Российская Академия Наук
Сибирское Отделение
Институт Систем Информатики

На правах рукописи

ЧЕРЕМУШКИН Евгений Сергеевич

АЛГОРИТМЫ И ПРОГРАММНЫЕ СИСТЕМЫ ДЛЯ АНАЛИЗА РЕГУЛЯТОРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

Специальность 05.13.11 –
Математическое и программное обеспечение вычислительных машин, комплексов, систем и 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

Научный руководитель: Мурзин Федор Александрович,
кандидат физико –
математических наук

Официальные оппоненты: Поройков Владимир Васильевич
доктор физико – математических
наук, профессор

Воробьев Юрий Николаевич
доктор биологических наук,
кандидат физико –
математических наук, профессор

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

Защита состоится 23 июня 2006 г. в 14 ч. 30 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 _____ 2006 г.

Ученый секретарь
Диссертационного совета,
к.ф.–м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Биоинформатика – это наука о компьютерных методах решения биологических задач. В настоящее время наблюдается активизация деятельности в биоинформатике, что связано, прежде всего, с появлением в молекулярной биологии и генетике очень больших объемов данных, обработку которых нужно автоматизировать.

Исследования в биоинформатике и создание соответствующего программного обеспечения является актуальным в связи с решением прикладных задачи: изучением болезней, в том числе наследственных, созданием высокотехнологичных лекарственных средств и др.

Одной из актуальных задач является задача разработки алгоритмов распознавания сайтов связывания с транскрипционными факторами (ССТФ). Специфические белки, называемые транскрипционными факторами, осуществляют регуляцию экспрессии генов. ССТФ определенных типов связываются с промоторными районами генов и стимулируют транскрипцию (производство РНК) этих генов.

Несмотря на разнообразие подходов, проблема построения точных алгоритмов распознавания ССТФ в настоящее время не может считаться окончательно решенной. Причина этого состоит в большом разнообразии контекстных, физико-химических и конформационных особенностей ССТФ; механизмов ДНК-белковых взаимодействий между ССТФ и транскрипционными факторами; специфичности контекста, окружающего ССТФ, степени консервативности нуклеотидного контекста в эволюции.

Технология анализа данных генетической информации требует создания и сопровождения сложных программных средств, а также алгоритмов, обеспечивающих предсказание и достоверность выводов.

В данной области применяются специальные процессы проектирования и анализа алгоритмов и программ, специальные форматы данных, редакторы генетических данных, базы данных и знаний, графические человеко-машинные интерфейсы.

Ввиду комплексной структуры активно исследуемых в настоящее время заболеваний, таких как рак и др. задача распознавания сайтов связывания с ТФ становится еще более актуальной. Эти заболевания нарушают регуляторную функцию большого количества генов, которая может быть исправлена с

помощью воздействия одного или нескольких транскрипционных факторов.

Для понимания, какие транскрипционные факторы вовлечены в регуляторный процесс необходимо создание алгоритмов и программ для распознавания соответствующих сайтов.

В последнее время стали появляться новые типы биологических данных, таких как микрочипы, однонуклеотидные полиморфизмы и др. Эта информация наряду с последовательностью ДНК может быть использована для распознавания ССТФ и таким образом улучшить его.

Поэтому, в частности, является актуальной разработка новых алгоритмов и программных средств для анализа микрочиповых данных.

Цель работы

Целью данной работы являлась разработка новых и улучшение имеющихся алгоритмов и программ для приближенной идентификации подцепочек в последовательностях ДНК, называемых цис-элементами или сайтами связывания транскрипционных факторов с ДНК (ССТФ). Разрабатываемые алгоритмы в каждом конкретном случае ориентированны на специфическую информацию, которой обладает биолог.

В результате был разработан комплекс алгоритмов предварительной фильтрации и затем последующей идентификации цис-элементов и объектно-ориентированная среда, реализующая эти алгоритмы.

Все алгоритмы, рассмотренные в работе, разбиваются на три большие группы: алгоритмы предварительной обработки ДНК, алгоритмы последующей обработки и алгоритмы визуализации. В ряде алгоритмов осуществляется переход от нуклеотидного уровня анализа ДНК на уровень анализа сигналов.

Методы исследования

Методы объектно-ориентированного программирования, проектирования и анализа алгоритмов и программ, разработки человеко-машинных интерфейсов; методы обработки сигналов, специального вида и приближенной идентификации подцепочек.

Также при разработке программно-аппаратных систем учитывалось требование платформенной независимости. В связи с тем, что системы создавались для работы проведения исследований

биологами-экспериментаторами, работающими на различных вычислительных системах. Большое внимание уделялось графическому представлению результатов анализа.

Научная новизна

Проведены исследования, направленные на изучение возможностей применения для анализа ДНК различных алгоритмов обработки сигналов. В частности, изучались корреляционные функции между сигналами, ассоциированными различными методами с ДНК, и сигналами, построенными на основе некоторых замечательных кодовых последовательностей.

В результате проведенных исследований был реализован ряд программных систем полезных для исследования генетической информации, базирующихся на алгоритмах приближенной идентификации подцепочек в последовательностях ДНК. Как результат, разработан набор алгоритмов поиска цис-элементов в регуляторных последовательностях ДНК которые используют экспериментальные биологические данные различных типов.

Предложена библиотека классов, функций и структур для обработки генетической информации: промоторов генов, цис-элементов, весовых матриц, промоторных моделей и др. На ее основе реализована программная система GRESA, нашедшая применение на практике.

Для анализа данных экспрессии генов и построения промоторной модели разработана программная система ExPlain. Система использует некоторую формализованную модель регуляторных генетических процессов в клетке.

Практическая ценность

Создан ряд алгоритмов, которые переданы отечественным и зарубежным заказчикам и применяются в коммерческих приложениях. В частности, программный продукт ExPlain внедрен и используется немецкой компанией Biobase.

По результатам работы была написана глава в книге “Analytical Tools for DNA, Genes and Genomes”, изданной в издательстве “DNA Press”.

Апробация работы

Результаты работы докладывались на различных конференциях: ECCB’2003 (Париж, Франция); Pacific Symposia on Biocomputing

(Гаваи, США); "Genome Informatics", (Cold Spring Harbor Laboratory); на Дне молодых ученых Samsung (Новосибирск); Конференции естественных вычислений ICNC'05 (Чаньша, Китай); Немецкой конференции по биоинформатике GCB'05 (Гамбург, Германия); конференции «Технологии Майкрософт в информатике и программировании» в 2004 – 2006 г.г. (Новосибирск).

Автором по теме диссертации опубликовано более 36 печатных работ.

Структура и объем работы

Диссертационная работа состоит из введения, трех глав и списка литературы. Объем диссертации – 142 стр. Список литературы содержит 38 наименований. Работа включает 56 рисунка и графика, полученных в результате расчетов на ЭВМ а также 10 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Целью данной работы явилась разработка новых и улучшение имеющихся алгоритмов и программ для приближенной идентификации подцепочек в последовательностях ДНК, называемых цис-элементами или сайтами связывания транскрипционных факторов с ДНК (ССТФ).

В первой главе рассматриваются алгоритмы идентификации подцепочек, соответствующих различным кодам: Кодам Баркера, Голея, Фрэнка, Хэмминга. Так же рассматриваются подходы по анализу ДНК с помощью автокорреляционной функции.

Нами было изучено применение теории шумоподобных сигналов к анализу регуляторных последовательностей ДНК. Аналогичных исследований других научных коллективов в мире автором не обнаружено.

Так же нами было изучено применение вейвлетов к анализу последовательностей ДНК.

Алгоритмы предварительной обработки ДНК основываются на принципах обработки шумоподобных сигналов и другими алгоритмами обработки информации. Шумоподобными сигналами (ШПС) называют такие сигналы, у которых произведение ширины спектра на длительность много больше единицы. В системах связи с ШПС ширина спектра ШПС всегда много больше ширины спектра передаваемого сообщения. Предположительно аналогичная ситуация наблюдается и в ДНК: некоторые сообщения в ДНК «теряются» в шуме. Последовательность ДНК преобразуется в сигнал, протяженный

во времени. Затем этот сигнал анализируется несколькими алгоритмами: алгоритм поиска сигналов Баркера, сигналов Фрэнка, алгоритм декодирования кодами Голея, алгоритм вейвлет-преобразования.

В данной главе отмечены некоторые свойства различных типов последовательностей ДНК в терминах автокорреляционной функции. Показано, что, низкая АКФ участка указывает на то, что этот участок несет функциональную нагрузку. Из этого следует, что природе «выгодно» поддерживать общее однообразие ДНК и только функционально важные участки имеют свой «уникальный» паттерн.

В работе рассматриваются регуляторные последовательности ДНК. Приведем биологическое описание функции ДНК.

Для того, чтобы РНК-полимераза закрепилась на промоторном районе и началась транскрипция, необходимо, чтобы на промоторном районе сформировался определенный комплекс регуляторных белков, называемых транскрипционными факторами. Набор этих белков, стимулирующих в конечном счете транскрипцию, различен для разных генов.

В различных тканях организма присутствуют различные транскрипционные факторы, которые в свою очередь тоже считываются с ДНК. Эти белки запускают работу разных специфичных наборов генов. Таким образом, обеспечивается разнообразие тканей организма.

Связывание транскрипционного фактора и ДНК происходит в определенном участке ДНК, называемом сайтом связывания с транскрипционным фактором (сайтом, цис-элементом, ССТФ). Сайты связывания с транскрипционными факторами составляют информационный состав промоторов.

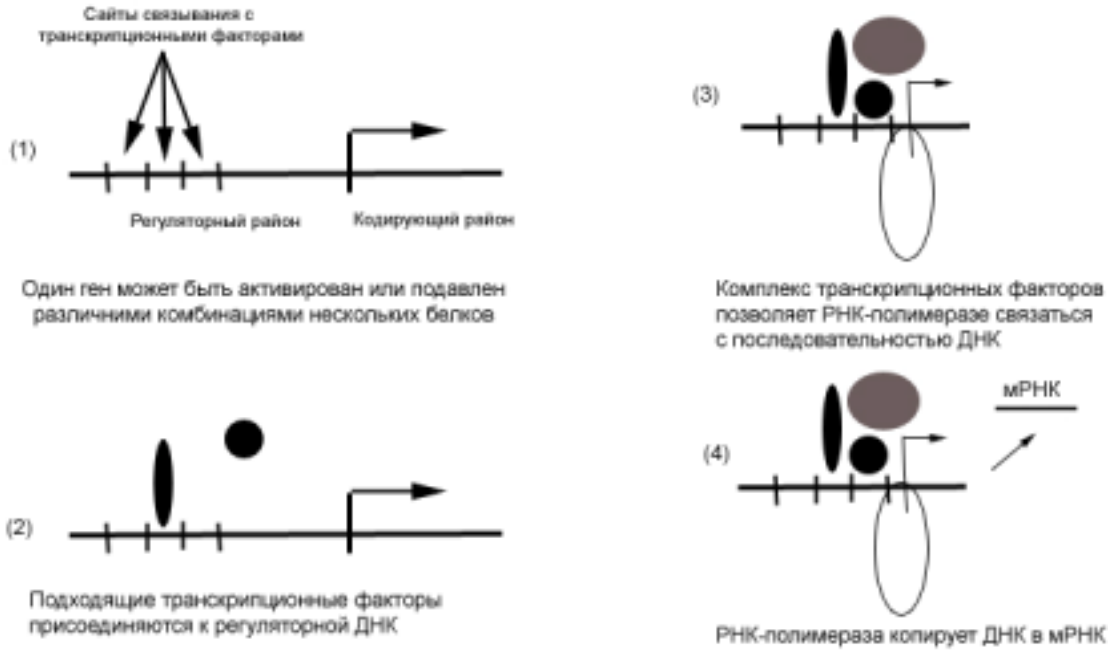

Рис. 1 Схема регуляции транскрипции

Здесь линией схематично представлена последовательность ДНК разделенная на два фрагмента. С правого фрагмента (кодирующий район) считывается белок, а левый фрагмент содержит специфические подцепочки, являющиеся сайтами связывания. К этим подцепочкам присоединяются специфические белки (транскрипционные факторы). После присоединения этих белков начинается считывание белка с кодирующего района (транскрипция).

Сигналом x назовем элемент пространства C_N циклических N-периодичных последовательностей. $x \in C_N$. Базовое множество для элементов последовательностей x будем выбирать Z,R или C соответственно.

Сигналу x сопоставим автокорреляционную последовательность $R_x(k) = \sum_{j=0}^{N-1} x(j+k)\overline{x(j)}$. Черта означает комплексное сопряжение.

Сигнал Фрэнка, так же называемый полифазным сигналом – это комплексная последовательность $\varphi(j)$ длины N=n*n задаваемая коэффициентами матрицы Фурье $F_n[j_1, j_0] = \omega_n^{-j_1 j_0}$ следующим образом: $\varphi(j_1 n + j_0) = F_n[j_1, j_0]$, где $\omega_n = e^{-2\pi i/n}$. Основное свойство сигнала Фрэнка состоит в том, что его циклические сдвиги образуют ортогональный базис в пространстве дискретных N-периодичных сигналов.

Введем **обобщенный сигнал Фрэнка**. Пусть $H_n[j_1, j_0] = \frac{1}{n} b(j_0) F_n[< j_1 + p(j_0) >_n, \pi(j_0)]$, где $j_0, j_1 = 0,1,\ldots,n-1$

Здесь b(j_0) комплексные коэффициенты, $p(j_0) \in \{0,1,...,n-1\}$ и π некоторая перестановка чисел 0,1,…,n-1. Обобщенный сигнал Фрэнка задается следующим образом:

$\psi(j_1 n + j_0) = H_n[j_1, j_0]$, $j_0, j_1 = 0,1,\ldots,n-1$.

Для того, чтобы циклические сдвиги такого сигнала задавали ортогональный базис необходимо и достаточно, чтобы $|b(j_0)| = 1$ при $j_0 = 0,1,\ldots,n-1$.

Преобразуем последовательность ДНК S к последовательности полифазного шумоподобного сигнала S'. Использовано преобразование A->1, T->-1, C->1i, G->-1i.

Для каждой позиции i в последовательности S получим семейство значений КФ $R_{(N,i,b,p,\pi)} = R_\psi(S,i) = \sum_{j=0}^{N-1} \psi(j)\overline{S(i+j)}$. Будем полагать, что p=1, $b(j_0) = 1$. Тогда в каждой позиции i получим семейство значений АКФ $R(i,\pi)$. Будем считать, что сигнал, определяемый подстановкой π, обнаружен в позиции i, если $R(i,\pi) > c$

Оценим статистические различия в количестве найденных сигналов в последовательностях ДНК различных типов.

Из проведенных исследований можем заключить, что некоторые сигналы значительно недопредставленны или перепредставленны в промоторах по отношению к случайным последовательностям. Это обозначает функциональную значимость данных сигналов в генетических последовательностях.

Во второй главе рассмотрены алгоритмы идентификации цис-элементов (подцепочек в регуляторных районах ДНК) с использованием известной заранее информации об этих подцепочках. Рассмотрен алгоритм весовых матриц.

Рассмотрен новый алгоритм распознавания двойных сайтов, разработанный автором. Предложенный нами алгоритм более специфичен, так как рассчитан только на сайты с двойным ядром. Поэтому алгоритм поиска двойных сайтов рассматривается авторами как наиболее применимый для данной задачи.

Авторами исследован существующий и предложен новый модифицированный алгоритм филогенетического футпринта, реализующий лучший по сравнению с существовавшими ранее методами поиск сайтов в гомологичных регуляторных последовательностях.

Нами предложен новый алгоритм анализа данных по экспрессии генов (микрочипов) с целью определения регуляторной промоторной модели гена. Так же предложен алгоритм выявления регуляторных свойств однонуклеотидных полиморфизмов.

Рассмотрен классический алгоритм весовых матриц. Этот алгоритм состоит в последовательном сопоставлении подцепочки с весовой матрицей и вычислении веса этой подцепочки. Сопоставление производится в режиме скользящего окна. Далее рассматривается метод построения весовой матрицы на основе алгоритма Гиббс Самплинга, являющегося приближенным алгоритмом поиска локального максимума.

Биологическая задача состоит в распознавании подцепочек (сайтов связывания с транскрипционными факторами) на основе выборки известных подцепочек (сайтов). Вес характеризует вероятность того, что эта подцепочка обладает заданным биологическим свойством. Свойство – это связывание с этим сайтом определенного белка.

В алгоритме филогенетического футпринта рассматривается алгоритм выравнивания двух цепочек символов с помощью схемы динамического программирования. В диссертации предложены модификации штрафов на вставку делеций (гэпов) в выравнивание и штрафов на замены. Проведены исследования, показывающие то, что с помощью модифицированного алгоритма распознавание подцепочек происходит лучше, чем при использовании классического.

Алгоритм филогенетического футпринта – это алгоритм, учитывающий эволюционные особенности развития организма. Идея метода основана на предположении, что цис-элементы в промоторах должны эволюционировать существенно медленнее, чем другие районы, которые не производят какой-либо консервативной функции. Таким образом, потенциальные ССТФ, которые находятся в эволюционно-консервативных районах промоторов с большей вероятностью можно считать “настоящими”. Таким образом, производится фильтрация сайтов с целью выделить только сходные сайты на выровненных последовательностях.

Автором проведено исследование алгоритма филогенетического футпринта и предложена модификация штрафных функций веса выравнивания с учетом особенностей регуляторных районов ДНК.

Для использования этого алгоритма сначала строится выравнивание двух (или в общем случае нескольких) последовательностей S_1 и S_2.

Эволюционно консервативные некодирующие последовательности (КНП или CNS) могут служить хорошими базисными элементами в геноме для поиска функционально значимых некодирующих участков ДНК. Были рассмотрены такие участки с целью поиска сайтов связывания с транскрипционными факторами между человеком и мышью.

В диссертации были использованы результаты группы “Berkley Genome Pipeline” глобального сравнения геномов человека и мыши. Каждый КНП представляет собой выравнивание последовательностей человека и мыши, взятых напрямую из генома. Это выравнивание показывает глобальное соответствие человеческого генома геному мыши.

Был рассмотрен набор из 17117 КНП общей длиной 2 418 267 элементов. Применив набор из 240 весовых матриц из библиотеки TRANSFAC 5.3 мы обнаружили 58 106 консервативных сайтов связывания. Эти сайты были записаны в базу данных с интерфейсом доступа через веб-сервер.

В итоге был разработан новый улучшенный алгоритм филогенетического футпринта, предназначенный для поиска сайтов на гомологичных регуляторных последовательностей. Качество работы алгоритма было проверено на искусственных и реальных данных.

В другой разработке использованы алгоритмы минимизации весовой функции, позволяющие избежать полного перебора. Использован генетический алгоритм и алгоритм метрополиса. Минимизируемая функция представляет собой вес, учитывающий свойства нормальности, критерии Стьюдента разделения двух выборок, критерия минимизации ошибок перепредсказания и недопредсказания, штрафы на количество параметров модели.

Биологическая задача состоит в определении набора ключевых молекул (транскрипционных факторов), характеризующих некорректную работу клетки в случае некоторого заболевания. В последствии, разрабатываемое лекарственное средство должно воздействовать на эти ключевые молекулы.

Регуляция каждого гена осуществляется несколькими регуляторными комплексами. Необходимо, чтобы присутствовал некоторый набор комплексов и причем некоторые из них могут быть заменены другими.

Мы воспользовались предположением, что связывание регуляторного комплекса с промотором – стохастический процесс. Следовательно, вес w связывания модуля K характеризует вероятность связывания модуля K. Поэтому, после проведенных исследований, была выбрана модель нечеткой логики для описания регуляторной модели промотора гена.

В третьей главе описаны программные системы, реализующие все описанные выше алгоритмы. В ходе работы были разработаны несколько программных продуктов, объединяющих все исследованные в рамках данной работы алгоритмы. История разработки этих программных продуктов содержит несколько

экспериментальных версий, которые были использованы для апробирования набора классов, реализованных в окончательной версии.

Реализован алгоритм филогенетического футпринта (Язык Си), создана база данных консервативных некодирующих последовательностей (Perl+MySql). Пакет программ SNPResearch (PHP), производящий анализ однонуклеотидных полиморфизмов в последовательностях ДНК. Далее разработана интегральная система GRESA (C++), включающая различные алгоритмы поиска ССТФ: алгоритм поиска сайтов ядерных рецепторов, алгоритм поиска ССТФ при имеющихся данных по экспрессии генов, алгоритм поиска сайтов в наборе последовательностей промоторов одного вида, различающихся по фенотипу, алгоритм анализа однонуклеотидных полиморфизмов, улучшенный алгоритм филогенетического футпринта.

На базе системы GRESA разработан пакет cissearch (C++, MFC) для поиска ССТФ в регуляторных последовательностях ДНК с использованием различной генетической информации. Проект по разработке программной системы cissearch получил поддержку фонда Бортника по программе «СТАРТ». На разработку этой системы зарегистрирована карта НИОКР номер 0120.0 503994. В настоящий момент пакет cissearch находится на этапе внедрения.

С использованием элементов пакета GRESA разработана интеллектуальная система анализа данных по экспрессии генов “ExPlain”.

Различные программные продукты реализованы на языках C, C++, Java, Perl, PHP.

Система GRESA (Gene REgulation and Sequence Analysis). Сейчас GRESA содержит порядка 68 000 строк кода.

Несколько слов о структуре системы GRESA. Классы **Site и SiteSet** представляют сайты на последовательности. Классы **Sequence, SequenceSet** представляют последовательность и набор последовательностей. Для использования только выбранных матриц в библиотеке GRESA существует профиль, реализуемый классами **Profile и ProfileEntry**. Профиль содержит набор имен матриц и пороги для поиска. Сайт матрицы Mj s_{ijk} считается распознанным на последовательности S_k, если $w_{ijk}>c_j$ и $cw_{ijk}>cc_j$, где c_j и cc_j порог сайта и порог ядра сайта, взятые из профиля.

Классы **Factor и FactorSet** отражают транскрипционный фактор и набор транскрипционных факторов соответственно. Транскрипционные факторы и весовые матрицы связаны следующим образом: Одна матрица может быть создана для нескольких транскрипционных факторов, а один транскрипционный фактор может быть представлен несколькими матрицами.

Так же в пакет «Ядро» входит класс **Alignment**, реализующий выравнивание по алгоритму Смита-Васермана и другие классы.

Набор классов, реализующий обработку микрочипов содержит несколько классов, описывающих композиционный модуль. Это **Complex** – базовый класс для композиционного модуля, имеющий виртуальную функцию для вычисления веса комплекса на наборе последовательностей, виртуальные функции мутации, рекомбинации (для генетического алгоритма) и некоторые общие переменные. **SimpleComplex** это композиционный модуль, состоящий из нескольких сайтов без учета их положения. **DistanceConplex** это комплекс из сайтов, которые расположены на последовательности в окне определенной длины. И, наконец, **BooleanComplex** реализует функциональность композиционного модуля, описанного в предыдущей главе. В этих реализациях базового комплекса реализованы методы мутации, скрещивания и вычисления веса.

Объекты типа комплекс используются в генетическом алгоритме реализованном в классе **GeneticAlgorithm** являющимся потомком класса **GresaAlgorithm**. Так же имеется реализованный алгоритм метрополиса, разработанный как альтернативный алгоритм поиска. Класс GeneticAlgorithm содержит массив объектов типа «комплекс» которые подлежат скрещиванию, мутациям и отбору. Алгоритм метрополиса, называемый так же Simulated Annealing, это недетерминистический алгоритм вычисления приближенного решения.

Для расширенной реализации алгоритма филогенетического футпринта были созданы классы **Footprint, FootprintView и FootprintModelTester**. Эти классы реализуют сам алгоритм футпринта и анти-футпринта, алгоритм визуализации результатов и тестирования на модельных данных.

Алгоритм сводится к выбору тех сайтов, которые присутствуют во всех последовательностях в одном и том же участке выравнивания. Тестирование алгоритма на искусственных модельных данных сводится к множественной генерации выравниваний и последующей проверке.

Среда GRESA постоянно дополняется и развивается. Разработка среды по технологии Экстремального программирования (Extreme

Programming) дает возможность постоянно поддерживать рабочую версию. Стабильность, при довольно большой и распределенной группе разработчиков, поддерживается за счет большого количества автоматизированных тестов. Жизненный цикл отдельного приложения состоит из этапов, когда приложение находится в стадии экспериментальной разработки, затем переходит в стабильную стадию.

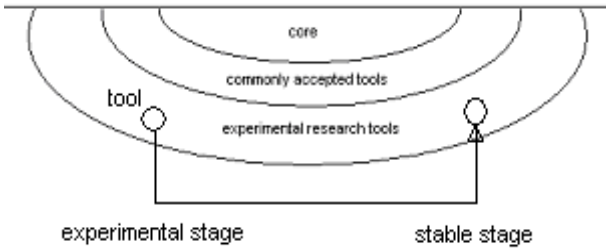

Рис. 2 Жизненный цикл отдельного приложения состоит из этапов, когда приложение находится в стадии экспериментальной разработки, затем переходит в стабильную стадию.

Далее оно может перейти в набор общепринятых инструментов. Примечательно, что любой член группы разработчиков может вносить изменения в любой класс, главное – сохранить успешное выполнение тестов.

Каждый алгоритм имеет интерфейс командной строки, принимающий входные файлы в качестве параметров. Для генерализованной обработки параметров командной строки реализован класс **ParamContainer**, содержащий функции разбора командной строки в массив параметров. Функциональность включает вложенные параметры, короткую и длинную запись параметров (-p param_value или –param_name=param_value).

Реализовано несколько графических интерфейсов. В том числе как серверные приложения (perl+mysql, php), так и локальные приложения (C++ MFC).

Для комплексного анализа данных по экспрессии генов (микрочипов) с целью распознавания регуляторных молекул, задействованных в регуляции экспериментов, проведенных пользователем, нашей группой разработана объединенная информационная система ExPlain – система по анализу результатов микрочипов с целью выявления функционально важных молекул.

Алгоритмы реализованные в системе ExPlain включают:

- Поиск сайтов связывания с помощью алгоритма весовых матриц match
- Поиск ключевых молекул - поиск молекул, которые оказывают влияние на работу некоторой выборки генов

- Поиск регуляторной модели генов данного эксперимента

Все алгоритмы интегрированы друг с другом, что позволяет пользователю производить детальный и комплексный анализ его данных.

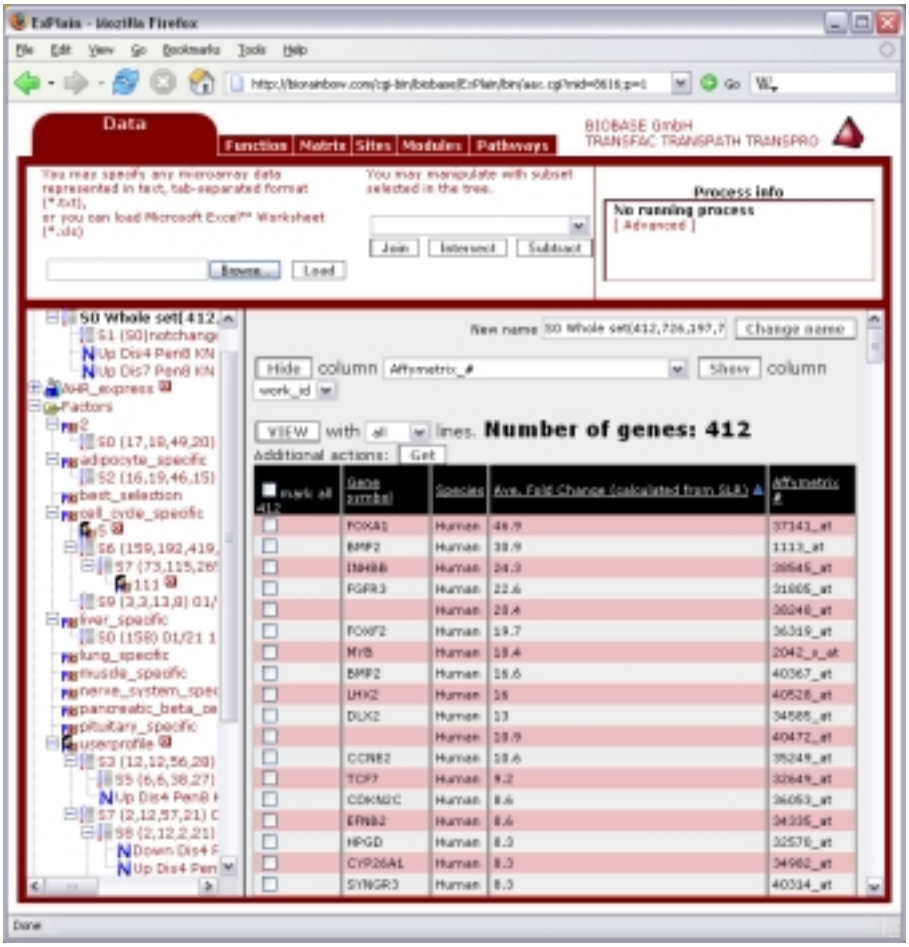

Рис. 3 Интерфейс программы ExPlain. Отображение входных данных

О ЛИЧНОМ ВКЛАДЕ АВТОРА

Реализация описанных выше алгоритмов и программ является довольно трудоёмкой задачей и выполнялась коллективно. Автором лично проведено проектирование представленных в диссертации программ и реализованы наиболее сложные с алгоритмической точки зрения вычислительные модули. Заслуга автора с точки зрения математики заключается в проработке наиболее сложных, из описанных выше алгоритмов. В частности выбор способов применения шумоподобных сигналов к анализу ДНК, разработка алгоритма предсказания сайтов ядерных рецепторов, разработка алгоритмов пакета cissearch.

Полностью начиная с анализа подходов и разработки алгоритмов, заканчивая реализацией и тестированием автором были

реализованы алгоритмы и программные системы реализующие алгоритм филогенетического футпринта, алгоритм поиска ядерных рецепторов, алгоритмы поиска шумоподобных сигналов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проведены комплексные исследования, позволившие разработать и реализовать ряд новых алгоритмов и усовершенствовать имеющиеся алгоритмы обработки сигналов с целью применения их для анализа регуляторных последовательностях ДНК.

2. Разработан набор алгоритмов поиска цис-элементов в регуляторных последовательностях ДНК которые используют экспериментальные биологические данные различных типов, такие как данные по экспрессии генов, данные об однонуклеотидных полиморфизмах, данные о гомологичных последовательностях.

3. Создана программная система GRESA, содержащая набор классов для обработки генетической информации: промоторов генов, цис-элементов, весовых матриц, промоторных моделей, предоставляющая широкие возможности для анализа генетической информации.

4. Разработана программная система ExPlain для анализа данных экспрессии генов и построения промоторной модели в соответствии с предложенной формализованной обобщенной регуляторной моделью гена.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **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, October 2005; Chapter 9, pp.185-214
2. **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. // Proceedings of the German Conference on Bioinformatics (GCB'05), Hamburg, Germany, Oct 5-7, 2005; 8 pp
3. **Ravi Shankar, Amit Chaurasia, Biswaroop Ghosh, Dmitry Chekmenev, Evgeny Cheremushkin, Alexander Kel and Mitali Mukerji** Genomic divergence vs. functional constraints: Comparative analysis of non-coding regions in neuronal and

housekeeping genes of human and chimpanzee // BMC Genomics В публикации

4. **A. Kel, T. Konovalova, T. Waleev, E. Cheremushkin, O. Kel-Margoulis and E. Wingender** Composite Module Analyst: A fitness-based tool for identification of transcription factor binding site combinations // Bioinformatics 10.02.2006 http://bioinformatics.oxfordjournals.org/cgi/reprint/btl041?ijkey=7F1eujtikqrnBcy&keytype=ref
5. **Тараскина А. С., Коновалова Т. Г., Валеев Т. Ф., Штокало Д.Н., Черемушкин Е. С.** Графическое представление результатов анализа в пакете программ по поиску регуляторных фрагментов в ДНК // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 23, 2006; стр.142-143
6. **Черемушкин Е.С.** Исследование последовательностей ДНК с помощью некоторых совершенных кодов // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 23, 2006; стр.145-146
7. **Черемушкин Е.С.** Обобщенные сигналы фрэнка в применении к анализу последовательностей ДНК// Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 23, 2006; стр.147-148
8. **Konovalova T., Valeev T., Cheremushkin E., Kel A.** Composite Module Analyst: Tool for Prediction of DNA Transcription Regulation. Testing on Simulated Data // Advances in Natural Computation, part 2, Springer, Germany, 2005 (LNCS 3611); pp.1202-1205 Proceedings of the First International Conference on Natural Computations (ICNC'05), Changsha, China, Aug 27-29, 2005
9. **Черемушкин Е.С.** Анализ различных участков ДНК с помощью автокорреляционной функции // Методы и инструменты конструирования и оптимизации программ, Новосибирск, 2005;
10. **Штокало Д.Н., Черемушкин Е.С.** Построение программного комплекса “Regulatory Sequences Analyser” для распознавания цис-элементов в последовательностях ДНК // Методы и инструменты конструирования и оптимизации программ, Новосибирск, 2005;
11. **Е.С. Черемушкин Е.С.** Исследование днк с применением теории шумоподобных сигналов // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 22-24, 2005; стр.140-142

12. **Е.С. Черемушкин, Т.Г. Коновалова, Т.Ф. Валеев** Разработка пакета программ по анализу регуляторных областей днк // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 22-24, 2005; стр.142-143
13. **Коновалова Т., Валеев Т, Черёмушкин Е.** Поиск композиционных промоторных модулей, регулирующих экспрессию генов эукариот // Тезисы конференции-конкурса "Технологии Микрософт в информатике и программировании" 22-24 февраля 2005; с.121-122
14. **Черёмушкин Е., Коновалова Т., Валеев Т.** Разработка пакета программ по анализу регуляторных областей ДНК // Тезисы конференции-конкурса "Технологии Микрософт в информатике и программировании" 22-24 февраля 2005; с.142-143
15. **Коновалова Т., Валеев Т, Черёмушкин Е.** Весовые матрицы и поиск композиционных промоторных модулей, регулирующих экспрессию генов эукариот. // Тезисы XLIII Международной Научной Студенческой Конференции, 11-14 апреля 2005; с. 123-124
16. **Черёмушкин Е., Коновалова Т., Валеев Т.** Программный комплекс для анализа регуляторных областей. // Тезисы XLIII Международной Научной Студенческой Конференции, 11-14 апреля 2005; с. 142-143
17. **Черемушкин Е.С.** Институт систем информатики им. А.П. Ершова. Шумоподобные сигналы и исследование ДНК // Тезисы XLIII Международной Научной Студенческой Конференции, 11-14 апреля 2005;
18. **Evgeny Cheremushkin, Alik Dunaev, Fedor Murzin** System of statistical comparison of methods of search of cis-elements. // Proc. of Samsung Young Scientist Day, Novosibirsk, 2004
19. **Черемушкин Е. С., Коновалова Т. Г., Мурзин Ф. А., Кель А. Э.** Система распознавания цис-элементов на последовательностях ДНК // Программные средства и математические основы информатики. - Новосибирск, 2004. - С. 255-269.
20. **Черемушкин Е.С., Кель А.Е., Лобив И.В., Мурзин Ф.А., Половинко О.Н.** Визуализация последовательностей днк посредством трансформаций цветового куба // Тезисы конференции ИВТН-2004 с. 28

21. **Лобанова М.В., Коновалова Т.Г., Черемушкин Е.С.** Интернет-инструмент для анализа snp в некодирующей ДНК // Тезисы конференции ИВТН-2004 с.33
22. **Бесчастнов Е., Лобанова М., Коновалова Т., Черемушкин Е.** Программная система поиска ЦИС-элементов // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 21-23, 2004; с. 90-91
23. **Черемушкин Е. С.** Филогенетический футпринт Новый метод для выравнивания промоторов // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 21-23, 2004; с. 133-134
24. **Черемушкин Е. С.** Система статистического сравнения методов поиска ЦИС-элементов // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 21-23, 2004; с. 134-135
25. **Черемушкин Е. С., Половинко О. Н., Лобив И. В., Дунаев А. А.** Визуализационные методы идентификации подцепочек в регуляторных последовательностях ДНК// Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 21-23, 2004; с. 136-137
26. **Черемушкина Е. Н., Черемушкин Е. С., Чекменев Д., Кель О.** Метод идентификации сайтов ядерных рецепторов // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 21-23, 2004; с. 137-139
27. **Коновалова Т., Бесчастнов Е., Лобанова М., Черемушкин Е.** GRESA DEVELOPMENT TOOLS: объединенная среда разработки и тестирования приложений в области анализа регуляторных последовательностей ДНК // Тезисы конференции-конкурса «Технологии Microsoft в информатике и программировании» Февраль 21-23, 2004; с. 107-109
28. **Черемушкин Е.С., Половинко О.Н., Лобив И.В., Дунаев А. А.** Визуализация и идентификация подцепочек в регуляторных последовательностях ДНК // Межвузовская научно-практическая студенческая конф. МНСК, 18-25 апреля, 2004 г., НГУ. Новосибирск 2004. с. 197-199
29. **Черемушкин Е.С.** Статистическое сравнение методов поиска цис-элементов // Межвузовская научно-практическая студенческая конф. МНСК, 18-25 апреля, 2004 г., НГУ. Новосибирск 2004. с. 199-201
30. **Черемушкин Е. С.** Филогенетический футпринт и выравнивание промоторов // Межвузовская научно-

практическая студенческая конф. МНСК, 18-25 апреля, 2004 г., НГУ. Новосибирск 2004. с.201-202

31. **Черемушкина Е.Н., Черемушкин Е. С., Чекменев Д. Кель О.** Алгоритмы идентификации сайтов ядерных рецепторов // Межвузовская научно-практическая студенческая конф. МНСК, 18-25 апреля, 2004 г., НГУ. Новосибирск 2004. с. 202-204
32. **Дунаев А. А., Кель А. Э., Лобив И. В., Мурзин Ф. А., Половинко О. Н., Черемушкин Е. С.** Визуализация генетической информации // Новые информационные технологии в науке и образовании. - Новосибирск, 2003. - С. 147-156.
33. **Cheremushkin, E. and Kel, A.** Whole genome human/mouse phylogenetic footprinting of potential transcription regulatory signals. // Proceedings of the Pacific Symposia on Biocomputing, 2003; p.291-302. PMID: 12603036
34. **Alexander Kel, Klaus Hornischer, Olga Kel-Margoulis, Evgeniy Cheremushkin and Edgar Wingender** Phylogenetic footprints of Composite Regulatory Elements in human, mouse and rat genomes. // Proceedings of the meeting: "Genome Informatics", Cold Spring Harbor Laboratory, May 7-May 11, 2003, p.66
35. **Kel, A.E., Goessling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E.** MATCH(TM): a tool for searching transcription factor binding sites in DNA sequences // Nucleic Acids Res. V.31, p.3576-3579. PMID: 12824369
36. **Konovalova (Ivanova), T., Cheremushkin, E., Beschastnov, E., and Kel, A.** Applicating of the metropolis algorithm to reveal composite modules in promoters of eukaryotic genes // Proceedings of the European Conference on Computational Biology, (ECCB'2003), Paris, France, Sept. 27-30, 447-448 (2003).

Черемушкин Е.С.

АЛГОРИТМЫ И ПРОГРАММНЫЕ СИСТЕМЫ ДЛЯ АНАЛИЗА РЕГУЛЯТОРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

Автореферат

Подписано в печать Объем 1,1 уч.-изд. л.
Формат бумаги 60 × 90 1/16 Тираж 100 экз.
Отпечатано на ризографе "AL Group"
630090, г. Новосибирск, пр. акад. Лаврентьева, 6