Российская Академия Наук
Сибирское Отделение
Институт Систем Информатики

УДК 004.42, 004.6, 004.8,
510.52, 519.233, 519.254, 519.6

На правах рукописи

ВАЛЕЕВ Тагир Фаридович

АЛГОРИТМЫ И ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ ИССЛЕДОВАНИЯ ПРОЦЕССОВ ГЕННОЙ РЕГУЛЯЦИИ

Специальность 05.13.11 –
Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Новосибирск 2006

Работа выполнена в Институте систем информатики
имени А.П. Ершова СО РАН

Научные руководители: Мурзин Федор Александрович,
кандидат физико – математических наук.

Кель Александр Эдуардович,
кандидат биологических наук.

Официальные оппоненты: Воробьев Юрий Николаевич,
доктор физико-математических наук.

Молородов Юрий Иванович,
кандидат физико-математических наук.

Ведущая организация: Институт математики
имени С.Л. Соболева СО РАН

Защита состоится 15 декабря 2006 г. в 15 ч 00 мин на заседании диссертационного совета К003.032.01 в Институте систем информатики имени А. П. Ершова Сибирского отделения РАН по адресу:

630090, г. Новосибирск, пр. Акад. Лаврентьева, 6.

С диссертацией можно ознакомиться в читальном зале библиотеки ИСИ СО РАН (пр. ак. Лаврентьева, 6)

Автореферат разослан 8 ноября 2006 г.

Ученый секретарь
диссертационного совета К003.032.01,
к.ф.–м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

В последние десятилетия процесс разработки новых лекарств приобретает всё большую наукоёмкость. При этом многие этапы получения прототипа выполняются посредством вычислительной техники с помощью достижений биоинформатики. Благодаря этому, вероятность получения хорошего прототипа, который действительно станет лекарством, а не будет отброшен в ходе клинических испытаний, возрастает на один-два порядка. Как правило, это справедливо для лекарств, подавляющих деятельность болезнетворных микроорганизмов в организме человека. Однако для лечения многих вирусных и генетических заболеваний необходимо чёткое понимание регуляторных процессов в клетке и их нарушений, поиск общих законов генной регуляции и получение знаний о процессах, происходящих в определённых клетках в определённый момент.

К настоящему времени накоплен некоторый запас знаний о генной регуляции, к сожалению, недостаточный для её глубокого понимания. Однако имеется возможность создать модель регуляторных процессов в клетке с использованием этих знаний, и на основе экспериментальных данных проанализировать характер регуляции в определённых условиях, а именно, исследовать отличия в процессах регуляции между клетками больного и здорового организмов, изменение регуляции во времени и т. д.

При исследовании генной регуляции применяются особые методы проектирования и анализа алгоритмов и программ, специальные форматы данных, базы данных и знаний большого объёма, требующие специальных методик взаимодействия с ними, визуальные человеко-машинные интерфейсы.

Ввиду использования больших объёмов данных и высокой сложности методов их обработки, особенно актуальным является оценка качества работы создаваемых программ и разработка систем автоматического тестирования. Так как расчёты могут выполняться в течение значительного времени (несколько суток), потеря результатов расчётов в результате ошибки операционной системы или сбоя электропитания крайне нежелательна. Поэтому необходимо применять принципы построения программных систем, обеспечивающие устойчивость к сбоям.

Цель работы

Цель работы заключается в разработке и совершенствовании математических и программных методов для анализа работы регуляторной системы при взаимодействии с промоторами генов в общем случае и в отдельных частных экспериментах. Основные задачи в рамках достижения этой цели включают:

- Изучение и формализацию знаний, накопленных в области генной регуляции, описание поведения регуляторной системы в виде алгоритмов.
- Построение параметризованной модели регуляторного комплекса, включающей в себя набор транскрипционных факторов и характер их взаимодействия.
- Реализацию способа оптимизации (подбора параметров) этой модели для достижения наибольшего соответствия модели экспериментальным данным.
- Разработку программной системы, упрощающей и автоматизирующей обработку экспериментальных данных, связанных с генной регуляцией.
- Создание программного инструментария, который выполняет процесс оптимизации и представляет подробные результаты, а также позволяет гибко управлять видом модели и процессом её оптимизации и оценить качество результата.
- Проектирование и реализация методов тестирования программных систем на искусственных и экспериментальных данных для оценки качества полученных результатов.

Методы исследования

Для решения поставленной задачи нами использовались различные алгоритмы и математические методы. Было выбрано представление регуляторной системы, как процессора, который принимает в качестве входного потока данных последовательность нуклеотидов промотора, а на выходе выдаёт числовое значение экспрессии. В итоге задача была сведена к определению внутренней структуры этого процессора для известных наборов входных и выходных данных. В качестве основы для внутренней структуры процессора использовалась нечёткая логика. Для оптимизации комплекса был выбран генетический алгоритм. В качестве целевой функции применяются различные компоненты такие, как: линейная регрессия, оптимизация ошибок перепредсказания и недопредсказания, критерий нормальности, t-тест Стьюдента и др. Для оценки качества результата использовались также дополнительные статистические методы такие, как метод складного ножа. Для моделирования промотора при тестировании на искусственных данных использовались цепи Маркова. Для описания текстовой записи получаемых комплексов использовался формализм Бэкуса-Наура.

Научная новизна

В рамках работы формализованы и представлены в виде алгоритмов знания о генной регуляции. При этом взаимодействия наиболее сложного характера (композиционные элементы, взаимосвязи булева типа и т. д.) были формализованы впервые. Такие знания другими исследователями ранее не учитывались, что ограничивало возможности исследований. Некоторые из предложенных способов оценки качества модели и входных данных также ранее не рассматривались.

В работе предложен оригинальный способ выполнения генетического оператора «кроссовер» на основании функции сходства для объектов с достаточно сложной внутренней структурой непостоянной длины. Потомки, получаемые в результате кроссовера, не только содержат часть информации от каждого родителя, но и гарантировано наследуют их свойства, не получая при этом каких-либо свойств, отсутствующих у родителей. Для таких комплексных объектов нет общепризнанного хорошего способа скрещивания. Тем не менее, использованный нами способ хорошо зарекомендовал себя на практике и может быть использован для других задач.

Предложен способ построения функции сходства двух комплексов обобщённого класса, который также представляет некоторый научный интерес. Аналогичным образом можно построить функции для других объектов. Вычисление надёжности в рамках мультизапуска было придумано автором и может использоваться для других задач, решаемых методами неполной оптимизации.

Наконец, проделанная работа имеет некоторую ценность для биологов, т. к. способствует пониманию генной регуляции в клетке. Одновременно, это может помочь создать новые алгоритмы для решения сложных задач, базирующиеся на использованных природой идеях (как это было, к примеру, с генетическим алгоритмом). Также работа связана с исследованиями по построению компьютеров на биологической основе, биокомпьютеров. Понимание генной регуляции способствует прогрессу в данной области.

Практическая ценность

В результате работы созданы программные продукты CMA, ExPlain и Cissearch. Все они предназначены для обработки экспериментов, связанных с изучением генной регуляции, и могут использоваться для проектирования лекарств и лечения трудноизлечимых болезней. Изучение генной регуляции наиболее актуально для анализа работы иммунной системы, процессов апоптоза (естественной смерти клеток), что непосредственно связано с исследованием таких заболеваний как рак и СПИД и поиском путей их лечения. Исследования изменений в организме, происхо-

дящих из-за генетических заболеваний, также облегчаются с использованием разработанных программных систем.

Разработанные программные средства успешно использовались различными специалистами для обработки данных по генной регуляции включая представителей международных биотехнологических и фармацевтических компаний (например, Serono Group); научных институтов и центров (например, Германского центра изучения рака, DKFZ); компании, занимающейся выпуском биологических баз данных BIOBASE GmbH.

Апробация работы

Результаты работы докладывались на международных научных конференциях, включая 1st Intl. Conf. on Natural Computations (ICNC'05) в г. Чанша, Китай; German Conf. on Bioinformatics (GCB'05) в г. Гамбург, Германия; 3rd Annual RECOMB Satellite Workshop on Regulatory Genomics в г. Сингапур, Сингапур; 3rd Intl. Conf. "Genomics, Proteomics, Bioinformatics and Nanotechnologies for Medicine" в г. Новосибирск и др. Работа была представлена на рабочем семинаре «Наукоёмкое программное обеспечение» конференции памяти академика А. П. Ершова «Перспективы систем информатики», на различных встречах, семинарах. Система ExPlain демонстрировалась на пленарных докладах международных конференций, на встречах с представителями свыше десятка биотехнологических и фармацевтических компаний.

Автором опубликовано 23 печатные работы, из них по теме диссертации – 16 работ.

Структура и объем работы

Диссертационная работа состоит из введения, четырёх глав, заключения и списка литературы. Объем диссертации — 163 стр. Список литературы содержит 81 наименование. Работа включает 26 рисунков и графиков, полученных в результате расчётов на ЭВМ, в том числе с использованием разработанного программного обеспечения.

СОДЕРЖАНИЕ РАБОТЫ

Во введении описано современное положение в области производства лекарств и обоснована необходимость изучения генной регуляции. После этого вкратце описан программный инструментарий, разработанный в рамках работы (программы CMA, ExPlain и CisSearch), приведена структура диссертации и описан используемый в тексте язык описания алгоритмов. Также отмечены возможные применения разработок в дру-

гих областях: изучение новых вычислительных технологий; построение алгоритмов и программ, а также создание биокомпьютеров.

Первая глава посвящена первичному анализу регуляции генов; описанные в ней методы не учитывают взаимодействия транскрипционных факторов между собой. В разделе 1.1. введён ряд определений, которые требуются в дальнейшем. Нуклеотидом мы называем элемент алфавита $Q' = \{\mathrm{A,C,G,T,N}\}$, последовательностью нуклеотидов — слово в этом алфавите $w \in Q'^{*}$, а подцепочкой — тройку (w,s,d), задающую последовательность w, стрэнд (принадлежность к одной из двух цепочек ДНК) s и положение начала подцепочки на геноме d, характеризующееся в свою очередь названием гена и смещением подцепочки относительно старта транскрипции (позиции в гене, откуда начинается считывание РНК). Отметим, что регуляторные области, которые мы в основном изучаем, это промоторы, они расположены выше старта транскрипции (то есть смещение подцепочки отрицательно).

Далее вводится понятие транскрипционного фактора как белка, активирующего или подавляющего работу гена и понятие композиционного элемента как функционального модуля, состоящего из двух таких факторов, определённым образом расположенных относительно друг друга. Сайт связывания транскрипционного фактора — это подцепочка в регуляторной области, где происходит непосредственное взаимодействия фактора с ДНК.

Наличие в регуляторной области гена тех или иных сайтов обуславливает возможность взаимодействия этого гена с определёнными факторами и может рассматриваться как сигнатура гена. Несколько факторов обычно действуют в комплексе. Данный комплекс мы рассматриваем, как процессор, который анализирует сигнатуры генов, поступающие на вход, и выдаёт уровень активности (экспрессии) соответствующего гена.

Как правило, точную структуру комплекса и характер взаимодействия между отдельными факторами в нём экспериментально установить невозможно, однако можно измерить значения экспрессии генов. Поэтому решаемая задача сводится к построению модели комплекса и сравнению выдаваемых моделью значений экспрессии с полученными в результате эксперимента.

В разделе 1.2. описана задача предсказания сайтов связывания и возможные пути её решения. Экспериментально известно достаточно мало сайтов, однако замечено сходство последовательностей в сайтах, которые связываются с определёнными факторами. Это даёт возможность объединить похожие сайты в общий класс, построить их модель и на её основании предсказывать сайты, которые не были открыты экспериментально.

Модель представляет собой весовую матрицу $4 \times L$, где L — длина сайта; каждой позиции внутри сайта для нуклеотидов A, C, G, T соответствует вероятность появления данного нуклеотида в данной позиции сайта с некоторой перенормировкой. Имея весовую матрицу, для любой последовательности длины L можно вычислить степень соответствия (вес) между последовательностью и моделью; в случае, если вес превышает некоторый порог, последовательность можно считать предсказанным сайтом, соответствующим данной матрице.

Матрицы созданы заранее в полуавтоматическом режиме и доступны в биологических базах данных, обычно одному фактору соответствует несколько матриц. Для предсказания сайтов по заданному набору матриц (например, матриц, соответствующих факторам, связанным с работой иммунной системы позвоночных) и с заданными порогами, используются профайлы: списки матриц с порогами. Предсказанный сайт характеризуется своей подцепочкой, весовой матрицей, по которой он был предсказан, и его весом. В разделе рассмотрены алгоритмы и математический аппарат для формирования матриц и предсказания сайтов, их работа проиллюстрирована на примерах.

В разделе 1.3. приведён простой алгоритм для поиска факторов, действующих в данном эксперименте по предсказанным сайтам и значениям экспрессии для заданного набора промоторов. Промоторы с низкой экспрессией помещаются в категорию ‘No’, а промоторы с высокой — в категорию ‘Yes’, после чего находится отношение плотности предсказанных сайтов в этих категориях для каждой матрицы. Подобный алгоритм, но на уровне факторов, а не отдельных матриц, применяется в пакете Cis-Search. В разделе приведён пример экспериментальных данных, где для ряда матриц некоторых факторов такое отношение весьма заметно (то есть предсказанные сайты этих матриц встречаются в ‘Yes’ значительно чаще, чем в ‘No’), это даёт возможность предположить, что в данном эксперименте эти факторы действительно регулировали активность.

Вторая глава посвящена построению модели комплекса с учётом различной информации из предметной области, а также методам её оптимизации. В разделе 2.1. описана общая постановка задачи, решаемой программой CMA. Вводится понятие модели, как параметризованной функции $RS : Sl \to [0, 1]$, где Sl — множество предсказанных сайтов для некоторого промотора. Модель задаёт набор правил и определяет, насколько данный промотор (а точнее — список сайтов, предсказанных на нём) соответствует этим правилам. Приводится пример простейшей модели с одним параметром:

$$RS_X(Sl) = \begin{cases} 1, \textit{ если } \exists \textit{ сайт матрицы } X \textit{ в } Sl \\ 0, \textit{ иначе} \end{cases}$$

Здесь параметр X — имя матрицы. Такой модели соответствуют промоторы, на которых присутствует как минимум один сайт матрицы X. Комплексом мы называем модель с подставленными параметрами. К примеру, в данной модели заменив X на имя конкретной матрицы, мы получим комплекс. Далее говорится о степени соответствия комплекса некоторым экспериментальным данным. Для каждого промотора i определено значение экспрессии ξ_i и принадлежность категории c_i (обычно категорий две — 'Yes' и 'No'). Степень соответствия определяется с помощью целевой функции Z.

В разделах 2.2. и 2.5. описаны различные виды моделей, объединённые в классы. Класс моделей представляет комплекс в ещё более общем виде и имеет ряд параметров, задание которых превращает класс в конкретную модель. Задание параметров класса выполняется пользователем. В работе рассмотрены три класса, реализованные в программе CMA от простого к сложному: оконный, булев и обобщённый.

В оконном классе параметрами класса является количество матриц в модели N и размер окна l (в нуклеотидах). Параметрами модели являются идентификаторы N матриц $X_1, X_2, ..., X_N$ и соответствующие им пороги $C_1, C_2, ..., C_N$. Значением комплекса является делённое на N максимальное количество предсказанных сайтов различных матриц из набора $X_1, X_2, ..., X_N$ расположенных в окне ширины l, причём таких, что их вес превышает соответствующий порог. Рассматриваются детали реализации подсчёта такого комплекса и вопросы оптимизации. Модель оконного класса учитывает то что для работы комплекса необходимо, чтобы сайты связывания были расположены близко друг к другу.

Модель булева класса представляет собой булеву формулу вида

$$\neg\left(\bigcup_j E_{0j}\right) \cap \left(\bigcap_i \left(\bigcup_j E_{ij}\right)\right).$$

Модель принимает значения 0 или 1 в зависимости от истинности этого выражения. Здесь E_{ij} — предикат, принимающий истинное значение, если на данном промоторе есть сайт матрицы X_{ij} с порогом выше, чем C_{ij}. Модель позволяет описать такие явления, как наличие факторов, подавляющих активность гена и взаимозаменяемость некоторых факторов. Параметры класса определяют, насколько сложной может быть такая формула, а параметрами модели являются X_{ij} и C_{ij}.

Модель обобщённого класса наиболее сложная. Она объединяет возможности булевой и оконной модели. Кроме того, она позволяет учесть сайты как отдельных матриц, так и композиционных элементов, вклад нескольких сайтов одной матрицы в работу комплекса (наличие нескольких сайтов повышает вероятность связывания). Параметры класса

позволяют не фиксировать строго количество матриц в модели, а задавать диапазоны для матриц и композиционных элементов.

Модель представляется в виде набора подкомплексов, объединяемых таким же выражением, как и в булевом классе, однако здесь уже E_{ij} — нечёткая логическая величина, характеризующая соответствие промотора одному из подкомплексов. Подкомплекс содержит набор матриц и композиционных элементов. Для каждой матрицы в качестве параметров модели определен идентификатор X, порог C и параметр ν, показывающий, сколько сайтов данной матрицы учитывается.

Для каждого композиционного элемента определены две матрицы X_1, X_2; соответствующие им пороги C_1, C_2, диапазон допустимых расстояний между сайтами этих матриц $d_{\min}, d_{\max}$, допустимая взаимная ориентация сайтов в паре δ_1, δ_2 и параметр ν. Вес подкомплекса определяется по вкладу всех учитываемых сайтов, причём вклад зависит от весов сайта, а в случае сайтов композиционных элементов — ещё и от расстояния между сайтами. Все сайты, учитываемые в подкомплексе, должны быть расположены в окне ширины l, которая является параметром класса. Другие параметры класса могут так или иначе ограничивать структуру модели. Важными частными случаями являются модели, ограниченные одним подкомплексом, а также модели, подкомплексы которых могут содержать ровно одну матрицу или ровно один композиционный элемент.

Поиск оптимального комплекса происходит подбором параметров модели с помощью генетического алгоритма. В разделе 2.4. обоснован выбор этого алгоритма и описаны возможные альтернативы. В разделах 2.4. и 2.5. описано устройство генетических операторов создания, мутации и кроссовера для всех классов, а также некоторые детали их программной реализации.

Особый интерес представляют мутация и кроссовер для обобщённого класса моделей. Генетические операции основаны не на побитовом изменении комплексов, а учитывают их внутреннюю структуру. В процессе мутации с некоторой вероятностью происходит одно из нескольких событий: добавление или удаление матрицы, композиционного элемента или подкомплекса; объединение двух матриц в композиционный элемент; распад композиционного элемента на две матрицы; рекомбинация матрицы и композиционного элемента; изменение одного из параметров матрицы или композиционного элемента (порога, ν, ориентации или расстояний $d_{\min}, d_{\max}$).

Кроссовер реализован иерархически. В программе имеются операции для кроссовера двух матриц, матрицы и композиционного элемента, двух композиционных элементов, двух подкомплексов. При кроссовере двух подкомплексов выбираются наиболее похожие матрицы и компози-

ционные элементы и скрещиваются между собой, и этот процесс повторяется, пока в родительских подкомплексах имеется хотя бы одна матрица или композиционный элемент. Аналогичным образом можно реализовать кроссовер комплексов через кроссовер подкомплексов.

Целевая функция Z описана в разделах 2.3. и 2.6. Она состоит из ряда компонент, которые характеризуют: соответствие профиля вычисленных значений комплексов RS экспериментальным значениям экспрессии (Z_R); степень различия значений RS для промоторов категорий 'Yes' и 'No' по критерию Стьюдента (Z_T); количество ошибок перепредсказания и недопредсказания (Z_E); соответствие распределения значений RS для категории 'Yes' нормальному (Z_N); степень сложности комплекса (Z_P). В качестве значения Z_R используется величина R^2 отклонения зависимости между RS и ξ_i от линейной:

$$\beta = \frac{\sum \rho_i^2 \sum \xi_i - \sum \rho_i \sum \rho_i \xi_i}{n \sum \rho_i^2 - \left(\sum \rho_i\right)^2};\ \alpha = \frac{\sum \rho_i \xi_i - \beta \sum \rho_i}{\sum \rho_i^2}$$

$$Z_R = 1 - \frac{\sum \left(\alpha \rho_i + \beta - \xi_i\right)^2}{\sum \xi_i^2 - \left(\sum \xi_i\right)/n}$$

Здесь $\rho_i = RS\left(Sl_i\right)$, а n — общее число промоторов. Значение Z_T вычисляется по двухвыборочному t-тесту Стьюдента с разными дисперсиями. При вычислении Z_E находится значение RS^*, разделяющее полученные RS таким образом, что количество промоторов из 'No', для которых $RS < RS^*$, сложенное с количеством промоторов из 'Yes', для которых $RS \geq RS^*$, максимально возможно. После нормализации эта сумма и используется в качестве Z_E. Подсчёт Z_N основан на статистике, введённой д'Агостино, которая учитывает соответствие скоса $\frac{1}{\sigma^3 n_{Yes}} \sum \left(\rho_i - \overline{\rho_{Yes}}\right)^3$ и эксцесса $\frac{1}{\sigma^4 n_{Yes}} \sum \left(\rho_i - \overline{\rho_{Yes}}\right)^4$ значениям, характерным для нормального распределения (0 и 3 соответственно). Компонента Z_P введена для того, чтобы при прочих равных условиях предпочтение отдавалось более простому комплексу: значение этой компоненты падает при увеличении количества матриц и композиционных элементов в комплексе. Пользователь может задавать степень влияния каждой из компонент, либо отключать некоторые из них.

Помимо компонент, в целевую функцию может быть введён также ряд предикатов-ограничений. При несоблюдении любого из них, целевая функция принимает значение 0. Ограничения позволяют требовать обязательного действия комплекса (то есть достаточно большого значения RS) на некоторое подмножество промоторов, наличия в комплексе матриц из

определённого поднабора и т. д. Они отражают некоторые дополнительные данные, известные по эксперименту. К примеру, точно известно, что в данном процессе работал некоторый фактор, и представляет интерес узнать, какие ещё факторы работали с ним вместе, и каков характер взаимодействия между ними.

Третья глава посвящена анализу результатов, получаемых в результате работы генетического алгоритма, описанной во второй главе. Раздел 3.1. посвящён описанию вывода результатов, а также возможности вычисления целевой функции для заданного пользователем комплекса. В качестве результата помимо итогового значения целевой функции для найденного наилучшего комплекса выводится также вся популяция после последнего поколения, значения компонент целевой функции для лучшего комплекса, расчёт RS для каждого промотора, распределение значений RS и экспрессии (в графическом виде), гистограммы распределения RS в категориях ‘Yes’ и ‘No’ и прочее.

В разделе 3.2. описано три метода для анализа результатов. Первый метод — это так называемый мультизапуск: возможность запустить генетический алгоритм с одинаковыми параметрами несколько раз и сравнить результаты, а также оценить надёжность. Для сравнения результатов используется функция схожести $S(c_1, c_2)$, которая позволяет сравнить два комплекса c_1 и c_2 одной и той же модели. Она симметрична и принимает значения в диапазоне $[0, 1]$, причём $S(c_1, c_2) = 1 \Leftrightarrow c_1 = c_2$. Тогда если в n запусках получились оптимальные комплексы $c_1, c_2, ..., c_n$, то надёжность будет выражаться, как

$$R = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} S(c_i, c_j).$$

В разделе описывается вид функции схожести для комплексов различных классов и доказывается, что введённые функции удовлетворяют заявленным свойствам. Затем приведён пример использования мультизапуска для определения оптимальных параметров запуска генетического алгоритма: при недостаточном размере популяции и количестве поколений получается низкая надёжность. Также мультизапуск может быть полезен для уточнения параметров класса: в приведённом примере использована модель оконного класса с $N = 3$, и результаты позволяют предположить, что с модель с $N = 4$ даст более высокое значение целевой функции.

Второй метод — это запуск с кластеризацией набора промоторов. Идея заключается в том, что иногда несколько комплексов действуют в эксперименте одновременно, и можно попытаться найти их все. Для этого после нахождения наилучшего комплекса из всего набора промоторов P

формируется подмножество-кластер C промоторов, экспрессия которых лучше соответствует модели:

$$C = \left\{ i \in P : \left(\alpha\rho_i + \beta - \xi_i\right)^2 \le \frac{1}{n}\sum_{j\in P}\left(\alpha\rho_j + \beta - \xi_j\right)^2 \right\}.$$

После этого для оставшихся промоторов $P' = P \setminus C$ снова ищется наилучший комплекс и т. д., пока в кластера не войдёт как минимум 90% промоторов. Найденные комплексы и кластера выводятся пользователю.

Третий подход к анализу результатов основан на статистическом методе складного ножа. С его помощью в частности можно отследить факт переобучения. Выборка разбивается на два поднабора случайным образом, и после поиска оптимального комплекса на одном из них, он применяется ко второму и результат сравнивается. Операция повторяется заданное количество раз, и результаты выводятся пользователю.

Кроме этого, в третьей главе, в разделе 3.3 описан способ оценки значимости отдельных матриц, входящих в комплекс. Способ основан на следующем. Вкладу учитываемых сайтов приписывается различный весовой коэффициент (зависящий от матрицы сайта) таким образом, чтобы компонента целевой функции Z_R принимала максимально возможное значение. Полученные коэффициенты отражают важность матриц.

В четвёртой главе описаны некоторые детали программной реализации CMA и ExPlain, а также приведены результаты тестирования CMA на искусственных и экспериментальных данных. Программа CMA реализована в виде кроссплатформенного приложения командной строки на C++ (около 0.5 МБ кода). Результаты тестирования на искусственно сгенерированных промоторах, в которые были внедрены сайты определённых факторов и композиционных элементов, показали, что CMA успешно находит внедрённые факторы. Результаты тестирования на двух экспериментальных наборах данных, для которых правильный результат частично известен, показали, что комплекс, находимый CMA, хорошо согласуется с известными данными.

Система ExPlain представляет собой кроссплатформенное веб-приложение на языке программирования Perl (около 1.0 МБ кода). Это — оболочка, которая объединяет в себе различные виды анализа регуляторных процессов на основании экспериментальных данных. В частности, в неё включена возможность запуска CMA. Обрабатывать данные в ExPlain значительно удобнее, так как данные, полученные из эксперимента, можно непосредственно загрузить в ExPlain, используя идентификаторы генов из любых популярных баз данных. ExPlain он преобразует их, найдёт промоторы, соответствующие генам, и извлечёт соответствующие им последовательности из базы TRANSPRO, после чего запустит CMA и представит результаты в удобном виде с использованием средств разметки

HTML. Кроме того, ExPlain предоставляет широкие возможности для предварительной фильтрации данных, конструирования профайлов и комплексной обработки данных различными методами.

В заключении описаны основные результаты работы, вклад автора в проделанную работу и идеи дальнейшего развития проекта.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

- Проведены комплексные исследования известных на сегодняшний день механизмов генной регуляции, предложены их формализованные модели в виде набора алгоритмов.
- Выполнены исследования по установлению соответствия между предложенными моделями и экспериментальными данными. Для оценки степени этого соответствия введена целевая функция на основе различных статистических методов. Предложены и реализованы алгоритмы для поиска оптимального соответствия, методы оценки качества полученных моделей и средства для тестирования алгоритмов поиска.
- Реализована программа CMA, представляющая собой инструмент для гибкого анализа регуляторной модели на основании данных по экспрессии генов в различных экспериментах. Проведено тестирование программы на искусственных и экспериментальных данных; результаты тестирования подтвердили согласие предсказываемых CMA регуляторных комплексов с экспериментальными данными.
- Реализована программная система ExPlain, упрощающая процесс обработки экспериментальных данных по генной регуляции. Она поддерживает популярные форматы файлов, используемые для хранения результатов экспериментов, связывает их с широко распространенными базами данных биологической информации, используемыми биотехнологическими компаниями, позволяет выполнять различные виды анализа и предоставляет графический человеко-машинный интерфейс.

О ЛИЧНОМ ВКЛАДЕ АВТОРА

Работа осуществлялась группой специалистов различного профиля. Автором выполнена большая часть работы по формализации достаточно расплывчатых биологических знаний в виде конкретных математических моделей.

Структура и алгоритмы генетических операторов для созданных моделей были практически полностью разработаны автором. Некоторые оригинальные идеи, положенные в основу генетических операторов, хо-

рошо проявили себя при тестировании на искусственных данных. В разработке целевой функции вклад автора оценивается им самим в 30%: общая структура целевой функции была придумана другими участниками проекта, однако затем автор подверг её значительной переработке. Автору принадлежит идея и математическое описание оценки качества с помощью мультизапуска и функции сходства.

Программная реализация CMA (не включая использованную библиотеку GRESA) выполнена автором приблизительно на 80-85%, в частности продумана общая структура приложения; полностью реализованы булев и обобщённый класс моделей, средства проверки результата; значительно переработана и оптимизирована целевая функция и оконный комплекс; написан изначально цикл генетического алгоритма (который впоследствии менялся другими разработчиками); реализованы классы для обработки параметров, вывода результата и отображения графической информации.

Участие автора в программной реализации системы ExPlain составляет около 25%. Автор продумал общую структуру приложения, внёс много идей по интерфейсу, создал часть структуры базы данных и реализовал, отладил и протестировал множество отдельных функций.

Кроме этого, автор выполнил часть работы по тестированию, в том числе с использованием экспериментальных данных.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **Waleev, T., Shtokalo, D., Konovalova, T., Voss, N., Cheremushkin, E., Stegmaier, P., Kel-Margoulis, O., Wingender, E., Kel, A.** Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. // Nucleic Acids Research. — 2006. — Vol. 34(Web Server issue). — P. 541–545.

2. **Cheremushkin, E., Konovalova, T., Valeev, T., Kel, A.** Methods for search of gene regulatory elements binding sites. // Analytical Tools for DNA: Genes and Genomes: Nuts & Bolts. — DNA Press, 2005. — P. 185–214.

3. **Kel, A., Konovalova, T., Waleev, T., Cheremushkin, E., Kel-Margoulis, O., Wingender, E.** Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. // Bioinformatics. — 2006. — Vol. 22(10). — P. 1190–1197.

4. **Konovalova, T., Valeev, T., Cheremushkin, E., Kel, A.** Composite Module Analyst: Tool for Prediction of DNA Transcription Regulation. Testing on Simulated Data. // Proc. of the First International Conference on Natural Computations (ICNC'05), Changsha, China, Aug. 27–29, 2005. — Advances in Natural Computation, part 2 (LNCS 3611). — Springer, 2005. — P. 1202–1205.

5. **Kel, A., Konovalova, T., Valeev, T., Cheremushkin, E., Kel-Margoulis, O., Wingender, E.** Composite Module Analyst: A Fitness-Based Tool for Prediction of Transcription Regulation. // Proc. of the German Conference on Bioinformatics (GCB'05), Hamburg, Germany, Oct. 5–7, 2005. — Lecture Notes in Informatics. — 2005. — Vol. P–71. — P. 63–76.

6. **Cheremushkin, E., Konovalova, T., Valeev, T., Shtokalo, D., Taraskina, A.** CisSearch: Software Package For Complex Analysis Of Gene Regulatory Sequences. // Proc. of the 3rd Annual RECOMB Satellite Workshop on Regulatory Genomics, Singapore, Jul. 17–18, 2006. — Singapore, 2006. — P. 100–108.

7. **Cheremushkin, E., Konovalova, T., Valeev, T., Shtokalo, D., Taraskina, A.** Software Package for Complex Analysis of Gene Regulatory. // Proc. of the 3rd International Conference “Genomics, Proteomics, Bioinformatics and Nanotechnologies for Medicine”, Novosibirsk, Jul. 12-16, 2006. — P. 97.

8. **Валеев Т. Ф.** Сравнительный анализ методов поиска регуляторных модулей в последовательностях ДНК, использующих данные микроэррэев. // Методы и инструменты конструирования и оптимизации программ. — Новосибирск, 2005. — С. 21–28.

9. **Валеев Т. Ф.** Генетический алгоритм как альтернатива для решения некоторых NP-полных задач. // Тез. докл. конференции-конкурса «Техно-

логии Microsoft в информатике и программировании», Новосибирск, 22–24 февраля 2005. — С. 112–113.

10. **Коновалова Т. Г., Валеев Т. Ф., Черёмушкин Е. С.** Поиск композиционных промоторных модулей, регулирующих экспрессию генов эукариот. // Тез. докл. конференции-конкурса «Технологии Microsoft в информатике и программировании», Новосибирск, 22–24 февраля 2005. — С. 121–122.

11. **Черёмушкин Е. С., Коновалова Т. Г., Валеев Т. Ф.** Разработка пакета программ по анализу регуляторных областей ДНК. // Тез. докл. конференции-конкурса «Технологии Microsoft в информатике и программировании», Новосибирск, 22–24 февраля 2005. — С. 142–143.

12. **Голосов К. В., Валеев Т. Ф., Коновалова Т. Г. и др.** Интегральная система анализа генетической информации ExPlain. // Тез. докл. конференции-конкурса «Технологии Microsoft в теории и практике программирования», Новосибирск, 22–24 февраля 2006. — С. 171–172.

13. **Тараскина А. С., Коновалова Т. Г., Валеев Т. Ф., Штокало Д. Н., Черёмушкин Е. С.** Графическое представление результатов анализа в пакете программ по поиску регуляторных фрагментов в ДНК. // Тез. докл. конференции-конкурса «Технологии Microsoft в теории и практике программирования», Новосибирск, 22–24 февраля 2006. — С. 223–225.

14. **Тараскина А. С., Коновалова Т. Г., Валеев Т. Ф., Штокало Д. Н., Черёмушкин Е. С.** Пакет программ CisSearch по поиску регуляторных фрагментов в ДНК. // Тез. докл. XIII Международной научной конференции «Ломоносов», Москва, 12-15 апреля 2006. — Т. IV. — С. 48–49.

15. **Черёмушкин Е. С., Валеев Т. Ф., Коновалова Т. Г., Штокало Д. Н., Голосов К. В., Кель А. Э.** ExPlain: программная система по анализу

микрочипов и поиску ключевых молекул. // Тез. докл. Шестой международной конференции «Перспективы систем информатики», рабочий семинар «Наукоёмкое программное обеспечение», Новосибирск, 28–29 июня 2006. — С. 106–110.

16. **Черёмушкин Е. С., Коновалова Т. Г., Валеев Т. Ф., Штокало Д. Н., Тараскина А. С.** Пакет программ CisSearch для анализа регуляторных последовательностей ДНК. // Тез. докл. Шестой международной конференции «Перспективы систем информатики», рабочий семинар «Наукоёмкое программное обеспечение», Новосибирск, 28–29 июня 2006. — С. 111–114.

Валеев Т. Ф.

АЛГОРИТМЫ И ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ МОДЕЛИРОВАНИЯ ПРОЦЕССОВ ГЕННОЙ РЕГУЛЯЦИИ

Автореферат

Подписано в печать Объем 1,1 уч.-изд. л.
Формат бумаги 60 × 90 1/16 Тираж 100 экз.
Отпечатано в ЗАО РИЦ «Прайс-курьер»
630090, г. Новосибирск, пр. Ак. Лаврентьева, 6, тел. 334-22-02
Заказ №135