На правах рукописи

Никулова Анна Алексеевна

Регуляторные модули в эукариотах: предсказание, анализ структуры и консервативности

03.01.09 - математическая биология, биоинформатика

Автореферат

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2012

Работа выполнена на факультете биоинженерии и биоинформатики Федерального государственного бюджетного образовательного учреждения высшего профессионального образования Московского государственного университета имени М.В. Ломоносова.

Научный руководитель:	Андрей Александрович Миронов кандидат физико-математических наук, доктор биологических наук
Официальные оппоненты:	Ройтберг Михаил Абрамович доктор физико-математических наук, Федеральное государственное бюджетное учреждение науки Институт математических проблем биологии Российской академии наук, заведующий лабораторией
	Спирин Сергей Александрович кандидат физико-математических наук, Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского Федерального государственного бюджетного образовательного учреждения высшего профессионального образования Московского государственного университета им. М.В. Ломоносова, старший научный сотрудник
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук

Защита диссертации состоится «29» октября 2012 года в 14:00 на заседании диссертационного совета Д.002.077.04 на базе Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича РАН по адресу: 127994, Москва, ГСП-4, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича РАН

Автореферат разослан «_____» сентября 2012 г.

Ученый секретарь диссертационного совета Д.002.077.04
доктор биологических наук, профессор
Рожкова Г.И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы. Одной из важнейших задач биоинформатики является выявление и изучение участков ДНК, участвующих в регуляции транскрипции генов. Эта задача стала особенно актуальной в последнее время в связи с появлением огромного количества новых геномных последовательностей, нуждающихся в функциональной аннотации.

Регуляторные участки ДНК, участвующие в регуляции транскрипции генов, представляют собой сайты связывания транскрипционных факторов (ТФ), специфически связывающихся с ДНК и влияющих на уровень транскрипции соответствующих генов. Основными трудностями при идентификации сайтов связывания транскрипционных факторов (ССТФ) в геномах эукариот являются сравнительно небольшая длина (5-12 пар нуклеотидов, пн) и значительная вырожденность ССТФ. К тому же сайты связывания могут располагаться довольно далеко (до 60 тыс. пн) от регулируемого гена. Таким образом, даже при наличии известной модели ССТФ (например, позиционно-весовой матрицы, ПВМ) поиск сайтов связывания дает огромное количество ложно-положительных предсказаний.

Однако известно, что в геномах эукариот ССТФ часто организованы в группы (кластеры, цис-регуляторные модули), покрывающие участки ДНК протяженностью несколько сотен пар оснований. По-видимому, эти модули координируют белок-белковые взаимодействия, тем самым регулируя уровень транскрипции генов. До сих пор не до конца понятно, как они устроены. Большинство исследователей обращают внимание именно на тип и близкое расположение ССТФ, однако было показано, что во многих случаях важным фактором является порядок расположения ССТФ и расстояния между ними [Makeev *et al.* 2003, Hallikas *et al.* 2006, Matys *et al.* 2006, Papatsenko *et al.* 2009], то есть ***структура (грамматика) регуляторных модулей***. Знание структуры регуляторных модулей могло бы не только значительно повысить качество предсказания программ для поиска регуляторных элементов, но также позволило бы предсказывать совместную работу ТФ.

Свидетельством функциональной важности структуры регуляторных модулей может служить сохранение грамматики модулей в процессе эволюции даже при значительной дивергенции геномных последовательностей. С другой стороны, структура регуляторных модулей сходно регулируемых генов, по-видимому, также должна быть похожа. Таким образом, анализ регуляторных участков ортологичных и/или ко-регулируемых генов позволит определить функционально важные закономерности расположения сайтов связывания.

Актуальным является разработка алгоритмов, позволяющих выявлять закономерности расположения ССТФ, характерные для набора сходно функционирующих регуляторных модулей, и использовать информацию о

выявленной структуре для повышения качества предсказаний регуляторных модулей в геномах эукариот. Предсказание регуляторных модулей, характеризующихся сходной структурой, позволит выявлять ко-регулируемые гены. Кроме того, выявленные закономерности взаимного расположения ССТФ могут быть использованы для описания работы данной регуляторной системы.

Цели и задачи работы. Целью данной работы является разработка эффективных алгоритмов, методов и программных приложений для предсказания и анализа регуляторных транскрипционных модулей и их структуры в геномах эукариот и применение разработанных методов к различным эукариотическим системам. В ходе работы были поставлены следующие задачи:

- предложить способ описания структуры регуляторных модулей, включающей частоты встречаемости сайтов связывания разных типов, предпочтение следования сайтов и характерные распределения расстояний между соседними сайтами;
- разработать алгоритм выявления структуры регуляторных модулей, содержащихся в наборе геномных последовательностей;
- разработать метод для поиска регуляторных модулей с учетом их структуры;
- разработать метод полногеномного поиска ко-регулируемых генов на основе наличия и консервативности предсказанных регуляторных модулей;
- применить разработанные алгоритмы для поиска регуляторных модулей и ко-регулируемых генов к ряду биологических систем; провести сравнение результатов, полученных с помощью разработанных алгоритмов, и результатов других программ, применяющихся в данной области;
- провести анализ структуры регуляторных модулей для ряда биологических систем, и сравнить сделанные наблюдения с содержащейся в литературе информацией о совместной работе транскрипционных факторов.

Научная новизна и практическая ценность. Научная новизна работы состоит в разработке новой вероятностной модели регуляторных модулей эукариот, описывающей их структуру, включающую частоты встречаемости сайтов, предпочтение следования сайтов связывания и характерные распределения расстояний между ними, а также применении этой модели для выявления структуры регуляторных модулей ортологичных и/или ко-регулируемых генов. Применение обобщенных скрытых Марковских моделей позволяет эффективно моделировать любые распределения расстояний между сайтами в регуляторных модулях. Обучение параметров модели на наборе ортологичных последовательностей позволяет учитывать эволюционную консервативность регуляторных модулей без использования выравнивания последовательностей, что делает алгоритм не зависимым от степени дивергенции последовательностей.

Разработанный и реализованный метод поиска регуляторных модулей в геномах

эукариот (свидетельство о регистрации в Государственном фонде алгоритмов и программ №2012610082) может быть использован для аннотации геномных последовательностей, изучения механизмов регуляции транскрипции и эволюции регуляторных модулей, поиска ко-регулируемых генов, а также для исследования генетических заболеваний, связанных с регуляцией экспрессии генов.

Апробация работы. Основные положения диссертации были представлены на международных конференциях: 3rd Int. Moscow Conference on Computational Molecular Biology MCCMB'07 (Москва, июль 2007), 4th Int. Moscow Conference on Computational Molecular Biology MCCMB'09 (Москва, июль 2009), 5th Int. Moscow Conference on Computational Molecular Biology MCCMB'11 (Москва, июль 2011), 30-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'07 (Звенигород, сентябрь 2007), 31-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'08 (Геленджик, октябрь 2008), 32-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'09 (Бекасово, декабрь 2009), 33-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'10 (Геленджик, сентябрь 2010), 34-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'11 (Геленджик, октябрь 2011), 4th International Conference on Bioinformatics Models, Methods and Algorithms BIOINFORMATICS'2012 (Виламура, Португалия, февраль 2012) и на научных встречах международной учебно-научной группы «Regulation and Evolution of Cellular Systems (RECESS)».

Объем и структура диссертации. Диссертационная работа изложена на 125 страницах и состоит из введения, четырех глав, выводов и списка цитированной литературы. Глава 1 содержит обзор литературы по теме диссертации. Глава 2 содержит описание и тестирование разработанного алгоритма для предсказания регуляторных моделей в геномах эукариот. Глава 3 содержит описание и тестирование алгоритма полногеномного поиска ко-регулируемых генов, разработанного в данной работе. Глава 4 содержит описание применения разработанного алгоритма для выявления структуры регуляторных модулей и обсуждение полученных результатов в контексте литературных данных. Список литературы включает 149 наименований. Работа содержит 23 рисунка, 5 таблиц и 2 приложения.

СОДЕРЖАНИЕ РАБОТЫ

Первая глава посвящена обзору литературы по теме диссертации. В первой части обзора приведены общие сведения о регуляции транскрипции генов, особенности строения цис-регуляторных модулей в геномах эукариот, а также краткая характеристика методов экспериментального изучения регуляции транскрипции. Вторая часть посвящена биоинформатическим методам изучения регуляторных элементов в геномах эукариот. В ней рассмотрены способы описания и поиска сайтов

связывания факторов транскрипции (ССТФ) и приведены основные подходы, используемые для поиска цис-регуляторных модулей в геномах эукариот. Практически все алгоритмы используют предпосылку о кластеризации ССТФ вдоль цепи ДНК. Также многие методы используют межвидовое сравнение, или же сравнение регуляторных модулей ко-регулируемых генов, для повышения качества предсказания. Некоторые методы привлекают дополнительную информацию помимо последовательностей геномов и моделей ССТФ, например, информацию об экспрессии генов. Поскольку в данной работе применяется подход, основанный на скрытых Марковских моделях (СММ), обзор литературы также содержит краткое описание общих идей СММ и основных алгоритмов для обучения параметров СММ и декодирования последовательности состояний.

Во **второй главе** приведено описание разработанного в данной работе алгоритма поиска регуляторных модулей в геномах эукариот на основе набора известных моделей сайтов (позиционно-весовых матриц, ПВМ), учитывающего консервативную структуру регуляторных модулей. Здесь же приведены результаты оценки качества предсказания регуляторных модулей для системы генов, специфически экспрессирующихся в мышечной ткани позвоночных, и системы раннего развития *Drosophila*.

В основе алгоритма лежит модель последовательности, содержащей в себе регуляторные модули. Модель описывает в том числе структуру регуляторных модулей, а именно частоты сайтов разных типов, из которых состоят модули, предпочтения в порядке следования сайтов и в расстоянии между ними. Закономерности взаиморасположения сайтов выявляются в результате обучения модели на регуляторных участках сходно регулируемых, то есть ортологичных и/или ко-регулируемых, генов.

Регуляторный модуль моделируется как кластер непересекающихся ССТФ, причем начало модуля совпадает с началом первого сайта кластера, а конец — с концом последнего сайта. То есть регуляторный модуль состоит из сайтов и разделяющих их последовательностей — спейсеров (рис.1).

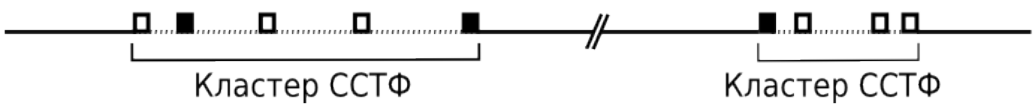

Рисунок 1. Схематическое изображение двух кластеров сайтов, окруженных фоновой последовательностью (непрерывные отрезки горизонтальной линии). Каждый кластер состоит из сайтов (показаны квадратами разных типов), разделенных спейсерами (пунктирные отрезки горизонтальной линии).

Для моделирования регуляторных модулей, окруженных фоновой последовательностью, использовалась скрытая Марковская модель (СММ), схема которой изображена на рисунке 2.

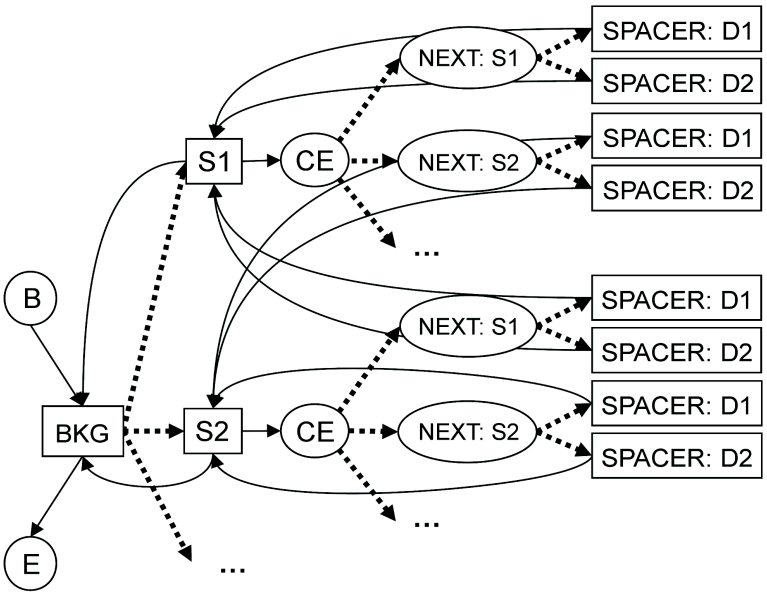

Рисунок 2. Схема СММ. Порождающие состояния изображены в виде прямоугольников, молчащие состояния — в виде овалов. Разрешенные переходы между состояниями показаны стрелками. Вероятности переходов, помеченных пунктиром, изменяются в процессе обучения модели по алгоритму Баума-Велча.

Архитектура СММ отражает наше представление о том, как устроены регуляторные модули в геномах эукариот. СММ, используемая в данной работе, содержит три основных типа порождающих состояний, соответствующих трем типам последовательности: фоновая последовательность между кластерами сайтов (состояние BKG), сайты (состояния S1, S2 и т.д.) и участки между сайтами в кластере — спейсеры (состояния SPACER:D1 и SPACER:D2). Количество состояний, порождающих сайты, равно количеству типов сайтов. Количество типов сайтов, в свою очередь, в два раза больше, чем количество ПВМ, используемых для построения модели, поскольку сайты, расположенные на разных цепях последовательности ДНК считаются сайтами разного типа.

Каждое порождающее состояние генерирует последовательность нуклеотидов, длина которой определяется распределением, характерным для данного состояния. СММ с таким типом архитектуры называют обобщенной СММ [Rabiner *et al.* 1989, Kulp *et al.* 1996, Lukashin *et al.* 1998]. Ее преимуществом является возможность использовать любое заданное распределение длин порождаемых последовательностей.

Распределение эмиссионных вероятностей для каждого порождающего состояния может быть описано следующим образом:

$$P_{state}(sequence) = P_{state}(sequence|L)\,P_{state}(L) \quad ,$$

где $P_{state}(sequence|L)$ - это вероятность породить определенную нуклеотидную последовательность в данном состоянии при условии, что длина последовательности равна L, а $P_{state}(L)$ - вероятность породить любую нуклеотидную последовательность

длины L в данном состоянии.

Для порождения последовательности нуклеотидов заданной длины в состоянии BKG используется локальная Марковская цепь первого порядка. Длины последовательностей, порождаемых в этом состоянии, распределены согласно геометрическому распределению со средним $1/p_{open}$ (где p_{open} - это вероятность открытия кластера):

$$P_{BKG}(L)=(1-p_{open})^{L-1}\cdot p_{open} \quad .$$

Состояния S1, S2, …, SN (N — количество типов сайтов) порождают последовательности нуклеотидов (сайты) согласно соответствующим ПВМ. В состояниях типа SPACER нуклеотиды генерируются в соответствии с той же локальной Марковской моделью, которая использовалась для порождения фоновых последовательностей. Однако распределения длин последовательностей, порождаемых в этих состояниях, могут быть любыми наперед заданными. Именно состояния типа SPACER определяют распределения расстояний между соседними сайтами в кластере. В данной работе используются всего два типа состояний SPACER: SPACER:D1 и SPACER:D2, характеризующиеся, соответственно, распределениями D1 и D2 (рис. 3): D1 - геометрическое распределение со средним m, отражающее кластеризацию сайтов без каких-либо предпочтительных расстояний между ними, D2 - синусоидальное затухающее распределение с периодом 10.5 нуклеотидов, которое соответствует ситуации, когда взаимодействующие белки связывают спираль ДНК с одной и той же стороны. Аналогичные распределения расстояний между сайтами, с расстояниями между пиками равными длине спирали ДНК, были ранее описаны в литературе [Makeev *et al.* 2003, Papatsenko *et al.* 2009, Fickett *et al.* 1996].

Каждое из состояний, порождающих сайты, имеет только два возможных перехода: обратно в состояние BKG (с вероятностью p_{close}), что соответствует закрытию кластера, или в молчащее (то есть не порождающее никаких символов) состояние CE (от «CLUSTER ELONGATION»), соответствующее продолжению кластера. Таким образом, среднее количество сайтов в кластере контролируется величиной параметра p_{close}.

Предлагаемая в данной работе СММ позволяет учитывать предпочтения в расположении сайтов в кластере, если таковые имеются. ССТФ определенных типов могут чаще находиться рядом друг с другом, нежели с сайтами других типов, например, потому что соответствующие им ТФ взаимодействуют друг с другом в момент связывания ДНК. Для того, чтобы учесть такую возможность, в СММ после каждого состояния типа CE вводится набор молчащих состояний типа NEXT (NEXT:S1, NEXT:S2, …, NEXT:SN), определяющих тип сайта, следующего за только что порожденным сайтом в кластере. Количество состояний в каждом из таких

наборов равно количеству типов сайтов, поскольку модель учитывает все возможные пары типов сайтов. Распределение переходных вероятностей из состояния CE в состояния типа NEXT может варьировать в зависимости от типа только что порожденного сайта в кластере, таким образом определяя предпочтения в порядке следования сайтов различных типов.

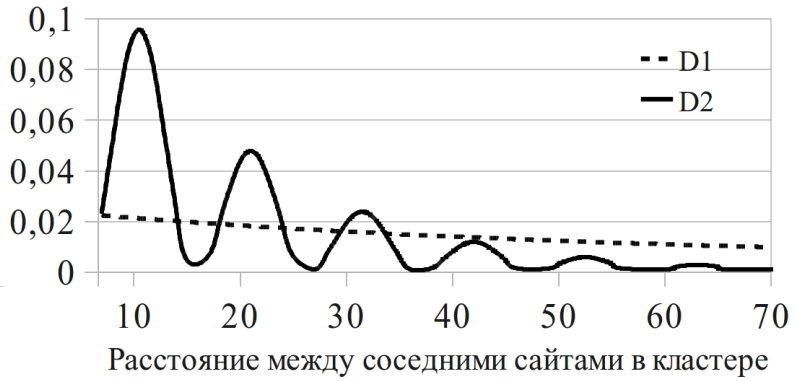

Рисунок 3. Распределения расстояния между соседними сайтами в кластере, использовавшиеся в данной работе.

СММ также позволяет учитывать предпочтения в выборе распределения расстояний между соседними сайтами. Из каждого состояния типа NEXT возможен переход в одно из состояний типа SPACER (SPACER:D1 или SPACER:D2), отличающихся между собой распределениями длин порождаемых последовательностей. Таким образом, распределение вероятностей переходов в состояния типа SPACER для каждого состояния типа NEXT определяет предпочтения в выборе распределения расстояния между сайтами для каждой пары типов сайтов.

Выявление структуры регуляторных модулей происходит в результате обучения параметров модели на наборе последовательностей, которые предположительно содержат регуляторные модули с похожей организацией. Для обучения параметров использовался алгоритм Баума-Велча [Baum *et al.* 1972, Durbin *et al.* 1998]. Поскольку целью обучения параметров является выявление структуры регуляторных областей, при обучении изменяются только параметры, определяющие структуру регуляторных модулей (переходы, выделенные пунктиром на рисунке 2).

Каждый путь в графе СММ, построенном для данной нуклеотидной последовательности, соответствует разметке этой последовательности на кластеры сайтов и фоновую последовательность. **Поиск оптимальной разметки последовательности**, соответствующей модели наилучшим образом, осуществляется по алгоритму, описанному в работе [Fariselli *et al.* 2005]. Этот алгоритм представляет собой комбинацию алгоритма «forward-backward» и алгоритма Витерби и демонстрирует более высокое качество предсказания, чем каждый из этих алгоритмов по отдельности.

Результатом работы этого алгоритма является набор кластеров сайтов,

найденных в последовательности. Для оценки значимости (веса) каждого найденного кластера используется логарифм отношения правдоподобия, который вычислялся как логарифм отношения апостериорной вероятности того, что данный участок последовательности был порожден моделью регуляторного участка, к апостериорной вероятности того, что данный участок последовательности был порожден фоновой моделью последовательности. Эти две вероятности вычисляются как вероятности отрезков путей в графе СММ, порождающих данный кластер и фоновую последовательность соответственно.

В случае поиска регуляторных модулей в группах последовательностей, относящихся к ортологичным (или ко-регулируемым) генам, алгоритм вычисляет *значение консервативности*, отражающее качество и консервативность состава предсказанных модулей для каждой группы последовательностей. Значение консервативности затем может использоваться в качестве дополнительного аргумента в пользу верности найденных регуляторных модулей, если мы уверены, что последовательности в группе действительно родственны друг другу или действительно содержат регуляторные модули ко-регулируемых генов. Или же для оценки того, насколько вероятно, что последовательности в группе действительно относятся к ко-регулируемым генам.

Для **оценки консервативности регуляторных модулей** для данной группы ортологичных генов рассчитывается величина (значение консервативности), которая отражает наличие предсказанных регуляторных областей в окрестностях значительного числа ортологичных генов и степень сходства регуляторных модулей, найденных в областях этих генов. Мера учитывает только наборы сайтов (количество сайтов каждого типа) в предсказанных регуляторных модулях, но не порядок следования сайтов. В процессе вычисления значения консервативности учитываются только предсказанные регуляторные модули с весом больше заданного порога.

В целях ясности изложения способа расчета значения консервативности, введем понятие ряда соответствующих регуляторных модулей. Предположим, что дана группа из N ортологичных генов. Для m из них были предсказаны регуляторные модули (причем для каждого гена может быть найдено один и более модулей). Предположим, что известно, какие из этих модулей соответствуют друг другу (в том смысле, что они состоят из похожих наборов сайтов), и что каждому модулю соответствует не более одного модуля в другом организме. Тогда можно говорить о ряде соответствующих регуляторных модулей, представленном в подмножестве данных ортологичных генов (рис. 4).

Сила (консервативность) ряда соответствующих регуляторных модулей вычисляется следующим образом. Для каждой пары регуляторных модулей (i и j) в ряду рассчитывается величина сходства между ними q_{ij} (пары сочетаний показаны пунктиром на рис. 4). Мера сходства между парой кластеров учитывает только состав

модуля (то есть количество сайтов каждого типа):

$$q_{ij} = \frac{n_i + n_j}{2} \cdot \frac{\cap_{ij}}{\cup_{ij}} \ ,$$

где n_i и n_j — количества сайтов в модулях i и j, $\cap_{ij}$ - размер пересечение наборов сайтов в модулях i и j, $\cup_{ij}$ — размер объединение наборов сайтов в модулях i и j (набор сайтов понимается как мультимножество). Сила ряда соответствующих регуляторных модулей рассчитывается как сумма q_{ij} по всем парам в ряду ($i < j$), нормированная на размер ряда (количество модулей, входящих в данный ряд). Тогда значение консервативности для данной группы генов равно суммарной силе всех рядов соответствующих модулей, найденных для этой группы генов:

$$r = \sum_k \sum_{i<j<N_k} \frac{q_{ij}}{N_k} \ ,$$

где k — индекс ряда соответствующих модулей, N_k — размер k-го ряда.

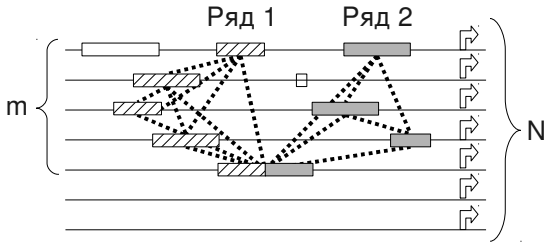

Рисунок 4. Пояснение к описанию вычисления значения консервативности предсказанных модулей для группы ортологичных генов. Горизонтальные линии соответствуют разным геномам. Стрелками показаны старты ортологичных генов. Прямоугольники обозначают регуляторные модули, при этом модули, формирующие один ряд, обозначены заливкой одного типа (см. объяснение в тексте).

Но поскольку в реальности неизвестно, какие модули соответствуют друг другу, для каждого предсказанного регуляторного модуля в каждом организме формируется свой ряд, путем выбора в остальных организмах наиболее похожих на него модулей (используя ту же меру сходства, что описана выше). Таким образом, количество рядов соответствующих модулей равно количеству модулей, найденных для данного набора генов. Итоговое значение консервативности регуляторных областей генов рассчитывается как суммарная сила всех рядов соответствующих модулей, нормированная на количество геномов, в которых были найдены модули (m).

В качестве дополнительного фильтра из рассмотрения были исключены группы ортологичных генов, для которых количество генов, для которых были найдены регуляторные модули с весом выше порога, было меньше 3 ($m < 3$) или меньше половины количества генов в данной ортологичной группе (то есть $m/N < 0.5$).

Описанная модель регуляторных последовательностей эукариот, а также алгоритмы обучения параметров модели и поиска регуляторных модулей, были реализованы на языке программирования Java в виде **программы CORECLUST**.

Тестирование разработанного подхода поиска регуляторных модулей осуществлялось на системе генов позвоночных, специфически экспрессирующихся в

мышечной ткани, и системе раннего развития *Drosophila.* Также было проведено сравнение результатов работы CORECLUST с результатами предсказания других известных программ. На обеих системах CORECLUST продемонстрировал способность предсказывать регуляторные модули с достаточно высокими показателями качества.

Для **тестирования программы на системе генов позвоночных, специфически экспрессирующихся в мышечной ткани,** использовалась выборка генов и регулирующих их ТФ, изначально составленная Wasserman и Fickett [Wasserman *et al.* 1998]. Эта выборка часто используется для оценки качества предсказания регуляторных областей. Выборка содержит 24 последовательности (со средней длиной 850 пн) из геномов человека, мыши, крысы, быка и цыпленка, содержащие в себе известные регуляторные модули, участвующие в регуляции экспрессии генов в мышечной ткани. Набор ПВМ содержит матрицы для 5 ТФ, участвующих в регуляции мышечных генов: Mef2, Myf, Sp1, Srf и Tef. Выборка была взята из материалов к статье [Klepper *et al.* 2008], авторы которой также использовали данную выборку для оценки качества ряда программ.

Качество предсказания регуляторных модулей программой CORECLUST оценивалось с помощью инструмента, разработанного Klepper и коллегами [Klepper *et al.* 2008]. Авторы этой статьи создали удобный инструмент для разносторонней оценки аккуратности предсказания регуляторных модулей, который позволил нам оценить и сравнить с другими программами предсказания, сделанные CORECLUST. Этот инструмент оценивает не только точность предсказания локализации модулей в последовательности, но также и способность программы правильно определять набор типов ССТФ, входящих в предсказанные модули.

Для всесторонней оценки качества предсказания авторы предлагают использовать сразу несколько мер соответствия предсказанных и известных регуляторных модулей: коэффициент корреляции (*CC*), чувствительность (*Sn*), специфичность (*Sp*), предсказательную ценность положительного результата (*PPV*), коэффициент эффективности (*PC*) и среднюю эффективность (*ASP*):

$$CC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}, \qquad Sn = \frac{TP}{TP+FN}, \qquad Sp = \frac{TN}{TN+FP},$$

$$PPV = \frac{TP}{TP+FP}, \qquad PC = \frac{TP}{TP+FP+FN}, \qquad ASP = \frac{Sn+PPV}{2},$$

где *TP* — это количество нуклеотидов, правильно предсказанных как входящие в регуляторный модуль, *TN* — количество нуклеотидов, правильно отнесенных к фоновой последовательности. *FN* — количество нуклеотидов, ошибочно отнесенных к фоновой последовательности, и *FP* — количество нуклеотидов ошибочно предсказанных, как принадлежащие регуляторному участку. Аналогичные величины использовались для оценки качества на уровне типов ССТФ (мотивов).

Результаты оценки предсказаний, сделанных CORECLUST, и сравнение с качеством предсказания 8 других известных программ (CMA [Kel *et al.* 2006], CisModule [Zhou *et al.* 2004], ModuleSearcher [Aerts *et al.* 2003], Stubb [Sinha *et al.* 2003], MSCAN [Johansson *et al.* 2003], MCAST [Bailey *et al.* 2003], Cister [Frith *et al.* 2001] и Cluster-Buster [Frith *et al.* 2003]) представлены на рисунке 5.

Сравнение на нуклеотидном уровне показало, что CORECLUST опережает другие программы почти по всем мерам. Особенно примечательно, что CORECLUST показывает себя наилучшим образом по мерам *CC*, *PC* и *ASP*, которые учитывают ошибки как первого, так и второго рода. Не очень высокая чувствительность предсказания CORECLUST может объясняться тем, что, несмотря на принадлежность к одной системе, некоторые гены, представленные в выборке, возможно, отличаются по структуре своих регуляторных модулей от большинства. Поскольку обучение программы осуществляется на всех последовательностях выборки, то модель регуляторных модулей отражает структуру, наиболее представленную во всей выборке.

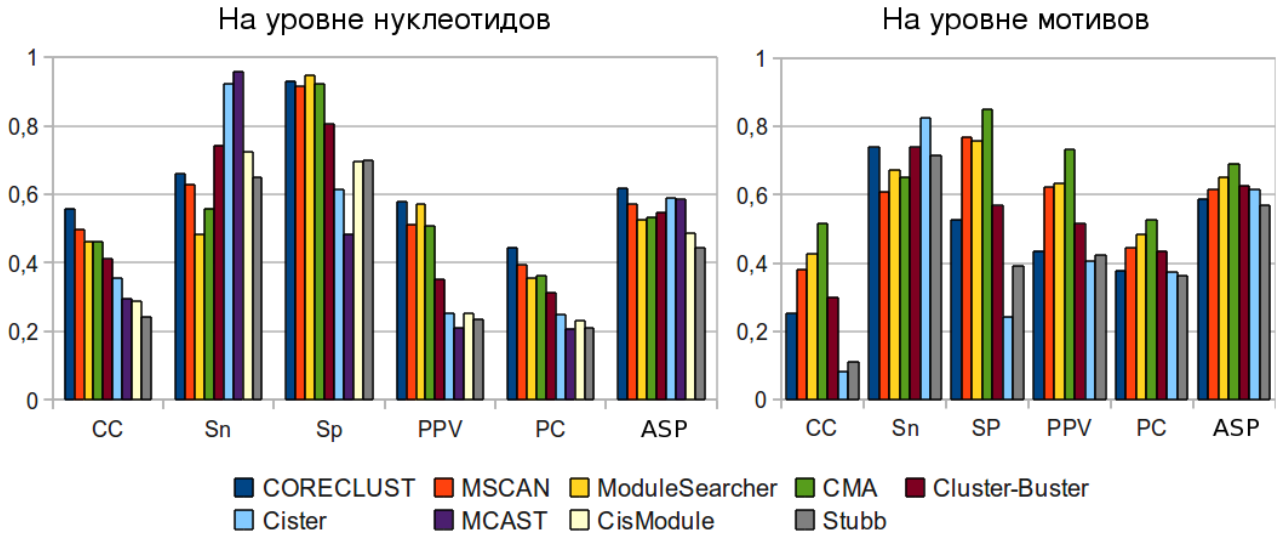

Рисунок 5. Сравнение качества предсказания программ на нуклеотидном уровне и уровне мотивов для системы генов, специфически экспрессирующихся в мышечной ткани.

Предсказания CORECLUST на уровне мотивов, напротив, имеют довольно высокую чувствительность, однако по значениям специфичности, PPV и других характеристик CORECLUST показывает средние результаты. Однако, стоит помнить, что специфичность, как и *PPV*, в данной области предсказаний всегда недооценивается, поскольку экспериментальные данные по участию ТФ в регуляции могут быть не полными.

Тестирование CORECLUST на системе развития *Drosophila* осуществлялось на 17 генах из генома *D. melanogaster*, имеющих экспериментально показанные регуляторные модули. Гены развития плодовой мушки имеют довольно обширные

регуляторные области, часто располагающиеся на расстоянии до 15-20 тысяч пар нуклеотидов от начала гена. Кроме того, в публичных базах данных доступно 12 аннотированных геномов рода *Drosophila*. Эти обстоятельства дают возможность обучать модель регуляторных модулей на последовательностях из каждого ортологичного ряда по отдельности.

Таблица 1. Сравнение качества (*CC*) предсказания программ для генов системы раннего развития *Drosophila*. Каждый ряд соответствует предсказаниям для одного гена *D. melanogaster*. Ряд TOTAL содержит значения *CC*, вычисленные для всех генов из набора. Жирным выделено максимальное значение *CC* в каждой строке. *p-значение, вычисленное с помощью одностороннего T-критерия Уилкоксона, отражающее значимость утверждения, что предсказания CORECLUST имеют более высокое значение *CC*, чем предсказания соответствующей программы.

Ген	CORECLUST	Stubb	MOPAT	Cluster-Buster
eve	**0,73**	0,56	0,54	0,58
h	**0,69**	0,17	0,26	0,49
btd	0,45	0,27	0,31	**0,47**
Kr	0,45	0,24	0,29	**0,64**
kni	0,43	0,22	0,27	**0,45**
gt	0,41	**0,48**	0,27	0,40
slp1	0,35	0,34	**0,44**	0,35
hb	0,32	**0,33**	0,17	0,22
ftz	0,31	**0,36**	0,32	0,27
fkh	**0,31**	0,28	0,27	-0,02
tll	**0,26**	0,15	0,09	0,17
prd	**0,26**	0,14	0,13	0,17
salm	**0,23**	0,07	-0,01	0,17
bowl	**0,20**	0,10	-0,01	0,17
run	0,08	**0,17**	0,07	0,11
ems	-0,02	**0,15**	-0,01	-0,02
cad	-0,03	**0,17**	-0,02	-0,04
TOTAL	**0,32**	0,20	0,20	0,29
медиана	**0,31**	0,22	0,26	0,22
стд.откл.	0,21	0,13	0,17	0,21
P-value*		< 0,05	< 0,0007	< 0,02

Поэтому обучение модели и поиск регуляторных модулей осуществлялся для каждого из генов отдельно, с использованием всех доступных ортологичных последовательностей. Для каждого гена рассматривался участок последовательности [-20000 пн,+20000 пн] относительно начала гена. Набор ПВМ [Kulakovskiy *et al.* 2010] содержал матрицы для 7 ТФ, участвующих в регуляции формирования передне-задней оси у плодовой мушки: Bcd, Hb, Cad, Kr, Kni, Tll и Gt.

Поскольку экспериментально подтвержденные модули известны только для генома *D. melanogaster*, качество предсказания оценивалось только на основе

регуляторных модулей, найденных в последовательностях из этого генома. Координаты известных модулей были взяты из базы данных REDFly [Gallo *et al.* 2011]. В качестве меры качества предсказания был выбран коэффициент корреляции между предсказанными и известными регуляторными модулями (*CC*, см. выше), поскольку эта мера учитывает ошибки обоих родов. При этом качество предсказания оценивалось только на нуклеотидном уровне в виду отсутствия достаточного количества данных о ССТФ, входящих в состав известных регуляторных модулей. Результаты предсказания CORECLUST сравнивались с результатами предсказания трех программ: Stubb [Sinha *et al.* 2003], MOPAT [Hu *et al.* 2008] и Cluster-Buster [Frith *et al.* 2003].

Сравнение качества предсказания программ (табл. 1) показало, что согласно T-критерию Уилкоксона, предсказания CORECLUST имеют более высокое значение *CC*, чем программы Stubb [Sinha *et al.* 2003] (*p-value* < 0.05), MOPAT [Hu *et al.* 2008] (*p-value* < 0.0007) и Cluster-Buster [Frith *et al.* 2003] (*p-value* < 0.02).

Таким образом, тестирование CORECLUST на системе мышечных генов позвоночных и системе раннего развития плодовой мушки показало, что программа применима к различным системам и организмам и может быть успешно использована для решения стандартной задачи поиска регуляторных модулей для набора системо-специфичных ТФ.

Преимущество учета корреляций между ССТФ для предсказания регуляторных областей в эукариотах было продемонстрировано в значительном количестве публикаций [Sinha *et al.* 2003, Erives *et al.* 2004, Hu *et al.* 2008]. Включение же в модель различных распределений расстояний между соседними сайтами в регуляторном модуле — новая черта нашего алгоритма. Для оценки преимущества, которое дает CORECLUST учет структурных особенностей регуляторных модулей, было проведено тестирование модифицированных версий программы CORECLUST с рабочими названиями CORECLUST-F и CORECLUST-FC. Отличия этих версий программы заключаются в том, что при обучении модели учитываются не все аспекты структуры регуляторных модулей. В CORECLUST-F учитываются только частоты сайтов разных типов, но не корреляции между типами сайтов и не предпочтения в расстояниях между сайтами, а в CORECLUST-FC учитываются частоты сайтов и корреляции между ними, но не предпочтения в расстояниях. Оценка качества предсказания CORECLUST-F и CORECLUST-FC осуществлялась аналогично тому, как это было описано выше. Сравнение трех версий программы, CORECLUST, CORECLUST-FC и CORECLUST-F показало, что в целом учет структурных особенностей регуляторных модулей (то есть корреляций между типами сайтов связывания и предпочтительных распределений расстояний между соседними сайтами в модуле) повышает чувствительность и немного снижает точность предсказания (рис. 6).

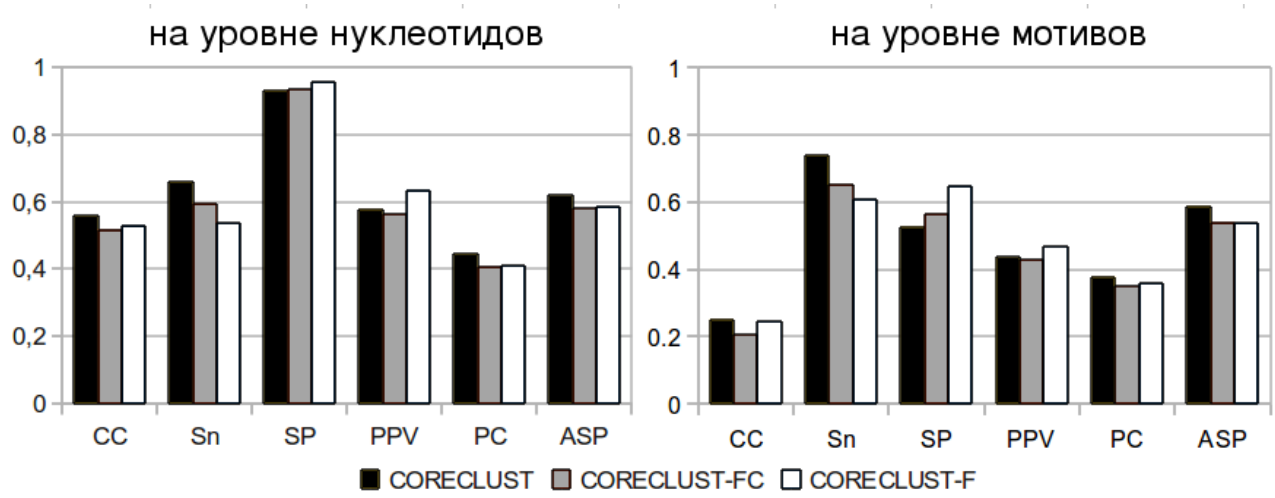

Рисунок 6. Сравнение качества предсказания CORECLUST для системы мышечных генов позвоночных с учетом разных аспектов структуры регуляторных модулей.

По-видимому, этот феномен можно объяснить тем, что при учете предпочтений следования сайтов и расстояний между ними, ССТФ со сравнительно небольшим весом, но при этом располагающиеся в “правильном” порядке и на “правильном” расстоянии друг от друга, получают возможность быть включенными в регуляторный модуль, тем самым повышая чувствительность предсказания. Такое объяснение хорошо соотносится с моделью кооперативной работы факторов транскрипции, позволяющей добиваться тонкой регуляции транскрипции в эукариотах.

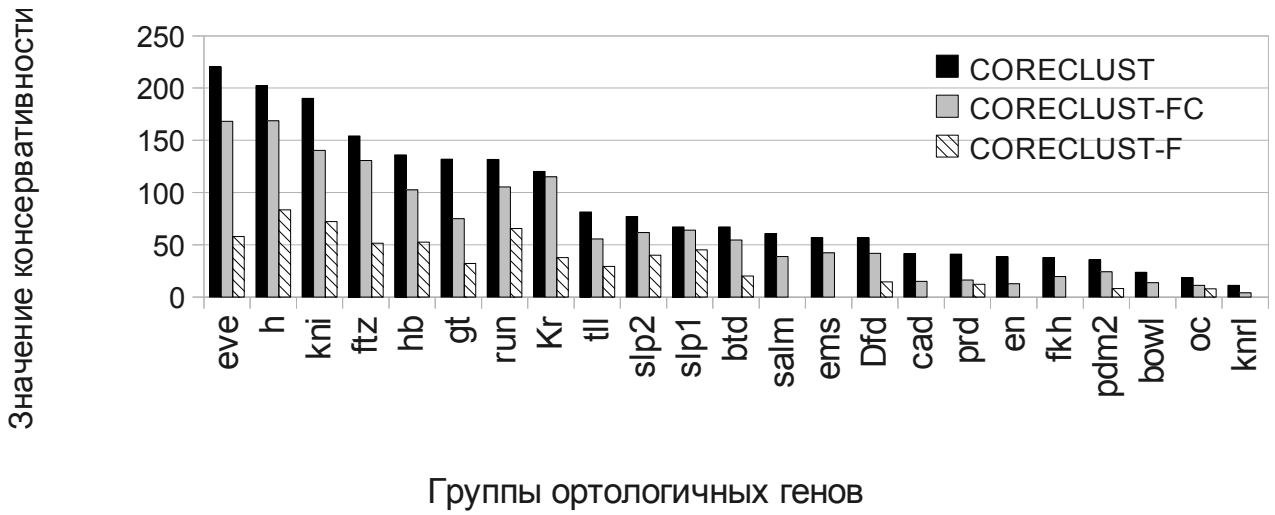

Рисунок 7. Сравнение консервативности регуляторных модулей, предсказанных CORECLUST, CORECLUST-FC и CORECLUST-F, для генов из системы раннего развития *Drosophila*. Расчет значения консервативности был выполнен для каждой группы ортологичных генов по отдельности. Группы ортологичных генов названы в соответствии с именами генов из генома *D. melanogaster.*

Если смотреть на консервативность состава регуляторных модулей, найденных для генов из одной ортологичной группы, оказывается, что модули, предсказанные

CORECLUST, то есть с учетом всех аспектов регуляторной структуры, намного консервативнее модулей, предсказанных CORECLUST-FC и CORECLUST-F (рис. 7). Это наблюдение может служить аргументом в пользу того, что предсказания полной версии программы CORECLUST надежнее, чем предсказания двух других версий.

Третья глава посвящена описанию применения разработанного алгоритма для полногеномного поиска регуляторных модулей с заданной структурой и выявления ко-регулируемых генов.

CORECLUST может быть применен для полногеномного поиска регуляторных модулей, характеризующихся структурой, похожей на структуру обучающих регуляторных модулей, и таким образом для идентификации генов, ко-регулируемых с данным геном или набором генов. Другими словами, если исследователь знает набор транскрипционных факторов, регулирующих экспрессию генов интересующей системы, и хотя бы один или несколько генов, которые с большой вероятностью регулируются этими факторами, то предлагаемый подход может быть использован для поиска других генов, регулируемых похожим на исходные гены образом.

Полногеномный поиск ко-регулируемых генов с использованием программы CORECLUST состоит из трех основных этапов (рис. 8):

1. *Обучение параметров модели.* Для обучения модели необходимо выбрать стартовый (обучающий) ген (или набор генов) в главном (референсном) геноме, регулируемый данным набором ТФ, и определить интервал относительно старта выбранного гена, в котором предположительно содержатся регуляторные модули, например, участок [-20000 пн, +20000 пн] относительно старта гена. Набор обучающих последовательностей формируется из соответствующих участков последовательностей ([-20000 пн, +20000 пн] относительно старта гена) для всех доступных ортологичных генов из интересующих геномов.
2. *Полногеномный поиск регуляторных модулей во всех исследуемых геномах.* Обученная на предыдущем этапе модель используется для поиска регуляторных модулей в окрестностях всех генов референсного генома и их ортологов из интересующих геномов.
3. *Выбор наиболее достоверных предсказаний, используя предположение о консервативности функциональных регуляторных модулей.* Поскольку функциональные регуляторные модули должны быть достаточно консервативными, на данном этапе отбираются такие группы ортологичных генов, для которых алгоритм нашел достаточно сильные регуляторные модули для значительной части генов из группы (в данной работе, хотя бы для половины генов) и при этом модули, найденные для ортологичных генов, похожи по составу сайтов. Для этого для каждой группы ортологичных генов вычисляется значение консервативности найденных регуляторных модулей, которое отражает силу (вес) и консервативность состава

предсказанных модулей. Затем группы ортологичных генов сортируются по значению консервативности. Гены референсного генома, принадлежащие к первым ортологичным группам из полученного списка, могут рассматриваться как наиболее вероятные кандидаты на роль генов, ко-регулируемых со стартовым геном.

Рисунок 8. Общая схема поиска ко-регулируемых генов. Горизонтальными линиями показаны участки перед ортологичными генам, старты генов обозначены стрелками. Внизу схемы изображен пример выдачи программы — регуляторные модули (прямоугольники), найденные в участках перед ортологичными генами. Высота прямоугольника отражает вес данного регуляторного модуля.

Таким образом, после полногеномного поиска регуляторных модулей применяется оценка эволюционной консервативности предсказанных модулей, что позволяет отобрать наиболее достоверные предсказания.

Для **тестирования** возможности применения описанного подхода для поиска ко-регулируемых генов, стартуя всего с одной группы ортологичных генов (то есть обучая параметры СММ на последовательностях, относящихся к этим генам), был осуществлен ряд полногеномных поисков генов, ко-регулируемых с генами из системы развития *Drosophila*.

Для построения модели использовался тот же набор ПВМ, что и для тестирования CORECLUST на данной системе. Поиск регуляторных модулей осуществлялся в 12 геномах рода *Drosophila* [Tweedie *et al.* 2009]. Область поиска для каждого гена определялась как участок последовательности [-20000 пн, +20000 пн] относительно старта гена (старт гена определялся как координата начала гена в базе

данных FlyBase [Tweedie *et al.* 2009]). При попадании соседних генов в этот интервал область поиска обрезалась по ближайшим границам соседних генов. Таким образом, поиск осуществлялся только в области, не выходящей за рамки межгенных областей, окружающих данный ген. При этом поиск происходил в последовательностях, ориентированных согласно направлениям соответствующих генов. Участки, содержащие повторы, выявленные программой RepeatMasker [http://www.repeatmasker.org], были исключены из рассмотрения.

Поиск ко-регулируемых генов был осуществлен для каждого из 22 генов системы раннего развития *Drosophila* (*h*, *eve*, *hb*, *kni*, *Kr*, *ftz*, *gt*, *run*, *prd*, *cad*, *slp1*, *slp2*, *tll*, *btd*, *salm*, *bowl*, *knrl*, *fkh*, *ems*, *Dfd*, *pdm2* и *en*). В каждом случае обучение модели происходило на областях [-20000 пн, +20000 пн] (относительно старта гена) всех доступных ортологичных генов. После этого обученная модель использовалась для поиска регуляторных модулей для всех генов во всех двенадцати геномах *Drosophila*. Результат каждого поиска представлял из себя список генов, для которых были найдены сильные и консервативные регуляторные модули, похожие на регуляторные модули обучающего гена, то есть генов, предположительно ко-регулируемых с обучающим геном. Во всех 22 случаях в результирующем списке генов были перепредставлены гены, участвующие в раннем развитии *Drosophila*. Например, для обучающего гена *hairy* (*h*), который является первичным pair-rule геном, вовлеченным в установление сегментов зародыша на 4-6 стадиях развития плодовой мушки, было найдено 45 предположительно ко-регулируемых генов. Шесть из них попали в результирующий список только потому, что их области поиска пересекались с областями поиска хорошо известных генов раннего развития плодовой мушки. Для ясности изложения эти гены были убраны из рассмотрения. Оставшиеся 39 генов характеризуются сильными консервативными предсказанными регуляторными модулями и могут рассматриваться как вероятные кандидаты на роль генов, ко-регулируемых с геном *h*. Анализ полученного списка генов с помощью программы GOStat [Beissbarth *et al.* 2004], показал, что в этом списке значительно перепредставлены GO-категории [Ashburner *et al.* 2000], связанные с ранним развитием плодовой мушки и в частности с сегментацией бластодермы (табл. 2). При этом список содержал все шесть pair-rule генов.

Более того, большинство генов из начала списка (табл. 3) являются известными участниками формирования передне-задней оси у плодовой мушки и действительно имеют экспериментально подтвержденные регуляторные модули, регулирующие транскрипцию этих генов на 4-6 стадиях развития *Drosophila*. Среди них всего три гена не являются известными участниками процесса раннего развития плодовой мушки (*CG13713*, *CG5103* и *Cyp6v1*). Тем не менее, они все равно являются хорошими кандидатами на роль генов, участвующих в раннем развитии *Drosophila*, поскольку, в соответствии с данными по иммуно-преципитации хроматина [Li *et al.*

2008], ТФ, регулирующие экспрессию генов развития, связываются с участками ДНК вблизи этих генов на 4-6 стадии развития.

Таблица 2. GO-категории, наиболее представленные в списке генов, предположительно ко-регулируемых с геном *h*. Значимость перепредставленности GO категорий оценена с помощью программы GOStat [Beissbarth et al. 2004].

Категория GO	# предсказанных генов, принадлежащих к категории	# генов в категории	P-Value
Blastoderm segmentation	14	137	2.98E-17
Embrionic pattern specification	14	176	5.51E-16
Segmentation	14	181	5.51E-16
Periodic partitioning by pair rule gene	6	6	2.18E-14
Posterior head segmentation	7	15	2.31E-13
Embrionic development	16	532	1.78E-12

Таблица 3 Первые 15 генов из списка генов, предположительно ко-регулируемых с геном *h.* Значение консервативности показывает степень консервативности регуляторных модулей, предсказанных для данного гена.

Ген	Значение консерва-тивности	Функция гена
h	194,8	pair-rule gene, TF; open tracheal system development, nervous system development
ftz	45,4	pair rule gene, TF; gonadal mesoderm development
eve	42,7	pair-rule gene, TF; regulation of axonogenesis and cardioblast cell fate specification
kni	32,1	gap gene, TF; dendrite morphogenesis, muscle organ and epidermis development
hb	28,3	gap gene, TF; torso signaling pathway, terminal region and neuroblast fate determination
slp1	27,0	pair-rule and segment polarity gene, TF; specification of segmental identity, head
run	20,0	pair-rule gene, TF; axon guidance, dendrite morphogenesis, eye morphogenesis
CG13713	17,4	regulation of localization (predicted)
slp2	16,5	pair-rule and segment polarity gene, TF
Kr	12,6	gap gene, TF; neuroblast fate determination, axon guidance
CG5103	11,5	transketolase (predicted)
Cyp6v1	10,1	cytochrome P450
pdm2	9,5	gap gene, TF; neuroblast development
gt	9,3	gap gene, TF; torso signaling pathway; terminal region determination

Для систематической оценки качества поиска ко-регулируемых генов было проведено сравнение списков ко-регулируемых генов, найденных нашей программой, со списком генов, полученным с помощью базовой программы для предсказания регуляторных модулей, Cluster-Buster [Frith *et al.* 2003]. Программа Cluster-Buster была выбрана для сравнения потому, что кроме набора ПВМ, она не использует информацию ни о структуре регуляторных областей, ни о их консервативности, но

при этом показывает очень хорошие результаты предсказания.

Сначала была сформирована положительная выборка генов, которые, с большой вероятностью участвуют в раннем развитии плодовой мушки. Для этого к списку хорошо известных из литературы генов раннего развития были добавлены гены, которые аннотированы термином «embryonic pattern specification» (GO:0009880) в базе данных GO [Ashburner *et al.* 2000] и экспрессируются на 2-4 стадии развития *D.melanogaster* [Tomancak *et al.* 2002]. В результате в «положительную» выборку вошли 115 генов.

Поскольку программа Cluster-Buster не выдает итоговый список генов, были применены 2 простые меры для отбора генов, с наиболее сильными и многочисленными регуляторными участками, предсказанными этой программой. После полногеномного поиска кластеров сайтов программой Cluster-Buster гены были отсортированы по:

1) максимальному весу предсказанных для них регуляторных модулей;
2) суммарному весу предсказанных регуляторных модулей.

Затем для каждого из 22 стартовых (обучающих) генов было проведено по три гипергеометрических теста для оценки статистической значимости перепредставленности генов положительной выборки в каждом из трех списков генов:

1) в итоговом списке ко-регулируемых генов, предсказанных программой CORECLUST, содержащем m генов;
2) в первых m генах, из списка предсказаний Cluster-Buster, отсортированного по максимальному весу модулей;
2) в первых m генах, из списка предсказаний Cluster-Buster, отсортированного по суммарному весу модулей.

Результаты сравнения, представленные на рисунке 9, демонстрируют, что список генов, найденных с применением предлагаемого подхода, учитывающего структуру регуляторных модулей и их консервативность, значительно лучшее совпадает с положительной выборкой, чем предсказания, сделанные программой Cluster-Buster. С другой стороны, сравнение качества предсказаний, выполненных с использованием модели, обученной на разных генах, показывает, что не все гены одинаково хорошо подходят для обучения модели. Возможно, это можно объяснить тем, что регуляторные участки некоторых из них содержат недостаточное количество сайтов связывания для того, чтобы хорошо обучить параметры модели.

Четвертая глава посвящена описанию выявленной структуры регуляторных модулей для системы генов позвоночных, специфически экспрессирующихся в мышечной ткани, и системы раннего развития *Drosophila*.

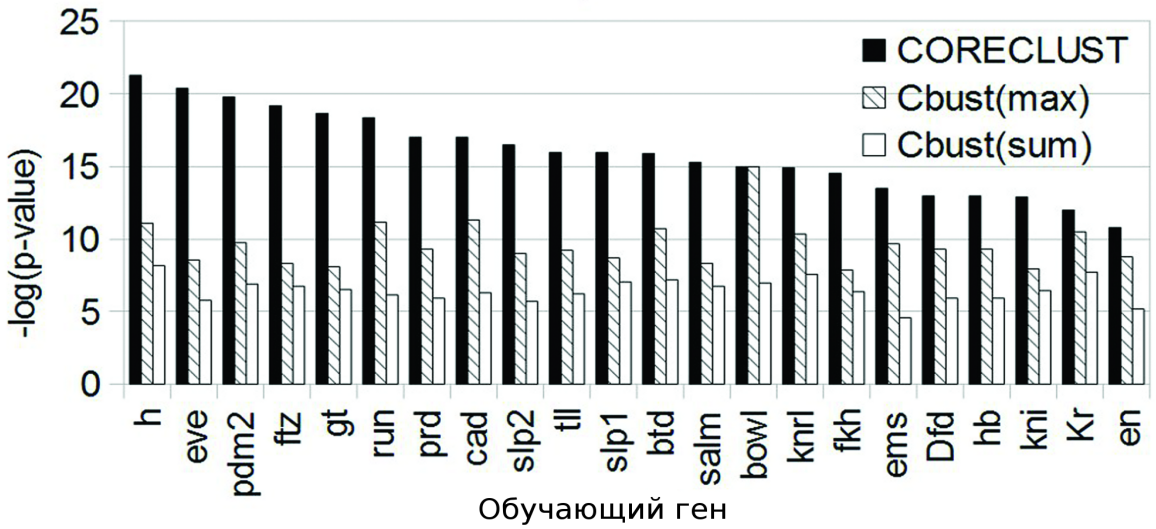

Рисунок 9. Оценка качества поиска генов, относящихся к системе раннего развития *Drosophila*. Гистограмма представляет сравнение значимости представленности генов из положительной выборки в наборах генов, найденных программами CORECLUST (предсказания представлены для разных обучающих генов) и Cluster-Buster (гены отсортированы по максимальному (Cbust (max)) и суммарному (Cbust (sum)) весу модулей).

В процессе обучения модели CORECLUST выявляет структурные особенности, а именно предпочтение следования сайтов и предпочтительное распределение расстояний между ними, которые разделяют сходно функционирующие регуляторные модули. Можно предположить, что выявленные закономерности расположения сайтов несут функциональную нагрузку, важную для корректной работы регуляторной системы, и поэтому сохраняются в процессе эволюции, а также наблюдаются у ко-регулируемых генов. Таким образом, CORECLUST может применяться не только для поиска регуляторных модулей, но также и для выявления и анализа предпочтительного взаимного расположения сайтов связывания, свойственного данной регуляторной системе.

Структура регуляторных модулей, содержащихся в обучающих последовательностях, описывается параметрами обученной модели. Для каждой пары типов сайтов *i* и *j* в модели содержится условная вероятность $P(j|i)$ наблюдать сайт типа *j* следом за сайтом типа *i*. Эта вероятность отражает частоту наблюдений пары сайтов соответствующих типов, располагающихся рядом в регуляторных модулях в обучающих последовательностях. Достаточно высокое значение вероятности $P(j|i)$ может говорить о том, что сайты данных типов часто располагаются рядом друг с другом, что в свою очередь может означать, что ТФ, связывающие эти сайты, взаимодействуют друг с другом во время регуляции транскрипции генов.

Анализ параметров модели, обученной на регуляторных участках генов, специфически экспрессирующихся в мышечной ткани, показал, что наиболее

вероятными парами ТФ, чьи сайты связывания часто располагаются рядом, являются Mef2-Myf, Sp1-Srf и Sp1-Sp1. Для всех трех пар в литературе описаны наблюдения, говорящие в пользу того, что эти белки действительно взаимодействуют друг с другом в процессе регуляции транскрипции соответствующих генов. Так, расположение сайтов связывания Mef2 и Myf рядом друг с другом, причем на расстоянии, кратном шагу спирали ДНК, было замечено J.W. Fickett [Fickett 1996] еще в 1996 году в процессе анализа известных регуляторных модулей мышечных генов. Интересно, что в соответствии с параметрами нашей модели, сайты связывания этих факторов также предпочтительно располагаются на расстоянии кратном шагу спирали ДНК, то есть расстояние между ними описывается синусоидальным распределением расстояний, использующимся в модели. Синергия между сайтами связывания Mef2 и Myf также зафиксирована в базе данных TransComplel (C00120) [Kel-Margoulis *et al.* 2002]. Взаимодействие факторов SRF и Sp1 косвенно подтверждается экспериментальными работами [Biesiada *et al.* 1999] и [Madsen *et al.* 1997]. Синергическая активация транскрипции фактором Sp1 была показана *in vivo* в статье [Anderson *et al.* 1991].

Другие, чуть менее предпочтительные, пары типов сайтов, выявленные в результате анализа модели, также имеют подтверждения в литературе и базе данных TransCompel: Tef-Mef2 [Maeda *et al.* 2002]; Mef2-Sp1 [Grayson *et al.* 1998]; Myf-Sp1 (TransComplel: C00027, C00028).

В соответствии с параметрами модели, обученной на регуляторных участках генов раннего развития плодовой мушки, регуляторная система раннего развития *Drosophila* характеризуется гомотипичесткими взаимодействиями между ТФ, что согласуется с наблюдениями, описанными в литературе [Lebrecht *et al.* 2005, Lifanov *et al.* 2003, Makeev *et al.* 2003]. Интересно, что в некоторых случаях сайты в таких гомотипических парах (например, Kr-Kr и Hb-Hb) имеют тенденцию быть сонаправленными. Более того, некоторые из гомотипических пар (например, Hb-Hb и Bcd-Bcd) характеризуются выбором синусоидального распределения в качестве предпочтительного распределения расстояний между сайтами. Похожий характер расстояний между сайтами связывания этих факторов наблюдался также в работе [Papatsenko *et al.* 2009]. Такое распределение расстояний между сайтами связывания одного ТФ хорошо согласуется с моделью кооперативного связывания молекул одного и того же фактора с одной стороны спирали ДНК, которое как раз должно приводить к формированию гомотипических кластеров сайтов связывания, преимущественно располагающихся на расстоянии, кратном длине спирали ДНК.

Анализ расположения ССТФ в регуляторных модулях, предсказанных около генов раннего развития *Drosophila*, выявил несколько интересных распределений расстояний между сайтами, не соответствующих распределениям расстояний, включенным в модель. Например, почти во всех наблюденных (15 из 18) парах сайтов Gt>Gt> (два сайта связывания фактора Gt, расположенные на прямой цепи ДНК)

расстояние между сайтами равно 51-58 нуклеотидов. Распределение расстояний между сайтами пары Kni>Kni< (сайты связывания фактора Kni, расположенные на разных цепях ДНК) отличается пиком на расстоянии 135-138 нуклеотидов (рис. 10), что довольно необычно и, возможно, говорит о связывании фактора Kni с компактизованной ДНК.

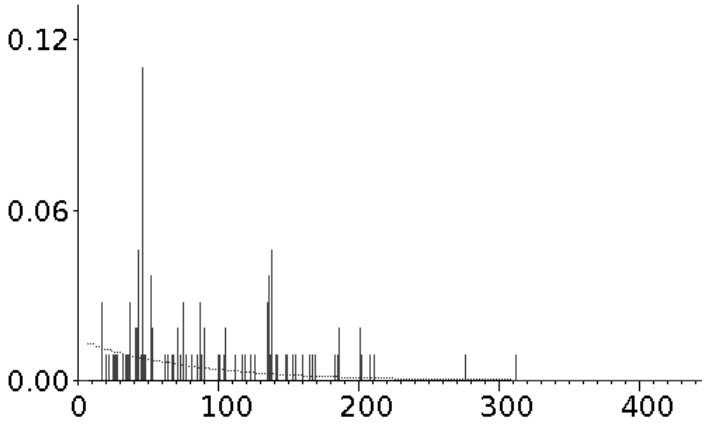

Рисунок 10. Распределение расстояний между сайтами пары Kni>Kni< в регуляторных модулях, найденных около генов раннего развития *Drosophila*. Расстояние измеряется между стартами ССТФ. Коэффициент корреляции между сайтами в паре равен 0.51, всего было наблюдено 109 пар. Фоновое распределение расстояний между сайтами показано пунктирной линией.

Таким образом, предложенный алгоритм выявления регуляторной структуры интересующей системы генов позволят делать биологически обоснованные предположения о совместной работе и характере взаимодействия ТФ.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Разработана вероятностная модель регуляторных модулей эукариот, описывающая их структуру, а именно частоты встречаемости сайтов связывания факторов транскрипции, предпочтение следования сайтов и характерные распределения расстояний между соседними сайтами.

2. Разработан и реализован метод определения параметров модели, описывающих структуру регуляторных модулей (закономерностей взаиморасположения сайтов связывания).

3. Разработан и реализован метод поиска регуляторных модулей в геномах эукариот для набора системо-специфичных ТФ, в основе которой лежит вероятностная модель регуляторных модулей. Показано, что разработанный метод позволяет эффективно искать регуляторные модули для набора системо-специфичных ТФ для генов из различных регуляторных систем и организмов.

4. Разработан метод полногеномного поиска ко-регулируемых генов на основе анализа предсказанных регуляторных модулей и оценки консервативности их структуры. Эффективность разработанного подхода продемонстрирована на примере системы раннего развития *Drosophila*.

5. Проведен анализ структуры регуляторных модулей для генов позвоночных, специфически экспрессирующихся в мышечной ткани, и генов раннего развития

Drosophila. Продемонстрирована возможность применения разработанного подхода к изучению совместной работы ТФ, а также выявлен ряд ранее неизвестных особенностей распределений расстояний между сайтами, позволяющих предполагать различные механизмы взаимодействия комплекса ТФ с ДНК, в том числе связывание комплекса с компактизованной ДНК.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в научных журналах

1. Anna A. Nikulova, Alexander V. Favorov, Roman A. Sutormin, Vsevolod J. Makeev, Andrey A. Mironov. CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation. Nucleic Acids Research (2012); doi: 10.1093/nar/gks235.

2. А. А. Никулова, М. С. Полищук, В. Г. Туманян, В. Ю. Макеев, А. А. Миронов, А. В. Фаворов. Корреляции кластеров сайтов связывания и экспериментальных данных по связыванию белков с ДНК позволяют предполагать структуру регуляторных модулей. Биофизика (2012) 57(2): 212–214.

Тезисы конференций

1. Nikulova A.A., Mironov A.A. Computational prediction and analysis of transcriptional regulatory modules in mammals. Proceedings of the 3rd International Moscow Conference on Computational Molecular Biology (MCCMB’07), 2007, pp. 228-229.

2. Никулова А.А., Миронов А.А. Поиск и анализ кластеров сайтов связывания транскрипционных факторов в геномах млекопитающих. Труды 30-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'07, 2007, с. 284-285.

3. Никулова А.А., Сутормин Р.А., Фаворов А.В., Миронов А.А. Построение HMM, основанной на правилах взаиморасположения сайтов связывания транскрипционных факторов, и ее применение для поиска корегулируемых генов в геномах рода Drosophila. Труды 31-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'08, 2008, с. 352-353.

4. Nikulova A., Mironov A. Prediction of regulatory elements in Drosophila genomes using hidden Markov model based on the arrangement of transcription factor binding sites. Proc. 4th International Moscow Conference on Computational Molecular Biology (MCCMB’09), 2009, pp. 261-262.

5. Nikulova A.A., Favorov A.V., Sutormin R.A., Mironov A.A. Prediction and Comparative Analysis of Transcriptional Regulatory Regions in Drosophila Genomes. Труды 32-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'09, 2009, с. 290-291.

6. Никулова А.А., Фаворов А.В., Миронов А.А. Предсказание и анализ

консервативных транскрипционных регуляторных областей в геномах рода *Drosophila*. Труды 33-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'10, 2010, с. 422-426.

7. A.A. Nikulova, A.V. Favorov, A.A. Mironov. An approach to predict cis-regulatory modules and identify conserved regulatory grammar in eukaryotic genomes. Proceedings of International Moscow Conference on Computational Molecular Biology (MCCMB'11), 2011, p. 253.

8. A.A. Nikulova, A.V. Favorov, A.A. Mironov. CORECLUST: prediction of cis-regulatory modules and revealing their internal structure. Труды 34-й конференции молодых ученых и специалистов ИППИ РАН ИТиС'11, 2011, с. 64-69.

9. A.A. Nikulova, A.V. Favorov, V.Yu. Makeev and A.A. Mironov. A generalized hidden Markov model for prediction of cis-regulatory modules in eukaryote genomes and description of their internal structure. Proceedings of 3rd International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS'2012), pp. 34-41.

Государственная регистрация программы

1. Никулова А.А., Сутормин Р.А., Фаворов А.В., Макеев В.Ю., Миронов А.А. Программа для поиска кластеров регуляторных сигналов в геномах эукариот. Свидетельство №2012610082.

Автор выражает глубокую благодарность своему научному руководителю Андрею Александровичу Миронову за руководство и помощь при выполнении диссертации, а также искреннюю признательность Александру Фаворову, Роману Сутормину и Михаилу Сергеевичу Гельфанду.