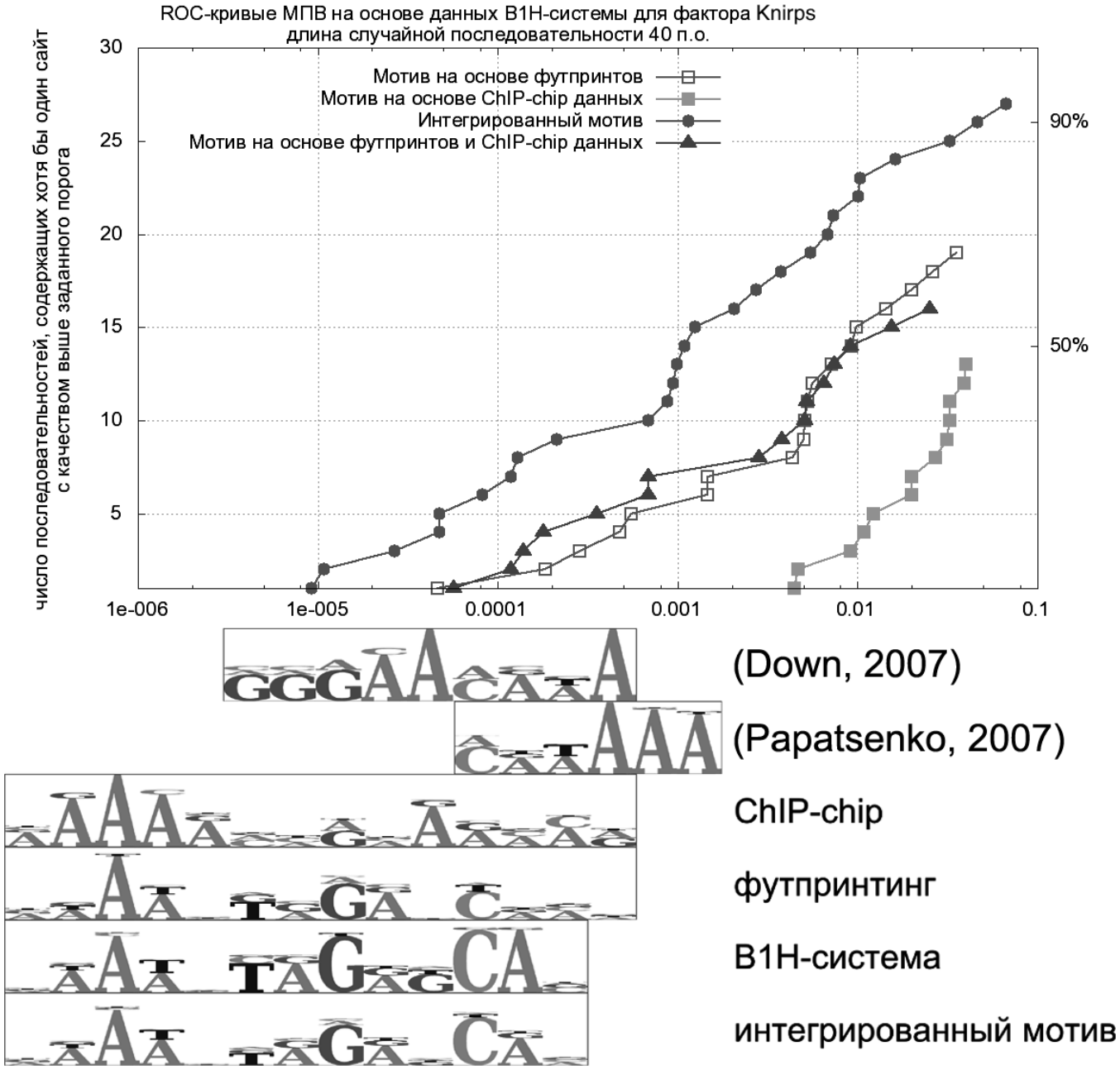

Рисунок 3. ROC-кривые, полученные на базе результатов B1H-системы в качестве тестового набора последовательностей. Мотивы построены с помощью алгоритма Chipmunk.

При визуализации ДИС используется для масштабирования колонок выравнивания. Мотивы, представленные в коллекциях (Down) и (Papatsenko), соответствуют неверным сигналам.

Выводы

1. Разработан метод построения оптимальной модели ССТФ с использованием экспериментальных данных, полученных традиционными экспериментальными методами. Метод реализован в виде вычислительного алгоритма Bigfoot. Показано, что при использовании данных ДНК футпринтинга для построения мотивов, распознаваемых ССТФ, необходимо использовать участки генома, прилегающие к картированным футпринтам. Предложен алгоритм, реализующий учет информации, содержащейся в геномных фланках футпринтов, при построении ОБЛВ.
2. Разработан метод построения оптимальной модели ССТФ путем интеграции данных различных экспериментальных методов, включая современные высокопроизводительные техники на базе иммунопреципитации хроматина. Метод реализован в виде вычислительного алгоритма Chipmunk.
3. Создан набор программных средств, реализующих предложенные алгоритмы. Созданные программы позволяют на базе различных вычислительных платформ анализировать данные, полученные с использованием широкого спектра экспериментальных методик. Создан единый вычислительный конвейер, интегрирующий новые алгоритмы и существующие программные инструменты AhoPro и SeSiMCMC.
4. Создана коллекция мотивов связывания факторов регуляции транскрипции *Drosophila melanogaster*, содержащая мотивы, полученные с использованием практически всей экспериментальной информации, представленной в открытых источниках. Показано, что разработанные методы позволяют выявить мотивы связывания, превосходящие по своим характеристикам известные мотивы связывания, для широкого набора белков-регуляторов транскрипции.
5. Разработанные программные средства и построенные коллекции уточненных мотивов доступны в сети Интернет по адресам: http://line.imb.ac.ru/DMMPMM, http://line.imb.ac.ru/iDMMPMM, http://line.imb.ac.ru/Chipmunk.

Список принятых сокращений

БЛВ	безделеционное локальное выравнивание
ДИС	дискретное информационное содержание
ДО	двухэтапная оптимизация
МП	максимизация правдоподобия
МПВ	матрица позиционных весов
МПП	матрица позиционных подсчетов
ОБЛВ	оптимальное безделеционное локальное выравнивание
ССТФ	сайты связывания транскрипционных факторов
ТФ	транскрипционный фактор
ChIP	chromatin immunoprecipitation
SELEX	systematic evolution of ligands by exponential enrichment

Список работ, опубликованных по теме диссертации

Статьи в реферируемых журналах

1. I.V. Kulakovskiy, A.V. Favorov, V.J. Makeev (2009) Motif discovery and motif finding from genome-mapped DNAse footprint data. *Bioinformatics*, **25**(18), 2318-2325

2. И.В. Кулаковский, В.Ю. Макеев (2009) Интеграция данных, полученных различными экспериментальными методами, для определения мотивов в последовательностях ДНК, распознаваемых факторами, регулирующими транскрипцию. *Биофизика*, **54**(6), 965

Тезисы конференций

1. I.V. Kulakovskiy, V.J. Makeev (2007) Constructing PWM from unaligned TFBS footprints. *Proceedings of the 3-rd Moscow Conference on Computational Molecular Biology,* 167-168, Moscow

2. А.В. Фаворов, М.С. Гельфанд, А. Герасимова, Д.А. Равчеев, И. Кулаковский, А. Миронов, В. Макеев (2007) Алгоритм SeSiMCMC для поиска участков специфического связывания белков-регуляторов транскрипции. *Сборник трудов конференции «Информационные технологии и системы» ИТиС'07*, 334-337, Звенигород

3. I.V. Kulakovskiy (2008) Integrated tool for analysis of DNA-protein binding data. *Helmholtz Russian-German Workshop on Systems Biology*, 45, Moscow

4. И.В. Кулаковский, А.А. Белостоцкий, А.В. Фаворов, В.А. Боева, Д.Б. Малько, В.Ю. Макеев (2008) Интеграция различных типов экспериментальных данных для анализа последовательностей регуляторных районов эукариот. *Сборник материалов IV съезда Российского общества биохимиков и молекулярных биологов*, 278, Новосибирск

5. A. Heinzel, I.V. Kulakovskiy, V.J. Makeev (2008) Comparison of ChIP-chip Sp1 binding location data for human chromosome 21, 22 with PWM hits. *The Sixth International Conference on Bioinformatics of Genome Regulation and Structure*, 97, Novosibirsk

6. I.V. Kulakovskiy, A.V. Favorov, V.J. Makeev (2008) Incorporating different types of experimental data on DNA-protein binding into the single *in silico* model. *The Sixth International Conference on Bioinformatics of Genome Regulation and Structure,*129, Novosibirsk

7. Y.A. Medvedeva, M.V. Fridman, N.J. Oparina, D.B. Malko, E.O. Ermakova, I.V. Kulakovskiy, V.J. Makeev (2008) Non-5' CpG islands in the human genome: probable involvement in transcription regulation. *Сборник трудов конференции «Информационные технологии и системы» ИТиС'08*, 298-299, Геленджик

8. I.V. Kulakovskiy, V.A. Boeva, A.V. Favorov, V.J. Makeev (2009) Chipmunk: a fast DNA motif finder for ChIP data and its application to data integration from different experimental sources. *Proceedings of the 4-th Moscow Conference on Computational Molecular Biology*, 194, Москва