Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича РАН

УДК 004.9, 577.214, 577.29, 575.89

№ госрегистрации 01201174374

Инв. № 004

УТВЕРЖДАЮ

Директор ИППИ РАН

академик РАН, проф.

_____А. П. Кулешов

«30» октября 2012 г.

ОТЧЁТ О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

«Разработка алгоритмов и программ для сравнительного анализа геномов позволяющие определять функциональные, в частности регуляторные сегменты геномов»

Этап № 4

Обобщение и оценка результатов исследований.

(заключительный)

шифр 2011-1.4-514-008-001

Руководитель работы

Ведущий научный сотрудник

ИППИ РАН Миронов А.А.

Доктор биол.наук, проф. « 30» октября 2012 г.

Москва

2012

Список исполнителей

Руководитель работы В.н.с. ИППИ РАН, д.б.н., к.ф.-м.н., проф.	(подпись, дата)	А.А. Миронов 30.10.2012
Зав. УНЦ «Биоинформатика» ИППИ РАН, д.б.н., к.ф.-м.н. проф.	(подпись, дата)	М.С. Гельфанд 30.10.2012
Стажер-исследователь ИППИ РАН, аспирант	(подпись, дата)	А.В. Артемов (2.1.2) 30.10.2012
Стажер-исследователь ИППИ РАН, аспирантка	(подпись, дата)	Н.А. Быкова (2.2) 30.10.2012
М.н.с. ИППИ РАН	(подпись, дата)	Д.В. Виноградов (1)
Стажер-исследователь ИППИ РАН, аспирантка	(подпись, дата)	30.10.2012 С. В. Виноградова (1)
Стажер-исследователь ИППИ РАН, аспирантка	(подпись, дата)	30.10.2012 С.К. Гарушянц (Заключение)
Н.с. ИППИ РАН, к.б.н.	(подпись, дата)	30.10.2012 Е.О. Ермакова (2.3.5.5)

Стажер-исследователь ИППИ РАН, аспирантка	(подпись, дата)	А.А. Никулова (2.3.1) 30.10.2012
Н.с. ИППИ РАН, к.ф.-м.н.	(подпись, дата)	Е.Д. Ставровская (2.3.5) 30.10.2012
М.н.с. ИППИ РАН	(подпись, дата)	Р.А. Сутормин (2.3.3) 30.10.2012
Стажер-исследователь ИППИ РАН, аспирантка	(подпись, дата)	Е.Е. Храмеева (2.3.1.3) 30.10.2012
Аспирант ИППИ РАН	(подпись, дата)	Р.А.Солдатов (2.3.2) 30.10.2012
Нормоконтролер	(подпись, дата)	Е.Д. Ставровская 30.10.2012

Реферат

Отчёт 108 стр., 1 часть, 22 рис., 5 табл., 38 источников, 1 приложение.

КЛЮЧЕВЫЕ СЛОВА: алгоритм, предсказание функции, эволюция, геном, транскрипция, регуляция, экспрессия генов, сравнительная геномика.

Объектом исследования являются геномная информация и регуляция транскрипции в геномах высших эукариот.

Целью работы является создание обоснованных методов сравнительно-геномного анализа и разработка алгоритмов и программ для поиска регуляторных модулей в геномах эукариот с использованием сравнительно-геномного анализа.

Результаты работы:

— Проведено обобщение результатов исследований

— Проведено сопоставление анализа научно-информационных источников и результатов теоретических и экспериментальных исследований;

— Проведена оценка эффективности полученных результатов в сравнении с современным научно-техническим уровнем;

— Проведен анализ выполнения требований технического задания на НИР;

— Проведена оценка полноты решения задач и достижения поставленных целей НИР.

— Разработаны рекомендации по использованию результатов проведенных НИР в реальном секторе экономики, а также в дальнейших исследованиях и разработках, в том числе:

— Проведена технико-экономическая оценка рыночного потенциала полученных результатов;

— Разработаны рекомендации и предложения по использованию результатов проведенных НИР в реальном секторе экономики;

—Реализованы мероприятий по достижению технико-экономических показателей (п. 8.2 технического задания).

СОДЕРЖАНИЕ

Нормативные ссылки

В настоящем отчете использованы ссылки на следующие нормативные документы:

ГОСТ 7.32-2001 «Отчет о научно-исследовательской работе. Структура и правила оформления»

Обозначения и сокращения

ТФ	Транскрипционный фактор
ТФСС	Сайт связывания транскрипционного фактора
ДНК	Дезоксирибонуклеиновая кислота
РНК	Рибонуклеиновая кислота
СММ	Скрытая Марковская модель
ПВМ	Позиционная весовая матрица
РМ	Регуляторный модуль
ЦРМ	Цис-регуляторный модуль
ПК	Программный комплекс

1 ВВЕДЕНИЕ

Сравнительная геномика и биоинформатика – быстро развивающаяся область современной биологии. Огромные, экспоненциально растущие объемы данных делают применение компьютерных методов анализа необходимой составной частью современных молекулярно-биологических исследований и, особенно, больших проектов по систематическому анализу геномов и метагеномов, транскриптомов, протеомов. Рост количества данных, а также появление новых экспериментальных методов ставит перед биоинформатикой новые задачи по построению алгоритмов анализа этих данных. К сожалению наши знания о работе клетки достаточно ограничены, поэтому большинство биоинформатических методов основаны на достаточно грубых моделях. С другой стороны, экспериментальные данные также имеют значительный уровень шума. Отсюда следует, что практически все методы анализа имеют достаточно высокий уровень ошибок. Для подавления ошибок применяют сравнительно-геномный подход. В основе этого подхода лежит представление о том, что действительно важные особенности должны воспроизводиться в различных организмах.

В основе большинства биоинформатических подходов лежит вычисление некоторой функции от последовательности и сравнение полученного значения с некоторым порогом. Простейшим примером такого подхода является вычисление веса позиционной весовой матрицы и сравнение полученного значения с порогом. Если вес превышает пороговое значение, то считается, что предсказанный сайт действительно является настоящим сайтом связывания. Как правило, пороговое значение выбирается достаточно произвольно в зависимости от задачи. Можно использовать два порога — верхний и нижний. Если вычисленное значение больше верхнего порога, то мы предсказываем биологическую функцию, а если ниже нижнего порога, то мы предсказываем отсутствие биологической функции. При этом остается некоторая «серая» зона, где мы не можем дать никакого ответа. Наличие серой зоны на самом деле отражает наше не полное знание и поэтому не точную модель. К сожалению в биологии нет возможности строить все более точные модели и по

другой причине. Увеличение точности модели приводит к увеличению количества параметров модели. При этом у нас, как правило, нет достаточного количества экспериментальных данных, которые позволили бы хорошо настроить модель. Проблема недостаточного количества данных приводит к необходимости использовать Байесовский подход и использовать некоторые априорные распределения вероятностей.

Для предсказания биологической функции используют сравнительно-геномный анализ, в основе которого лежит представление о давлении отбора на функцию. Основным методом сравнительно-геномного анализа в настоящее время является экспертная оценка, когда исследователь привлекает сравнительную информацию, а также целый ряд дополнительных соображений.

Для проведения сравнительно-геномного анализа в настоящей работе мы разработали новые подходы, основанные на анализе эволюции последовательностей под давлением отбора и в нейтральной модели.

Исследование регуляции экспрессии генов является важнейшим направлением современной молекулярной биологии. Для этого используется целый ряд экспериментальных методов. Однако, клетка может находиться в очень большом количестве состояний, многие из которых не поддаются экспериментальной проверке. Поэтому задача предсказания регуляторных модулей остается актуальной. Кроме того, особенности структуры регуляторных модулей представляет большой интерес для понимания особенностей функционирования регуляторных систем.

1.1 Основание, исходные данные и обоснование необходимости проведения работы

Основанием для проведения работ является Государственный контракт № 07.514.11.4007 от « 13 » июля 2011 г. «Разработка алгоритмов и программ для сравнительного анализа геномов позволяющие определять функциональные, в частности регуляторные сегменты геномов» , шифр заявки 2011-1.4-514-008-001

1.1.1 Исходные данные.

Исходными данными для поиска регуляторных модулей являются набор ПВМ для ТФ, регулирующих экспрессию генов интересующей биологической системы, и набор последовательностей, предположительно содержащих регуляторные модули ортологичных и/или ко-регулируемых генов этой системы. Для полногеномного анализа входными данными являются родственные геномы и набор матриц ПВМ. Для сравнительно-геномного анализа дополнительными входными данными являются эволюционные деревья последовательностей.

1.1.2 Результаты

Проект состоял из четырех этапов.

На первом этапе проекта проведена аналитическая работа — патентного поиск, аналитического обзора информационных источников и выбор направлений исследований.

На втором этапе проекта построены эффективные алгоритмы уточнения предсказаний с учетом дерева, основанные на оригинальной идее скрытых Марковских моделей на деревьях. Разработана методика оценки статистической значимости сравнительно-геномного анализа. Разработана архитектура программного комплекса ЭО ПК CORECLUST для предсказания регуляторных модулей.

На третьем этапе разработан программный комплекс (ПК) для предсказания регуляторных модулей в геномах эукариот. Для реализации ПК CORECLUST разработана программа оптимизации параметров вероятностной модели регуляторных модулей. ПК CORECLUST, который имеет две реализации – локальную и клиент-серверную. Произведено тестирование ПК, разработана программная документация с учетом результатов тестирования. Применение ПК CORECLUST к реальным биологическим данным позволило получить интересные результаты. Проанализирована структура предсказанных регуляторных модулей генов раннего развития плодовой мушки и генов мышечной системы позвоночных; выявлены необычные распределения расстояний между сайтами, характеризующие пиками на расстояниях порядка 150-200 нп.

На четвертом этапе произведено обобщение и оценка полученных результатов

исследований. Произведено сопоставление анализа научно-информационных источников и результатов теоретических и экспериментальных исследований, сделана оценка эффективности полученных результатов в сравнении с современным научно-техническим уровнем. Осуществлен анализ выполнения требований технического задания на НИР, произведена оценка полноты решения задач и достижения поставленных целей НИР. Разработаны рекомендации по использованию результатов проведенных НИР в реальном секторе экономики, а также в дальнейших исследованиях и разработках.

По результатам этапов представлены отчеты:

Отчет по научно-исследовательской работе за этап №1, инвентарный номер 001

Отчет по научно-исследовательской работе за этап №2, инвентарный номер 002

Отчет по научно-исследовательской работе за этап №3, инвентарный номер 003

Отчет по научно-исследовательской работе за этап №4, инвентарный номер 004

2 ОСНОВНАЯ ЧАСТЬ

2.1 Уровень разработки, актуальность и новизна

В данной работе была создана обобщенная скрытая Марковская модель регуляторных модулей эукариот, описывающая регуляторную структуру, включающую частоты сайтов, предпочтения следования сайтов и предпочтительные распределения расстояний меду сайтами. Впервые рассмотрена возможность учета расстояний между сайтами в модели регуляторных модулей. Разработан и реализован метод поиска регуляторных областей в геномах эукариот, который может быть использован при изучении механизмов регуляции транскрипции, эволюции регуляторных участков, а так же для поиска ко-регулируемых генов. Проанализирована структура предсказанных регуляторных модулей генов раннего развития плодовой мушки и генов мышечной системы позвоночных; выявлены необычные распределения расстояний между сайтами, характеризующие пиками на расстояниях порядка 150-200 нп и показано, что структура регуляторных модулей важна для их функционирования. На основе разработанной модели создан программный комплекс (ПК), который имеет две реализации – локальную и клиент-серверную. Произведено тестирование ПК, разработана программная документация с учетом результатов тестирования.

Впервые разработан метод сравнительно-геномного анализа, основанный на Скрытых Марковских моделях на деревьях. Этот метод позволяет предсказывать вероятности состояний биологических систем на узлах деревьев (в том числе и на листьях) в условиях не полной определенности наблюдений на листьях. Метод реализован в виде программы CMP_GENOMICS. Предложенный метод не имеет аналогов даже на уровне постановки задачи.

Для исследования статистической значимости наблюдений параметров (например, энергии вторичной структуры РНК, или силы сайтов связывания транскрипционных факторов) на современных последовательностях был впервые разработан подход, основанный на модели диффузии параметров в пространстве последовательностей. Построены статистики для оценки значимости давления отбора на исследуемые параметры, исследовано поведение

указанных статистик в различных модельных примерах.

Все разработанные подходы, методы и программные реализации важны для понимания особенностей эволюции последовательностей, а также для предсказания функциональных, в частности регуляторных сегментов геномов.

2.2 Обобщение и оценка полученных результатов

2.2.1 Обобщение результатов исследований

Цель настоящего исследования была разработка методов предсказания функциональных участков генома с использованием сравнительно-геномного подхода. В настоящей работе были разработаны два подхода к проблеме сравнительно-геномного анализа с учетом эволюционной модели. Сравнительно-геномный анализ был применен для построения вероятностной модели, описывающей регуляторные модули в геномах многоклеточных организмов.

2.2.1.1 Применение уравнений диффузии к сравнительно-геномному анализу

Наши подходы к сравнительно-геномному анализу основывались на следующих представлениях. Молекулярная эволюция происходит на уровне последовательностей, а давление отбора происходит на уровне фенотипа. При этом мы рассматриваем молекулярные фенотипы, такие как наличие или отсутствие регуляции того или иного гена, наличие или отсутствие сигнального пептида, наличие или отсутствие функциональной вторичной структуры РНК и пр. Однако в большом числе случаев прямые экспериментальные данные по этим фенотипам либо отсутствуют, либо существуют в весьма ограниченном объеме. Более того, экспериментальные данные, как правило, отвечают некоторым специальным условиям, которые часто бывают весьма далекими от тех условий, в которых функционирует клетка в естественной среде, или в организме.

С другой стороны, биоинформационный анализ последовательностей, как правило, основан на построении модели биологического явления. Построенная модель биологического явления лишь до определенной степени отражает реальную ситуацию в клетке. Это связано,

прежде всего, с тем, что механизмы функционирования клеточных систем достаточно сложны и взаимосвязаны. Кроме того, развитие современной биологии показывает, что наши представления об этих механизмов весьма ограничены. При построении моделей обычно приходится решать проблему — слишком подробная модель имеет много параметров и при небольшом размере обучающей выборки может оказаться «переобученой», а слишком грубая модель имеет небольшую точность. После того, как построена модель и подобраны параметры, ее применяют для предсказаний биологических свойств. При этом обычно поступают так: для последовательности или ее фрагмента с помощью модели вычисляют число — вес. Затем сравнивают этот вес с некоторым порогом. Если вес превышает заданное значение, то принимают решение о наличии биологического свойства. Эту технологию можно представить как отображение пространства последовательностей в пространство весов (с помощью модели), и затем отображение пространства весов в булево пространство (рис.1).

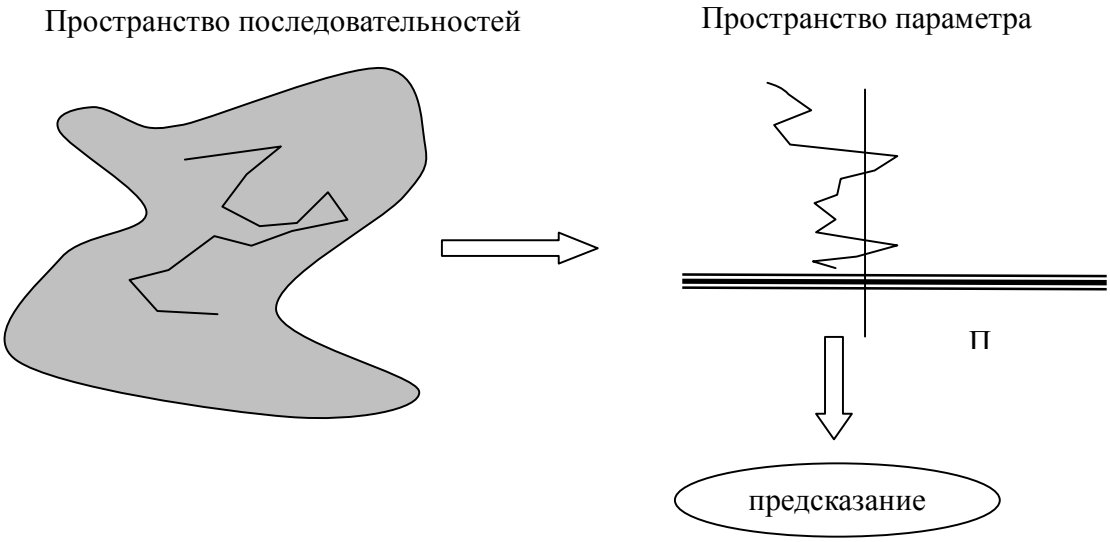


Рисунок 1. Отображение пространства последовательностей на пространство параметра.

Сравнительно-геномный анализ основан на идее давления отбора на важные

биологические свойства. В наивной формулировке это означает, что если мы видим в двух ортологичных (имеющих общего предка) последовательностях некоторое свойство, то наше предсказание получает дополнительную поддержку. Однако, при оценке вероятностей в таком анализе используется предположение о независимости наблюдений. Это предположение оправдано на больших эволюционных расстояниях, но не применимо при малых расстояниях. В предельном случае сравниваемые последовательности идентичны и поэтому наблюдения также идентичны. Ситуация осложняется еще тем, что эволюционные расстояния на дереве — разные и поэтому степень зависимости наблюдений также разная.

Возникает задача — как оценить вероятность наблюдений (чисел, полученных в результате применения той или иной биоинформатической модели) при нулевой гипотезе. В качестве нулевой гипотезы используется нейтральная модель эволюции, либо модель эволюции под давлением отбора на какие-либо другие свойства последовательности, отличные от исследуемого свойства (условно-нейтральная эволюция). Нейтральная, или условно-нейтральная эволюция может рассматриваться как диффузионный процесс в пространстве последовательностей.

При сравнительно-геномном анализе в центре исследования находится как правило некоторое свойство последовательностей, например, GC состав, способность кодировать белок, энергия вторичной структуры РНК и пр. Такое свойство выражается как некоторая функция от последовательности. При наличии наблюдаемых значений функции для ортологичных последовательностей, возникают задачи оценки значимости наблюдений, предсказания функциональных и эволюционных особенностей. Для решения этих задач мы предлагаем использовать модель эволюции функции основанную на диффузионном процессе или уравнении Фоккера-Планка [1].

Биологические процессы нередко в своем описании используют аналогии из статистической физики. Такие понятия как энтропия [2], информация [3], потоки [4] перенесены в эволюционную биологию. Одним из важных явлений как в статистической физике, так и в эволюционной биологии является диффузия. Например, изменение частот аллелей под действием различных сил --- отбора, мутаций, генетического дрейфа, рассматривалось как диффузионный процесс [5]. Такое приближение, строго оформленное

Кимурой [6], легло в основу разработки новой теории нейтральной эволюции [7]. При таком подходе эволюцию можно представлять на многомерной адаптивной поверхности [8]. При другом подходе исследуются возможные эволюционные пути между двумя промежутками времени, в которые частоты аллелей фиксированы. Тогда говорят об ансамбле путей и распределении вероятностей по путям. При этом центральным понятием становится fitness flux [4]. Еще один подход --- изучение количественных характеристик, например, фенотипа [9]. При этом распределение таких величин в популяции часто представляют «волной», двигающейся в сторону оптимального значения [11], а само оптимальное значение может изменяться [11].

Представим свойство как функцию над последовательностью $x: S \to R$ (S --- пространство последовательностей, R --- вещественные числа). Если на свойство не действует отбор, то изменение x описывается диффузионным процессом, а плотность вероятности соответственно уравнением Фоккера-Планка. При этом распределение этих значений по геному (или полученное симуляцией в рамках модели эволюции) порождает силовое поле, в котором происходит диффузия.

Таким образом, имея ряд наблюдений и вероятностную модель эволюции возникают задачи оценки их значимости, предсказания «необычных» событий. Для этого в работе вводится ряд статистик, на основе которых проводится статистический анализ данных.

Нейтральный эволюционный процесс можно представить как свободное блуждание в пространстве последовательностей. Наблюдение, или измеряемая характеристика (кодирующий потенциал, оптимальная энергия сруктуры РНК, сила сайта связывания и пр.) является функцией последовательности $x: S \to R$. Будем предполагать функцию $x(s)$ в некотором смысле непрерывной --- малые изменения последовательности приводят к малым изменениям функции. Тогда случайные эволюционные блуждания в пространстве последовательностей будет отображаться на случайное блуждание в пространстве наблюдений. Если текущее значение x сильно необычно (малое число последовательностей с заданным x), то переход в «более обычное» значение (число последовательностей с новым значением увеличивается) будет более вероятен, чем переход в менее вероятное значение. Таким образом, поведение x можно описать диффузионным процессом:

$$dx = a(x)dt + b(x)dB_t \quad (1)$$

Здесь: x — наблюдаемый параметр, $b(x)$ — коэффициент диффузии, ответственный за скорость эволюции, $a(x)$ — коэффициент сноса, B_t — нормальное распределение. Природа коэффициента сноса связана с тем, что некоторые значения параметра (веса) более вероятны, или достигаются на большом количестве последовательностей, а другие — менее вероятны.

Стационарное (фоновое) распределение $p(x)$ значений x можно получить симуляцией в рамках модели эволюции, это может быть Бернуллиевская или Марковская симуляция, нарезка из генома и т.п. Так как $p(x)$ --- стационарное распределение для x, получаем уравнение связывающее снос и диффузию:

$$-a(x)p(x) + \frac{1}{2}\frac{d(b^2(x)p(x)}{dx} = 0 \quad (2)$$

Уравнения (2) недостаточно для определения функций $a(x)$ и $b(x)$, поэтому они получаются приближенно с помощью симуляций. В дальнейшем будем считать, что $p(x) = N(0,1)$ (если свойство x имеет другое стационарное распределение, то преобразуем её к нормальному распределению).

Существуют подходы, при которых эволюционные изменения количественных характеристик и фенотипов представляются в виде диффузионных процессов [12]. В таких случаях, как правило, модель броуновского движения ($a(x)=0$, $b(x)$=const) применяется для описание нейтральной эволюции, а процесс Орнштейна-Уленбека ($a(x)=\alpha(x-\beta)$, $b(x)$=const) для учета как отбора, так и дрейфа [13]. Процесс Орнштейна-Уленбека может представлять хорошее приближение к реальности [14], при этом оставаясь математически доступным для анализа. Произвольный диффузионный процесс (1), как правило, не решаем аналитически, поэтому мы используем приближение процесса Орнштейна-Уленбека. В то же время функции $a(x)$ и $b(x)$ вычислены для трех биологических примеров и приведены дальше. При сравнительно-геномном подходе рассматривают группу последовательностей, эволюционировавших от общего предка. В этом случае необходимо учитывать их расположение на филогенетическом дереве.

Уравнения на деревьях

Пусть последовательности $s_1,...,s_n$ эволюционировали от общего предка s. Обозначим $x_i=x(s_i)$, $y=x(s)$. Рассмотрим простейший случай, когда они эволюционировали независимо (рис.2).

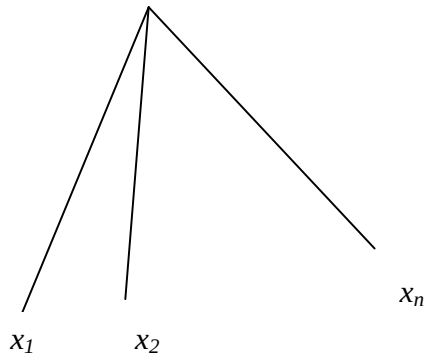


Рисунок 2. Независимая эволюция последовательностей.

Обозначим $\rho(x_1,...,x_n \mid y)$ --- вероятность появления значений x_i, $i=1...n$, при условии, что предковое значение было y. Вероятности $\rho(x_1,...,x_n \mid y)$ получаются из уравнения (1).

$$\rho(x1,...,xn \mid y) = \prod_i \rho(xi \mid y) \quad (3)$$

отсюда

$$\rho(x1,...,xn) = \int_{-\infty}^{+\infty} \rho(x1,...,xn \mid y)\, p(y)\, dy \quad (4)$$

Более сложным является простейшее ветвящееся дерево (рис.3).

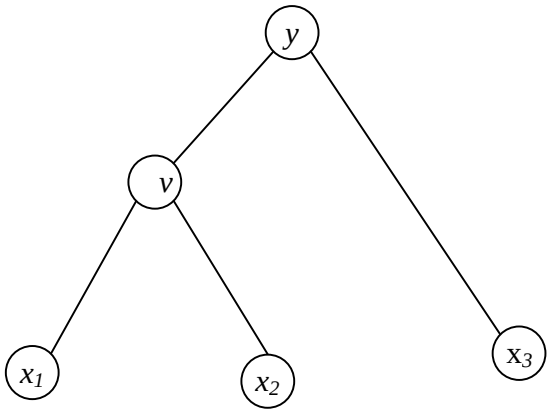


Рисунок 3. Дерево с тремя листьями.

Для этого случая плотность вероятностей примет вид:

$$\rho(x1, x2, x3) = \rho(x1, x2 \mid y) \cdot \rho(x3 \mid y) = \\ = \rho(x3 \mid y) \cdot \int_{-\infty}^{+\infty} \rho(x1, x2 \mid v)\, p(v \mid y) dy \quad (4)$$

отсюда:

$$\rho(x1, x2, x3) = \\ = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho(x1 \mid v)\rho(x2 \mid v)\, p(v \mid y)\rho(x3 \mid y)\, p(y) dy \quad (5)$$

Для общего вида дерева единую формулу для $\rho(x_1,...,x_n \mid y)$ выписать не удастся, однако принцип её получения ясен и возможно рекурсивное вычисление.

Таким образом, для наблюдаемых на деревьях получены формулы, отражающие вероятности этих событий.

Биологические примеры

Представим анализ диффузионного процесса для трех биологических примеров. Особый интерес представляет вид функций сноса *a(x)* и диффузии *b(x)*. Это позволит качественно понять насколько приближение процессом Орнштейна-Уленбека (*a(x)*---

линейная, *b(x)*---постоянная) адекватно. Для каждого примера продемонстрированы графики функции сноса *a(x)* и диффузии *b(x)* (см. (1)).

Результаты получены симуляцией в рамках бернулиевской модели эволюции последовательностей (замена равновероятна для каждой позиции последовательности и равновероятна на каждый нуклеотид). Все вычисления проводятся не для самой функции *x(s)*, а для такого ее преобразования, что стационарное распределение становится стандартным нормальным. Функций *a(x)* и *b(x)* нормированы так, что *b(0)=1*.

Первый и простейший пример описывает эволюцию нуклеотидного состава. Отклонение нуклеотидного состава от равномерного опишем как Евклидово расстояние υ (рис.4):

$$\nu(s) = \sqrt{(\pi_a - 0.25)^2 + (\pi_c - 0.25)^2 + (\pi_g - 0.25)^2 + (\pi_t - 0.25)^2} \quad (6)$$

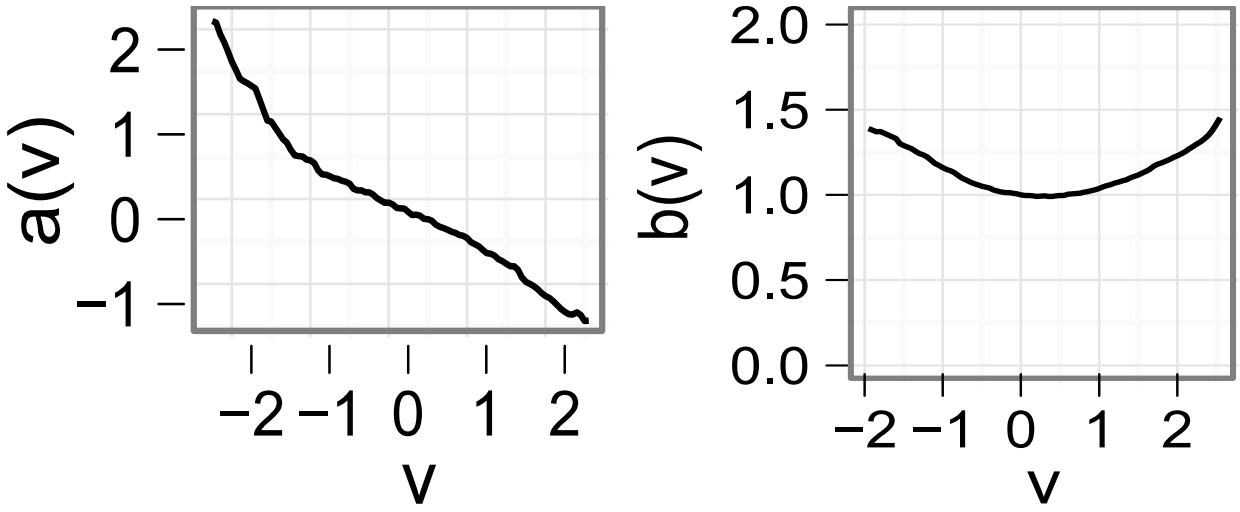


Рисунок 4. Зависимость коэффициентов a(x) и b(x) для модели частот нуклеотидов.

Второй пример (рис.5) функции от последовательности --- минимальная энергия *e(s)* вторичной структуры РНК (для вычисления используется программа RNAfold [15]).

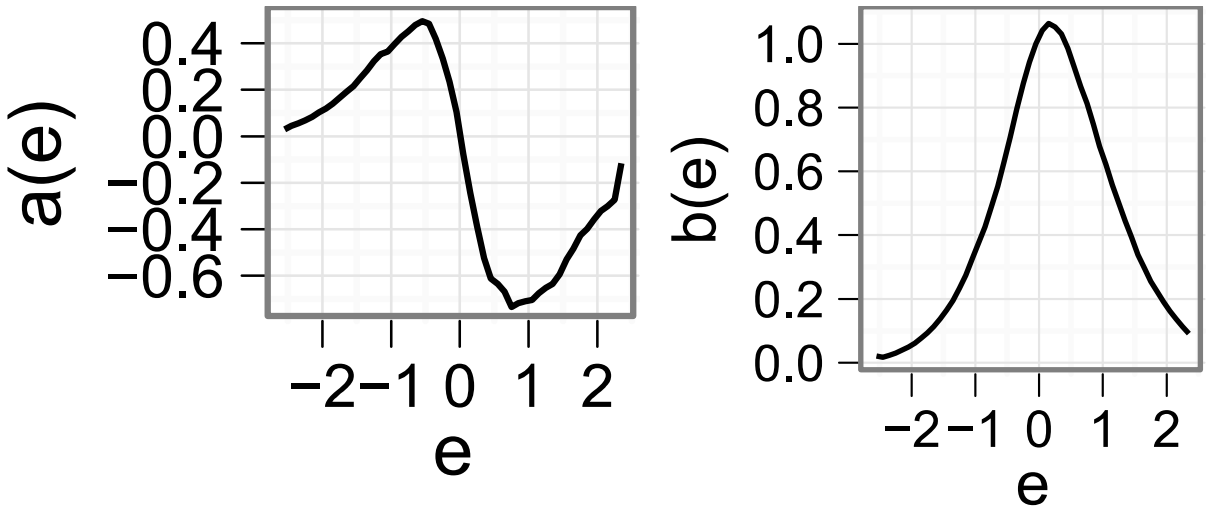


Рисунок 5. Зависимость коэффициентов a(x) и b(x) энергии вторичной структуры РНК.

Третий пример (рис.6) связан с поиском сайтов связывания транскрипционных факторов с помощью позиционной весовой матрицы (PWM). Функция от последовательности определяется как максимальный вес PWM:

$$w(s) = \max_i w_i \qquad w_i = \sum_{j=i}^{i+n} w(j, s_i) \tag{7}$$

где $w(j, s_i)$ позиционная весовая матрица размера $n \times 4$.

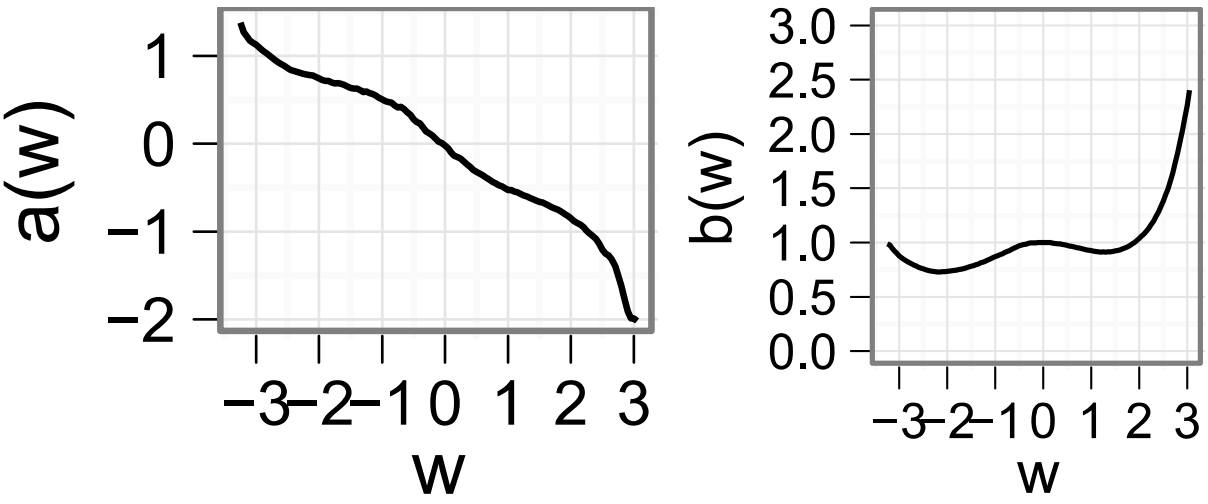


Рисунок 6. Зависимость коэффициентов a(x) и b(x) для модели максимального веса

весовой матрицы поиска сайтов связывания.

Отметим некоторые важные свойства функций *a(x)* и *b(x)*, которые можно видеть на рисунках:

- Функция *a(x)* похожа на линейную По-крайней мере в некоторой достаточно большой окрестности нуля
- Поведение коэффициента диффузии b(x) имеет достаточно ограниченную вариацию для функции частот и силы сайта связывания По-крайней мере в промежутке значений от -2 до 2
- Отношение *a(x)/b(x)* имеет хорошее приближение к линейной функции для всех примеров (рис. 7).

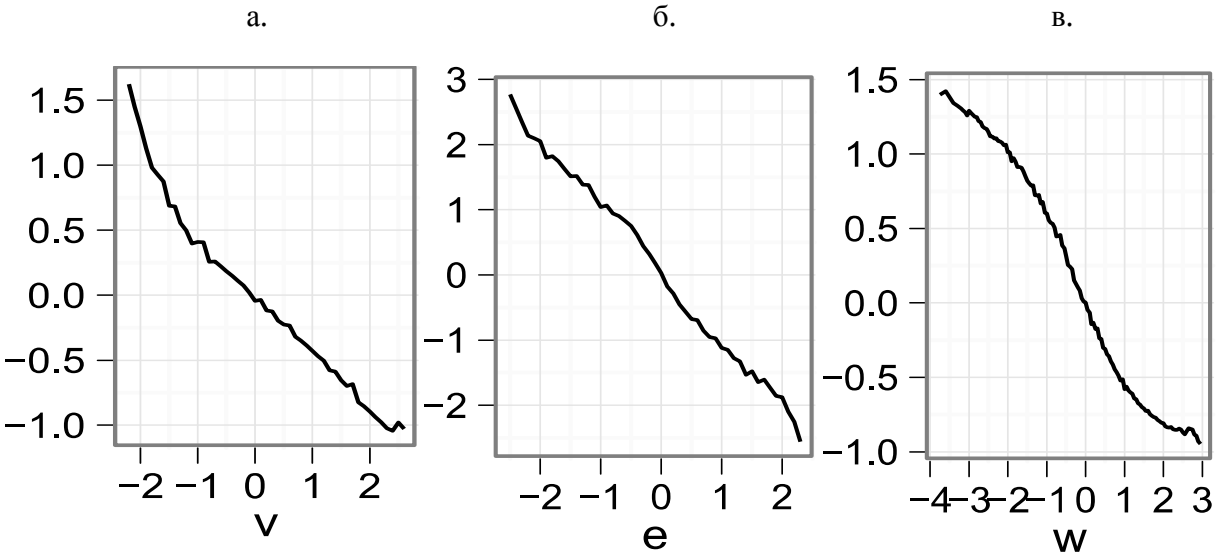


Рисунок 7. Зависимость отношения *a(x)/b(x)* для модели частот нуклеотидов (а), энергии вторичной структуры РНК (б), максимального значения силы сайта связывания (в).

Статистический анализ

Будем считать, что функция *x(t)* подчиняется процессу Орнштейна-Уленбека. В этом

случае плотность $\rho(x,t|y)$ имеет нормальное распределение. Предположим, нам известны наблюдения x1,...,xn и филогенетическое дерево. В этом случае плотность $\rho(x1,\ldots,xn)$ вычисляется и имеет многомерное нормальное распределение:

$$\rho(x1,...xn) = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \bar{x}^T A \bar{x} \right) \quad (8)$$

где $\bar{x} = (x1,..., xn)$, A --- матрица параметров многомерного нормального распределения. Параметры распределения вычисляются рекурсивным алгоритмом.

Для предсказания функциональных особенностей геномов и эволюционной специфичности на основе модели введем две статистики.

Статистики

Для оценки того, насколько наблюдения (значения x) на реальных последовательностях отличаются от ожидаемых значений при нейтральной модели эволюции мы вводим две статистики. Первая статистика характеризует степень отклонения от ожидаемых значений:

$$r^2 = \bar{x}^T A \bar{x} \quad (9)$$

Отметим, что $r^2 \approx \chi^2(n-1)$. Чем больше r^2, тем более значимы наблюдения.

Нам известна априорная плотность вероятности в предке $p(y)$. Также можно вычислить апостериорную плотность:

$$p(y \mid x1,..., xn) = \frac{\rho(x1,..., xn)\, p(y)}{\rho(x1,...xn)} \quad (10)$$

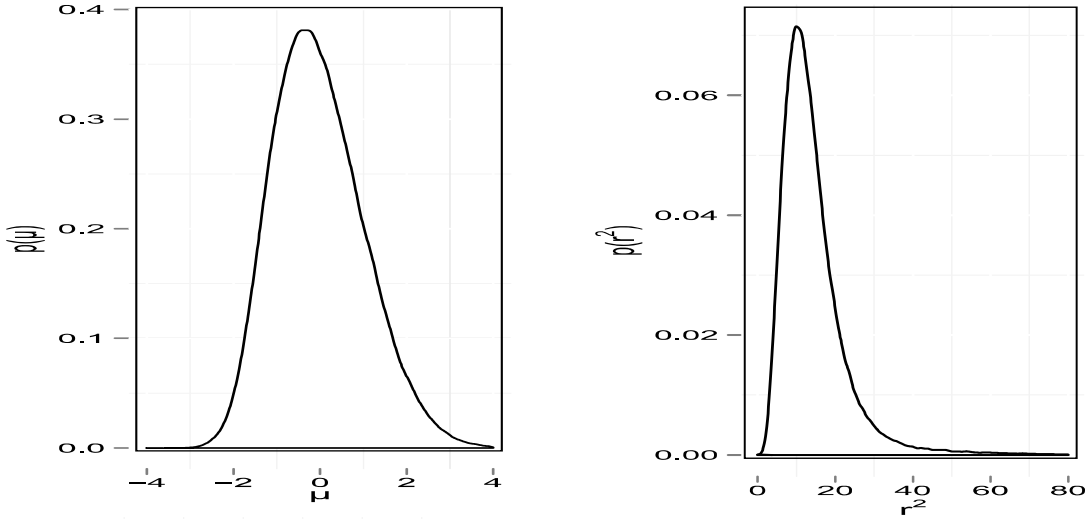
Можно построить статистику, основанную на апостериорном распределении значений в предке. Апостериорная плотность нормально распределена. Обозначим её параметры (μ,σ), они зависят от структуры дерева, длин веток и значений наблюдаемых. Вторая статистика --- среднее μ, которое можно интерпретировать как влияние «сноса» наблюдаемых на предковую последовательность:

$$\mu = \sum a_i x_i \quad (11)$$

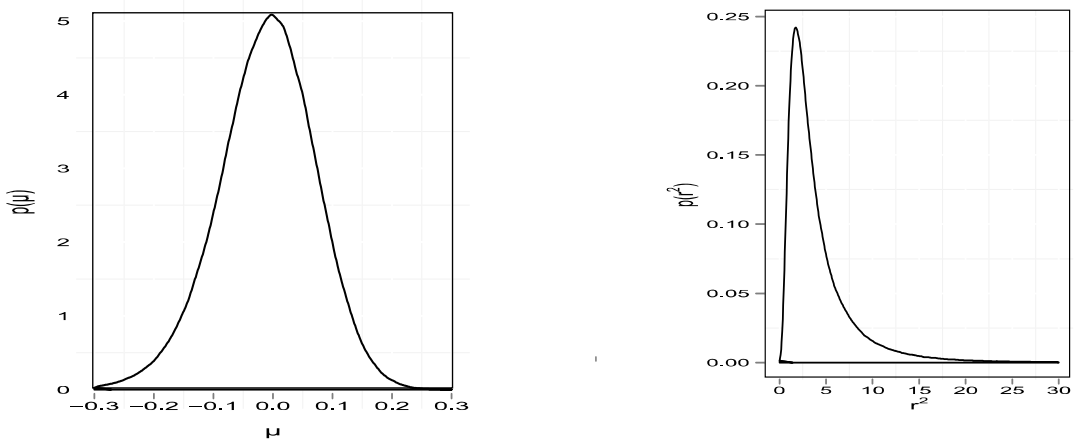
где коэффициенты *ai* алгоритмически вычислимы. Отметим, что эта статистика имеет нормальное распределение со средним 0.

При этом, так как эволюция реальных биологических функций не удовлетворяет процессу Орнштейна-Уленбека, статистики перестают описываться известными распределениями. Распределения статистик для рассмотренных фукнций показаны на рис. 8.

а)



б)



в)

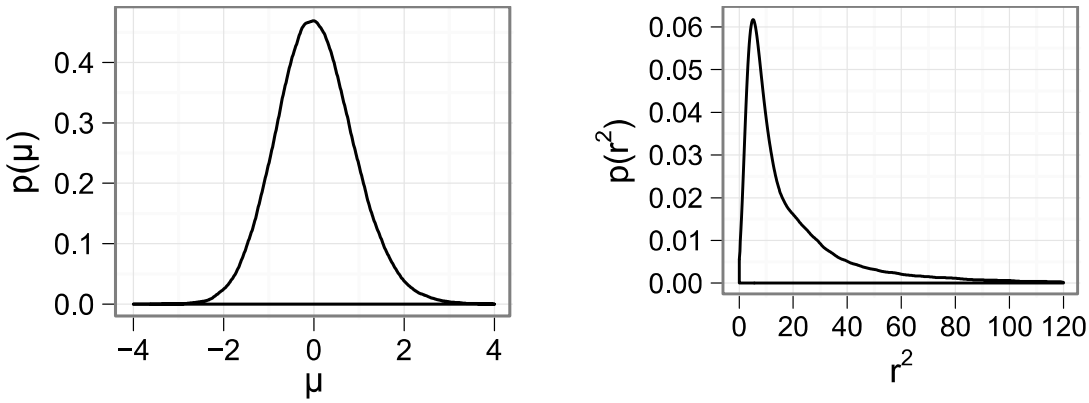


Рисунок 8. Распределение статистик для частот нуклеотидов (а), энергии вторичной структуры (б), позиционной весовой матрицы (в)

Применение статистик

Примем нулевую гипотезу --- отбор не действует на функцию *x*. Зададим пороги значимости r^2 и μ (например 5\%-ые квантили). Если статистика $r^2(x1,...xn) \geq r_0^2$, то наблюдения считаются значимыми. Если при этом $\mu(x1,...,xn) \geq \mu_0$, то нулевая гипотеза отвергается. Считается, что действует отбор и соответствующие наблюдения функциональны. Если же $\mu(x1,...,xn) < \mu_0$, то гипотеза не отвергается. Причины появления значимого значения $r^2(x1,...xn)$ требуют дальнейшего исследования (вероятные причины --- отбору подверглась только часть наблюдений, данные или дерево не соответствует действительности). При этом статистики r^2 и μ связаны: если одна статистика «значима», то очень вероятно, что и вторая статистика «значима». Пример их связи для энергии вторичной структуры РНК на рис.9

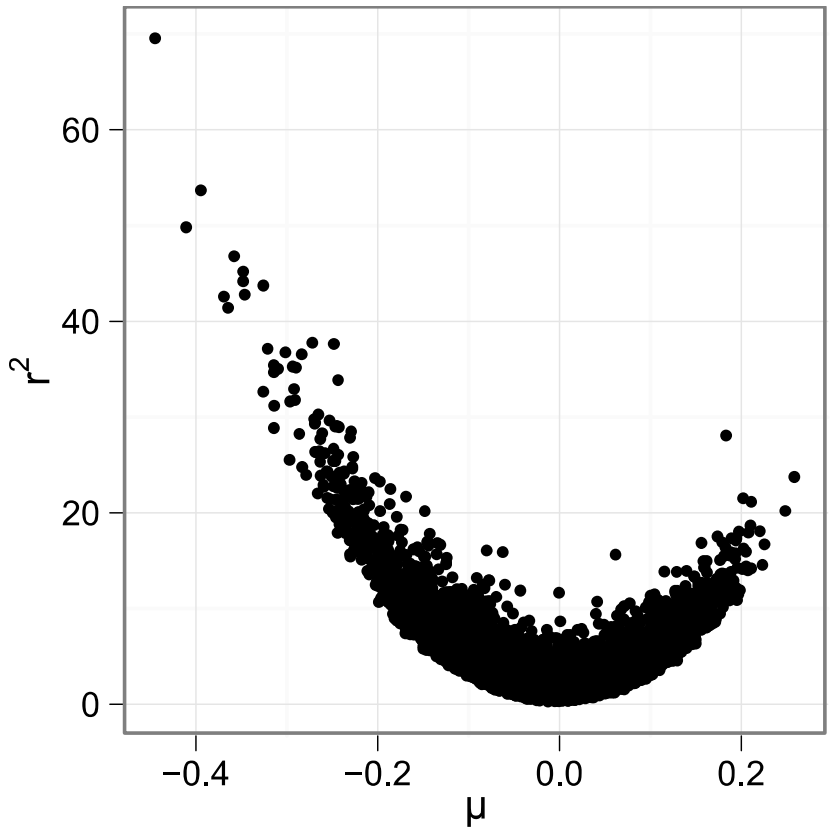


Рисунок 9. Зависимость статистик r^2 и μ для случая энергии вторичной структуры.

Компьютерный эксперимент

В основе статистического анализа лежит нулевая гипотеза: отбор не действует на биологическое свойство (функцию). Таким образом, статистики должны давать значимые значения в случае, если отбор есть. Проведем компьютерный эксперимент:

Рассмотрим функцию $x(s) = \nu(s)$ --- отклонение частот встречаемости нуклеотидов от равномерных (6). Через $Q(\pi_a, \pi_c, \pi_g, \pi_t)$ обозначим бернулиевскую модель симуляции последовательностей со стационарными частотами встречаемости $(\pi_a, \pi_c, \pi_g, \pi_t)$. Пусть есть отбор на частоты встречаемости, это реализуется в изменениях частот π. Все параметры для функции *x(s)* вычисляем в рамках модели *Q*(0.25,0.25,0.25,0.25), а отбор моделируем в виде симуляции в рамках модели $Q(0.25+ε,0.25+ ε,0.25- ε,0.25- ε), где *ε*>0. Ожидаем, что чем больше брать значения *ε*, тем сильней действует отбор и тем более «значимы» значения

статистик («значимость» статистик измеряется величиной p-value). Результаты продемонстрированы на Рис. 10 и согласуются с ожиданиями, показывая, что статистики «ловят» эффект отбора.

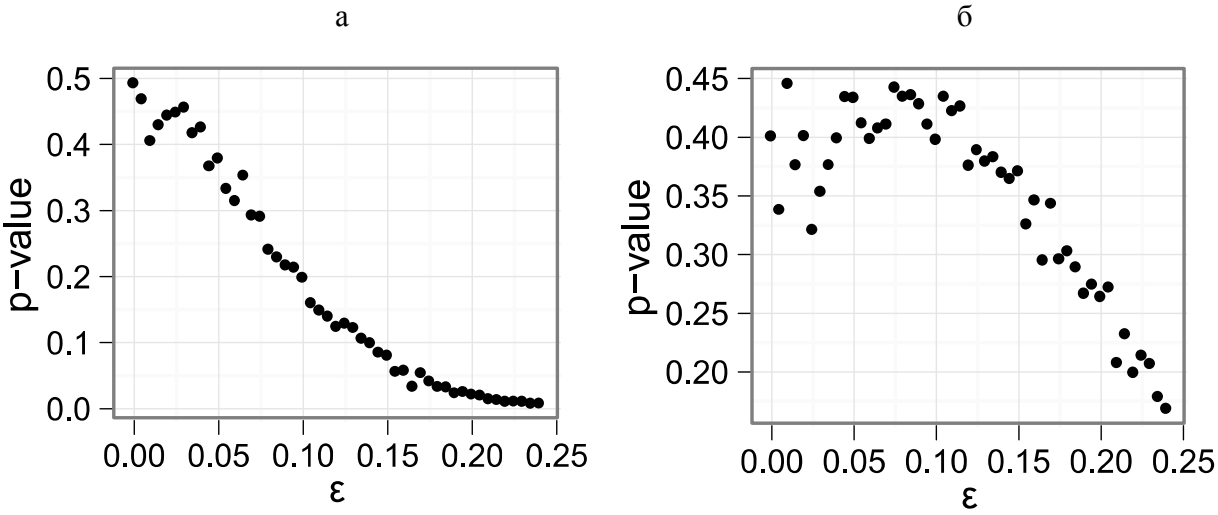


Рисунок 10. Зависимость p-значения (p-value) от ε. а) – статистика r^2 ; б) – статистика μ

Выводы

Диффузия является хорошим приближением для описания эволюции непрерывных характеристик последовательностей при отсутствии отбора. Дальнейшее приближение процессом Орнштейна-Уленбека не точно отражает эволюцию биологических функций --- что видно из сравнения функций сноса *a(x)* и диффузии *b(x)*. Тем не менее, данный подход является мощным инструментом для предсказания функциональной специфичности на основе введенных статистик, как продемонстрировал компьютерный эксперимент.

2.2.2 Скрытые Марковские модели для исследования эволюции и предсказания функций в условиях не полной определенности

Разработанный метод оценки статистической значимости наблюдений с точки зрения эволюционной теории показывает только наличие давления отбора. При этом он не дает

информации о биологическом значении сделанных наблюдений. Кроме того, описанный подход предполагает, что все ветви дерева находятся под одинаковым давлением стабилизирующего отбора. Для того, чтобы выделить клады, отвечающие разным давлениям отбора необходимо провести расчеты для каждого узла. Рекурсивная процедура вычисления апостериорной вероятности (наподобие представленной в формуле (4)) производит такие вычисления. Однако по самой постановке задачи мы можем только определить наличие давления отбора на соответствующий параметр. Для того, чтобы картировать на дереве эволюционные события, необходимо расширить модель.

Введем следующие понятия. *Наблюдением* будем называть вычисленные числа на листьях дерева. *Биологическим состоянием* или просто *состоянием* будем называть реальную биологическую природу. Например, веса сайтов связывания у ортологичных генов в существующих геномах являются *наблюдениями*, а сам факт регуляции – биологическим *состоянием*. Обычно целью биоинформацоинного анализа является предсказание именно биологического состояния, основываясь на сделанных наблюдениях. При этом интерес представляют не только состояния у существующих геномов, но также эволюционный сценарий, т.е. биологические состояния у предков.

Существующие подходы [16-20], как правило, используют определенные состояния на листьях. При этом рассматривается задача о предсказании состояний в промежуточных узлах.

Для решения этих задач в условиях неопределенности состояний на листьях предложен подход, основанный на скрытых Марковских моделях (см., напр.[21]). В формализме скрытых Марковских моделей мы будем считать биологические состояния скрытыми состояниями, а наблюдения – эмиссиями. Для того, чтобы полностью описать скрытую Марковскую модель необходимо определить эмиссионные и переходные вероятности, т.е. вероятности порождения наблюдаемых значений и вероятности переходов между состояниями. Переходные вероятности мы определим из модели пуассоновского процесса, когда вероятность смены состояний экспоненциально зависит от длины веток. Для случая двух состояний в соответствии с [22] переходные вероятности будут:

$$P(t) = p^{ij}(t) = \begin{pmatrix} \lambda_0 + \lambda_1 e^{-\mu t} & \lambda_1 - \lambda_0 e^{-\mu t} \\ \lambda_0 - \lambda_1 e^{-\mu t} & \lambda_1 + \lambda_0 e^{-\mu t} \end{pmatrix} \quad (12)$$

$$\mu = \alpha + \beta; \quad \lambda_0 = \frac{\alpha}{\alpha + \beta}; \quad \lambda_1 = \frac{\beta}{\alpha + \beta}$$

Где α и β – скорости переходов из состояния 1 в состояние 2 и наоборот, t – эволюционное время, пропорциональное длинам веток.

На рис. 11 представлено сравнение подходов – стандартного, когда на листьях определены состояния, и нашего, когда на листьях определены только некоторые значения, как-то связанные с реальными биологическими состояниями.

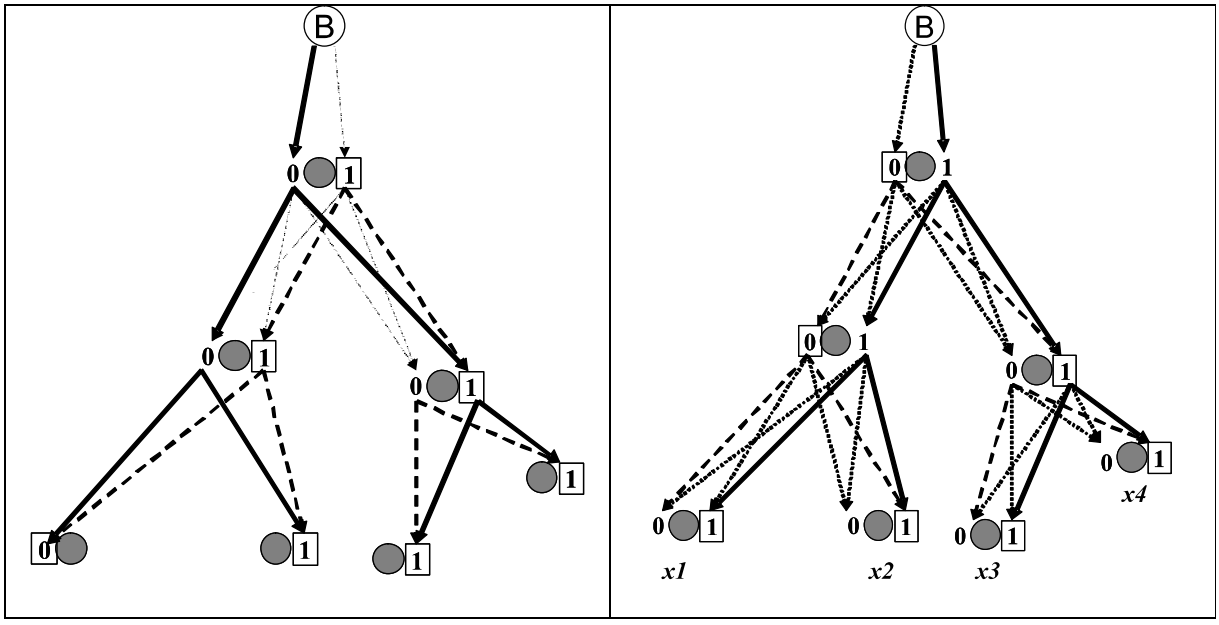


Рисунок 11. Модели для реконструкции предковых состояний. Сплошные линии соответствуют переходам для оптимальной реконструкции; пунктирные линии линии для оптимизации на узлах; точечные линии неоптимальных пути. Прямоугольники обозначают предсказанные состояния; B --- начальное состояние. Слева базовая модель с определенными состояниями на листьях; Справа HMM модель. Предсказание, основанное на пороге, определяет состояние 0 на левом листе, в то время как HMM- анализ исправляет прогноз на

этом листе и изменяет реконструкцию на других узлах.

HMM подход зависит от следующих параметров. а) распределение априорных вероятностей; б) Параметры переходных вероятностей в уравнении (12). Первый параметр отражает наши априорные представления о биологической проблеме у будут обсуждаться позже. Переходные вероятности можно оценить с помощью обычной процедуры максимизации правдоподобия (ML). Следуя [18] мы используем Байесову технику для апостериорных распределений вероятностейс помощью Монте-Карло моделирования Марковских цепей (MCMC, Markov Chain Monte Carlo). Алгоритм выбора параметров непосредственно связан с алгоритмами предсказания состояний и апостериорных вероятностей и будет обсуждаться ниже.

Алгоритмы восстановления состояний и вычисления апостериорных вероятностей аналогичны соответствующим алгоритмам в теории линейных скрытых Марковских моделей.

Алгоритм Витерби для скрытых Марковских моделей на деревьях.

Когда состояния на всех узлах определены, вероятность наблюдаемых значений может быть рассчитана как:

$$P(x1,\ldots,xn \mid S1,\ldots Sr) = \prod_{j\in\{1,\ldots,2n-1\};Yk=parent(Yj)} P(Sk \to Sj) \bullet \prod_{i\in\{1,..,n\}} P(xi \mid Si) \qquad (13)$$

Здесь xk – наблюдения на листьях, Si – состояния на узлах, P(Si→Sj) – переходные вероятности в соответствии с формулами (11), p(x_k | S_k) --- эмиссионная вероятность для xk быть в состоянии Sk. Полная вероятность наблюдать данные есть сумма вероятностей по всем возможным наборам состояний на листьях:

$$P(x1,\ldots,xn) = \sum_{S1,\ldots,Sr} P(x1,\ldots,xn \mid S1,\ldots,Sr) \qquad (14)$$

В соответствии с общим подходом максимального правдоподобия (ML), наиболее вероятный набор состояний обеспечивает максимальную вероятность наблюдать данные. Для нашей вероятностной модели функция задается в уравнении (14) и ML может быть оценена с помощью модификации алгоритма Витерби. Витерби V_Y^i переменные в каждом узле Y и состоянии i соответствуют максимальной вероятности данных на листьях в поддереве узла начинающегося с Y в состояние i. Вероятности перехода от начальной точки B равны

априорным вероятностям состояний ω_i. Рекурсии Витерби для этого случая может быть записана следующим образом:

$$V_Y^i = \begin{cases} \max_j \left(p^{ij}(t_M) \cdot V_M^j\right) \max_k \left(p^{ik}(t_L) \cdot V_L^k\right), & M, L = child(Y) \\ V_Y^i = \rho^i(x_Y), \qquad child(Y) = 0 \end{cases} \tag{15}$$

Оптимальные состояния восстанавливаются с помощью рекурсии

$$\pi_Y^{iR} = \arg\max V_R^j p^{ij}(t_R), \ \pi_Y^{iL} = \arg\max V_L^j p^{ij}(t_L), \ \pi_B = \arg\max V_{root}^j \omega_i \tag{16}$$

Рекурсия начинается на листьях и распространяется к корню. Когда рекурсия закончится, обратным проходом с использованием переменных π восстанавливаются состояния на узлах, в том числе на листьях. Восстановление состояний на листьях на самом деле является предсказанием состояния на листьях.

Апостериорное декодирование для скрытых Марковских моделей на деревьях.

Аналогично алгоритму апостериорного декодирования (forward-backward) в линейных Марковских моделях, можно построить алгоритм апостериорного декодирования для деревьев --- алгоритм Up-down.

Вероятность состояния i в узле Y есть отношение полной вероятности наблюдений при фиксированном состоянии в узле к полной вероятности наблюдений.

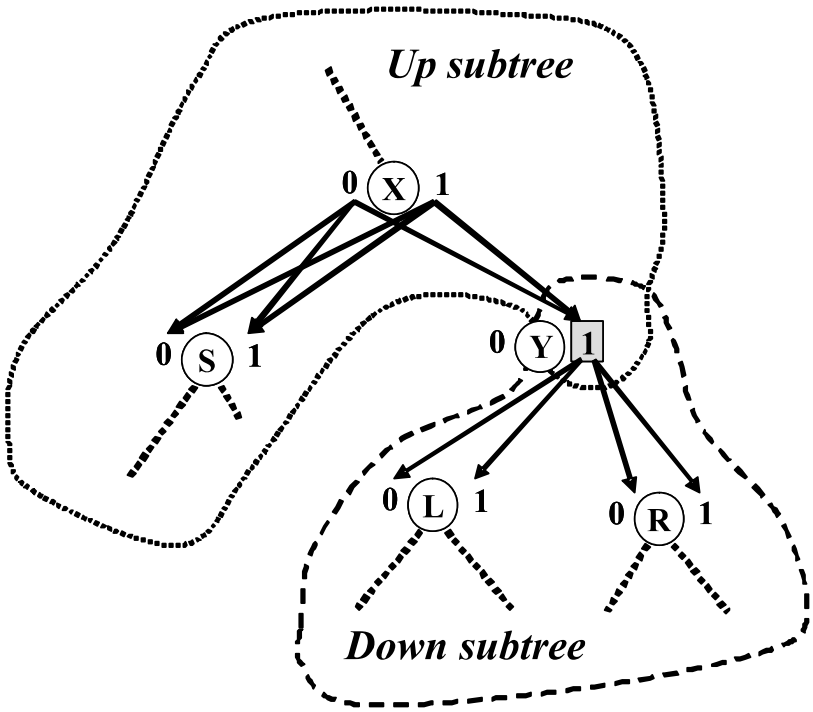


Рисунок 12. Алгоритм Up-Down. Разбиение дерева относительно состояния 1 в вершине *Y* показано пунктирными линиями.

Полная вероятность состояния *i* в узле *Y* может быть записана в виде произведения двух сомножителей --- вероятности поддерева, где вершина *Y* является корнем и состояние в этой вершине *i*, и вероятности поддерева где *Y* является листом и состояние в этом узле *i*:

$$P_Y^i = \frac{\sum_S P(x1,...,xn \mid S1,...,i,...,Sr)}{P(tree)} = \frac{P(upTree_Y, i) \cdot P(downTree_Y \mid i)}{P(tree)} = \\ = \frac{u_Y^i \cdot d_Y^i}{\sum_i d_{root}^i \omega_i} \qquad (17)$$

Здесь u_Y^i --- **up** - переменная — вероятность наблюдения {x} в верхнем поддереве относительно вершины *Y* в качестве листа и состояния в этой вершине *i*; d_Y^i --- **down** - переменная — вероятность наблюдения {x} в нижнем поддереве с вершиной *Y* в качестве корня при состоянии *i*. Верхняя и нижняя переменные вычисляются рекурсивно:

$$\begin{cases} d_Y^i = \sum_j \left(p^{ij}(t_R) \cdot d_R^j \right) \cdot \sum_k \left(p^{ik}(t_L) \cdot d_L^k \right), & R, L = child(Y) \\ d_Y^i = \rho(x_Y), & child(Y) = 0 \end{cases}$$
$$\begin{cases} u_Y^i = \sum_j \left(p^{ij}(t_Y) \cdot u_X^j \right) \cdot \sum_k \left(p^{ik}(t_S) \cdot d_S^k \right), & X = parent(Y), S = sister(Y) \\ u_{root}^i = \omega_i \end{cases} \quad (18)$$

Здесь *Y* --- текущая вершина; *L* и *R* --- левая и правая дочерние вершины; *X* --- родительская вершина; *S* --- сестринская вершина. Нижние переменные *d* вычисляются снизу вверх от листов к корню; Верхние переменные *u* вычисляются сверху вниз от корня к листьям. Полная вероятность наблюдений есть:

$$P(tree) = d_B \quad (19)$$

Апостериорное декодирование позволяет не только предсказать вероятность состояний на узлах, но и оценить вероятности переходов между состояниями, то есть вероятность эволюционных событий. Апостериорная вероятность событий X→ Y на ветке Эволюционного дерева может быть рассчитана по формуле:

$$P(X^i \to Y^j) = \frac{u_X^i d_Y^j}{P(tree)} \quad (20)$$

Оценка параметров модели

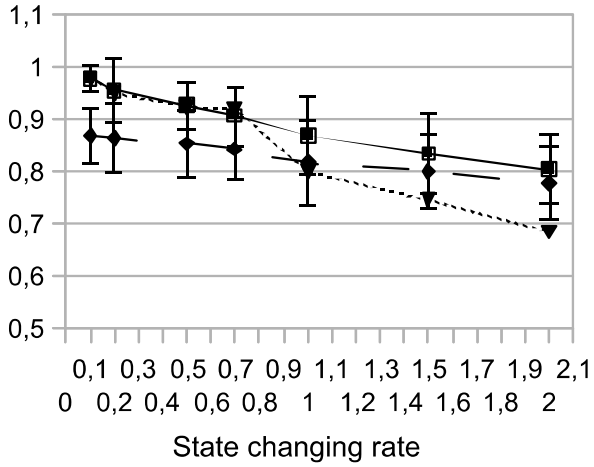
Оценка параметров модели (скоростей переходов) производится с помощью алгоритма Митрополиса-Гастингса с целью максимизировать полную вероятность дерева. В наших модельных вычислениях использованы следующие параметры МСМС: Длина Марковской цепи =10^6, количество различных стартовых значений = 5000, Шаг выборки = 50. Для получения окончательных Up-Down результатов производилось усреднение по всем полученным наборам параметров. Параметры же переходов α и β являются нашими априорными представлениями об относительной частоте событий.

Монте-Карло моделирование

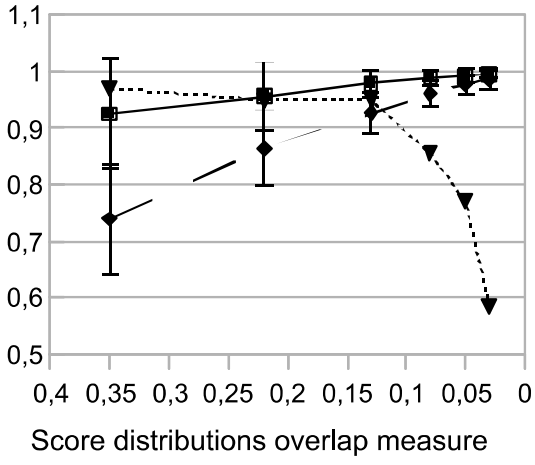
Разница между pHMM и базовой моделью очевидна: в то время как базовая модель не позволяет реконструировать меньше событий, чем парсимония, pHMM обеспечивает

возможность коррекции предсказаний на листьях. Эта особенность представляется важной в случае, когда программа предсказания дает большую погрешность, и веса принадлежат серой зоне. Действительно, при значениях веса вблизи порога решений, дискретный прогноз состояний, которое используется в основном алгоритме реконструкции, выдает почти случайное предсказание, что может привести к переоценке количества событий, скоростей переходов, и, наконец, неправильное назначение состояний. Для тестирования способности pHMM к реконструкции правильных состояний были проведены случайные тесты. Генерировались случайные деревья, а на узлах генерировались состояния в соответствии с формулой (11). Наблюдаемые значения на листьях (веса) выбирались из априорных распределений весов для состояний (задавались как два бета-распределения). Далее применялась процедура декодирования, и предсказывались состояния. Известные и предсказанные состояния сравнивались. Точность предсказания определялась как доля правильно предсказанных реконструкций. На рис. 13 показаны результаты тестирования.

а



б



в

г

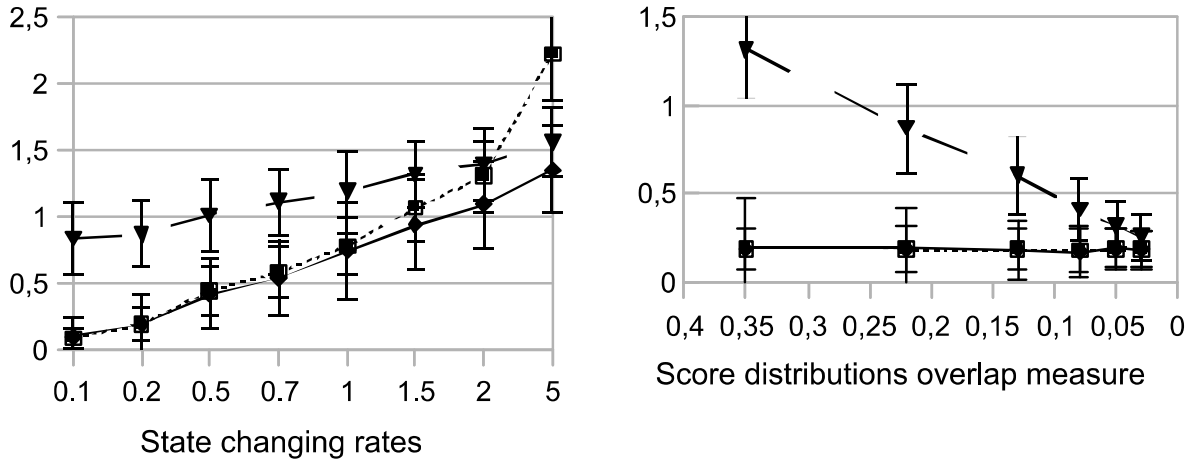


Рисунок.13 Сравнение базовой модели и pHMM. точечные линии показывают долю числа случаев, когда наш метод работает лучше. а) Точность предсказаний при заданных распределениях весов с перекрытием = 0.22 при разных скоростях эволюции. б) Точность предсказаний при фиксированной скорости эволюции = 0.2 и различном уровне перекрытия распределений весов. в) Среднее количество реконструированных событий (по отношению к размеру дерева) при фиксированном распределении весов с перекрытием распределений = 0.22 и различных скоростях эволюции. г) Среднее количество реконструированных событий (по отношению к размеру дерева) рпи фиксированной скорости эволюции = 0.2 и различном уровне перекрытия распределений весов. На рис а) и б) Сплошные линии соответствуют точности нашего метода, пунктирные линии --- точности базового метода, на рис в) и г) Сплошные линии изображают реальное число событий, пунктирная линия событий число реконструированных основной процедурой реконструкции, пунктир --- нашим методом.

Применение к реальным данным

Для случая сигнальных пептидов мы использовали простую модель из двух биологических состояний: состояние N --- отсутствие сигнального пептида и состояние SP --- наличие сигнального пептида. Наблюденные значения --- Dscores рассчитанные программой SingalP3.0-NN. Распределение Dscore на всем наборе данных был представлен как смесь двух бета распределений. Одно из распределений соответствовало состоянию SP, другое ---

состоянию N (non-signal). Результаты pHMM и базового метода (когда состояния на листьях определялись независимо от эволюционного дерева, основываясь только на результатах программы SingalP) примененного к кластеру кластера амидазы (PRK07056). Базовый метод демонстрирует, 6 эволюционных событий --- 5 приобретений и 1 потерю сигнального пептида в то время как pHMM показывает только одно событие --- потерю. Причиной переоценки числа событий для основного подхода связана с тем, что многие листья принадлежит к серой зоне. pHMM также корректирует предсказания на 7 листьях. Предковые состояния в реконструкциях отличается узле Burkholderia. Если использовать предустановленный порог программы SignalP, то количество событий будет даже больше --- 8 (5 приобретений и 3 потери). Таким образом, можно сделать вывод, что, хотя существует возможность потерять какие-то недавние события, предложенный метод может быть использован для выбора наиболее значимых событий и подавить шум программы предсказания, что приведет к большей надежности предсказаний. Сравнение апостериорной и априорной вероятностей состояний на терминальных узлах показывает, что в 83 % случаев pHMM усилил первоначальное предсказание, в 14 % ослабил и в 3 % случаев изменил предсказание.

Заключение

Обобщая представленные данные, а также множество других проведенных тестов, можно с уверенностью сказать, что предложенные алгоритмы дают существенное улучшение предсказаний, заметно сужая серую зону. Узким местом в представленном подходе является задание априорных распределений весов на последовательностях, имеющих и не имеющих заданное биологическое свойство. Для решения этой проблемы возможны различные подходы, в зависимости от характера задачи. Один из таких подходов заключается в следующем. Мы предполагаем виды распределений (например, считаем их нормальными, либо принадлежащими семейству бете-распределений). Кроме того, из характера задачи мы, как правило, имеем априорное представление об относительной представленности в исследуемой выборке последовательностей, имеющих искомое свойство и без оного. На основе таких знаний мы можем использовать известные методы разделения распределений. После такого разделения мы получим параметры искомых распределений.

2.2.2.1 Предсказание регуляторных модулей

Задача изучения регуляции экспрессии генов является одной из важнейших зедач современной молекулярной биологии. Как биоинформатическая задача изучения экспрессии генов возникла практически с самого возникновения биоинформатики. Тогда был разработан ряд подходов, которые и сейчас лежат в основе компьютерного анализа экспрессии генов. Ярким примером таких подходов является техника позиционных весовых матриц, которая до сих пор используется в широком круге исследований в качестве составной части.

В настоящее время накоплено множество данных по экспрессии генов в различных организмах. Есть несколько источников такого рода данных. Во-первых, данные по экспрессии, полученные с помощью микрочипов. Технология микрочипов возникла в середине девяностых годов и получила большое развитие в начале нашего века. Во-вторых, это данные по секвенированию РНК с помощью технологии секвенирования нового поколения. Экспрессия генов определяется целым рядом факторов. Одним из ведущих факторов является связывания белков (так называемых факторов транскрипции) со специфическими участками ДНК — сайтами связывания. Эти сайты связывания также могут быть определены экспериментально с помощью методов иммунопрецепитации хроматина [23]. Однако все получаемые экспериментальные данные относятся к клеточным линиям, что достаточно далеко от реальных условий, в которых работает клетка, либо к нативным клеткам в специфических условиях.

Пространство условий, в которых функционирует реальная клетка весьма широко и не может быть покрыто (по крайней мере, пока) экспериментально. Поэтому задача биоинформатического предсказания экспрессии генов по-прежнему остается актуальной. По-видимому, именно регуляция экспрессии генов в большой степени определяет индивидуальные особенности особей. Структура регуляторных областей у бактерий устроена достаточно просто. Каждый ген регулируется одним-двумя (редко больше) факторами транскрипции. Регуляторные области бактерий достаточно короткие. У грибов, в частности дрожжей, система регуляции несколько сложнее, но по-прежнему достаточно просто устроена. Другое дело – многоклеточные организмы. Они имеют весьма сложную и

разветвленную систему регуляции экспрессии, определяющие тканевую специфичность, морфологию организма, взаимодействие клеток. В частности многие раковые заболевания обусловлены именно нарушениями в регуляции экспрессии генов. В частности, выход из-под контроля генов, важных для эмбрионального развития приводят к раковому перерождению клеток. В эукариотах многие гены регулируются десятком и более факторами транскрипции. При этом распознавание реальных сайтов связывания ТФ затруднено еще и тем, что эти сайты связывания достаточно вырождены. Поэтому простейшие традиционные биоинформационные подходы здесь работают достаточно плохо. К счастью, сайты связывания ТФ не работают независимо, а образуют кластеры, которые обеспечивают кооперативный эффект. Такие кластеры называют регуляторными модулями. Поиску кластеров сайтов связывания посвящено множество работ [24-27]. Регуляторные модули могут располагаться достаточно далеко от регулируемых генов.

Кооперативность сайтов связывания накладывает определенные ограничения на структуру регуляторных модулей. 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 [28-31], 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 делать предположения о совместной работе ТФ.

Можно предположить, что если структура регуляторных модулей действительно важна для регуляции транскрипции гена, то она будет сохранятся в процессе эволюции. Несмотря на достаточно большую дивергенцию по последовательности, регуляторные участки ортологичных генов могут иметь схожую организацию (структуру). С другой стороны, регуляторные участки сходно регулируемых генов, по-видимому, также должны быть похожи. Таким образом, изучая регуляторные участки ортологичных и/или ко-регулируемых генов, можно попытаться выделить закономерности расположения сайтов, сохраняющиеся для большинства из этих генов, а значит, вероятно, важные для правильной регуляции

транскрипции.

Актуальным является разработка алгоритмов, позволяющих выявлять закономерности расположения ССТФ, характерные для набора сходно функционирующих регуляторных модулей, и использовать информацию о выявленной регуляторной структуре для повышения качества предсказаний регуляторных участков 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.

Вероятностная модель регуляторных модулей

Разработан алгоритм поиска регуляторных модулей в геномах эукариот на основе набора известных моделей сайтов (позиционно-весовых матриц, ПВМ), учитывающего консервативную структуру регуляторных модулей. 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.14).

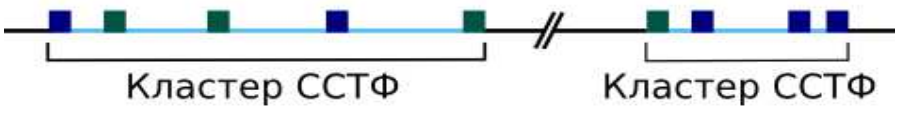


Рисунок 14. Схематическое изображение двух кластеров сайтов, окруженных фоновой последовательностью (черные отрезки горизонтальной линии). Каждый кластер состоит из сайтов (синие и зеленые квадраты), разделенных спейсерами (голубые отрезки горизонтальной линии).

Для моделирования регуляторных модулей, окруженных фоновой последовательностью, использовалась скрытая Марковская модель (СММ), схема которой изображена на рисунке 15.

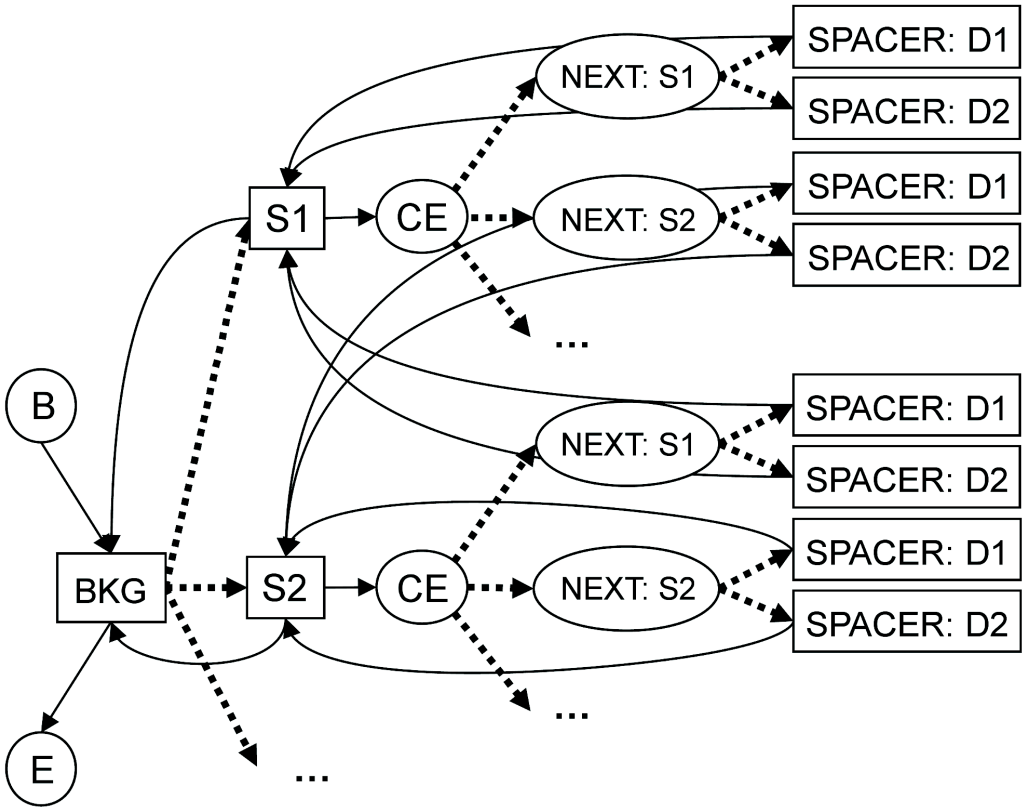


Рисунок 15. Схема СММ. Порождающие состояния изображены в виде прямоугольников, молчащие состояния — в виде овалов. Разрешенные переходы между состояниями показаны стрелками. Вероятности переходов, помеченных пунктиром, изменяются в процессе обучения модели по алгоритму Баума-Велча.

Архитектура CMM отражает наше понимание о том, как устроены регуляторные модули в геномах эукариот. CMM, 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. CMM с таким типом архитектуры называют обобщенной CMM [32-34]. Ее преимуществом является возможность использовать любое заданное распределение длин порождаемых последовательностей.

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

$$P_{state}(sequence) = P_{state}(sequence \mid L)P_{state}(L) \quad (21)$$

Где $P_{state}(sequence \mid L)$ – 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} – вероятность открытия кластера):

$$P_{BKG}(L) = (1 - p_{open})^{L-1} \cdot p_{open} \quad (22)$$

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. 15): 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 [28, 31, 35].

Каждое из состояний, порождающих сайты, имеет только два возможных перехода: обратно в состояние BKG (с вероятностью pclose), 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 pclose.

Предлагаемая в данной работе CMM 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 CMM 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 16. Распределения расстояния между соседними сайтами в кластере, использовавшиеся в данной работе.

CMM 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.

Обучение модели

Выявление структуры регуляторных участков происходит в результате обучения параметров модели на наборе последовательностей, которые предположительно содержат

регуляторные модули с похожей организацией. Для обучения параметров использовался алгоритм Баума-Велча [36, 21]. Поскольку целью обучения параметров является выявление структуры регуляторных областей, при обучении изменяются только параметры, определяющие структуру регуляторных модулей (переходы, выделенные пунктиром на рисунке 15).

Распознавание регуляторных модулей

Каждый путь в графе CMM, 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 [37]. Этот алгоритм представляет собой комбинацию алгоритма апостериорного декодирования «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 веса каждого найденного кластера используется логарифм отношения правдоподобия, который вычислялся как логарифм отношения апостериорной вероятности, 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 CMM, порождающих данный кластер и фоновую последовательность соответственно.

Учет консервативности

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. 17).

Сила (консервативность) ряда соответствующих регуляторных модулей вычисляется следующим образом. Для каждой пары регуляторных модулей (i и j) в ряду рассчитывается величина сходства между ними qij (пары сочетаний показаны пунктиром на рис. 17). Мера сходства между парой кластеров учитывает только состав модуля (то есть количество сайтов каждого типа):

$$q_{ij} = \frac{n_i + n_j}{2} \cdot \frac{\bigcap_{ij}}{\bigcup_{ij}} \quad (23)$$

где ni и nj — количества сайтов в модулях i и j, Ωij - размер пересечение наборов

сайтов в модулях i и j, Uij — размер объединение наборов сайтов в модулях i и j. Сила ряда соответствующих регуляторных модулей рассчитывается как сумма qij 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.

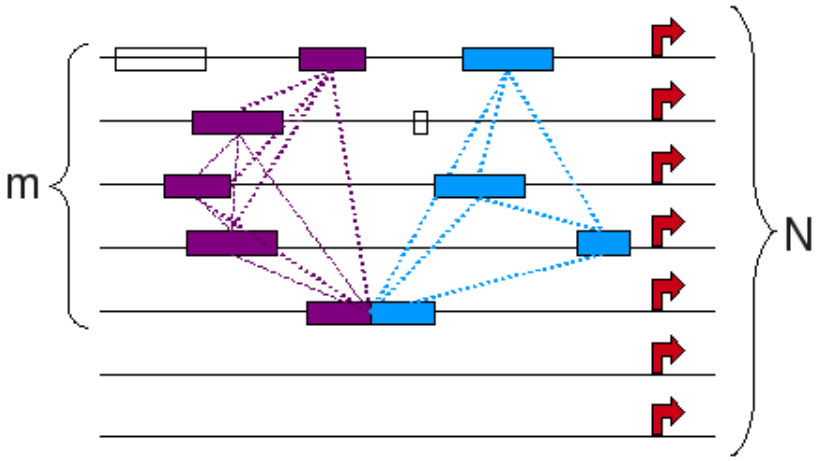


Рисунок 17. Пояснение к описанию вычисления значения консервативности предсказанных модулей для группы ортологичных генов. Горизонтальные линии соответствуют разным геномам. Красными стрелками показаны старты ортологичных генов. Прямоугольники обозначают предсказанные регуляторные модули, при этом одним цветом обозначены модули, формирующие один ряд (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 $m/N < 0.5$).

Свойства регуляторных модулей

Был проведен ряд тестов на различных наборах родственных геномов и разичных известных регуляторных систем. В результате было установлено наличие предпочтения следования сайтов связывания и выбора расстояний между сайтами. На рис.18 показаны вероятности следования сайтов связывания для системы мышечных промоторов.

	mef2>	mef2<	myf>	myf<	sp1>	sp1<	srf>	srf<	tef>	tef<
mef2>	0,00 / 0,49	0,00 / 0,50	0,00 / 0,53	0,96 / 1,00	0,00 / 0,63	0,00 / 0,60	0,00 / 0,50	0,00 / 0,50	0,00 / 0,50	0,01 / 0,77
mef2<	0,00 / 0,50	0,00 / 0,81	0,17 / 0,00	0,01 / 0,89	0,00 / 0,33	0,41 / 0,01	0,00 / 0,45	0,18 / 0,01	0,00 / 0,39	0,22 / 0,51
myf>	0,25 / 0,83	0,10 / 0,00	0,00 / 0,22	0,00 / 0,79	0,16 / 1,00	0,00 / 0,51	0,11 / 0,79	0,00 / 0,42	0,37 / 1,00	0,00 / 0,56
myf<	0,00 / 0,50	0,00 / 0,61	0,22 / 0,99	0,00 / 0,52	0,45 / 0,98	0,22 / 0,10	0,00 / 0,53	0,09 / 0,77	0,00 / 0,50	0,01 / 0,82
sp1>	0,00 / 0,50	0,34 / 0,01	0,00 / 0,52	0,01 / 0,84	0,01 / 0,39	0,59 / 0,99	0,01 / 0,72	0,02 / 0,77	0,02 / 0,95	0,00 / 0,47
sp1<	0,16 / 0,01	0,00 / 0,64	0,34 / 0,43	0,00 / 0,84	0,00 / 0,51	0,06 / 0,94	0,39 / 0,68	0,00 / 0,08	0,00 / 0,63	0,04 / 0,99
srf>	0,11 / 1,00	0,00 / 0,52	0,00 / 0,52	0,00 / 0,54	0,00 / 0,48	0,11 / 0,00	0,07 / 0,99	0,38 / 0,21	0,32 / 0,00	0,00 / 0,50
srf<	0,00 / 0,51	0,00 / 0,51	0,00 / 0,51	0,15 / 0,00	0,00 / 0,42	0,65 / 0,00	0,19 / 0,99	0,00 / 0,17	0,00 / 0,52	0,00 / 0,52
tef>	0,01 / 0,06	0,02 / 0,01	0,00 / 0,45	0,01 / 0,95	0,28 / 0,00	0,52 / 0,00	0,02 / 0,15	0,00 / 0,49	0,14 / 1,00	0,00 / 0,50
tef<	0,00 / 0,50	0,50 / 1,00	0,00 / 0,48	0,33 / 0,81	0,13 / 0,01	0,01 / 0,61	0,01 / 0,46	0,00 / 0,50	0,01 / 0,18	0,00 / 0,51

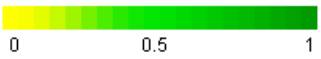


Рисунок 18. Параметры модели, обученной на мышечной выборке последовательностей. Каждая ячейка (i, j) таблицы содержит значения параметров модели для данной пары типов сайтов: условную вероятность наблюдать сайт типа j следом за сайтов типа i (первое число) и вероятность выбора синусоидального распределения расстояний между сайтами этих типов (второе число). Цвет ячейки соответствует значению условной вероятности для данной пары. Направление сайта относительно регулируемого гена обозначено символами '>' (положительная цепь) и '<' (комплиментарная цепь).

Анализ параметров обученной модели выявил несколько пар типов сайтов с достаточно высоким значением P(j|i). Согласно параметрам модели (рис. 18), наиболее вероятными парами ТФ, чьи сайты связывания часто располагаются рядом, являются Mef2-

Myf, Sp1-Srf и Sp1-Sp1. Для всех трех пар в литературе существуют наблюдения 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 длине спирали ДНК, было замечено исследователем Fickett еще в 1996 году в процессе анализа известных регуляторных модулей мышечных генов [35]. Интересно, что в соответствии с параметрами нашей модели, сайты связывания этих факторов также предпочтительно располагаются на расстоянии кратном длине спирали ДНК (рис. 18), то есть описывается синусоидальным распределением расстояний, использующемся в модели (см. описание модели в главе 2). Синергия между сайтами связывания Mef2 и Myf также зафиксирована в базе данных TransComplel (идентификатор записи – C00120).

Взаимодействие факторов SRF и Sp1 косвенно подтверждается экспериментальными работами, в которых было показано, что наличие сайтов связывания обоих ТФ в регуляторных модулях необходимо для корректной регуляции транскрипции гена альфа-актина сердечных мышц человека и гена тяжелой цепи миозина гладких мышц крысы. Синергическая активация транскрипции фактором Sp1 была показана in vivo.

Анализ распределения расстояний между сайтами связывания также показал, что сайты связывания распределены не случайно и существуют предпочтения расстояний. Поскольку поиск регуляторных модулей происходил с помощью обученной модели, параметры которой обсуждались выше, то наличие значительной доли гомотипических пар сайтов среди всех перепредставленных пар было вполне ожидаемо (рис. 19). Более того, среди всех возможных гомотипических пар сайтов для ТФ Hb и Kr перепредставленными оказались только пары Hb<Hb< и Kr<Kr< (а также пара Kr>Kr>, количество наблюдений которой оказалось чуть ниже порога, и поэтому формально не включенная в список перепредставленных пар), то есть пары сонаправленных сайтов, что также соответствует параметрам модели. Наблюдение перепредставленных пар с определенной ориентацией сайтов относительно друг друга соответствует идее, что ТФ, сайты связывания которых необычно часто располагаются рядом, работают кооперативно, и поэтому взаимная ориентация этих сайтов должна быть важна для регуляции транскрипции. Интересно, в

найденных регуляторных модулях наиболее многочисленными оказались пары сайтов только в определенной ориентации относительно регулируемого гена (например, пара сайтов Hb<Hb< была наблюдена 79 раз, а Hb>Hb> только 45, пара сайтов Kni<Cad< наблюдалась 96 раз, а комплементарная ей комбинация Cad>Kni> всего 22 раза. Можно предположить, что наблюдение несимметричного расположение пар сайтов относительно регулируемого гена имеет некий биологический смысл. Возможно, в некоторых случаях важна не только ориентация сайтов связывания относительно друг друга, но также и относительно регулируемого гена, поскольку комплекс белков, связывающих эти сайты, более эффективно влияет на уровень транскрипции, находясь в определенной ориентации относительно регулируемого гена.

Наблюденные распределения расстояний для пар сайтов, характеризующихся, в соответствии с моделью, синусоидальным распределением расстояний (Bcd>Bcd<, Bcd<Bcd< и Hb<Hb<), действительно имеют спадающие пики с интервалом в 10-11 нуклеотидов, в чем опять же нет ничего удивительного. А вот для тех пар, для которых вероятность выбора синусоидального распределения в модели была мала (Kr<Kr<, Kni>Kni< и Kni<Cad<), наблюденные распределения расстояний были довольно необычными (рис. 19 и 20). Особенно интересным кажется распределений для пары Kni>Kni<, поскольку оно имеет необычный и довольно выраженный пик на расстоянии 135-138 пн. Такие большие расстояния между сайтами довольно необычны и, по-видимому, говорят о связывании фактора Kni с компактизованной ДНК.

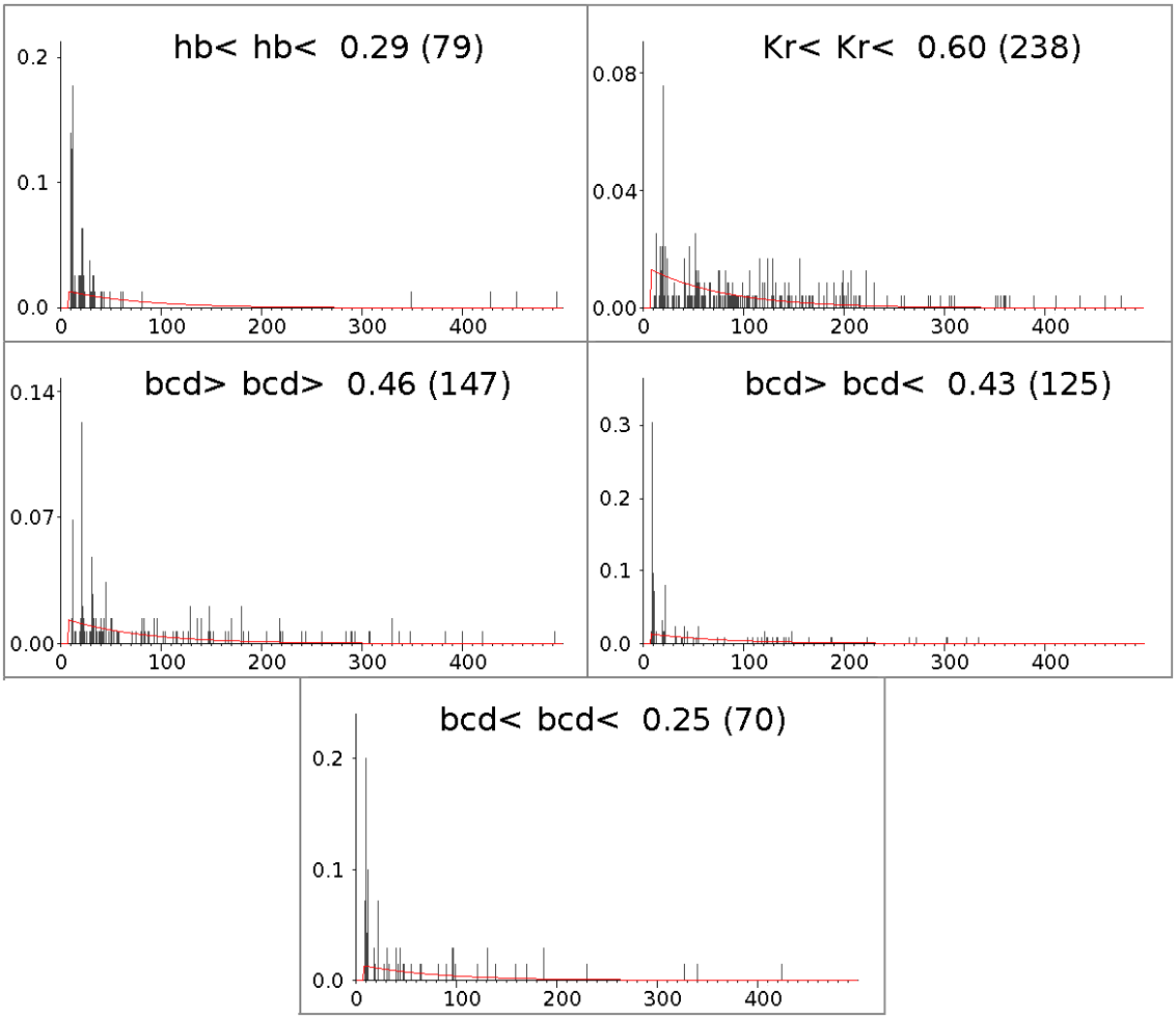


Рисунок 19. Распределения расстояний между сайтами гомотипических пар, перепредставленных в регуляторных модулях, найденных около генов развития Drosophila. Для каждого распределения указаны имена сайтов, составляющих пару (направление сайта указывается символами '>' для положительной цепи и '<' для комплементарной цепи), коэффициент корреляции между сайтами и количество наблюденных пар (дано в скобках). Фоновое распределение расстояний между сайтами показано красной линией.

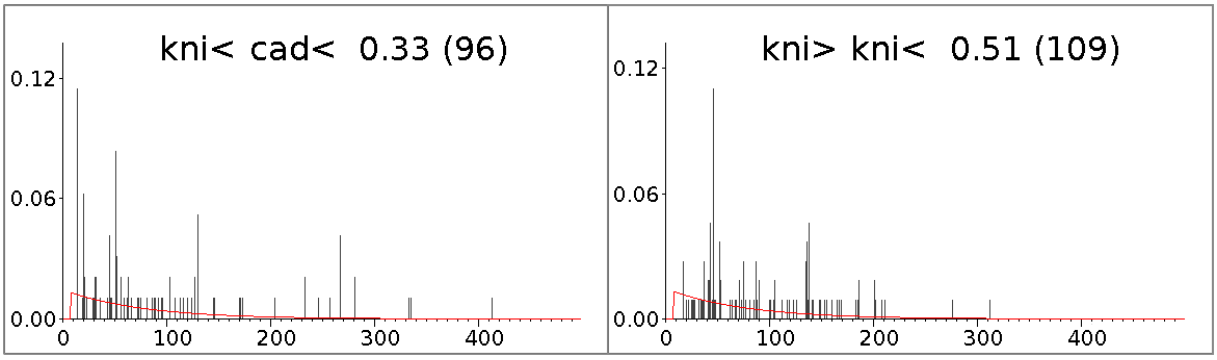


Рисунок 20. Распределения расстояний между сайтами гетеротипических пар, перепредставленных в регуляторных модулях, найденных около генов развития Drosophila. Обозначения как для рисунка 21.

Заключение

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.

Обобщая представленные данные, можно с уверенностью заключить, что нами разработана новая модель описания регуляторных модулей, учитывающая их структуру. Указанная модель позволяет предсказывать регуляторные модули в геномах эукариот. Кроме того, в результате обучения нашей модели можно получить представление о грамматике регуляторных модулей, т.е. об особенностях их структуры, в частности о характере зависимости следования сайтов связывания о направлении транскрипции, о характере распределения расстояний между модулями. Результатом работы программы является предсказание списка ко-регулируемых генов. Тестирование на известных системах показало, что предсказанные списки генов полностью соответствуют известным генам, и некоторое количество не аннотированных ранее генов также было предсказано.

2.2.3 Сопоставление анализа научно-информационных источников и результатов теоретических и экспериментальных исследований

Анализ научно-информационных источников, сделанный на первом этапе работы и

сопоставление его с результатами исследований показал, что:

– Описание эволюции параметров, как диффузионного процесса является расширением подхода, описанного Кимурой [6], однако в современных условиях этот подход может применяться не к фенотипам, как это было сделано в оригинальной работе, а свойствам генетических текстов. Такой подход к анализу эволюции геномов является новым. Нами впервые были исследованы особенности поведения диффузионного уравнения для ряда конкретных типов параметров последовательностей, в том числе таких сложных, как энергия вторичной структуры РНК. Разработанный нами подход позволит в дальнейшем проводить математически более обоснованный анализ эволюции последовательностей, а также оценивать статистическую значимость сравнительно-геномного анализа.

– Разработанный нами метод скрытых Марковских моделей на деревьях также является новым подходом. Ближайшим аналогом такого рода анализа являются Байесовы сети. Однако, разработанный нами подход намного проще и эффективнее в применении к задачам исследования эволюции и сравнительно-геномного анализа. Тестирование на искусственных и реальных примерах показал, что этот подход позволяет быстро и эффективно решать указанный круг задач (см. п.3.3).

– Для предсказания регуляторных модулей Скрытые Марковские модели использовались и ранее. Однако наше расширение модели с использованием информации о структуре регуляторных модулей позволило, во-первых, увеличить точность предсказаний, а, во-вторых, позволило в явной форме получать информацию о структуре регуляторных модулей. Работа с известными регуляторными системами позволила определить новые, ранее не известные свойства распределения расстояний и ориентации сайтов связывания. На настоящий момент наша программа предсказания модулей является наиболее точной.

2.2.4 Оценка эффективности полученных результатов в сравнении с современным научно-техническим уровнем

Поскольку исследование диффузионной модели эволюции параметров при эволюции последовательностей предложено нами впервые, то сравнивать с другими аналогичными

работами бессмысленно – таких работ просто не существует.

Оценка эффективности Скрытых Марковских моделей на деревьях. Нами было проведено сравнение качества работы нашего подхода с наиболее похожей программой Bayes-trait [38]. К сожалению нет возможности проводить испытания такого родо программ на реальных данных, поскольку реальные данные для предков не доступны, а молекулярные события не картированы во времени. Кроме того, реальные биологические состояния зачастую не определены с достаточной степенью достоверности. Поэтому испытания проводились методом Монте-Карло на искусственных данных, когда состояния и события полностью определены. Задача такого рода сравнительного анализа программ заключается в том, чтобы определить, насколько программа может предсказать правильные ответы при случайных данных, сгенерированных с помощью определенной модели. Испытания проводились по следующей схеме. Генерировались случайные эволюционные деревья. Длина ветвей выбиралась из распределения длин ветвей реальных деревьях. Затем задавались следующие параметры: скорость эволюции (вероятность смены состояний в зависимости от длины ветвей), распределения наблюдаемых значений при разных состояниях. С использованием этих параметров порождались события смены состояний на ветвях, и, тем самым, определялись состояния на узлах. Это те состояния, которые должна предсказать программа. С помощью заданных распределений вероятностей эмиссии наблюдаемых значений генерировались сами значения на листьях. Таким образом перед программой ставилась задача – при заданном дереве и наблюдениях на листьях определить (предсказать) состояния на узлах дерева и на листьях. Далее, предсказания сравнивались со значениями, которые были сгенерированы и определялись точность, чувствительность и специфичность работы программы. На рис. 21 показаны зависимость точности программ при различных скоростях эволюции и перекрытии распределений 19%. Оценка точности сделана по результатам 1000 испытаний для каждой точки.

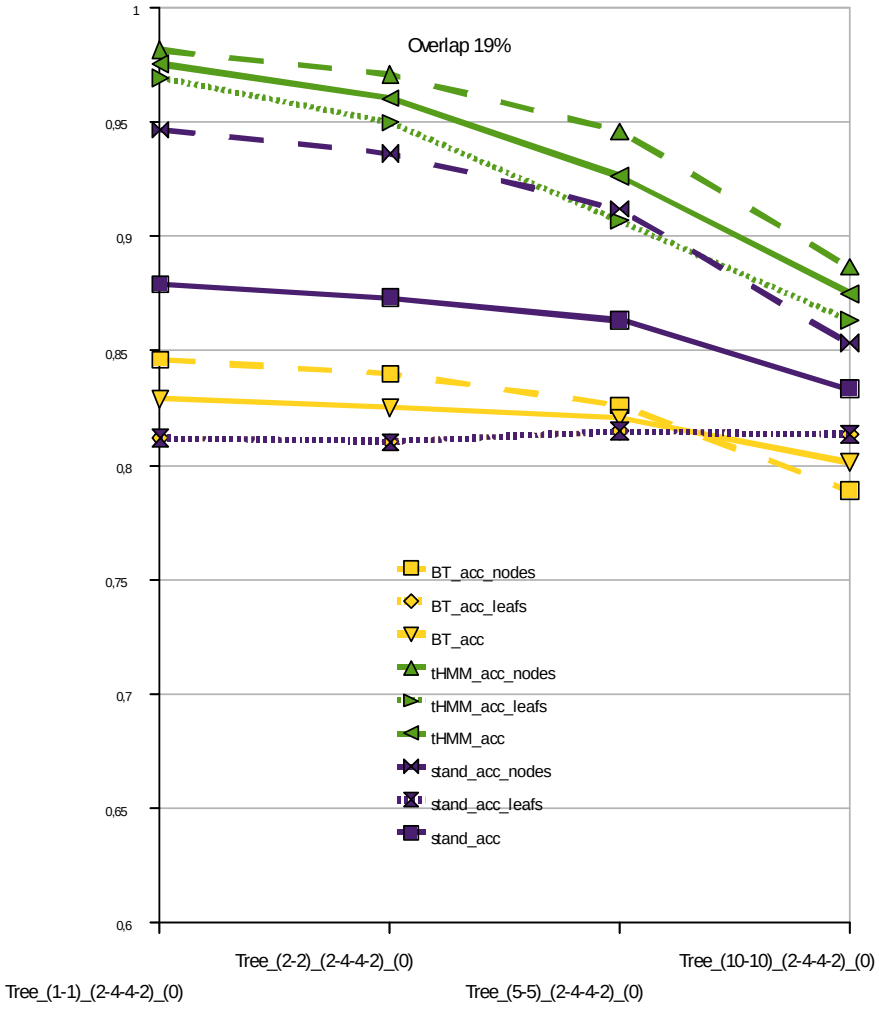


Рисунок 21. Зависимость точности (acc) от скорости эволюции. Обозначения: Bt – результаты программы BayesTraits, tHMM – результаты нашего подхода, stand – результаты при стандартном подходе; nodes – точность предсказания состояний на внутренних узлах, leafs – точность предсказания на листьях, без метки – точность предсказания эволюционных событий. По горизонтальной оси – скорость эволюции.

Из рис.21 очевидно, что наш подход существенно превосходит широко используемую программу BayesTraits по точности предсказания. Причина такого превосходства заключается в том, что наш подход существенно аккуратнее работает с вероятностями и при оптимизации

оценивает не только нижнее поддерево, но и верхнее поддерево. Было проведено еще множество испытаний при различных параметрах. Во всех случаях наш подход существенно превосходил по качеству существующие подходы.

Эффективность предсказания регуляторных модулей оценивалась с помощью публичного сервиса http://tare.medisin.ntnu.no/composite/composite.php. Это сервис вычисляет основные параметры, характеризующие качество предсказания – специфичность, чувствительность, предсказательную ценность положительного результата, коэффициент эффективности, среднюю эффективность. Все эти паказатели являются производными от основных – истинно-позитивные, ложно-позитивные, истинно-негативные и ложно-негативные предсказания. В качестве тестовой выборки использованы мышечные промоторы позвоночных. Для этих последовательностей есть достаточно большой набор экспериментальных данных. Однако при оценке абсолютных значений качества предсказаний надо иметь в виду, что существующий набор экспериментальных данных не совсем адекватно оценивает негативные результаты, поскольку отсутствие экспериментальных данных не означает отсутствия эффекта. На рис. 22 показаны результаты сравнения нашей программы с другими популярными программами.

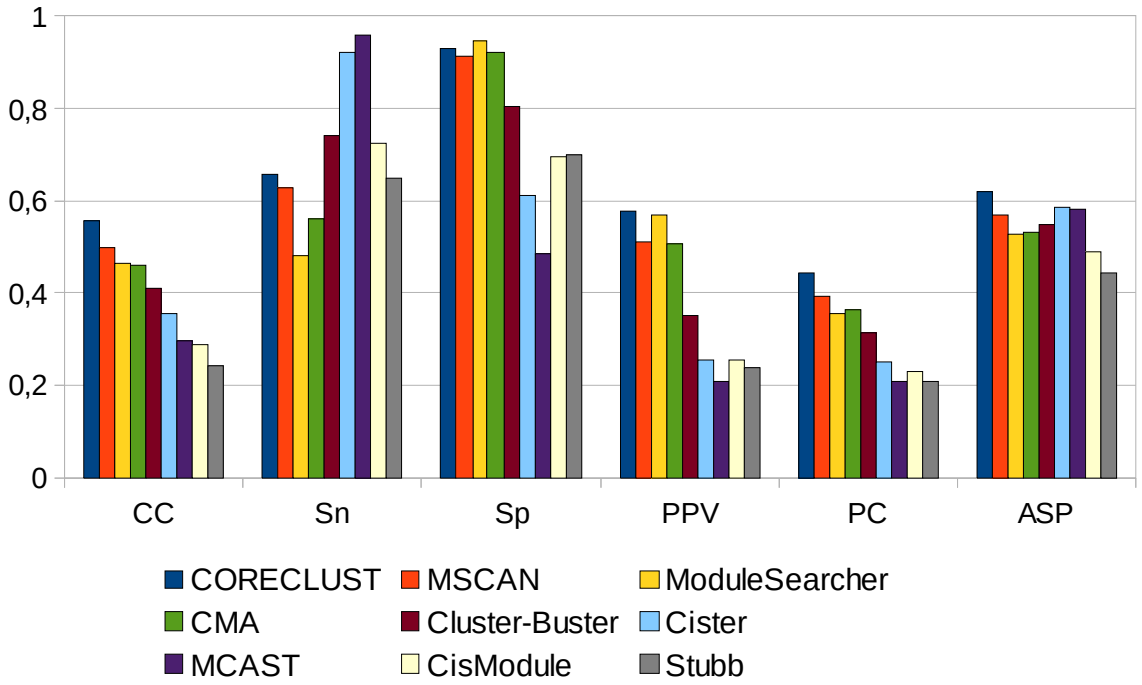


Рисунок 22. Результаты сравнения предсказательной силы различных программ на нуклеотидном уровне. Здесь CC – коэффициент корреляции, Sn – чувствительность, Sp – специфичность, PPV – предсказательная ценность положительного результата, PC – коэффициент эффективности ASP – средняя эффективность.

Сравнение на нуклеотидном уровне показало (рис. 10), 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 участков от большинства. Поскольку обучение программы осуществляется на всех последовательностях выборки, то модель регуляторных участков отражает структуру, наиболее представленную во всей выборке.

Таблица 1. Сравнение качества (CC) работы программ для генов системы раннего развития Drosophila. Ряд TOTAL содержит значения CC, вычисленные для всех генов из

набора. Жирным выделено максимальное значение CC в каждой строчке. *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

Сравнение качества предсказания программ (таблица 1) показало, что согласно T‑критерию Уилкоксона, предсказания CORECLUST имеют более высокое значение CC, чем программы Stubb (p-value < 0.05), MOPAT (p-value < 0.0007) и Cluster-Buster (p-value < 0.02). Согласно сравнению p-значений (p-value), Stubb, чья модель очень близка к модели CORECLUST, характеризуется наилучшим качеством предсказания для данной выборки

после CORECLUST. В то же время, Stubb показал не очень хорошие результаты предсказания на системе генов, специфически экспрессирующихся в мышечной ткани позвоночных, где он оказался хуже по всем мерам, чем CORECLUST и Cluster-Buster. Интересно, что Cluster-Buster, использующий очень простую модель, которая не учитывает ни корреляции между сайтами, ни сравнение между видами, показывает достаточно хорошие результаты на обеих выборках.

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 участков для набора системо-специфичных ТФ.

2.2.5 Оценка полноты решения задач и достижения поставленных целей НИР

Целью выполнения работы была разработка программного обеспечения, осуществляющего хранение/передачу, обработку и анализ данных о символьных последовательностях, полученных в рамках теоретической и прикладной геномики с целью повышения качества вычислительной обработки биологических данных, используемых в приложениях в области медицины, биотехнологии и сельского хозяйства. Было разработано программное обеспечение для предсказания и анализа регуляторных модудей, а также алгоритмы и программы для сравнительного анализа геномов. Показано, что качество предсказаний было улучшено. Поставленные цели можно считать достигнутыми.

Проведенные исследования полностью соответствуют требованиям Технического задания:

1. Аналитический обзор (отчет по этапу 1) содержит 117 литературных источников, в том числе 90 источников, опубликованных с 2005 года.
2. Разработаны новые подходы, основанные на анализе диффузии в пространстве параметров и на скрытых Марковских моделях. Эти подходы позволяют оценивать статистичскую значимость сравнительно-геномного анализа.

3. Разработана библиотека классов CMP_GENOMICS. Соответствующая программа зарегестрирована в Фонде Алгоритмов и программ
4. Разработаны вероятностные модели регуляторных модулей (участков ДНК, содержащих сайты связывания фактров транскрипции)
5. Разработаны программный комплекс ЭО ПК CORECLUST, предназначенный для определения функциональных сегментов геномов с помощью анализа текстовых сигналов, и позволяющий в частности определять регуляторные участки путем анализа кластеров сайтов ДНК
6. Разработаны программа и методика испытаний ЭО ПК CORECLUST. Проведены исследовательские испытания ЭО ПК CORECLUST. Комплекс полностью соответсвует требованиям ТЗ, а по скорости работы значительно превосходит требования.
7. Получены новые результаты относительно распределения расстояний между сайтами связывания и по предпочтению следования сайтов связывания.
8. Разработан проект ТЗ на ОКР по теме: «Разработка автоматизированной системы для сравнительно-геномного анализа регуляции экспрессии генов».

В техническом задании требовалось провести оценку применимости разработанных методов для анализа последовательностей небиологической природы. Поскольку разработанные методы основаны на эволюционном подходе, они применимы только для анализа биологических последовательностей.

2.2.6 Технико-экономическая оценка рыночного потенциала полученных результатов и рекомендации по их использованию

2.2.6.1 Технико-экономическая оценка рыночного потенциала полученных результатов

Расчеты и проведённые исследования, приведённые в отчёте, позволяют сделать вывод, что стоимость представленного на оценку объекта интеллектуальной собственности по состоянию на 31 октября 2012 года составляет

4 936 600,08 рублей

(Четыре миллиона девятьсот тридцать шесть тысяч шестьсот рублей 8 копеек)

2.2.6.1.1 Описание объекта оценки

I. Комплекс программ (ИПК) CORECLUST, предназначенный для поиска РМ в геномах эукариот с учетом их регуляторной структуры и консервативности, а также описания статистических свойств найденных РМ и оценки консервативности РМ, найденных в группах геномных последовательностей. ИПК содержит две части — серверную и клиентскую.

Серверная часть ИПК CORECLUST состоит из трех программных компонентов:

-- программного компонента CORECLUST-train для оценки параметров вероятностных моделей РМ, который решает задачу оптимизации параметров скрытой Марковской модели, описывающей последовательность, содержащую РМ, алгоритмом максимизации ожидания.

-- программного компонента CORECLUST-search для поиска РМ в геномных последовательностях, решающего задачу поиска РМ в данных геномных последовательностях, используя для этого предоставленную ему модель РМ.

-- библиотеки классов CMP_GENOMICS, решающей в том числе задачу оценки консервативности РМ, найденных в группе родственных последовательностей.

Клиентская часть включает:

- модуль доступа через веб-интерфейс, предназначенный для загрузки пользовательских данных на вход программе, запуска программы и визуализации результатов.

Все программные элементы и их компоненты прошли успешно тесты.

С помощью разработанной методики было исследована структура предсказанных регуляторных модулей генов раннего развития плодовой мушки. Сравнение с другими известными программами 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 участков от большинства. Поскольку обучение программы осуществляется на всех последовательностях выборки, то модель регуляторных участков отражает структуру, наиболее представленную во всей выборке. На уровне мотивов предсказания CORECLUST, напротив, показал довольно высокую чувствительность. 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 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 участков для набора системо-специфичных ТФ.

Тестирование разработанной программы на двух хорошо изученных биологических системах (системе генов позвоночных, специфически экспрессирующихся в мышечной ткани, и системе раннего развития Drosophila) показало, что программа применима к различным системам и организмам и может быть успешно использована для решения стандартной задачи поиска регуляторных участков для набора системо-специфичных ТФ с качеством, превышающим качество предсказания других известных программ, решающих аналогичную задачу. Продемонстрировано, что учет структурных особенностей регуляторных модулей позволяет выявлять правильные (то есть подтверждающиеся известными) регуляторные модули. Более того, регуляторные модули, найденные алгоритмом, учитывающим все рассмотренные аспекты регуляторной структуры, более консервативны, что может служить аргументом в пользу более высокой надежности этих предсказаний.

2.2.6.1.2 Выбор методов оценки

Технология оценки объектов ИС бывает:

– договорная (например, при определении вклада в уставной капитал, когда учредители

предприятия, без проведения каких-либо расчетов, договариваются между собой относительно стоимости ИС, образующей собой вклад);

– расчетная, или аналитическая (например, при оценке изобретения, когда искомая стоимость изобретения определяется расчетом по определенному алгоритму);
– экспертная (когда не работают или слишком дорого обходятся предыдущие технологии, например, при оценке многих объектов авторских прав).

Основной подход к оценке ИС, используемый в РФ и принимаемый Налоговой инспекцией — расчетный (аналитический), именно он и будет использован в данной оценке.

При расчетной оценке ОИС будут использоваться следующие подходы: затратный, сравнительный, доходный.

2.2.6.1.3 Расчёт рыночной стоимости объекта оценки методом исходных затрат в рамках затратного подхода

Сущность метода. Метод исходных затрат строится на использовании реальных ретроспективных данных о расходах, произведенных с целью создания оцениваемых результатов интеллектуальной деятельности. Его особенностью является обязательность индексации выявленных затрат с целью их приведения к уровню цен на товары и услуги, соответствующему дате оценки. Такая индексация должна проводиться с применением рассчитываемого Госкомстатом индекса потребительских цен. Допустимо также применение отраслевых индексов в случае доступности информации о них (в строительстве и смежных отраслях, например, возможно использование индексов, публикуемых фирмой «КО-Инвест»).

Рыночная стоимость ОИС по данному методу определяется в следующей последовательности:

– исследуется документы предприятия за тот период, в течение которого выполнялась работа по созданию ОИС и доведению его до готовности к использованию в

запланированных целях, при этом выявляются все фактические затраты, непосредственно связанные с созданием (приобретением) ОИС;

- строится календарный график фактического расходования средств;
- производится приведение выявленных фактических затрат к дате оценки с помощью индексов, учитывающих изменение цен за время, прошедшее с момента совершения затрат до времени выполнения оценки;
- скорректированные затраты суммируются и полученная сумма увеличивается на размер разумной предпринимательской прибыли (по ставке, не меньшей ставки рефинансирования ЦБ РФ);
- определяется в денежном выражении величина снижения стоимости (износа) объекта оценки, обусловленного устареванием, накопившимся с момента создания ОИС до даты оценки.

Расчёт стоимости. Определение фактически понесенных затрат на создание ОИС

Фактические затраты определены в соответствии с перечислениями по гранту к оценке разработанной методики и данных, полученных из средств массовой информации, которые сведены в следующую таблицу 1.

Определение индексов приведения фактически понесенных затрат к дате оценки. Поскольку фактические затраты были произведены с июля 2011 года по октябрь 2012 года, то для приведения значений стоимостей затрат к дате оценки индексирование производилось с применением рассчитываемого Госкомстатом индекса потребительских цен на июль 2011 года равный 100,0%.

Определение снижение стоимости объекта оценки. Износ на объект оценки не определялся по причине новизны, уникальности и актуальности объекта оценки.

Таблица 2. Смета затрат на создание ОИС

п/п	Статья затрат	Приведенная (округленная) стоимость затрат , руб.
	Заработная плата	3 817 162
	Социальные начисления	1 200 981
	Накладные расходы	855 000
	Затраты на проведение экспериментальной части	0
	Прочие прямые расходы	678 857
	ИТОГО затрат	6 552 000
	Индексация	6 552 000
	Износ	0
	Прибыль предпринимателя	540 540
	Рыночная стоимость объекта оценки	7 092 540

Прибыль предпринимателя (инвестора) принята равной ставке рефинансирования, которая составляет – 8,25%.

Таким образом, рыночная стоимость проекта на дату оценки 31 октября 2012г., полученная в рамках затратного подхода составляет ***7 092 540 рублей.***

Расчёт рыночной стоимости объекта оценки методом освобождения от роялти с капитализацией денежного потока. Чтобы определить рыночную стоимость объекта оценки, приносящего стабильную прибыль, следует умножить показатель доходности на специальный множитель (мультипликатор) М (см. (1))

$$V = M \times (\text{прибыль от использования объекта оценки за год}), \quad (1)$$

где,

V – рыночная стоимость объекта оценки, руб.

M – мультипликатор, определяется по формуле (2)

$$M = 1/r_0, = 100/(\text{ставка капитализации}), \quad (2)$$

где,

r_0 – ставка капитализации, рассчитывается для предприятия с учетом поправки на больший риск.

Последовательность действий при установлении рыночной стоимости исключительных прав на изобретение или промышленный образец методом освобождения от роялти состоит в следующем:

- Определить сферу использования изобретения (промышленного образца).
- Оценить объем использования. Определить примерный объем реализации продукции, изготавливаемой с использованием изобретения (промышленного образца);
- Определить по таблице отраслевых роялти наиболее вероятную ставку роялти в случае продажи лицензии на использование изобретения (промышленного образца).
- Определить ожидаемую среднегодовую прибыль (до налогообложения) от использования оцениваемого актива за период действия охранного документа. Как правило, эта прибыль совпадает со среднегодовыми поступлениями в виде роялти или меньше этих поступлений, если учесть затраты на поддержание патента в силе и т.п.
- Умножить среднегодовую прибыль от использования оцениваемого актива на M. Получаемый результат можно считать рыночной стоимостью оцениваемого актива на сегодняшний день.

2.2.6.1 Определение сферы использования объекта оценки

Разработанные подходы и программы могут быть использованы в персонализованной медицине, в научных исследованиях регуляции экспрессии генов (как в теоретических так и в

экспериментальных), а также в образовательном процессе.

2.2.6.2 Определение примерного объема денежного потока от реализации объекта оценки

Конкурентные продукты. Некоторые наши результаты не имеют аналогов в мире (например, использование диффузионной модели в сравнительной геномике), другие (предсказание регуляторных модулей) – напротив находятся в достаточно жесткой конкуренции. Использование Скрытых Марковских моделей на деревьях занимают промежуточное положение – есть конкурентные программы, но они дают намного худшие результаты.

При анализе коммерческой перспективы надо принимать во внимание то, что большинство программных продуктов в этой области находятся под лицензией открытого программного кода (GNU и GPL) и распространяются бесплатно.

Ситуация, когда большинство программных продуктов распространяются бесплатно, делает бесперспективными попытки выносить эти продукты на рынок. Кроме того, в период длительной рецессии, начавшийся в начале века, фармацевтические компании в основном отказались от создания принципиально новых лекарств, поскольку это требует гигантских долговременных инвестиций – они сосредоточились на модификации традиционных лекарственных препаратов. Поэтому основной коммерческий потребитель биоинформационных программ сокращает свои потребности в услугах биоинформатики. Все это является причиной того, что многие биоинформатические компании в настоящее время переживают не лучшие времена.

Способы использования продуктов. Однако, бизнес-модель открытых кодов предполагает широкое привлечение авторов программных продуктов для консалтинга и для кастомизации продуктов под конкретных коммерческих пользователей. Основными потребителями такого рода услуг в случае наших программных продуктов могут быть:

- исследовательские группы и лаборатории
- компании, занимающиеся разработкой диагностических методов

– с некоторыми оговорками фармацевтические компании.

В нашем случае наиболее перспективной и наименее рискованной является стратегия, ориентированная на предоставлении отдельных консалтинговых услуг с использованием разработанных программ, а также с использованием приобретенного в процессе разработке опыта.

Потенциальными потребителями результатов данного проекта являются академические и коммерческие центры генетической диагностики и исследовательские лаборатории. Важным потребителем разработанных техник являются структуры, ориентированные на медицину. В настоящее время разработками новых стратегий лечения (если не принимать во внимание откровенных жуликов и халтурщиков уровня гр. Петрика) являются разного рода государственные институты, в том числе институты академии Медицинских наук, Академии сельскохозяйственных наук, РАН. Основным источником финансирования этих структур пока является Государственный бюджет. Другие заказчики также по источникам финансирования связаны в основной своей части с государственным бюджетом. Это различные исследовательские группы, занимающиеся фундаментальными исследованиями. Наконец, третьим видом потребителей возможных услуг по консалтингу и кастомизации программного обеспечения являются фармацевтические и биотехнологические компании.

При анализе коммерческой перспективы следует обратить внимание на то, что продвижение наших, вполне современных и актуальных, но, в общем-то, ограниченных по области применения, продуктов требует значительных вложений.

Определение ставки роялти. Стандартная ставка роялти в % цены единицы продукции (суммы чистого дохода, объема реализации, экономии затрат) для средств фармацевтической и косметической промышленности составляет 2 - 5%. Поскольку реализация методики и набора кандидатов имеет большие риски, то принято нижнее значение интервала – 2%.

Определение величины соответствующей ставки капитализации. В общем случае ставка капитализации равна

<ставка капитализации> = <ставка дисконтирования> - <темп роста доходов> (3)

Норма или ставка дисконтирования является характеристикой риска, отражающей вероятность неполучения планируемых денежных потоков. Поскольку интеллектуальная собственность изолированно от бизнеса не создает денежных поступлений, норма дисконтирования для её оценки определяется так же, как в оценке действующих предприятий (бизнеса).

Определение ставки дисконтирования проводится по формуле

$$R=Rf+\beta(Rm-Rf)+S1+S2 \quad (4)$$

где:

R – требуемая инвестором ставка дохода (на вкладываемый капитал);

Rf – «безрисковая» ставка дохода;

β – коэффициент бета (характеризует доходность отрасли);

Rm – общая доходность рынка в целом;

S1 – премия для малых предприятий;

S2 – премия за риск, характерный для конкретного проекта (компании).

При расчетах в твердой валюте в качестве «безрисковой» ставки можно взять ставку ЛИБОР. Коэффициент β для отрасли, в которой предполагается использовать ОИС (реализовать проект), и показатель общей доходности рынка (среднерыночного портфеля ценных бумаг) Rm следует взять из опубликованных данных или запросить у специализированной фирмы (за плату). Последние два слагаемых определяются скорее на основе суждений, чем объективных данных. Проще всего их вообще не учитывать. Однако в результате отбрасывания этих двух слагаемых полученная ставка дисконта с большой вероятностью окажется заниженной, а установленная на ее основе инвестиционная стоимость нематериальных активов – завышенной. Поэтому в качестве премии для малых предприятий можно добавить 1% – 5% в зависимости от реальных масштабов проекта. Премия за риск,

характерный для конкретного проекта, также не должна превышать 5%. Для расчета значения последних двух показателей приняты равными – 5%.

Безрисковая ставка принята на уровне – 7 % (депозиты СБ РФ от 10 000 рублей). Величина Rm по мнению многих практикующих российских оценщиков близка к 17,00%. Коэффициент β различается по отраслям промышленности и колеблется вокруг единицы. Однако, учитывая повышенный риск инвестиций в нематериальные активы, эта величина существенно больше единицы. Строго говоря, ее следует определять индивидуально для каждого конкретного случая, но в среднем для оценки ИС в научно-технической сфере она скорее близка к двум, чем к единице. Дополнительным основанием для такого вывода может служить следующее очень приблизительное рассуждение. Типичные ставки дисконта при оценке нематериальных активов составляют 20% - 40% при средней по промышленности около 17%. Если принять S1+S2≈10% и Rf=7%, отсюда легко получается «среднее американское» значение β≈2 для нематериальных активов. При всей приблизительности этого рассуждения оно дает больше, чем отсутствие всяких данных. Принимая то же среднее значение β=2 для российской научно-технической сферы, получим примерное значение ставки дисконта.

R=Rf+β(Rm-Rf)+S1+S2=7,0%+2(17,0%-7,0%)+5%+5%=37,00 %

Поскольку доход от использования объекта оценки постоянный, то темп роста доходов принят равным 0,0 %. Таким образом, ставка капитализации составляет 37,00 %.

2.2.6.3 Расчет рыночной стоимости объекта оценки

Таблица 3. Расчет стоимости методики методом освобождения от роялти (модификация с капитализацией прибыли до налогообложения)

Наименование показателя	Значение
Чистый доход предприятия, млн. руб. в год	37,50
Ставка роялти, %	2,00
Доход авторов (разработчиков, исполнителей), млн. руб.	0,75

Ставка капитализации, %	37,00
Рыночная стоимость методик и алгоритмов ИПК на дату оценки, тыс. руб.	1 837,00

Таким образом, рыночная стоимость методики на дату оценки 31 октябрь 2012 г., полученная в рамках доходного подхода округленно составляет ***1 837 000,00 рублей.***

Таблица 4. Расчет стоимости программных пакетов методом освобождения от роялти (модификация с капитализацией прибыли до налогообложения)

Наименование показателя	Значение
Чистый доход предприятия, млн. руб. в год	15,00
Ставка роялти, %	2,00
Доход авторов (разработчиков, исполнителей), млн. руб.	0,3
Ставка капитализации, %	37,00
Рыночная стоимость программных пакетов ИПК на дату оценки, тыс. руб.	764,00

Таким образом, рыночная стоимость программных пакетов на дату оценки 31 октября 2012 г., полученная в рамках доходного подхода, округленно составляет ***764 000,00 рублей.***

Таким образом, рыночная стоимость результатов исследования на дату оценки 31 октября 2012 г. составляет

1 837 000,00 + 764 000,00 = 2 601 000,00 рублей

2.2.6.4 Итоговое заключение по оценке

Стоимость пакета ОИС, определенная в рамках настоящей работы, составляет:

по методу исходных затрат – 7 092 540 рублей;

по методу освобождения от роялти – 2 601 000 рублей.

Таблица 5. Расчет достоверности подходов

№ п/п	Критерии	Затратный подход, %	Сравнительный подход, %	Доходный подход, %
	Достоверность информации	70	0	40
	Полнота информации	40	0	40
	Способность учитывать действительные намерения покупателя и продавца	45	0	55
	Способность учитывать конъюнктуру рынка	45	0	55
	Допущения принятые в расчетах	60	0	50
	Сумма баллов (01+02+03+04+05)	260	0	240
	Удельные весовые показатели, % (06/5)	52	0	48

В общем случае сведение результатов производится по формуле:

$$V = V_1 Q_1 + V_2 Q_2 + V_3 Q_3 , \quad (5)$$

Где V – обоснованная рыночная стоимость объекта оценки, руб.;

V_1 - стоимость объекта, определённая затратным подходом, руб. ;

V_2 - стоимость объекта, определённая сравнительным подходом, руб.;

V_3 - стоимость объекта, определённая доходным подходом, руб. ;

Q_1 - средневзвешенное значение достоверности затратного подхода;

Q_2 - средневзвешенное значение достоверности сравнительного подхода;

Q_3 - средневзвешенное значение достоверности доходного подхода;

Таким образом, рыночная стоимость объекта оценки равна:

7 092 540 х 0,52 + 2 601 000 х 0,48 = 4 936 600,08

4 936 600,08 рублей

(Четыре миллиона девятьсот тридцать шесть тысяч шестьсот рублей 8 копеек)

2.2.6.5 Рекомендации и предложения по использованию результатов проведенной НИР в реальном секторе экономики.

В процессе работы над проектом были разработаны следующие продукты: метод оценки статистической значимости сравнительно-геномного анализа, метод предсказания биологических состояний на узлах эволюционного дерева, метод предсказания регуляторных модулей в геномах эукариот. Все разработанные методы основаны на изучении эволюции биологических свойств.

Разработанные подходы и программы имеют определенный потенциал в медицине, а точнее в персонализованной медицине. В настоящее время в медицине все большее развитие принимают методы, связанные с генотипированием. Программа предсказания регуляторных модулей может найти свое применение при анализе генотипов, в частности для локализации регуляторных мутаций, повышающих риск генетических полигенных заболеваний. С другой стороны, сравнительно-геномный анализ может позволить оценить адекватность использования тех или иных модельных организмов. Например, отличие структуры регуляторных модулей у некоторых генов в мыши может сделать этот модельный организм неадекватной моделью для изучения соответствующего заболевания, сопряженного с экспрессией соответствующего гена. Такого рода анализ необходимо проводить с использованием набора различных модельных организмов, с тем, чтобы выделить адекватные модельные организмы.

Разработанные подходы могут найти применение и в образовательном процессе. Использование достаточно продвинутых математических подходов, в частности для изучения эволюции, служит хорошим примером применения математики и информатики в исследовании биологических процессов, и может послужить стимулом к изучению математики у биологов и медиков.

Исследования геномов различных организмов предполагают аннотацию геномов.

Аннотация геномов предполагает в первую очередь предсказание генов. Однако, важным аспектом аннотации геномов является предсказание регуляции генов и определение факторов, которые эту регуляцию осуществляют. Предложенные подходы могут помочь понять взаимодействие белковых молекул, вовлеченных в регуляцию экспрессии генов. В последнее время все большее внимание уделяется изучению некодирующих РНК, которые называют «темной материей» генома. Эти некодирующие РНК экспрессируются в геноме, и при этом играют роль в регуляции экспрессии других генов на различных уровнях –структуры хроматина, сплайсинга, инициации трансляции и стабильности РНК. Изучение регуляции экспрессии генов некодирующих РНК предполагает поиск соответствующих регуляторных модулей, а также изучение консервативности их структуры и экспрессии.

В настоящее время происходит весьма интенсивное развитие экспериментальных процедур, позволяющих экспериментально определять все более точно различные особенности функционирования генома. При этом многие методы предсказания постепенно теряют свою актуальность. Сейчас невозможно предсказать, сколько времени еще задача предсказания регуляции экспрессии генов будет оставаться актуальной. Однако можно с уверенность сказать, что идеи, заложенные в разработанные нами методы сохранят свою актуальность в течение достаточно большого промежутка времени и можно надеяться, что некоторые наши разработки войдут в учебники биоинформатики.

ЗАКЛЮЧЕНИЕ

Краткие выводы по результатам выполнения этапа НИР

В результате проведенных работ были разработаны новые методы и программы анализа биологических последовательностей, а именно:

– Методы оценки статистической значимости сравнительно-геномного анализа с использованием диффузионной модели в пространстве параметров. Предложены статистики и проведены исследования их распределений на случайных последовательностях.

– Метод предсказания биологических состояний на основе наблюдений, основанный на анализе скрытых Марковских моделей. Разработаны оригинальные алгоритмы, значительно превышающие по точности предсказания состояний все существующие алгоритмы. Алгоритмы реализованы в виде программ и библиотек классов.

– Разработан новый метод предсказания регуляторных модулей в геномах эукариот. Метод реализован в виде программного комплекса CORRECLUST. Программный комплекс по качеству предсказания превышает существующие программы, а также позволяет выявлять особенности структуры (грамматики) регуляторных модулей. Программный комплекс реализован в виде клиент-серверной системы, а также в виде отдельной (локальной) программы.

Оценка полноты решений поставленных задач

Все цели и задачи, предусмотренные Техническим заданием и Календарным планом на этапе №4, решены в полном объеме.

Рекомендации по использованию результатов научно-исследовательской работы

Результаты НИР могут быть использованы при аннотации геномов, при

диагностических исследованиях, и других исследованиях в области молекулярной биологии. Разработанные подходы к сравнительно-геномному анализу могут лечь в основу решения широкого круга задач исследования биологических последовательностей. Учитывая сложность объекта исследований представляется наиболее перспективным использование разработанных программ а также накопленного опыта при предоставлении консалтинговых услуг биотехнологическим , медицинским и исследовательским организациям, а также в собственных исследованиях.

Оценка технико-экономической эффективности внедрения

Расчеты и проведённые исследования, приведённые в отчёте, позволяют сделать вывод, что стоимость представленного на оценку объекта интеллектуальной собственности по состоянию на 31 октября 2012 года составляет

4 936 600,08 рублей

(Четыре миллиона девятьсот тридцать шесть тысяч шестьсот рублей 8 копеек)

Список использованных источников

1. Risken..1989.The Fokker-Planck Equation Method of Solution and Applications. Springer-Verlag Berlin Heidelberg,.
2. Vladar,Barton.2009.Statistical mechanics and the evolution of polygenic quantitative traits..Genetics.181:997--1011,
3. Smith.2000.The concept of information in biology. Philosophy of Science.67(2):177--194
4. Lassig,. Mustonen.2010.Fitness flux and ubiquity of adaptive evolution.PNAS.107(9):4248--4253
5. Fisher.1930.The genetical theory of natural selection. Clarendon.
6. Kimura.1964.Diffusion models in population genetics. Journal of Applied Probability.1(2):177--232
7. Kimura.1985.The neutral theory of molecular Evolution. Cambridge University Press.
8. Wright.1988.Surfaces of selective value revisited. Am.Nat.131(1):115--123
9. Burger.1999.Evolution of genetic variability and the advantage of sex and recombination in changing environment.Genetics.153:1055--1069
10. Lande.1976.Natural selection and random genetic drift in phenotypic evolution. Evolution.30(2):314--334
11. Waxman, Broom, Tang.2003.Mathematical analysis of a model describing evolution of an asexual population in a changing environment. Mathematical Biosciences.186:93--108
12. Hansen.1997.Stabilizing selection and the comparative analysis of adaptation. Evolution.51:1341--1351
13. Butler, King.2004.Phylogenetic comparative analysis: A modeling approach for adaptive evolution. The American Naturalist.164:683--695
14. Haetl, Bedford.2009. Optimization of gene expression by natural selection.PNAS.106:1133--

1138

15. Fontana, Stadler, Bonhoeffer, Tacker, Schuster, Hofacker .1994. Fast folding and comparison of rna secondary structures. Chemical Monthly. 125:167—188

16. Stefan Ekman, Heidi L. Andersen and M and Wedin, The Limitations of Ancestral State Reconstruction and the Evolution of the Ascus in the Lecanorales (Lichenized Ascomycota), Syst. Biol. 57(1):141156, 2008

17. Stephen H Montgomery, Isabella Capellini, Robert A Barton, Nicholas I Mundy, Reconstructing the ups and downs of primate brain evolution: implications for adaptive hypotheses and Homo floresiensis, BMC Biology 2010, 8:9

18. Mark Pagel , Andrew Meade and Daniel Barker, Bayesian Estimation of Ancestral Character States on Phylogenies, Syst. Biol. 53(5):673684, 2004

19. Fredrik Ronquist, Bayesian inference of character evolution, Trends Ecol Evol. 2004 Sep;19(9):475-81.

20. Mark Pagel, The Maximum Likelihood Approach to Reconstructing Ancestral Character States of DiscreteCharacters on Phylogenies, Systematic Biology, Vol. 48, No. 3 (Sep., 1999), pp. 612-622

21. Дурбин Р., Эдди Ш., Крог А., Митчисон Г.Анализ биологических последовательностей. ISBN 5-93972-559-7.РХД, 2006 г., 480 стр.

22. Eyre-Walker A., Problems with parsimony in sequences of biased base composition, J Mol Evol. 1998 Dec;47(6):686-90.

23. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011 Dec 9;147(6):1408-19.

24. Frith, M.C., Li, M.C., Weng, Z., 2003. Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31, 3666–3668.

25. Sandelin, A., Wasserman, W.W., 2005. Prediction of nuclear hormone receptor response elements. Mol. Endocrinol. 19, 595–606.

26. Rajewsky, N., Vergassola, M., Gaul, U., Siggia, E.D., 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics 3, 30–30.

27. Noto, K., Craven, M., 2007. Learning probabilistic models of cis-regulatory modules that represent logical and spatial aspects. Bioinformatics 23, e156 –e162.

28. Makeev, V.J., Lifanov, A.P., Nazina, A.G., Papatsenko, D.A., 2003. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res 31, 6016–6026.

29. Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., Taipale, J., 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124, 47–59.

30. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E., 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34, D108–110.

31. Papatsenko, D., Goltsev, Y., Levine, M., 2009. Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res., 37, 5665–5677.

32. Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 77, 257–286.

33. Kulp, D., Haussler, D., Reese, M.G., Eeckman, F.H., 1996. A generalized hidden Markov model for the recognition of human genes in DNA. Proc Int Conf Intell Syst Mol Biol 4, 134–142.

34. Lukashin, A.V., Borodovsky, M., 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res., 4, 1107-1115.

35. Fickett, J.W., 1996. Coordinate positioning of MEF2 and myogenin binding sites. Gene 172, GC19–32.

36. Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in

the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist. 41(1), 164–171.

37. Fariselli, P., Martelli, P.L., Casadio, R., 2005. A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins. BMC Bioinformatics 6 Suppl 4, S12.

38. Barker D, Meade A, Pagel M. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics. 2007 Jan 1;23(1):14-20.

проект

УТВЕРЖДАЮ

Директор ИППИ РАН

академик РАН, проф.

_____А. П. Кулешов

« 1 » ноября 2012 г.

Техническое задание на ОКР по теме:

«Разработка автоматизированной системы для сравнительно-геномного анализа регуляции экспрессии генов».

Основание для проведения ОКР

Договор на выполнение ОКР № _____ от «___»_____ 201_.

Исполнитель ОКР

Определяется по итогам конкурса.

Цель выполнения ОКР

Разработка программного комплекса для предсказания и анализа регуляторных модулей в геномах эукариот с использованием сравнительно-генмоного анализа CORECLUST.

Назначение продукции

Программный комплекс предназначен для анализа полных геномов эукариот с целью поиска регуляторных участков.

Программный комплекс должен обеспечивать:

1. Поиск потенциальных регуляторных областей в геномах с использованием информации о структуре сайтов связывания с использованием различных моделей сайтов связывания.
2. Проводить сравнительно-геномный анализ предсказанных регуляторных областей и давать статистическую оценку значимости предсказаний.

Технические требования к программному комплексу

1 Состав продукции

1.1 В состав разрабатываемого Комплекса должны входить:

1.1.1	Программный модуль обучения параметров модели регуляторных модулей;
1.1.2	Программный модуль поиска регуляторных модулей в геноме.
1.1.3	Программный модуль сранительно-геномного анализа найденных регуляторных модулей

1.1.4	Вновь разрабатываемый программный модуль сбора, просмотра и анализа результатов расчетов;
1.1.5	Вновь разрабатываемая база данных для хранения геномов, дополнительной информации и результатов расчетов;
1.1.6	Эксплуатационная документация;

Программные модули 1.1.1-1.1.3, разработанные в рамках ГК №07.514.11.4007, являются законченными программными компонентами и не требуют доработки.

2 Требования к функциональным характеристикам

2.1 Требования к составу выполняемых функций

Программный комплекс должен обеспечивать возможность выполнения перечисленных ниже функций:

2.1.1. хранить в реляционной базе данных модели сайтов связывания для различных транскрипционных факторов;

2.1.2. строить модели сайтов связывания по данным exo-ChIP-seq;

2.1.3. искать потенциальные сайты связывания в заданных областях геномов (областях заданного размера перед аннотированными генами, экспериментально определенными транскриптами, в доступных областях хроматина, по всему геному, и пр.) с использованием моделей сайтов связывания.

2.1.4. проводить кластеризацию найденных сайтов связывания с использованием различных моделей регуляторных модулей – модель кластеризации с учетом грамматики регуляторных модулей и без учета грамматики модулей

2.1.5. оптимизировать параметры моделей регуляторных модулей.

2.1.6. проводить сравнительно-геномый анализ предсказаний с учетом эволюционного дерева

2.1.7. допускать использование дополнительной экспериментальной информации о структуре хроматина, в том числе данные о метилировании ДНК, модификациях гистонов, доступности хроматина.

2.1.8. импортировать информацию о геномах из публичных баз данных, в частности из UCSC Genome Browser.

2.1.9. хранить результаты своей работы в реляционной базе данных

2.1.10. генерировать визуальное представление результатов

2.1.11. обеспечивать режимы работы в виде Java-приложения (с расширенными возможностями) и в виде веб-интерфейса (с сокращенными возможностями)

2.1.12. база данных должна быть реализована на основе свободно распространяемого серверного ПО с открытым исходным кодом (типа MySQL, PostgreSQL);

2.1.13. веб-сервер должен обеспечивать одновременную работу не менее 5 сессий.

2.1.14. время работы программы предсказания на типичных входных данных (5 геномов, 10 факторов связывания) не должно превышать 24 часа на одном процессоре.

2.2 Требования к организации входных данных

Входными данными разрабатываемого ПК должны являться:

2.2.1. полные аннотированные геномы в формате gbk

2.2.2. наборы последовательностей в формате FASTA

2.2.3. позиционные весовые матрицы в формате TRANSFAC

2.2.4. таблицы ортологов в табличном формате

2.2.5. треки в формате Genome Browser

2.3 Требования к организации выходных данных

Выходными данными разрабатываемого ПК должны являться:

2.3.1. Списки ко-регулируемых генов в табличном представлении с указанием

статистической значимости

2.3.2. Графическое представление карт расположения предсказанных регуляторных модулей (экранное представление и графические файлы в формате png)

2.3.3. Карты расположения регуляторных модулей в формате треков Genome Browser

2.3.4. отчеты о диагностике состояния системы и сообщения о всех возникших ошибках

2.4 Требования к временным характеристикам

Разрабатываемый Комплекс должен обеспечивать следующее время выполнения:

		Время работы	
	Задача	Одно процессорная система	Кластер из 16 ядер
2.4.1	Поиск по 5 геномам млекопитающих с использованием базы данных в качестве хранилища геномов	24 ч	4 ч
2.4.2	Поиск по 5 геномам млекопитающих с использованием файлов качестве хранилища геномов	4 ч	0.5 ч
2.4.3	Обучение параметров по одному ортологичному ряду	180 с	60 с

2.5 Требования к информационным структурам и методам решения

Программный комплекс должен быть создан на платформе Java. Веб-интерфейс создается с использованием технологий jsp и gwt. Хранилища данных существуют в двух формах – отдельные файлы для обеспечения скорости работы программ поиска регуляторных модулей и реляционных баз данных для хренения геномов, сопровождающей информации (аннотаций), результатов предсказаний, дополнительной информации об экспериментальных данных. База данных работает под управлением MySQL. Используется многопоточный доступ к базе данных.

3 Требования к надёжности

3.1 Время восстановления после отказа

1) Время восстановления после отказа, вызванного сбоем электропитания технических средств (иными внешними факторами), не фатальным сбоем (не крахом) операционной системы, при условии соблюдения условий эксплуатации технических и программных средств, 60 минут, не больше.

2) Время восстановления после отказа, вызванного неисправностью технических средств, фатальным сбоем (крахом) операционной системы, не должно превышать времени, требуемого на устранение неисправностей технических средств и переустановки программных средств.

Надежное (устойчивое) функционирование программы должно быть обеспечено выполнением Заказчиком совокупности организационно-технических мероприятий, перечень которых приведен ниже:

а) Технические средства должны быть обеспечены системой бесперебойного питания с запасом времени работы, достаточным для корректного закрытия баз данных, файлов и операционной системы.

б) использованием лицензионного программного обеспечения;

в) регулярным выполнением рекомендаций Министерства труда и социального развития РФ, изложенных в Постановлении от 23 июля 1998 г. Об утверждении межотраслевых типовых норм времени на работы по сервисному обслуживанию ПЭВМ и оргтехники и сопровождению программных средств»;

г) регулярным выполнением требований ГОСТ 51188-98. Защита информации. Испытания программных средств на наличие компьютерных вирусов

д) контролем входной и выходной информации

е) работой конечного пользователя без предоставления ему административных привилегий

3.2 Критерии отказа и предельного состояния разрабатываемого программного комплекса:

1) Отказом разрабатываемого программного комплекса считают прекращение или некорректное выполнение Комплексом функций, заданных требованиями п. 5.2.1 настоящего технического задания;

2) Предельным состоянием разрабатываемого Комплекса считают :

а) Работу более, чем с 20 геномами

б) Использование в одной задаче более 20 транскрипционных факторов

3.3 Отказы из-за некорректных действий пользователей системы

Отказы программы вследствие некорректных действий пользователя при взаимодействии с программой через Веб интерфейс недопустимы.

4 Условия эксплуатации

4.1 Климатические условия эксплуатации

Климатические условия эксплуатации, при которых должны обеспечиваться заданные характеристики, должны удовлетворять требованиям, предъявляемым к техническим средствам в части условий их эксплуатации.

4.2 Требования к видам обслуживания

Виды и периодичность обслуживания разрабатываемого Комплекса должны соответствовать требованиям ГОСТ Р 51188-98 и рекомендациям Министерства труда и социального развития РФ, изложенным в Постановлении № 28 от 23 июля 1998 г. «Об утверждении межотраслевых типовых норм времени на работы по сервисному обслуживанию ПЭВМ и оргтехники и сопровождению программных средств».

4.3 Требования к квалификации и численности персонала

Минимальное количество персонала, требуемого для работы программы, должно составлять не менее 2 штатных единиц — системный администратор и конечный пользователь программы — оператор. Системный администратор должен иметь высшее

профильное образование и сертификаты компании-производителя операционной системы.
В перечень задач, выполняемых системным администратором, должны входить:

а) задача поддержания работоспособности технических средств;

б) задачи установки (инсталляции) и поддержания работоспособности системных программных средств — операционной системы;

в) задача установки (инсталляции) программы.

г) задача создания резервных копий базы данных.

5 Требования к составу и параметрам технических средств

В состав технических средств должен входить как минимум один сервер и одна рабочая станция.

Требования с серверу:

а) Не менее четырех четырехядерных процессоров 64 бит (x64) с тактовой частотой не менее 2 Ггц,;

б) оперативную память объемом, 4 Гигабайт, не менее;

в) 4 Тбайт дисковая система с зеркалированием

г) под управлением Linux-подобной операционной системой

д) установленный MySQL server

е) Установленные Java машина, веб-сервер Apache, Tomcat

Требования к рабочей станции:

а) Стандартный персональный компьютер IBM или Mac

б) Установленная Java машина (для использования в режиме Java-приложения

в) Веб-browser один из: Sefari, Firefox, Chrome, Opera для работы в режиме Веб-интерфейса. Browser должен допускать Cookies.

Состав и характеристики технических средств, необходимых для обеспечения функционирования разрабатываемого Комплекса, должны быть окончательно определены на этапе Технического проектирования.

6 Требования к информационной и программной совместимости

6.1 Требования к информационным структурам и методам решения

Программный комплекс должен быть создан на платформе Java. Веб-интерфейс создается с использованием технологий jsp и gwt. Хранилища данных существуют в двух формах – отдельные файлы для обеспечения скорости работы программ поиска регуляторных модулей и реляционных баз данных для хренения геномов, сопровождающей информации (аннотаций), результатов предсказаний, дополнительной информации об экспериментальных данных. База данных работает под управлением MySQL. Используется многопоточный доступ к базе данных.

Структура баз данных

Структура базы данных должна допускать импорт геномов, их аннотаций, моделей сайтов связывания, эволюционных деревьев, а также разнородных экспериментальных данных. Кроме того, структура базы данных должна обеспечивать быстрый экспорт геномных последовательностей для обработки.

Требования к запросам пользователей данных из базы

Пользователи и администраторы работают с базой данных через Веб интерфейс либо через специальные административные средства. Администраторы системы должны иметь возможность редактировать таблицы, создавать новые таблицы с заранее определенной структурой для хранения экспериментальных данных (треков). Программы обработки (поиска регуляторных модулей, сравнительно-геномного анализа) могут создавать временные таблицы в базе данных. Пользовательские таблицы, созданные программами обработки, хранятся в базе данных ограниченный период времени. Этот параметр настраивается. По умолчанию период хранения результатов обработки составляет две недели.

Требования к исходным кодам и языкам программирования

В качестве основной платформы используется Java. Дополнительная статистическая обработка осуществляется средствами R. Веб-интерфейс создается средствами jsp и gwt.

Требования к программным средствам, используемым программой

Системные программные средства, используемые программой на сервере, должны быть представлены Linux-подобной операционной системой. Серверная а также клиентская часть при работе в режиме приложения поддерживаются системой Java SE Development Kit версии не ниже 6. БД должна находится под управлением СУБД MySQL версии не ниже 5.5.

Требования к защите информации и программ

Требования к защите информации и программ не предъявляются.

6.2 Специальные требования

Программа должна обеспечивать одновременную работу не менее 5 ользователей в режиме Веб интерфейса либо Java-приложения.

7 Требования к упаковке и маркировке

7.1 Требования к упаковке

7.1.1 Упаковка носителей с программным комплексом должна проводиться в закрытых вентилируемых помещениях при температуре от плюс 15 до плюс 40 °С и относительной влажности не более 80 % при отсутствии агрессивных примесей в окружающей среде.

7.1.2 Подготовленные к упаковке носители с программным комплексом укладывают в тару, представляющую собой коробки из картона гофрированного (ГОСТ 7376-89 или ГОСТ 7933- 89) согласно чертежам предприятия-изготовителя тары. Коробка должна быть перевязана голубой лентой.

7.1.3 Эксплуатационная документация должна быть уложена в потребительскую тару вместе с носителями.

7.1.4 Габариты грузового места должны быть не более 200х200х100 мм. Масса НЕТТО - не более 1 кг. Масса БРУТТО - не более 1 кг.

7.2 Требования к маркировке

7.2.1 Носитель информации должен иметь маркировку с обозначением товарного знака компании-разработчика, наименования, номера версии, порядкового номера, даты изготовления.

7.2.2 Маркировка должна быть нанесена на коробку в виде наклейки, выполненной полиграфическим способом с учетом требований ГОСТ 9181-74.

8 Требования к транспортированию и хранению

8.1 Разрабатываемый программный комплекс должен транспортироваться в упаковке в салоне автомобильного, крытых вагонах или контейнерах железнодорожного или морского транспорта, а также в герметичных отсеках авиационного транспорта на расстояние:

– воздушным транспортом на любое расстояние;

– железнодорожным транспортом до 10000 км;

– автомобильным транспортом до 1000 км со скоростью не более 60 км/час по шоссейным дорогам с твердым покрытием и до 500 км со скоростью не более 20 км/час по грунтовым дорогам.

8.2 Условия транспортирования:

– температура окружающей среды: от минус 50 до 50 °С;

– относительная влажность до 95 % при температуре 30 °С;

– атмосферное давление от 84 до 107 кПа (от 630 до 800 мм рт. ст.);

– воздействие ударных нагрузок многократного действия с пиковым ускорением не более 15g (147 м/с^2) при длительности действия ударного ускорения 10–15 мс

8.3 Гарантийный срок хранения разрабатываемого программного комплекса в заводской упаковке в отапливаемом помещении - не менее 3 лет.

8.4 Допускается транспортировка программного комплекса по компьютерным сетям.

9 Требования по стандартизации и унификации

9.1 Разрабатываемые компоненты разрабатываемого Комплекса должны обеспечивать унификацию функциональных задач, операций и интерфейсов.

9.2 Экранные формы разрабатываемого ПК должны проектироваться с учетом требований унификации:

1) все экранные формы пользовательского интерфейса должны быть выполнены в едином графическом дизайне, с одинаковым расположением основных элементов управления и навигации;

2) для обозначения сходных операций должны использоваться сходные графические значки, кнопки и другие управляющие (навигационные) элементы. Термины, используемые для обозначения типовых операций (добавление информационной сущности, редактирование поля данных), а также последовательности действий пользователя при их выполнении, должны быть унифицированы;

3) внешнее поведение сходных элементов интерфейса (реакция на наведение указателя «мыши», переключение фокуса, нажатие кнопки) должны реализовываться одинаково для однотипных элементов. Система должна соответствовать требованиям эргономики и профессиональной медицины при условии комплектования высококачественным оборудованием (ПЭВМ, монитор и прочее оборудование), имеющим необходимые сертификаты соответствия и безопасности Росстандарта.

9.3 Разрабатываемые компоненты разрабатываемого Комплекса должны обеспечивать унификацию функциональных задач, операций и интерфейсов.

10 Требования к документации

10.1 Предварительный состав программной документации

10.1.1 Состав программной документации приведен в приложении 1 «Перечень технической документации, разрабатываемой в рамках государственного контракта»

10.1.2 Техническая (конструкторская, технологическая, программная, эксплуатационная, ремонтная) документация должна соответствовать требованиям стандартов ЕСКД, ЕСТД,

ЕСПД.

10.1.3 Перечень технической и другой отчетной документации, подлежащей оформлению и сдаче Исполнителем Заказчику на этапах выполнения работ, определяется требованиями настоящего технического задания и требованиями Заказчика.

10.1.4 Техническая и другая отчетная документация представляется Заказчику в электронном виде.

11 Специальные требования

11.1 Требования к проведению испытаний

11.1.1 На всех этапах разработки разрабатываемого Комплекса должна производиться оценка качества программных средств в соответствии с требованиями ГОСТ 28195-99.

11.1.2 Для подтверждения соответствия разрабатываемой продукции требованиям настоящего технического задания и нормативно-технической документации должны быть проведены следующие испытания опытного образца:

1) предварительные испытания с целью предварительной оценки соответствия опытного образца продукции требованиям настоящего ТЗ, а также для определения готовности опытного образца к приемочным испытаниям;

2) приемочные испытания с целью оценки всех определенных настоящим ТЗ характеристик продукции, проверки и подтверждения соответствия опытного образца продукции требованиям ТЗ в условиях, максимально приближенных к условиям реальной эксплуатации (применения, использования) продукции, а также для принятия решений о возможности промышленного производства и реализации продукции.

11.1.3 Для проведения испытаний должно быть изготовлено следующее количество опытных образцов разрабатываемого Комплекса:

1) для предварительных испытаний – 1 шт.;

2) для приемочных испытаний - 2 шт.

11.1.4 Предварительные испытания опытных образцов должны быть проведены по утвержденным программам и методикам исполнителя ОКР.

11.1.5 Приемо-сдаточные испытания должны проводиться на объекте Заказчика в оговоренные сроки. Приемо-сдаточные испытания программы должны проводиться согласно разработанной Исполнителем и согласованной Заказчиком Программы и методик испытаний. Ход проведения приемо-сдаточных испытаний Заказчик и Исполнитель документируют в Протоколе проведения испытаний

12 Технико-экономические показатели

12.1 Основные технико-экономические требования

12.1.1 Разрабатываемые программные средства должны обеспечить анализ регуляции экспрессии генов в геномах эукариот. Внедрение создаваемой в рамках ОКР научно-технической продукции должно обеспечивать следующие социально-экономические эффекты:

1. генерацию новых знаний, основанных на анализе геномной информации
2. поиск и изучение новых взаимодействий транскрипционных факторов и поиск новых маркеров генетически обусловленных заболеваний
3. в перспективе исследования использование настоящего продукта позволит строить персонализированную медицину

12.1.2 Разрабатываемая продукция должна быть ориентирована на применение в следующих областях:

1. исследования в области молекулярной биологии
2. исследования в области медико-генетической диагностики
3. исследования в области персонализированной медицины

13 Требования к патентной чистоте и патентоспособности

13.1 На этапах 1 и 4 должны быть проведены патентные исследования в соответствии с ГОСТ Р 15.011-96.

13.2 Патентная чистота на методы изготовления и конструктивные решения должна быть обеспечена в отношении Российской Федерации и стран, куда возможна поставка изделий, а также передача технической, информационной и другой документации.

14 Содержание, сроки выполнения этапов

14.1 Наименование этапов и выполняемые работы

Этап 1. Техническое предложение:

14.1.1.1 Разработка и согласование с Заказчиком "Комплектности технической документации, разрабатываемой в рамках договора" (далее Комплектность ТД).

14.1.1.2 Проведение патентных исследований в соответствии с ГОСТ Р 15.011

14.1.1.3 Разработка технического предложения, в том числе:

— проработка результатов предшествующих НИР;

— сравнительная оценка рассматриваемых вариантов;

— обоснование и выбор оптимального варианта (вариантов) технического решения (решений).

14.1.1.4 Разработка технической документации в соответствии с согласованной комплектностью.

14.1.1.5 Оформление документации технического предложения в соответствии с ГОСТ 19.102-77, его рассмотрение и утверждение на научно-техническом совете

Этап 2. Эскизный проект:

14.1.2.1 Разработка эскизного проекта, в том числе:

— обоснование и выбор оптимального варианта (вариантов) технического решения, принятие принципиальных решений;

— проведение расчетов и оценок;

— эскизное алгоритмирование;

— разработка и обоснование блок-схем и временных графиков функционирования;

— изготовление и испытания макетов;

14.1.2.2 Разработка технической документации в соответствии с согласованной комплектностью.

14.1.2.3 Оформление документации эскизного проекта в соответствии с ГОСТ 19.102-77, его рассмотрение и утверждение на научно-техническом совете.

Этап 3. Технический проект:

14.1.3.1 Разработка технического проекта, в том числе:

— разработка алгоритмов решения программных задач;

— разработка программных решений разрабатываемого Комплекса;

— уточнение структуры входных и выходных данных;

— определение форм представления входных и выходных данных;

— разработка структуры программных компонентов и Комплекса в целом;

— определение конфигурации технических (аппаратных) средств функционирования ПО (ПК);

— программная реализация и испытание макетов (прототипов);

— оценка технологичности изготовления.

14.1.3.2 Разработка технической документации в соответствии с согласованной комплектностью.

14.1.3.3 Оформление документации технического проекта в соответствии с ГОСТ 19.102-77, его рассмотрение и утверждение на научно-техническом совете.

Этап 4. Разработка рабочей программной документации и изготовление опытного образца:

14.1.4.1 Программная реализация программного комплекса CORECLUST.

14.1.4.2 Разработка программной документации на программного комплекса CORECLUST

14.1.4.3 Программная отладка разрабатываемого программного комплекса CORECLUST

14.1.4.4 Разработка проектов эксплуатационной документации

14.1.4.5 Экспертиза разработанной рабочей программной документации

14.1.4.6 Разработка программы и методик предварительных испытаний

14.1.4.7 Разработка технической документации в соответствии с согласованной комплектностью

14.1.4.8 Изготовление (компиляция) опытного образца

Этап 5. Проведение предварительных и приемочных испытаний:

14.1.5.1 Проведение предварительных испытаний опытного образца программного комплекса CORECLUST

14.1.5.2 Корректировка РПД, ЭД программного комплекса CORECLUST по результатам предварительных испытаний, присвоение РПД литеры "О".

14.1.5.3 Перекомпиляция опытного образца программного комплекса CORECLUST по результатам предварительных испытаний и корректировки РПД.

14.1.5.4 Разработка программы и методик приемочных испытаний.

14.1.5.5 Подготовка РПД опытного образца программного комплекса CORECLUST к приемочным испытаниям.

14.1.5.6 Проведение приемочных испытаний опытного образца программного комплекса CORECLUST

14.1.5.7 Проверка и оценка проектов ЭД

14.1.5.9 Перекомпиляция опытных образцов программного комплекса CORECLUST по

результатам приемочных испытаний и корректировки РПД.

14.2 Сроки исполнения и финансирование по этапам

Перечень документов, разрабатываемых на этапах выполнения ОКР, сроки исполнения и контрактная цена приведены в календарном плане (приложение № 2 к договору).

15 Порядок выполнения и приемки этапов ОКР

15.1 Работа должна выполняться в соответствии с требованиями ГОСТ 19.102-77 (ГОСТ Р 15.201-2000).

15.2 Место проведения предварительных и приемочных испытаний – определяется по результатам конкурса

Руководитель работы

Ведущий научный сотрудник

ИППИ РАН

Доктор биол.наук, проф.

Миронов А.А.

«15» ноября 2012 г.

Приложение 1 к ТЗ НИОКР

Перечень технической документации, разрабатываемой в рамках государственного контракта

1 техническое задание;

2 программа и методики испытаний;

3 руководство оператора (на комплекс в целом);

4 текст программы (на каждый модуль и на комплекс в целом).

5 Описание программы (на каждый модуль и на комплекс в целом)

6 Описание применения (на комплекс в целом)

7 Руководство программиста (на комплекс в целом)