УДК 004.4

Ю. Ю. Васькин [1], **И. В. Хомичева** [2], **Е. В. Игнатьева** [2], **Е. Е. Витяев** [3]

[1] Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

[2] Институт цитологии и генетики СО РАН
пр. Акад. Лаврентьева, 10, Новосибирск, 630090, Россия

[3] Институт математики СО РАН
пр. Акад. Коптюга, 4, Новосибирск, 630090, Россия

E-mail: vaskin90@gmail.com; khomicheva@bionet.nsc.ru;
eignat@bionet.nsc.ru; vityaev@math.nsc.ru

АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РЕГУЛЯТОРНЫХ РАЙОНОВ ГЕНОВ РЕЛЯЦИОННОЙ СИСТЕМОЙ EXPERT DISCOVERY, ВСТРОЕННОЙ В ПАКЕТ UGENE [*]

Задача автоматического восстановления иерархической структуры регуляторной области эукариотического гена поставлена на стыке биологии, математики и информационных технологий. Решение этой задачи предполагает понимание сложных механизмов регуляции генов эукариот и применение интеллектуальных технологий для многопараметрического анализа. В данной работе приведено описание интегрированной системы, которая реализует метод реляционного анализа биологических данных. Система позволяет учитывать доступную биологу априорную информацию об анализируемых данных, осуществлять анализ на каждом уровне организации регуляторного района, последовательно производить поиск решения от простой гипотезы к более сложной. Интеграция систем предоставляет удобную среду для проведения комплексных исследований и автоматизации работы эксперта-биолога.

Ключевые слова: комплексный сигнал, иерархический анализ, интегрированная система, реляционный подход, распознавание, регуляторные районы генов, аннотация.

Введение

Анализ регуляторных последовательностей генов и поиск структурно-функциональных закономерностей их организации представляет актуальную проблему биологии, которая далека от окончательного решения. Во многом это обусловлено сложностью строения регуляторных областей и многообразием механизмов регуляции транскрипции. Возникает необходимость анализировать разнородную информацию по физико-химическим, структурным, информационным свойствам регуляторных последовательностей генов, экспериментальных данных об их функционировании.

Одно из самых важных свойств гена – способность к экспрессии, процесс, в ходе которого на основе генетической информации (последовательности нуклеотидов ДНК, соответствующей некоторому гену) синтезируется определенное количество функционального продукта этого гена – РНК или белка. Экспрессия – это сложный и многостадийный процесс, первым этапом которого является транскрипция, у эукариот она проходит в ядре. Интенсивность

[*] Работа выполнена при финансовой поддержке РФФИ (грант № 11-07-00560-а), интеграционных проектов СО РАН (проекты № 47, 113, 115, 119), президиума РАН (программы № Б.25, Б.27, А.II.6), при частичной поддержке Минобрнауки РФ (госконтракты № П857 и 07.514.11.4003, 07.514.11.4023, 14.740.12.0819), Совета по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-3606.2010.1).

транскрипции каждого конкретного эукариотического гена подвержена гибкой регуляции в зависимости от клеточных условий (типа клеток и тканей, стадии развития организма, клеточного цикла, индукторам либо репрессорам, действующим на клетки) [1; 2]. Регуляция транскрипции генов осуществляется при участии большого количества регуляторных белков: транскрипционных факторов (ТФ), коактиваторов, корепрессоров, медиаторов [3]. Важнейшую роль в этом процессе играют ТФ, которые специфически взаимодействуют с определенными участками ДНК в регуляторных районах генов – сайтами связывания транскрипционных факторов (ССТФ). Помимо взаимодействия с ДНК ТФ участвуют в белок-белковых взаимодействиях с другими регуляторными белками, формируя сложные мультибелковые комплексы, активирующие либо подавляющие транскрипцию генов.

Возможность гибкой регуляции экспрессии генов эукариот обусловлена наличием довольно обширных регуляторных областей, имеющих блочно-иерархическую структуру [4–7].

Первому уровню иерархии регуляторных областей генов соответствуют ССТФ – короткие последовательности ДНК (10-20 нуклеотидов), являющиеся местом посадки транскрипционных факторов [8].

Следующим уровнем иерархии являются композиционные элементы, представляющие собой близко расположенные ССТФ, которые в результате белок-белковых взаимодействий между соответствующими транскрипционными факторами приобретают новые регуляторные свойства. Композиционные элементы синергичного типа обеспечивают в результате белок-белковых взаимодействий неаддитивно высокий уровень активации транскрипции. Композиционные элементы антагонистического типа включают перекрывающиеся либо очень близко расположенные ССТФ. В этой ситуации два белковых фактора конкурируют друг с другом за связывание с ДНК, благодаря чему возможна смена стимулирующего влияния фактора-активатора на ингибирующее влияние фактора-репрессора и, наоборот, в зависимости от клеточной ситуации [1; 2; 9]

Регуляторные единицы (промоторные районы, энхансеры, сайленсеры) являются следующим уровнем в системе иерархической организации регуляторных районов генов. Их регуляторные функции реализуются благодаря наличию в них ССТФ и композиционных элементов, взаимодействующих с регуляторными белками [1; 2]. Расположение регуляторных единиц относительно старта транскрипции генов и их протяженность варьирует существенным образом. Энхансеры и сайленсеры – регуляторные единицы, активирующие либо подавляющие транскрипцию конкретного гена и удаленные от его старта транскрипции на значительное расстояние (до 50 000 п. о). Энхансеры и сайленсеры могут находиться как в 5’- и 3’-фланкирующих областях генов, так и в интронах. Промоторные районы расположены непосредственно перед стартом транскрипции генов. Их размер, как правило, варьирует в пределах от 200 до 1 000 нуклеотидов [10].

Самый высший уровень иерархии строения регуляторных областей генов соответствует системе интегральной регуляции транскрипции [5], которая реализуется при участии сложных комплексов регуляторных белков, взаимодействующих со всей совокупностью регуляторных единиц и элементов конкретного гена. Состав мультибелковых комплексов определяется ДНК – белковыми взаимодействиями, основанными на суперпозиции разных кодов ДНК (линейных, конформационных) [11].

Разнообразие строения регуляторных районов генов велико, что определяется необходимостью реализовать индивидуальный способ регуляции каждого конкретного гена в соответствии с клеточной ситуацией. Например, по современным оценкам, в геноме человека закодировано около 1 500 транскрипционных факторов [12]. Можно ожидать, что регуляторные районы генов включают такое же большое количество разных типов ССТФ. Регуляторные районы каждого гена включает уникальную комбинацию ССТФ различных типов. По данным базы TRRD, регуляторные районы конкретного гена могут содержать более 20 различных ССТФ, функциональность которых подтверждена экспериментально, а, в свою очередь, вся система интегральной регуляции гена может включать десятки регуляторных единиц [7]. Разнообразие строения регуляторных районов выражается еще и в том, что, как отмечалось выше, и протяженность регуляторных единиц, и их локализация варьируются существенным образом.

В настоящее время широко применяются разнообразные методы компьютерного анализа строения регуляторных районов генов, каждый из которых соответствует определенному иерархическому уровню. Для распознавания ССТФ используются такие подходы, как метод весовых матриц [13; 14], метод SITECON [15], SiteGA [16] и ряд других. Все известные подходы имеют определенные недостатки, поскольку не учитывают взаиморасположение сайтов и характеризуются определенными уровнями перепредсказания (или недопредсказания) [17].

Задача, соответствующая одному из уровней иерархии строения регуляторных районов генов, состоит в обнаружении закономерностей расположения ССТФ [18; 19]. Однако поскольку регуляторные районы генов содержат уникальные комбинацию ССТФ, разрабатываемые методы сталкиваются с плохой репрезентативностью данных обучения, содержащих недостаточное число частных случаев более общего явления.

Задача анализа регуляторных единиц генов и всей системы интегральной регуляции транскрипции существенно превышает по сложности задачи, связанные с анализом ССТФ и их комбинаций. Это связано с огромным разнообразием строения регуляторных районов генов, которое, как было изложено ранее, обусловлено возможностью присутствия большого количества элементарных сигналов (ССТФ, конформационных и физико-химических свойств, нуклеосомного потенциала) в пределах регуляторных районов, а также вариабельностью протяженности и расположения самих регуляторных районов. С информационной точки зрения задача анализа регуляторных районов генов эукариот состоит в иерархическом анализе генетической информации.

Для решения этой проблемы необходимо применение современных компьютерных технологий, связанных с интеллектуальным анализом данных (Data Mining and Knowledge Discovery). На сегодняшний день ни один из известных методов не в состоянии полностью ее решить. В большинстве случаев эксперты-генетики вынуждены вручную анализировать огромное количество информации, часто противоречивой, чтобы добиться биологически значимого результата.

Автоматические методы анализа и распознавания регуляторных областей, в общем случае, должны учитывать различные контекстные, физические, химические и структурные особенности ДНК. Таким образом, создание интегрированного метода распознавания, который бы рассматривал сигналы различных типов, полученные как результат работы других методов распознавания, является актуальной задачей.

В настоящей работе описан метод интеллектуального анализа регуляторных областей, который основывается на интеграции взаимодополняющих инструментов: системы ExpertDiscovery, мощного инструмента для иерархического анализа регуляторных районов генов и мультиплатформенного биоинформационного пакета UGENE, объединяющего большое число алгоритмов для работы с генетической информацией [1] [20–23]. Система Expert Discovery интегрирована в программный пакет UGENE в виде модуля. Модули UGENE объединены общим интерфейсом и логикой работы. Таким образом, результаты работы различных модулей можно комбинировать, что раскрывает потенциал системы ExpertDiscovery по поиску комплексных закономерностей. С помощью UGENE можно осуществить распознавание сигналов на низких уровнях иерархии и передать эти результаты ExpertDiscovery, которая способна произвести более сложный иерархический анализ. В отличие от известных методов анализа регуляторных областей, которые фокусируются на выявлении отдельных сигналов, система ExpertDiscovery использует интегрированный подход для анализа генетической информации. Система ExpertDiscovery позволяет проводить исследования на любом уровне, от нуклеотидного до геномного, и дает возможность интеграции специфичных методов на каждом уровне.

Реляционный подход к обнаружению знаний

Система ExpertDiscovery применяет оригинальный реляционный подход (Relational Data Mining) для обнаружения знаний [2] [6]. Данный подход ранее применялся в системе Discovery

[1] Unipro UGENE: an open-source bioinformatics toolkit; http://ugene.unipro.ru

[2] Scientific Discovery Web Site, http://www.math.nsc.ru/AP/ScientificDiscovery

для решения большого ряда практических задач психофизики, диагностики раковых заболеваний и предсказания курсов ценных бумаг. В основе этой системы лежит семантический вероятностный вывод.

Этот метод представляет собой синтез логики и вероятности. И связан с утверждением: предсказание нельзя вывести, а можно только вычислить. Идея семантического подхода в программировании, выдвинутая Ю. Л. Ершовым, С. С. Гончаровым и Д. И. Свириденко, заключается в том, что процесс вычисления рассматривается как проверка истинности утверждения на некоторой модели (например, моделью могут быть данные, представленные некой многосортной системой). При таком подходе процедуру логического вывода можно обобщить, рассматривая более разнообразные взаимоотношения высказываний и модели: рассмотреть процесс вычисления как, например, определение наиболее вероятных, подтвержденных или нечетких высказываний на модели. Такой обобщенный вывод будем называть *семантическим* [24].

Идея извлечения новых знаний состоит в последовательном уточнении гипотезы таким образом, чтобы на каждом следующем шаге получались гипотезы с большей вероятностью и определенностью. При этом осуществляется проверка значимости полученного результата при помощи статистических критериев.

Под *семантическим вероятностным выводом* понимается такая последовательность правил C_1, C_2 , ... , C_n , что:

1) $C_i = \left(A_1^i \,\&\, ... \,\&\, A_{ki}^i \Rightarrow G \right), \quad i = 1,...,n;$

2) C_i – *подправило* правила C_{i+1}, т. е. $\left\{ A_1^i,..., A_{ki}^i \right\}$ содержит $\left\{ A_1^{i+1},..., A_{ki+1}^{i+1} \right\}$;

3) $\mathrm{Prob}\left(C_i\right) < \mathrm{Prob}\left(C_{i+1}\right), \quad i = 1,2,...,n-1,$ где *условная вероятность* правила C_i определяется следующим образом:

$$\mathrm{Prob}\left(C_i\right) < \mathrm{Prob}\left(G / A_1^i \,\&\, ... \,\&\, A_{ki}^i\right) = \mathrm{Prob}\left(G \,\&\, A_1^i \,\&\, ... \,\&\, A_{ki}^i\right) \Big/ \mathrm{Prob}\left(A_1^i \,\&\, ... \,\&\, A_{ki}^i\right);$$

4) C_i – *вероятностный закон,* т. е. для любого правила $C' = \left(A_1^i \,\&\, ... \,\&\, A_{ki}^i \Rightarrow G \right)$ правила C_i, $\left\{ A_1,..., A_j \right\}$ содержит $\left\{ A_1^i,..., A_{ki}^i \right\}$ выполнено неравенство $\mathrm{Prob}\left(C'\right) < \mathrm{Prob}\left(C_i\right)$;

5) C_n – *сильнейший вероятностный закон*, т. е. правило C_n не является подправилом никакого другого вероятностного закона.

Система Discovery реализует семантический вероятностный вывод с обнаружением знаний в виде множества вероятностных законов, сильнейших вероятностных законов и максимально специфичных законов.

Вариацией системы Discovery является ExpertDiscovery, настроенная на обнаружение знаний в выборках нуклеотидных последовательностей в соответствии с семантическим вероятностным выводом в виде комплексных сигналов с указанными параметрами.

Применение реляционного подхода к анализу регуляторных районов Система ExpertDiscovery

Введем ряд базовых для системы ExpertDiscovery понятий, необходимых для формализации и описания метода.

Как известно, генетическая информация представляется в виде символьных последовательностей в 4- или 15-буквенном коде.

Сигнал – система правил, определяющих свойства участков последовательностей ДНК. Элементарный сигнал – неделимый сигнал, который характеризуется именем и местами в последовательности, где он присутствует.

Гипотезы экспертов формулируются в виде комплексных сигналов (КС), определяемых рекурсивно на основе элементарных сигналов и операций над ними:

1) элементарный сигнал является КС;

2) результат воздействия на КС операций (подробное определение операций приведено ниже) «повтора» или принадлежности «интервалу» является КС;

3) результат воздействия на два КС операции «дистанция» между сигналами является КС.

Схематически КС может быть представлен в виде дерева (рис. 1). Элементарными сигналами являются буквы, обозначающие нуклеотиды в цепи ДНК, а операциями – «дистанция» (между буквами и КС) и «повторы» (двух КС). Как следует из определения, каждая буква и каждое отдельно взятое поддерево и есть КС.

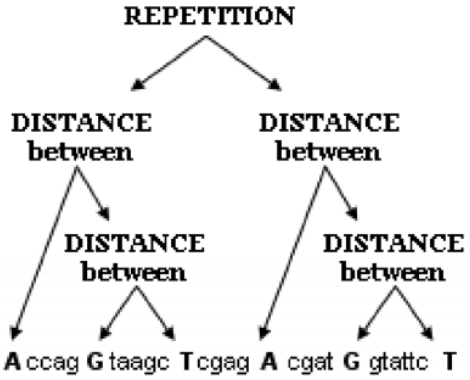

Рис. 1. Схематическое представление КС

Над КС определены следующие операции.

Дистанция между сигналами. На вход подаются два КС s_1 и s_2 и указывается, что дистанция между ними может изменяться от min до max, и имеет ли значение порядок. Полученный на выходе КС считается найденным на последовательности в некоторой позиции, если в этой позиции найден сигнал s_1, и на расстоянии от min до max символов от него найден сигнал s_2. В случае, если порядок не имеет значения, сначала может быть найден s_2, а потом s_1. Параметры min и max задаются экспертом.

Повтор сигнала. Указывает, что результирующий КС является повторением входного сигнала s от $N_{\min}$ до $N_{\max}$ раз, при этом расстояние между соседними повторами принадлежит диапазону от min до max. Параметры $N_{\min}$, $N_{\max}$ и min, max задаются экспертом.

Принадлежность сигнала интервалу. Указывает, что входной КС следует искать только в интервале от min до max. Здесь min и max абсолютные значения относительно первого символа последовательности. Эта операция осмыслена только для выровненных последовательностей. Параметры min и max задаются экспертом. При этом дистанция между двумя КС может быть измерена различными способами, такими как:

- от конца первого до начала второго;
- от начала первого до начала второго;
- от середины первого до начала второго.

Способ, которым следует измерять дистанцию, является параметром соответствующей операции и задается экспертом.

Задавая параметры операций, эксперт тем самым задает множество операций SetO, которые могут использоваться при задании КС как гипотез, а также множество SetKC всех КС, которые хочет проверить эксперт или которые надо обнаружить автоматически.

Пользователь определяет множество операций SetO, которые буду применяться к КС, тем самым формулируя гипотезы экспертов в виде КС и последовательно их уточняя. Также необходимо задать параметры, по которым производится отбор КС.

На первом шаге за начальную популяцию сигналов берутся элементарные сигналы. С увеличением шага КС текущей популяции уточняются. Для уточнения текущего КС делается следующее:

1. Выбирается один из элементарных сигналов T данного КС.
2. Из набора операций SetO берется одна из операций O и осуществляется замена T на O, примененную к некоторым другим элементарным сигналам.
3. У получившегося КС проверяется *критерий отбора* (см. далее):

а) если он выполнился, то данный КС записывается в результирующее множество ResCS;

б) иначе – проверяется *критерий ветвлений* (см. далее). В случае его выполнения сигнал переносится в следующую популяцию;

в) если ни один из предыдущих критериев не выполнен, то КС отсеивается.

Далее рассматривается следующий КС текущей популяции, когда сигналы в текущей популяции закончились, алгоритм переходит к следующей популяции. Цикл продолжается, пока популяция не опустеет. Результаты работы алгоритма – множество полученных КС ResCS. Следует отметить, что каждый полученный КС более значим и вероятен, чем любой его подсигнал.

Для проверки КС нужны две выборки – позитивная и негативная. Назовем их условно YES и NO. Выборка YES содержит последовательности, которые заведомо содержат некоторые сигналы. Последовательности выборки NO заведомо не содержат эти сигналы, либо эти

последовательности сгенерированы случайно и нужны для проверки статистических параметров сигналов.

В системе используются следующие *критерии отбора* КС:

- порог *условной вероятности* КС – минимальное значение условной вероятности, которое должен иметь сигнал. Также проверяется, что сигнал более вероятен, чем предыдущий подсигнал;
- порог *статистической значимости по критерию Фишера* для проверки 3 и 4 свойств семантического вероятностного вывода;
- если установлена *минимизация уровня значимости по критерию Фишера*, то проверяется, что сигнал более значим, чем предыдущий подсигнал;
- порог *статистической значимости по критерию Юла* [25];
- порог *покрытия позитивной выборки*;
- проверка на уникальность. На разных шагах могут быть найдены сигналы с одинаковой структурой. Можно выбирать между сохранением всех сигналов или только уникальных.

Для проверки *критерия ветвления*:

- порог *условной вероятности* КС. Также проверяется, что получившийся после ветвления сигнал более вероятен, чем исходный;
- порог *статистической значимости по критерию Фишера*;
- если установлена минимизация уровня значимости по критерию Фишера, то проверяется, что получившийся после ветвления сигнал более значим, чем исходный;
- *минимальная сложность* (количество входящих в его состав операций) КС;
- *максимальная сложность* КС;
- условия на *корреляцию аргументов операции «дистанция»* в КС.

При проверке получения результата или продолжения ветвления используются следующие критерии.

1. *Условная вероятность P* принадлежности данного КС выборке YES:

$$P = a_{11} / (a_{10} + a_{11}),$$

где

a_{11} – общее количество реализаций сигнала на выборке YES;

a_{10} – общее количество реализаций сигнала на выборке NO.

2. *Статистическая значимость сигнала по критерию Фишера* (точный критерий независимости Фишера для таблиц сопряженности [26]). Для вычисления уровня значимости f используются 4 величины:

t_{00} – количество негативных последовательностей, на которых не реализован сигнал;

t_{01} – общее число реализаций сигнала на позитивной выборке;

t_{10} – общее число реализаций сигнала на негативной выборке;

t_{11} – количество позитивных последовательностей, на которых не реализован сигнал:

$$f = (t_{00} + t_{01})!(t_{10} + t_{11})!(t_{00} + t_{10})!(t_{01} + t_{10})! / ((t_{00} + t_{01} + t_{10} + t_{11})! t_{00}! t_{01}! t_{10}! t_{11}!).$$

3. *Статистическая значимость сигнала по критерию Юла*.

4. *Покрытие позитивной выборки* в процентах (для последовательностей позитивной выборки, содержащих сигнал).

5. *Покрытие негативной выборки* в процентах (для последовательностей негативной выборки содержащих сигнал).

Для операции «дистанция» оценивается *уровень корреляции между аргументами*.

Система UGENE

Целью проекта UGENE является качественная интеграция различных алгоритмов анализа генетических данных в единой рабочей среде [23]. Среди таких алгоритмов: поиск шаблонов, локальное выравнивание (Smith-Waterman), HMMER, поиск сайтов рестрикции, выравнивание на геном (Bowtie, UGENE genome aligner), филогенетический анализ, множественное выравнивание (MUSCLE, KAlign) и т. д. Реализованы оригинальные конструкторы: для вычислительных схем (Workflow Designer [27]) и для комплексных запросов (Query Designer). Одним из основных преимуществ UGENE является адаптация алгоритмов для использования

общей внутренней модели данных при встраивании в пакет. Это позволяет различным модулям эффективно «общаться» друг с другом без дополнительных усилий на конвертацию данных. Также UGENE поддерживает запись и чтение порядка 20 распространенных форматов биологических данных. Некоторые алгоритмы оптимизированы для использования современной аппаратной базы (мультипроцессорные вычисления, NVIDIA CUDA, ATI Stream Technology и т. д.).

Большое внимание уделяется визуализации полученных результатов алгоритмов и эффективному взаимодействию пользователя с системой через удобный интерфейс. Созданы и отлажены средства для отображения последовательностей и их аннотаций, выравниваний, 3D-структур и филогенетических деревьев, сборок ДНК и т. д. Есть возможность взаимодействия с удаленными базами данных (NCBI, Genbank, PDB и др.).

UGENE реализована на Qt4, инструментарии разработки на языке C++, что обеспечивает системе поддержку всех популярных платформ: Win, *nix и Mac. Проект распространяется по лицензии GPLv2 с открытым исходным кодом.

Интегрированная система UGENE и ExpertDiscovery

Интегрированная система является достаточно мощным средством иерархического анализа регуляторных областей. Генерируя разметки разными методами, мы позволяем системе осуществлять распознавание на высоких уровнях иерархии, что не умеют делать другие программы. Таким образом, мы можем получить и исследовать модель сложной регуляторной области. И весь этот функционал доступен в контексте одного программного пакета.

В UGENE используется система подключаемых модулей (plugin'ов): каждый дополнительный функционал UGENE выделяется в отдельный модуль и может быть отключен или подключен пользователем, модули могут взаимодействовать друг с другом.

Алгоритмы системы ExpertDiscovery хорошо вписываются в концепцию UGENE, поэтому было решено интегрировать их в проект, оформив в виде отдельного модуля, который бы повторял и расширял возможности ExpertDiscovery.

ExpertDiscovery, встроенная в UGENE, обладает рядом преимуществ.

1. Мультиплатформенность.
2. Единая система:

а) реализация различных алгоритмов в рамках одного проекта, несомненно, дает больше возможностей, чем большое количество разрозненных узконаправленных приложений. Такой подход облегчает труд пользователя: достаточно запустить UGENE и получить доступ к широкому ряду алгоритмов (в том числе ExpertDiscovery) вместо того, чтобы запускать различные несвязанные программы;

б) модули UGENE объединены некоторым общим интерфейсом и логикой работы с пользователем. Пользователь, уже знакомый с UGENE, сможет быстрее освоить возможности нового модуля, находясь в единой интерфейсной и функционирующей по знакомым принципам среде. Так, ExpertDiscovery использует уже наработанные интерфейсные и визуальные решения (отображение последовательностей, аннотирование, запуск вычислительных заданий и т. д.);

в) появляются возможности по расширению и комбинации результатов. Например, на вход ExpertDiscovery, в качестве разметок, можно подать результаты алгоритмов, реализованных в UGENE (SITECON, Weight Matrix, Query Designer и т. д.);

г) форматы данных. В ExpertDiscovery можно загружать последовательности в любом поддерживаемом UGENE формате (FASTA, FASTAQ, Genbank, GFF, EMBL и др.).

UGENE расширяет расширяет библиотеку КС ExpertDiscovery. Во встроенной версии системы элементарными сигналами могут быть сигналы, представленные в табл. 1. Благодаря UGENE Query Designer можно создать разметку, содержащую результаты работы различных агоритмов (SITECON, PWM, Repeat Finder) (рис. 2), полное описание алгоритмов и их запуска в Query Deisigner можно найти в документации UGENE. Что касается иерархического анализа регуляторных областей, то средствами UGENE можно создавать разметки последовательностей элементарными сигналами, которые будут загружены в ExpertDiscovery для дальнейшего анализа и построения более сложных моделей регуляторных областей. Например, это можно сделать с помощью встроенного конструктора вычислительных схем (Workflow Designer), создав соответствующую схему.

Таблица 1

Элементарные сигналы ExpertDiscovery

Элементарный сигнал	Описание
Нуклеотиды, контекстные сигналы, любые слова в расширенном коде IUPAC [28]	Дополнительно к разметке нуклеотидами пользователь может загрузить произвольные контекстные сигналы, обнаруженные, например, с помощью программ поиска «по маске» [29]
Потенциальные ССТФ, распознанные традиционными подходами, методом весовых матриц, статистическими методами	Распознавание возможно на основе матриц из баз данных JASPAR (511 матриц) и UniPROBE (275 матриц). Кроме того, возможно распознавание и других ССТФ на основе выборки нуклеотидных последовательностей ССТФ (либо готовой матрицы), предоставленной пользователем
Потенциальные ССТФ, распознанные методом SITECON, основанном на анализе консервативных конформационных либо физико-химических свойств ДНК [15]	Возможно распознавание 44 сайтов связывания эукариот, для которых в UGENE имеются модели консервативных конформационных либо физико-химических свойств, выявленные методом SITECON. Кроме того, если пользователь имеет обучающую выборку других ССТФ, в среде UGENE возможно построение модели для данного типа сайтов и распознавание методом SITECON
Разметки, сгенерированные конструктором вычислительных схем (UGENE Workflow Designer)	Инструмент по простому созданию вычислительных схем позволяет генерировать разметки различными методами
Шаблоны, найденные с помощью конструктора запросов (UGENE Query Designer)	UGENE Query Designer позволяет выявлять сигналы с различной функциональной значимостью: открытые рамки считывания, повторы, сайты рестрикции, шаблоны, потенциальные ССТФ (найденные методами Weight Matrix и SITECON)
Любые аннотации последовательностей, загруженные в формате genbank *	Открытые базы данных содержат аннотации, в том числе и экспериментально подтвержденные данные о расположении ССТФ, в распространенном формате для записи последовательностей и их аннотаций
КС, обнаруженные системой ExpertDiscovery	Любой КС может быть добавлен в разметку и использоваться в качестве элементарного при выделении более сложных КС

* http://www.ncbi.nlm.nih.gov/collab/FT/#5.1

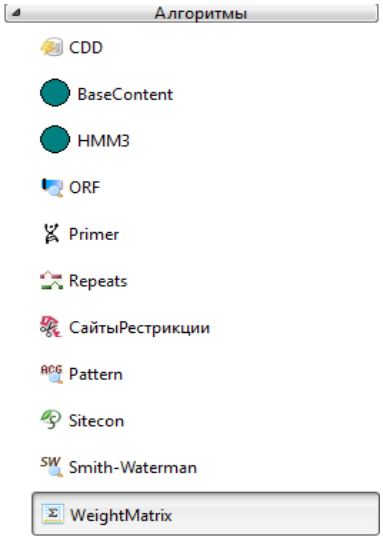

Рис. 2. Алгоритмы, доступные через UGENE Query Designer

На рис. 3 представлена вычислительная схема, обеспечивающая разметку последовательности ДНК ССТФ, найденными методом весовых матриц. Для решения этой задачи создана схема, содержащая элемент считывания последовательностей, считывания весовой матрицы, элемент, выполняющий распознавание и записывающий результат в файл. Результатом выполнения данной схемы является файл с аннотациями последовательностей сайтом связывания IRF. Далее этот файл можно загрузить в систему ExpertDiscovery в качестве разметки, и сайт IRF будет выступать в роли элементарного сигнала.

Отметим, что в качестве разметки для ExpertDiscovery может выступать любой файл с аннотациями в формате genbank. Например, можно получить разметку с помощью похожей вычислительной схемы, использовать метод SITECON вместо весовой матрицы, запустить поиск элементов с помощью UGENE Query Designer или других средств UGENE.

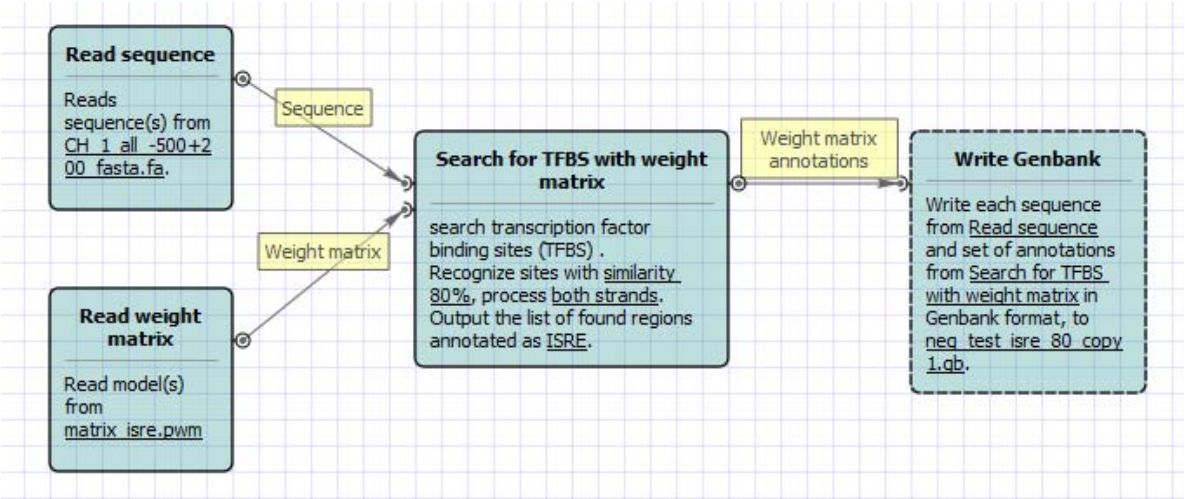

Рис. 3. Схема UGENE Workflow Designer по созданию разметок последовательностей методом весовых матриц

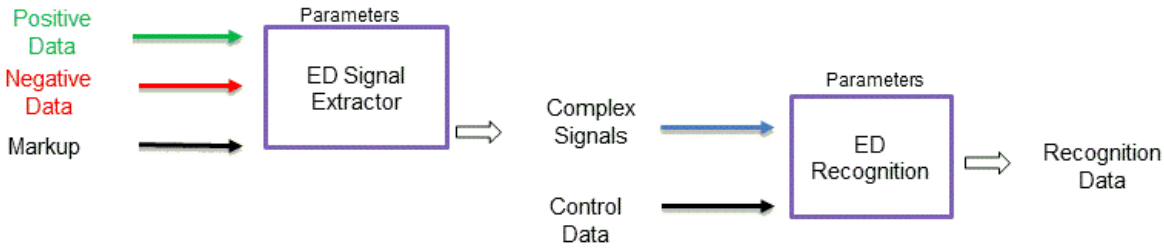

Рис. 4. Основная итерация работы пользователя системы ExpertDiscovery

Основной цикл работы программы ExpertDiscovery выглядит следующим образом (рис. 4). В системе ExpertDiscovery можно выделить логически две части: первая отвечает за построение КС (ED Signal Extractor); вторая – за распознавание КС на последовательностях (ED Recognition).

Эксперт загружает позитивную выборку последовательностей (Positive), содержащую интересующий регуляторный объект, и негативную выборку (Negative), которая этот объект не содержит. На основании этих двух выборок будет проходить обучение системы. Также необходимо загрузить разметки этих последовательностей элементарными сигналами (Markup), на основании которых будут построены комплексные закономерности, и установить параметры (Parameters) распознавания. В результате работы алгоритма выделения сигналов пользователь получает КС (Complex Signals). Далее он может распознать нужные ему КС на последовательностях контрольной выборки (Control). Пользователь устанавливает порог распознавания и, в итоге, получает данные распознавания (Recognition Data) в виде HTML-отчета или профиля распознавания.

ExpertDiscovery представлена в виде plugin'а к системе UGENE, запускается из меню Tools в главном окне UGENE (рис. 5).

Функционал по управлению документом, загрузке разметок, выделению сигналов и т. д. представлен в виде кнопок на панели инструментов в верхней части окна. Окно разделено на три области. Левая верхняя часть содержит иерархический список элементов системы: последовательностей (позитивных, негативных, контрольных), разметок, сигналов. Левая нижняя часть отображает свойства выбранного элемента. Справа – область отображения выбран-

ных последовательностей, выбранный сигнал отображается в виде аннотаций на этих последовательностях.

Загрузка данных. Для построения модели регуляторной области системе необходимы обучающие данные: позитивная и негативная выборка нуклеотидных последовательностей. Последовательности позитивной выборки содержат интересующий эксперта район. Это может быть выборка последовательностей, содержащая сайты связывания определенного типа, или выборка определенной группы генов.

В качестве негативной выборки, как правило, используют набор последовательностей, не содержащих исследуемый сигнал (либо набор сигналов). Однако зачастую сформировать такую выборку до выполнения этапа компьютерного анализа бывает достаточно трудно. Поэтому негативная выборка может включать так называемые «случайные последовательности», сгенерированные автоматически с сохранением частот встречаемости символов относительно позитивной выборки.

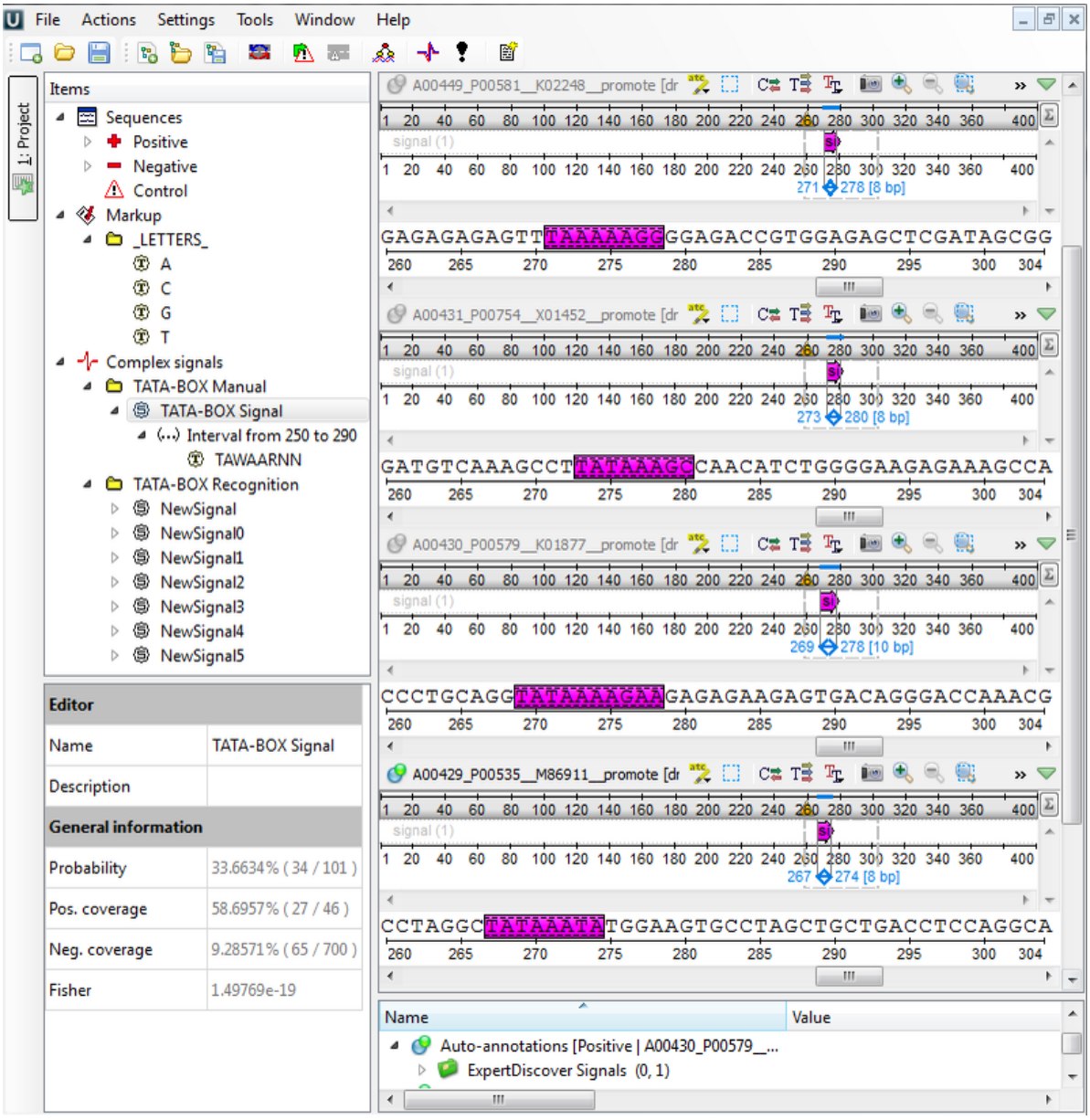

Рис. 5. Главное окно ExpertDiscovery

Диалоговое окно для загрузки выборок нуклеотидных последовательностей (рис. 6) вызывается с помощью кнопки «New ExpertDiscovery Document» на панели инструментов. Выбираются файлы в любом формате последовательностей, поддерживаемом UGENE.

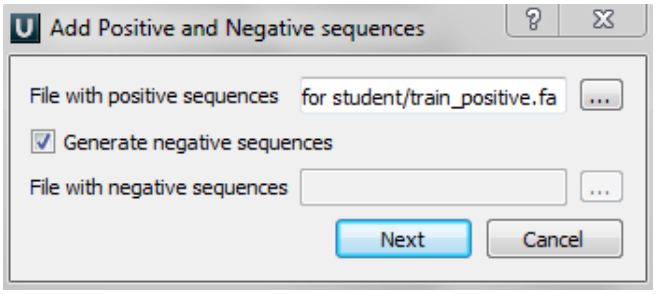

Рис. 6. Диалоговое окно для загрузки выборок последовательностей ДНК

Далее, необходимо загрузить разметку последовательностей элементарными сигналами, на основании которых будут построены КС. Можно сделать разметку нуклеотидами или загрузить полученный ранее файл. Как правило, разметка характеризуется местами в последовательности, в которых присутствует элементарный сигнал, и именем включающего его семейства.

С помощью кнопки «Load Markup», расположенной на панели инструментов, вызывается диалоговое окно загрузки (рис. 7). Если не указан флаг «Append to Current Markup», то старая разметка будет удалена.

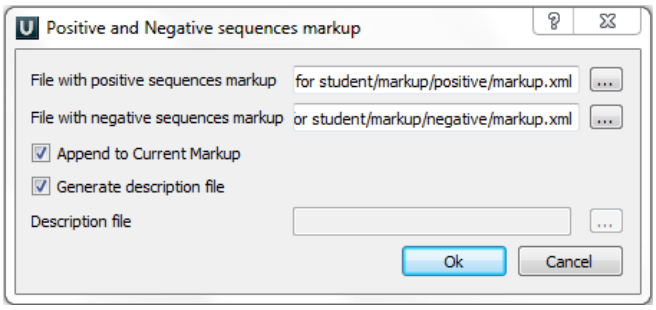

Рис. 7. Диалоговое окно для загрузки разметок нуклеотидных последовательностей

Редактирование КС. Для ручного создания КС нужно воспользоваться контекстным меню вкладки «Complex signals» в окне проекта «Items». Для удобства можно создавать группирующие папки.

КС представлен в виде иерархического дерева, узлами которого являются операции, а листьями – элементы разметок или слова.

Когда КС создан и выделен в области параметров, можно менять его структуру и отслеживать параметры. Доступными типами вершин дерева являются: операция «дистанция» (бинарная), операция «повтор», операция «интервал», элемент разметки, слово. КС считается определенным до конца, когда все его листья содержат терминальные символы – слова или элементы разметок.

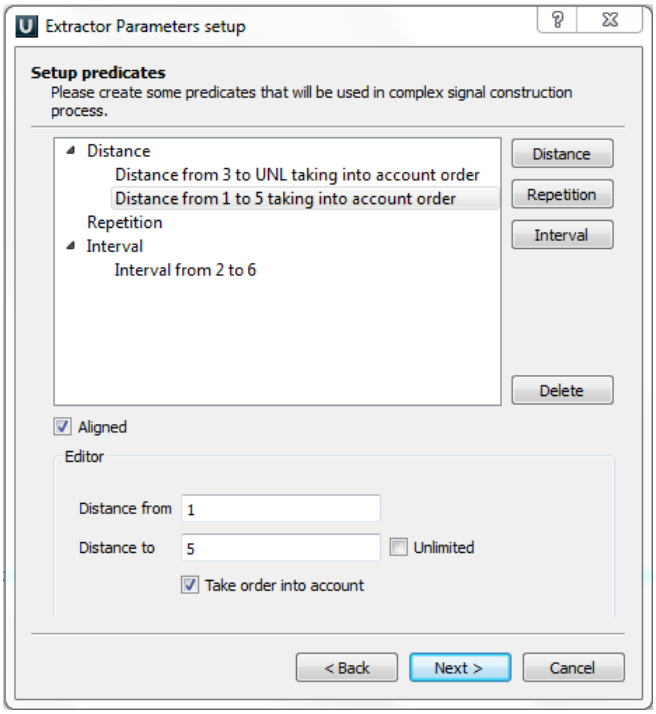

Рис. 8. Диалоговое окно создания условий для выделения КС

Автоматическое построение КС. На основании обучающих данных (позитивная и негативная выборки, разметки) система может восстановить структуру регуляторной области в виде КС. Мастер настройки процесса распознавания доступен через кнопку «Extract signals» на панели инструментов.

С помощью первого окна (рис. 8) задаются параметры распознавания (описано выше). На следующих этапах указываются операции, которые будут узлами КС, и осуществляется выбор папки проекта, куда по мере генерации будут добавляться КС.

Для того чтобы увидеть расположение КС на последовательности, необходимо выбрать последовательности для отображения. Это делается через контекстное меню последовательности. Далее можно выбрать любой КС, и он появится в виде автоаннотации на каждой отображенной последовательности. Более того, имеется возможность увидеть сразу несколько сигналов на последовательности, для этого через контекстное меню сигнала пользователь выбирает сигналы для группового отображения. Та же операция применяется при выборе сигналов для распознавания.

Распознавание КС на последовательностях. После автоматического выделения КС их можно распознать на любой последовательности. Выборку таких последовательностей можно загрузить дополнительно, как контрольную.

Для распознавания пользователь выбирает некоторое множество КС, каждый из которых применяется к последовательности. Затем к символу последовательности, где встретился сигнал, прибавляется вес, равный $-\log(1 - P)$, где P есть значение условной вероятности сигнала. Общий вес последовательности рассчитывается как суммарный вес всех ее символов. Последовательность считается распознанной, т. е. на ней присутствует выбранный КС, если ее вес больше выбранного порогового значения. По позитивным данным эксперт может подобрать это значение. Выбор порога осуществляется через диалоговое окно, вызываемое с помощью кнопки «Set recognition bound», в данном окне отображается ошибка первого и второго рода для выбранного значения.

Также для удобства пользователя можно сгенерировать отчет о распознавании в формате HTML. Отчет содержит статистические параметры и результат распознавания для каждой последовательности.

Заключение

UGENE вместе с модулем ExpertDiscovery представляет собой удобный и эффективный инструмент для автоматизации работы экспертов. Интегрированная система позволяет комбинировать результаты различных методов, что дает возможность обнаружения сложных иерархических закономерностей. Работа осуществляется в контексте одной системы, что в разы ускоряет продуктивность работы за счет удобного и быстрого взаимодействия между модулями и отсутствия необходимости конвертировать данные.

Разрабатываемый оригинальный подход к интеграции данных от различных методов и знаний экспертов показал свою эффективность на практике. Полученные результаты опубликованы в серии статей [20–22; 30; 31].

Список литературы

1. *Кель А. Э., Колчанов Н. А., Кель О. В., Ромащенко А. Г., Ананько Е. А., Игнатьева Е. В., Меркулова Т. И., Подколодная О. А., Степаненко И. Л., Кочетов А. В., Колпаков Ф. А., Подколодный Н. Л., Наумочкин А. А.* TRRD – база данных транскрипционных регуляторных районов генов эукариот // Молекулярная биология. 1997. Т. 31. С. 626–636.

2. *Кель О. В., Кель А. Э., Ромащенко А. Г., Вингендер Э., Колчанов Н. А.* Композиционные регуляторные элементы: классификация и описание в базе данных COMPEL // Молекулярная биология. 1997. Т. 31, № 4. С. 601–615.

3. *Lemon B., Tjian R.* Orchestrated Response: A Symphony of Transcription Factors for Gene Control // Genes Dev. 2000. Vol. 14. No. 20. P. 2551–2569.

4. *Arnone M. I., Davidson E. H.* The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems // Development. 1997. Vol. 124 (10). P. 1851–1864.

5. *Kolchanov N. A., Podkolodnaya O. A., Ananko E. A., Ignatieva E. V., Stepanenko I. L., Kel-Margoulis O. V., Kel A. E., Merkulova T. I., Goryachkovskaya T. N., Busygina T. V., Kolpakov F. A., Podkolodny N. L., Naumochkin A. N., Korostishevskaya I. M., Romashchenko A. G., Overton G. C.* Transcription Regulatory Regions Database (TRRD): Its Status in 2000 // Nucleic Acids Res. 2000. Vol. 28. No. 1. P. 298–301.

6. *Kovalerchuk B., Vityaev E.* Data Mining in Finance: Advances in Relational and Hybrid Methods. (Kluwer international series in engineering and computer science; SECS 547). Kluwer Academic Publishers, 2000. 308 p.

7. *Колчанов Н. А., Подколодная О. А., Ананько Е. А., Игнатьева Е. В., Степаненко И. Л., Хлебодарова Т. М., Меркулова Т. И., Меркулов В. М., Мищенко Е. Л., Ибрагимова С. С., Смирнова О. Г., Подколодный Н. Л., Ромащенко А. Г., Ощепков Д. Ю., Мигинский Д. С.* Регуляторные последовательности ДНК: описание в базах данных // Системная компьютерная биология / Под ред. Н. А. Колчанова, С. С. Гончарова, В. А. Лихошвай, В. А. Иванисенко. Серия: Интеграционные проекты. Новосибирск: Изд-во СО РАН, 2008. Вып. 14. С. 15–37.

8. *Nikolov D. B., Burley S. K.* RNA Polymerase II Transcription Initiation: A Structural View // Proc. Natl. Acad. Sci. USA. 1997. Vol. 94. P. 15–22.

9. *Kel O. V., Romaschenko A. G., Kel A. E., Wingender E., Kolchanov N. A.* A Compilation of Composite Regulatory Elements Affecting Gene Transcription in Vertebrates // Nucleic Acids Res. 1995. Vol. 23 (20). P. 4097–4103.

10. *Caley M., Smale S. T.* Transcriptional Regulation in Eukaryotes. Cold Spring Harbor; N. Y.: Cold Spring Harbor Laboratory Press, 2000. 640 p.

11. *Trifonov E. N.* Genetic Level of DNA Sequences Is Determined by Superposition of Many Codes // Mol. Biol. 1997. Vol. 31. P. 759–767.

12. *Fulton D. L., Sundararajan S., Badis G., Hughes T. R., Wasserman W. W., Roach J. C., Sladek R.* TFCat: The Curated Catalog of Mouse and Human Transcription Factors // Genome Biol. 2009. Vol. 10 (3). P. R29.

13. *Quandt K., Frech K., Karas H., Wingender E., Werner T.* MatInd and MatInspector: New Fast and Versatile Tools for Detection of Consensus Matches in Nucleotide Sequence Data // Nucleic Acids Res. 1995. Vol. 23 (23). P. 4878–4884.

14. *Stormo G. D.* DNA Binding Sites: Representation and Discovery // Bioinformatics. 2000. Vol. 16. P. 16–23.

15. *Oshchepkov D. Y., Vityaev E. E., Grigorovich D. A., Ignatieva E. V., Khlebodarova T. M.* SITECON: A Tool for Detecting Conservative Conformational and Physicochemical Properties in Transcription Factor Binding Site Alignments and for Site Recognition // Nucleic Acids Res. 2004. Vol. 32. P. 208–212.

16. *Levitsky V. G., Ignatieva E. V., Ananko E. A., Turnaev I. I., Merkulova T. I., Kolchanov N. A., Hodgman T. C.* Effective Transcription Factor Binding Site Prediction Using a Combination of Optimization, a Genetic Algorithm and Discriminant Analysis to Capture Distant Interactions // BMC Bioinformatics. 2007. Vol. 8 (1). P. 481.

17. *Kolchanov N. A., Merkulova T. I., Ignatieva E. V., Ananko E. A., Oshchepkov D. Y., Levitsky V. G., Vasiliev G. V., Klimova N. V., Merkulov V. M., Charles Hodgman T.* Combined Experimental and Computational Approaches to Study the Regulatory Elements in Eukaryotic Genes // Brief Bioinform. 2007. Vol. 8 (4). P. 266–274.

18. *Kel A., Kel-Margoulis O., Babenko V., Wingender E.* Recognition of NFATp/AP-1 Composite Elements within Genes Induced upon the Activation of Immune Cells // J. Mol. Biol. 1999. Vol. 288 (3). P. 353–376.

19. *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.

20. *Khomicheva I. V., Vityaev E. E., Shipilov T. I., Levitsky V. G.,* Transcription Factor Binding Sites Recognition by the ExpertDiscovery System Based on the Recursive Complex Signals // Proc. of the V International Conference on Bioinformatics of Genome Regulation and Structure (BGRS2006, 16–22 July, Novosibirsk, Russia). Novosibirsk, 2006 Vol. 1. P. 77–80.

21. *Khomicheva I. V., Vityaev E. E., Ananko E. A., Levitsky V. G., Shipilov T. I.* Hierarchical Analysis of the Eukaryotic Transcription Regulatory Regions Based on the DNA Codes of Transcription // Proc. of the III Moscow Conference on Computational Molecular Biology. Moscow, Russia, 2007. P. 142–144.

22. *Khomicheva I., Demin A., Vityaev E.* Transcription Factor Binding Site Discovery by the Probabilistic Rules // PKDD Proc. XI European Conference on Principles and Practice of Know-

ledge Discovery in Databases / Eds. J. N. Kok, J. Koronacki et al. Warsaw, Poland, 2007. P. 104–109.

23. *Okonechnikov K., Golosova O., Varlamov A., Fursov M.* Unipro UGENE: An Open Source Toolkit for Complex Genome Analysis // Proc. of the XII Annual Bioinformatics Open Source Conference. URL: http://www.oboedit.org/BOSC2011/BOSC2011-program.pdf

24. *Витяев Е. Е.* Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. Новосибирск, 2006.

25. *Yule U.* On the Association of Attributes in Statistics // Philosophical Transactions of the Royal Society of London. Ser. A. 1990. Vol. 194. P. 257–319.

26. *Кендал М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.

27. *Fursov M. Y., Oshchepkov D. Y., Novikova O. S.* UGENE: Interactive Computational Schemes for Genome Analysis // Proc. of the V Moscow International Congress on Biotechnology. 2009. Vol. 3. P. 14–15.

28. *Cornish-Bowden A.* Enzyme kinetics // Comprehensive Biotechnology. 1985. Vol. 1. P. 521–538.

29. *Grillo G., Licciulli F., Liuni S., Sbisa E., Pesole G.* PatSearch: A Program for the Detection of Patterns and Structural Motifs in Nucleotide Sequences // Nucleic Acids Res. 2003. Vol. 31 (13). P. 3608–3612.

30. *Khomicheva I. V., Vityaev E. E., Ananko E. A., Shipilov T. I., Levitsky V. G.* ExpertDiscovery System Application for the Hierarchical Analysis of the Eukaryotic Transcription Regulatory Regions Based on the DNA Codes of Transcription // Intelligent Data Analysis. 2008. Vol. 12. No. 5. P. 481–494.

31. *Khomicheva I. V., Vityaev E. E., Shipilov T. I.* Discovery of the Transcription Factor Binding Sites in the Aligned and Unaligned DNA Sequences // Proc. of the V International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008, 22-28 June, Novosibirsk, Russia). Novosibirsk, 2008. P. 116.

Материал поступил в редколлегию 02.12.2011

Yu. Yu. Vaskin, I. V. Khomicheva, E. V. Ignatieva, E. E. Vityaev

ANALYSIS OF REGULATORY REGIONS OF GENES BY EXPERTDISCOVERY RELATION SYSTEM, INTEGRATED INTO UGENE TOOLKIT

The task of automatic extraction of hierarchical structure of eukaryotic gene regulatory region is posed on the junction of the fields of biology, mathematics and information technologies. A solution of the problem implies understanding of the sophisticated mechanisms of eukaryotic gene regulation and applying data mining technologies for analysis with many features. The paper discusses the integrated system implementing a powerful relation mining of biological data. The system allows taking into account prior information about the analyzed data that is known by the biologist, performing the analysis on each hierarchical level, searching for a solution from a simple hypothesis to a complex one. The integration of the system provides a convenient environment for conducting complex research and automating the work of the biologist.

Keywords: complex signal, hierarchical analysis, integrated system, relation data mining, recognition, genes regulatory regions, annotation.