

**И. В. Хомичева**

Место работы: Институт цитологии и генетики СО РАН

Должность: м.н.с.

Научная степень и звание: магистр математики

Адрес: ул. Вознесенская, 2, Новосибирск, 630559, Россия

E-mail: [khomicheva@bionet.nsc.ru](mailto:khomicheva@bionet.nsc.ru)

**Е. Е. Витяев**

Место работы: Институт математики СО РАН

Должность: Старший научный сотрудник

Научная степень и звание: д.ф.-м.н.

Адрес: ул. Демакова, 18, Новосибирск, 630128, Россия

E-mail: [vityaev@math.nsc.ru](mailto:vityaev@math.nsc.ru)

**Игнатьева Е.В.**

Место работы: Институт цитологии и генетики СО РАН

Должность: с.н.с.

Научная степень и звание: к.б.н.

Адрес: пр.Ак.Лаврентьева, 10, Новосибирск, 630090, Россия

E-mail: [eignat@bionet.nsc.ru](mailto:eignat@bionet.nsc.ru)

**Ананько Е.А.**

Место работы: Институт цитологии и генетики СО РАН

Должность: с.н.с.

Научная степень и звание: к.б.н.

Адрес: пр.Ак.Лаврентьева, 10, Новосибирск, 630090, Россия

E-mail: [eananko@bionet.nsc.ru](mailto:eananko@bionet.nsc.ru)

**Т. И. Шипилов**

Место работы: Институт математики СО РАН

Должность: аспирант

E-mail: [tshipilov@gmail.com](mailto:tshipilov@gmail.com)

## **Применение программной системы ExpertDiscovery для поиска закономерностей структурно-функциональной организации регуляторных районов генов**

### **Аннотация**

Появление качественно новых экспериментальных технологий в таких областях современной биологии, как геномика, транскриптомика, протеомика, клеточная биология, нанобиоинженерия и др., привело к экспоненциальному росту объемов экспериментальных данных, требующих систематизации и осмысления. Новые методы интеллектуального анализа данных призваны решить задачу интеграции первичных экспериментальных данных, которые слабо связаны, плохо структурированы, имеют разную степень полноты и сами по себе не позволяют реконструировать полноценный портрет изучаемой биологической системы или процесса. Одной из таких сложных и не решенных задач является задача выявления закономерностей организации регуляторных районов генов. Для решения этой задачи нами разработан интегрированный метод извлечения знаний ExpertDiscovery, обнаруживающий комплексные закономерности организации регуляторных районов генов эукариот. В качестве элементарных сигналов

для построения комплексных сигналов система использует различные характеристики обнаруженные, например, другими методами извлечения знаний. Объединяя закономерности, обнаруженные на всех уровнях исследования, система ExpertDiscovery позволяет построить иерархическую модель регуляторных районов специфической группы генов.

### **Ключевые слова**

Комплексный сигнал, реляционный метод извлечения знаний, интеграционный подход, иерархический анализ, регуляторные районы генов, распознавание, сравнение оценок точности

## **1. Введение**

В основе создания медицинских препаратов нового поколения, предупреждения, профилактики наследственных заболеваний и др., лежит задача управления экспрессией генов эукариот. Экспрессия генов – сложный многостадийный процесс, первым этапом которого является транскрипция. У эукариотических организмов транскрипция осуществляется в ядрах клеток. В ходе транскрипции происходит синтез определенного количества продуктов генов - молекул РНК. Интенсивность транскрипции каждого конкретного гена подвержена очень точной регуляции в зависимости от клеточных условий (типа клеток и тканей, стадии развития организма, клеточного цикла, индукторам либо репрессорам, действующим на клетки).

Возможность гибкой регуляции транскрипции генов эукариот обеспечивается наличием протяженных регуляторных районов генов, имеющих сложную блочно-иерархическую структуру [Dyran, 1989; Arnone, Davidson, 1997].

Первый уровень иерархии включает сайты связывания различных транскрипционных факторов (ССТФ), короткие участки ДНК, служащие местом посадки для регуляторных белков (транскрипционных факторов) [Nikolov, Burley, 1997]. Встречаемость и расположение ССТФ в регуляторных районах генов отражает ткане- и стадие-специфичные особенности регуляции их экспрессии. Все известные к настоящему времени методы распознавания ССТФ имеют достаточно высокие уровни недопредсказания (либо перепредсказания). Причиной этому является большое разнообразие ДНК-белковых взаимодействий между сайтами и транскрипционными факторами, различные ткане-, стадиеспецифичные механизмы регуляции транскрипции, специфичность контекста, окружающего ССТФ в регуляторных районах.

Следующий иерархический уровень организации регуляторных районов генов соответствует упорядоченным сочетаниям ССТФ: композиционным элементам и цис-регуляторным модулям (CRM). Пара сближенных сайтов образует композиционный элемент, если совместный эффект взаимодействующих с ними регуляторных белков существенно отличается от эффекта, который мог бы получиться в результате их простого суммирования. При этом эффект может оказаться синергичным (неаддитивно высоким), либо, наоборот, антагонистичным (смена активации на репрессию).

Цис-регуляторные модули (CRM), включают устойчивые сочетания сайтов связывания факторов различных типов и других мотивов [Blanchette M. et al., 2006]. Их наличие характерно для регуляторных районов генов, экспрессирующихся тканеспецифическим образом и отражает наличие тканеспецифичного набора транскрипционных факторов, функционирующих в каждой конкретной ткани. Задача, соответствующая данному уровню иерархии строения регуляторных районов генов, состоит в обнаружении закономерностей

расположения ССТФ. Однако, поскольку каждый ген содержит уникальную комбинацию ССТФ в своем регуляторном районе, необходимую для регуляции экспрессии в определенных условиях, разрабатываемые методы сталкиваются с плохой репрезентативностью данных обучения, содержащих недостаточное число частных случаев более общего явления.

Высший уровень иерархии строения регуляторных районов генов соответствует системе интегральной регуляции транскрипции, основанной на суперпозиции разных кодов ДНК (линейных, конформационных), [Trifonov, 1997].

Анализ регуляторных последовательностей генов представляет собой актуальную проблему биологии и вызов для разработки новых методов извлечения знаний (Knowledge Discovery in Databases and Data Mining, KDD&DM).

Такие подходы KDD&DM, как нейронные сети, решающие деревья, генетические алгоритмы, Байесовские сети и т.д. эффективно применяются для решения широкого круга задач системной биологии. Несмотря на разнообразие математических подходов направления KDD&DM, использование различных парадигм обучения, гибкость подходов, огромную базу приложений в различных областях, методы KDD&DM не предлагают адекватного решения задачи распознавания регуляторных районов генов эукариот [Витяев и др., 2008]. Эти подходы, как правило, чувствительны только к конкретным характеристикам и, как следствие, дают хорошие результаты распознавания на одной группе последовательностей и низкую точность распознавания на другой.

Для решения задачи анализа и распознавания регуляторных районов генов эукариот в общем случае необходимо учитывать различные контекстные, физико-химические и конформационные особенности ДНК, таким образом, моделируя процесс распознавания регуляторных районов эукариотической транскрипционной машиной. Построение интегрированного метода распознавания, который бы объединял сигналы различных типов, полученные в результате применения других методов и таким образом, создавал модель регуляторного района, является актуальной задачей.

Интересным примером исследования в направлении интеграции является программа PromoterExplorer [Xie *et al.*, 2006]. На вход программе подается высоко-размерный вектор, компонентами которого являются существенные для анализа характеристики ДНК, такие как: (1) локальное распределение совершенных пентамеров, обладающих наибольшей апостериорной вероятностью (по Байесу); (2) потенциальные CpG острова; (3) оцифрованная последовательность ДНК. Далее происходит многоуровневое обучение программы отличать последовательности промоторов от других последовательностей, не принадлежащих анализируемому классу. Причем на каждом очередном уровне обучения решающий функционал усложняется.

## **2. Задача распознавания регуляторных районов генов эукариот.**

Задача анализа и распознавания регуляторных районов генов эукариот является достаточно сложной и в настоящее время не решенной до конца задачей. Для решения этой задачи мы применили реляционный подход к обнаружению знаний (Relational Data Mining) [Витяев, 2006; Vityaev, Kovalerchuk, 2008; Vityaev, Kovalerchuk, 2004; Kovalerchuk, Vityaev, 2000]. Этот подход и реализующая его система Discovery успешно применялись для решения ряда практических задач в различных областях знаний – психофизике, диагностике раковых заболеваний, предсказании курсов акций ценных бумаг и т.д. (смотри [Scientific Discovery Web Site]).

Реляционный подход состоит в следующем:

(1) из данных извлекается информация, интерпретируемая в онтологии Предметной Области (ПО). Для этого используется теория измерений и онтология ПО с целью символического представления (в логике первого порядка) содержащейся в данных информации. Разработана оригинальная методика такого преобразования [Kovalerchuk, Vityaev, 2008];

(2) символическое представление информации, содержащейся в данных, потребовало разработки логическо-вероятностного метода (в языке первого порядка) их анализа. Такой метод в виде системы Discovery, относящейся к классу методов обнаружения правил (rule-based) в языке первого порядка был разработан авторами статей [Витяев, 2006; Vityaev, Kovalerchuk, 2008; Vityaev, Kovalerchuk, 2004; Kovalerchuk, Vityaev, 2000]. В последнее время подобные методы стали разрабатываться в рамках направления Probabilistic Logic Programming. В отличие от этих методов система Discovery основана на определенном синтезе логики и вероятности [Evgenii Vityaev, 2006] в виде специального Семантического Вероятностного Вывода (СВВ). Идея СВВ состоит в последовательном уточнении гипотез таким образом, чтобы на каждом последующем шаге получались гипотезы с большей вероятностью и определённой. При этом осуществляется проверка статистической значимости полученного результата при помощи статистических критериев.

Формально под *семантическим вероятностным выводом* понимается такая последовательность правил  $C_1, C_2, \dots, C_n$ , что:

1.  $C_i = (A_1^i \& \dots \& A_{k_i}^i \Rightarrow G), i = 1, \dots, n$ ;
2.  $C_i$  - *подправило* правила  $C_{i+1}$ , т.е.  $\{A_1^i, \dots, A_{k_i}^i\} \subset \{A_1^{i+1}, \dots, A_{k_{i+1}}^{i+1}\}$ ;
3.  $\text{Prob}(C_i) < \text{Prob}(C_{i+1}), i = 1, 2, \dots, n-1$ , где *Условная Вероятность* правила  $C_i$  определяется следующим образом:  $\text{Prob}(C_i) = \text{Prob}(G/A_1^i \& \dots \& A_{k_i}^i) = \text{Prob}(G \& A_1^i \& \dots \& A_{k_i}^i) / \text{Prob}(A_1^i \& \dots \& A_{k_i}^i)$ ;
4.  $C_i$  – *Вероятностные Законы*, т.е. для любого подправила  $C' = (A_1 \& \dots \& A_j \Rightarrow G)$  правила  $C_i$ ,  $\{A_1, \dots, A_j\} \subset \{A_1^i, \dots, A_{k_i}^i\}$  выполнено неравенство  $\text{Prob}(C') < \text{Prob}(C_i)$ ;
5.  $C_n$  – *Сильнейший Вероятностный Закон*, т.е. правило  $C_n$  не является подправилом никакого другого вероятностного закона.

Система Discovery практически реализует семантический вероятностный вывод и обнаруживает знания в виде множества вероятностных законов, сильнейших вероятностных законов и максимально специфических законов [Витяев, 2006; Vityaev, Kovalerchuk, 2004; Evgenii Vityaev, 2006].

Реляционный подход был применен для решения задачи анализа регуляторных районов генов. Информацией, извлекаемой из данных (ДНК) являлись комплексные сигналы.

*Комплексные Сигналы* (КС) определяются рекурсивно по индукции на основе элементарных сигналов:

- *элементарный сигнал* является КС;
- *ориентация* КС (прямая, симметричная, инвертированная) является КС;
- *упорядоченная пара* КС с расстоянием между ними, варьирующимся в определенном интервале, является КС. Пользователем указывается, что дистанция между сигналами может варьировать от  $\min$  до  $\max$ , и при этом имеет значение порядок расположения сигналов.
- *принадлежность* КС некоторому интервалу относительно старта транскрипции (либо начала фазированной последовательности) является КС. Указывает, что

входной сигнал следует искать только в интервале от min до max, где min и max абсолютные значения относительно первого символа последовательности.

- *повторение* КС  $N$  раз ( $2 \leq N_{min} \leq N \leq N_{max}$ ) является КС. При этом расстояние между соседними копиями сигнала принадлежит заданному пользователем диапазону. Пользователь указывает  $N_{min}$ ,  $N_{max}$  и диапазон (от min до max) расстояний между соседними повторами.

КС можно представить иерархическим деревом. Пример такого дерева в случае, когда элементарными сигналами являются нуклеотиды, приведен на рис. 1. В этом примере нуклеотид G расположен между нуклеотидами A и T, причем расстояние между A и G, G и T варьируется в диапазонах, заданных пользователем. Этот комплексный сигнал повторяется как минимум дважды ( $N_{min}$ ), и максимум  $N_{max}$  раз в последовательности ДНК.

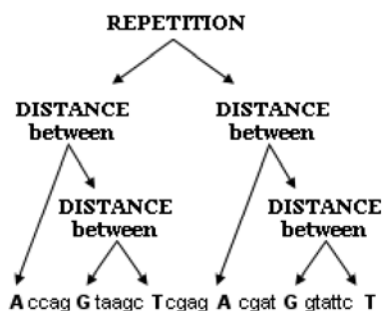


Рисунок 1. Иерархическое дерево комплексного сигнала.

Элементарными сигналами могут быть:

- (1) нуклеотиды, контекстные сигналы, любые слова в расширенном коде IUPAC (Cornish-Bowden, 1985); олигонуклеотиды [Vishnevsky, Kolchanov, 2005];
- (2) потенциальные функциональные сайты, предсказанные по гомологии (или посредством весовой матрицы) с аннотированными последовательностями из специализированных молекулярно-биологических баз данных;
- (3) участок с консервативными для сайтов связывания конформационными или физико-химическими характеристиками (углы двойной спирали между соседними нуклеотидами, температура плавления ДНК) [Oshchepkov *et al.*, 2004];
- (4) элемент вторичной структуры (Z-ДНК, шпилька РНК);
- (5) участок низкой сложности текста (политракт) [Orlov, Potapov, 2004];
- (6) сайты позиционирования нуклеосом [Levitsky *et al.*, 2005].

Элементарный сигнал — неделимый сигнал, который характеризуется именем и местами в последовательности, где он присутствует.

На основе информации, извлеченной из ДНК с помощью КС, был разработан вариант системы Discovery, оперирующей этой информацией. В результате была создана система ExpertDiscovery, позволяющая пользователю (эксперту-биологу) задавать используемые элементарные сигналы и обнаруживать с их помощью КС с параметрами, заданными пользователем.

Элементарные сигналы могут задаваться экспертом интерактивно, а также загружаться в систему в виде аннотации последовательностей ДНК. Они могут быть получены применением известных программ распознавания сигналов. Самым простым примером элементарного сигнала является буква. Более сложным примером является некоторое слово. Другие элементарные сигналы могут соответствовать физико-химическим и конформационным свойствам участков последовательности.

Система ExpertDiscovery, начиная с элементарных сигналов, конструирует КС путем *изменения ориентации* сигналов, взятия *пары сигналов* с некоторым расстоянием, *фиксации некоторого интервала* относительно старта транскрипции, в котором должен находиться сигнал и рассмотрением *повторов сигналов*. Обнаруженные КС используются для создания новых КС. Усложнение КС осуществляется в соответствии с семантическим вероятностным выводом, в котором применяются некоторые статистические критерии (см. §2) для проверки 3,4 свойств СВВ.

Таким образом, эксперт-биолог может автоматически обнаруживать КС, просматривать расположение этих сигналов в последовательностях ДНК и определять их статистические параметры на анализируемой контрастной выборке данных.

### 3. Система ExpertDiscovery.

Ключевым в алгоритме системы ExpertDiscovery является класс рассматриваемых гипотез и процесс их уточнения. Для работы алгоритма требуется определить множество SetO операций (*изменения ориентации*, взятия *пары сигналов*, *фиксации интервала*, *повторение сигналов*), которые будут использоваться для генерации КС, а также задать критерии отбора КС.

На первом шаге алгоритм рассматривает в качестве первой популяции КС все элементарные сигналы. На последующих шагах мы уточняем КС текущей популяции. Для уточнения рассматриваемого КС делается следующее:

- (1) выбирается один из элементарных сигналов данного КС;
- (2) из набора операций SetO берётся одна из операций, и осуществляется замена элементарного сигнала на эту операцию, примененную к некоторым другим элементарным сигналам;
- (3) у полученного КС проверяются *критерии отбора* (см. далее):
  - если они выполнены, то данный КС записывается в результирующее множество ResKC обнаруженных КС;
  - иначе проверяются *критерии ветвления* (см. далее). В случае их выполнения сигнал переносится в следующую популяцию.
  - если ни один из предыдущих критериев не выполнен, то КС отсеивается.

После этого алгоритм переходит к рассмотрению следующего КС текущей популяции. Когда все КС текущей популяции рассмотрены, алгоритм переходит к обработке следующей популяции.

Этот цикл продолжается до тех пор, пока не получится пустая популяция КС. Результатом работы алгоритма является совокупность ResKC обнаруженных КС.

Для вычисления критериев отбора и ветвления КС необходимы две выборки YES и NO. Выборка YES содержит последовательности, содержащие сигналы, выборка NO содержит последовательности некоторых других классов или случайно сгенерированные и используется для подсчёта статистических параметров сигнала.

В системе используются следующие *критерии*, используемые как при *отборе*, так и при *ветвлении* КС:

- *порог условной вероятности* КС – минимальное значение условной вероятности, которое должен иметь сигнал;

- *порог статистической значимости* по критерию Фишера для проверки свойств 3,4 СВВ

Кроме того, для критерия отбора используется *порог покрытия позитивной выборки*, а для критерия ветвления - *минимальная и максимальная сложность* КС, количество входящих в состав КС операций.

Используемые в критериях величины определяются следующим образом:

1. *условная вероятность*  $P = a_{11} / (a_{10} + a_{11})$  принадлежности КС выборке YES, где  $a_{11}$  – общее количество реализаций КС на выборке YES,  $a_{10}$  - общее количество реализаций КС на выборке NO.
2. *статистическая значимость* КС по критерию Фишера – точный критерий независимости Фишера для таблиц сопряженности [Кендал, Стьюарт, 1973]
3. *порог покрытия позитивной выборки* в % – минимальный процент последовательностей выборки YES, содержащих КС.

#### **4. Построение моделей и распознавание сайтов связывания транскрипционных факторов.**

Регуляторные районы генов содержат в своем составе сайты связывания транскрипционных факторов (ССТФ). Компьютерная аннотация ССТФ важна для понимания регуляции экспрессии гена. Вместе с тем, задача распознавания ССТФ, в настоящее время, не может считаться до конца решенной.

В процессе исследования была показана эффективность системы ExpertDiscovery в экспериментах по обнаружению КС и распознаванию как выровненных последовательностей ССТФ, так и не выровненных последовательностей ССТФ.

#### **Эксперимент 1. Выровненные выборки сайтов связывания транскрипционных факторов.**

В случае, когда элементарными сигналами для построения комплексных являются нуклеотиды, система ExpertDiscovery обнаруживает закономерности нуклеотидного контекста [Khomicheva *et al*, 2007а,б,в; Khomicheva *et al*, 2006].

Мы проанализировали последовательности ДНК сайтов связывания 5-ти семейств транскрипционных факторов, SF1 (steroidogenic factor-1), SREBP (sterol regulatory element binding protein), EGR1 (early growth response factor 1), CEBP (CCAAT enhancer-binding protein), HNF4 (Hepatocyte nuclear factor 4). Обучающие данные (расширенные последовательности ДНК для экспериментально подтвержденных ССТФ этих пяти типов) были извлечены из базы данных регуляторных районов транскрипции TRRD [Kolchanov *et al.*, 2002], и проверены экспертами-биологами.

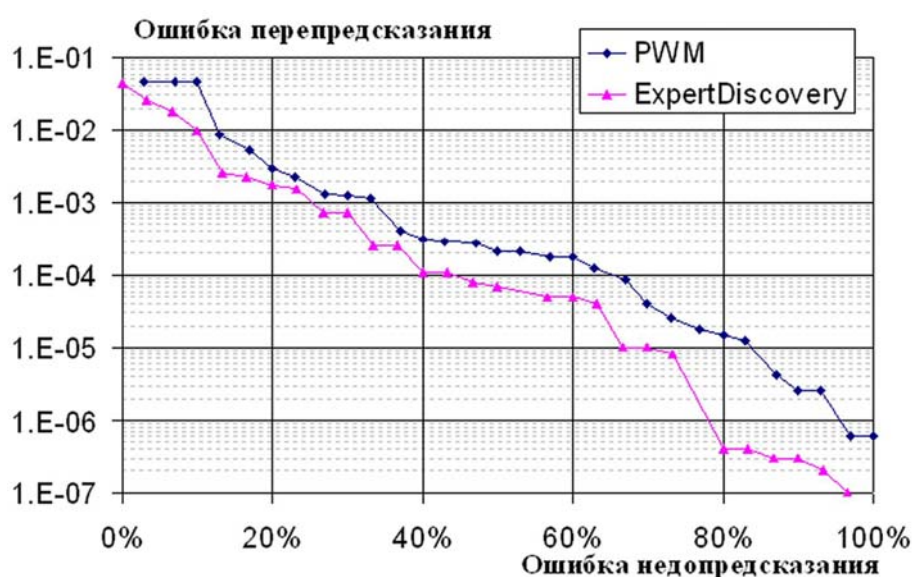
Выборки данных содержали 53 последовательности, соответствующие сайтам связывания SF1, 38 – сайтам SREBP (SRE типа), 22 - EGR1, 88 – CEBP и 30 - HNF4 (табл. 1). Каждая выборка данных содержала последовательности, между которыми установлено соответствие, т.е. осуществлено множественное выравнивание. Контрастные выборки последовательностей, негативное обучение, необходимое для построения метода ExpertDiscovery, и негативный контроль состояли из случайных последовательностей, сгенерированных с сохранением частот встречаемости нуклеотидов, как в выборках реальных последовательностей.

Точность распознавания ССТФ системой ExpertDiscovery оценивалась в сравнении с точностью распознавания ССТФ методом оптимизированной весовой матрицы (position weight matrix, PWM, [Stormo, 2000]). Сравнение оценок точности проводилось в соответствии со стандартными процедурами скользящего контроля и “складного ножа” (bootstrap и jackknife, [Efron, Gong, 1983]).

В таблице 1 приведены ошибки второго рода (перепредсказание) для методов ExpertDiscovery и PWM, соответствующие ошибке первого рода (недопредсказание) 50%. Рисунок 2 отражает результаты всей процедуры “складного ножа”, примененной для сравнения точностей распознавания ССТФ HNF4 методами ExpertDiscovery и PWM.

Таблица 1. Анализируемые последовательности ССТФ. Ошибка перепредсказания для системы ExpertDiscovery и метода PWM, полученная для контрольных последовательностей на фиксированном пороге, соответствующим ошибке недопредсказания равной 50%.

ССТФ	Объем выборки	Длина последовательности	Ошибка второго рода (перепредсказание)	
			ExpertDiscovery	PWM
SF1	53	13	5.01E-05	6.87E-05
SREBP	38	18	1.97E-04	8.32E-04
EGR1	22	10	8.09E-04	4.06E-03
CEBP	88	28	1.03E-04	5.12E-04
HNF4	30	13	7.00E-05	2.14E-04



**Рисунок 2.** Зависимость ошибки перепредсказания от ошибки недопредсказания ССТФ HNF4 для методов ‘ExpertDiscovery’ и PWM.

Сравнение показало, что на исследованных примерах система ExpertDiscovery улавливает закономерности нуклеотидного контекста и имеет точность, сравнимую, или превосходящую метод PWM. Значительное улучшение может быть достигнуто в случае адекватного размера обучающих данных, содержащих репрезентативную выборку ССТФ.

Метод PWM, наряду с другими методами распознавания ССТФ, основанными на выявлении консенсуса [Schneider, Stephens, 1990; Ulyanov, Stormo, 1995], использует упрощающее априорное предположение о независимом вкладе каждой позиции в формирование



комплекса ДНК/белок. Ряд работ [Benos и др., 2002; Man, Stormo, 2001; Barash и др., 2003; Udalova и др., 2002] указывает на то, что это предположение не верно, и противоречит биологическим принципам формирования комплекса и предпочтениям факторов определенных сочетаний нуклеотидов, которые совокупно вносят вклад в энергию связывания. В пределах ССТФ выделяются консервативные участки (коровые районы), разделенные вариабельными (спейсерами). Число коровых районов может быть от одного до нескольких. Наличие коровых районов связано с тем, что транскрипционные факторы имеют модульную структуру, и могут содержать несколько доменов или входящих в них субъединиц, выполняющих специфичные функции. В отличие от метода PWM система ExpertDiscovery обнаруживает зависимости между нуклеотидами, достаточно удаленными друг от друга в общем случае. Комплексные сигналы (КС) покрывают биологически осмысленные подгруппы последовательностей.

Для того, чтобы продемонстрировать на примере данное утверждение, рассмотрим некоторые КС, обнаруженные программой ExpertDiscovery при анализе нуклеотидных последовательностей сайтов связывания SF1. Матрица абсолютных нуклеотидных частот, приведенная в таблице 2, характеризует анализируемую выборку.

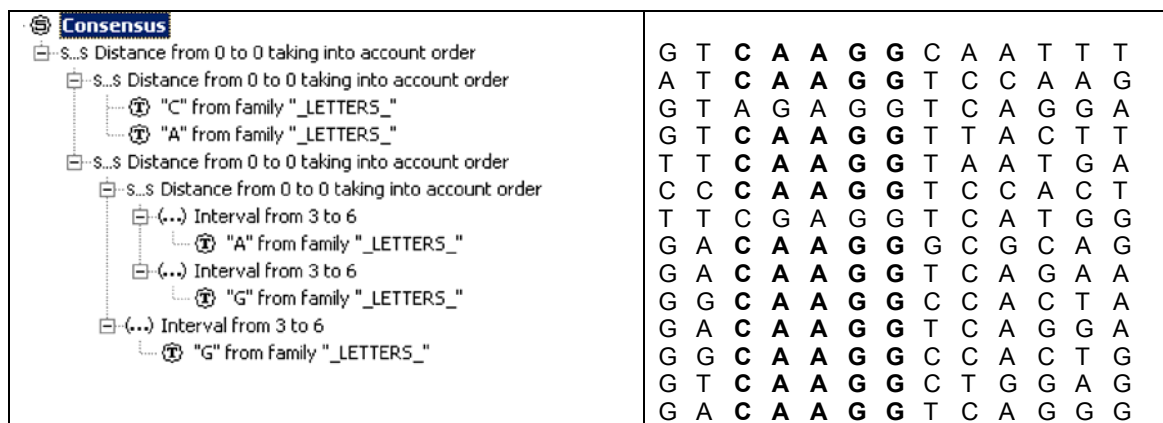
Инвариантные нуклеотидные позиции определяют консенсус сайта (табл. 2, заглавные буквы), или, так называемый, коровый район. Анализ КС, автоматически сгенерированных программой ExpertDiscovery, позволяет извлечь необходимое знание о биологической модели связывания в частном случае, по сравнению с информацией, полученной от PWM.

Таблица 2. Матрица абсолютных нуклеотидных частот сайта связывания SF1. Последняя строка содержит нуклеотиды, частота встречаемости которых в данной позиции максимальна. Предположительная последовательность кора выделена заглавными буквами.

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	7	8	3	<b>47</b>	<b>51</b>	0	0	3	3	36	9	15	17
T	10	25	1	0	0	0	0	34	10	2	10	13	9
G	27	5	6	6	1	<b>53</b>	<b>53</b>	1	2	7	18	14	21
C	9	15	<b>43</b>	0	1	0	0	15	38	8	16	11	6
	g	t	C	A	A	G	G	t	c	a	g	a	g

Системой ExpertDiscovery обнаружен КС “Consensus”, соответствующий коровому району ССТФ SF1 (Рис.3). Иерархическое дерево сигнала “Consensus” определяет пять соседних нуклеотидов, преобладающих в таблице данных. Эта закономерность выполнена для 39 объектов (74%) позитивного обучения и 45 объектов (0,3%) негативного обучения (Рис. 4, параметры сигнала).

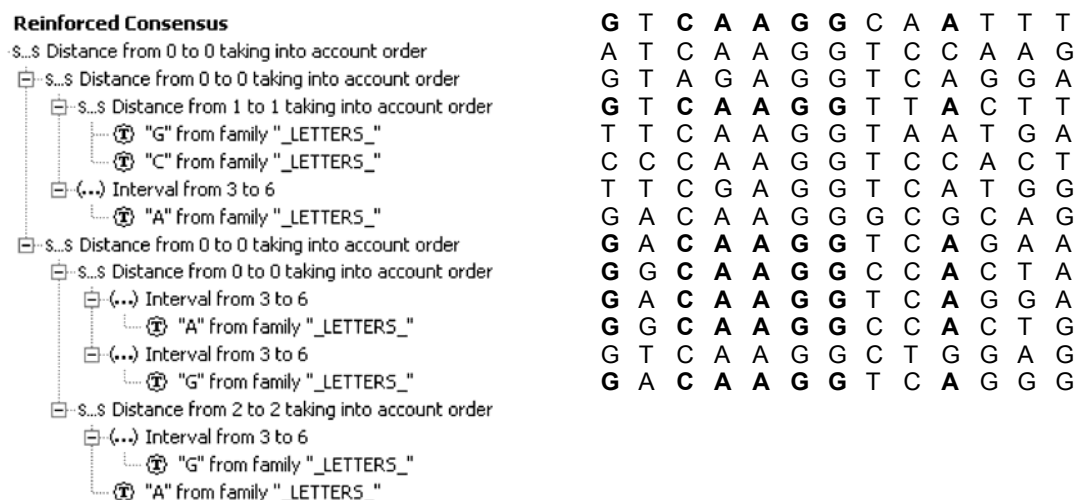
В соответствие с методологией Discovery система усложняет (уточняет) текущий КС, если условная вероятность уточненного КС увеличилась, а уровень значимости по критерию Фишера уменьшился. На рисунке 5 представлен КС, полученный как уточнение КС “Consensus”. Полученный КС определяет две нуклеотидные позиции, фланкирующие консенсус, вероятно, информативных для ДНК/белок связывающего механизма. Этот сигнал “Reinforced Consensus” является одним из самых значимых КС их полученной совокупности, его условная вероятность составила 93%. Закономерность выполнена для 14 ССТФ SF1, и только для 1 объекта негативного обучения (из 16 000, Рис. 6).



**Рисунок 3.** Иерархическое дерево комплексного сигнала, соответствующего коровому району ССТФ SF1s. Справа жирным выделены нуклеотиды, участвующие в закономерности. Запись ‘Distance from 0 to 0 taking into account order’ означает два соседних комплексных сигнала.

General information	
Probability	46.428571% ( 39 / 84 )
Pos. coverage	73.584906% ( 39 / 53 )
Neg. coverage	0.281250% ( 45 / 16000 )
Fisher	0.000000

**Рисунок 4.** Параметры комплексного сигнала, приведенного на рисунке 3. “Probability” – условная вероятность, “Pos./Neg. coverage” – уровень покрытия, число последовательностей, удовлетворяющих закономерности, “Fisher” – уровень значимости сигнала по критерию Фишера.



**Рисунок 5.** Иерархическое дерево наиболее значимого КС, соответствующего уточненному сигналу “Consensus” (Рис. 3). Справа жирным выделены нуклеотиды, участвующие в КС. Запись ‘Distance from 0 to 0 taking into account order’ означает два соседних комплексных сигнала.

General information	
Probability	93.333333% ( 14 / 15 )
Pos. coverage	26.415094% ( 14 / 53 )
Neg. coverage	0.006250% ( 1 / 16000 )
Fisher	0.000000

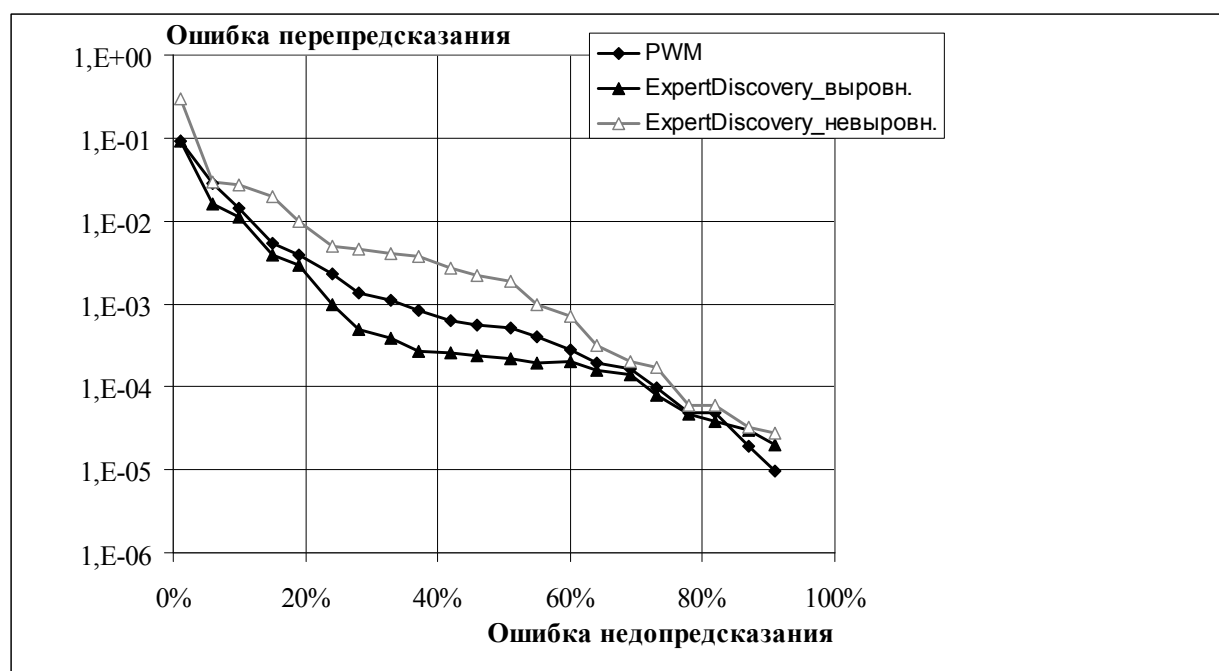
**Рисунок 6.** Параметры комплексного сигнала, приведенного на рисунке 5. “Probability” – условная вероятность, “Pos./Neg. coverage” – уровень покрытия, число последовательностей, удовлетворяющих закономерности, “Fisher” – уровень значимости сигнала по критерию Фишера.

## Эксперимент 2. Невыровненные выборки сайтов связывания транскрипционных факторов.

В случае, когда информация о выравнивании последовательностей не подается априори на вход системе ExpertDiscovery, система обнаруживает КС (1) статистически значимые, (2) иерархически усложняющиеся, (3) содержащие гэпы, без привязки к конкретным нуклеотидным позициям [Khomicheva *et al.*, 2008].

Нами проведен эксперимент по сравнению точностей распознавания ССТФ системой ExpertDiscovery и PWM. В качестве модельного объекта использовалась выровненная выборка ССТФ СЕВР, и выборка, которая не подвергалась предварительной процедуре выравнивания. Длина последовательностей в выборке составляла 50 нуклеотидов, объем выборки – 96 сайтов.

На рисунке 7 приведены результаты процедуры “складного ножа”, примененной для сравнения точностей распознавания ССТФ СЕВР двумя методами.



**Рисунок 7.** Зависимость ошибки перепредсказания от ошибки недопредсказания ССТФ СЕВР для методов ExpertDiscovery и PWM.

Сравнение показало, что метод ExpertDiscovery не уступает PWM во всей области принятия решения в случае выровненных последовательностей ССТФ, и имеет несколько худшую точность по сравнению с PWM в случае невыровненных последовательностей ССТФ.

Рассмотрим пример КС, обнаруженного программой ExpertDiscovery при анализе ССТФ СЕВР в случае невыровненных последовательностей сайтов. Консенсусом сайтов СЕВР представленной выборки является последовательность T(T/G)(A/G)NG(A/C)AA (Рис. 8)

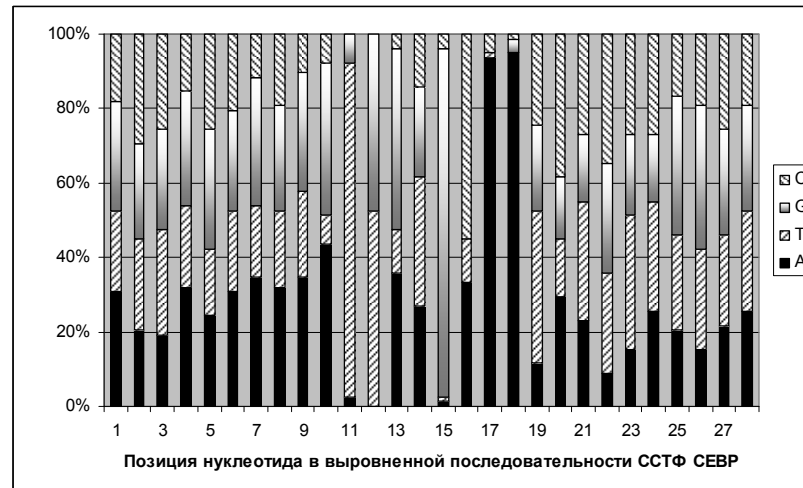


Рисунок 8. Диаграмма относительных частоты встречаемости нуклеотидов в анализируемой выборке ССТФ СЕБР.

КС, покрывающий консенсус сайтов СЕБР, обнаруживается, в общем случае, с некоторым смещением, что соответствует реальному расположению сайтов в ДНК последовательно-сти (Рис. 9).



Рисунок 9. Комплексный сигнал, обнаруженный системой ExpertDiscovery при анализе ССТФ СЕБР в случае невыровненных последовательностей. Слева представлена визуализация расположения сигнала в данных, справа представлено иерархическое дерево сигнала и параметры.

Следует отметить, что сайт связывания СЕБР имеет очень вырожденную структуру, и плохо поддается процедуре выравнивания, что сильно сказывается на качестве распознавания этого типа сайтов традиционными методами, например, с помощью PWM. Как показывает описанный выше пример (а также другие данные, не приведенные в статье), для сайтов с подобной вырожденной структурой метод ExpertDiscovery работает лучше традиционных методов распознавания. Он дает более качественное распознавание и к тому же не требует выравнивания обучающих данных.

## 5. Построение иерархических моделей регуляторных районов коэкспрессирующихся генов на основе данных о расположении потенциальных сайтов связывания транскрипционных факторов.

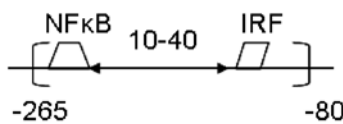
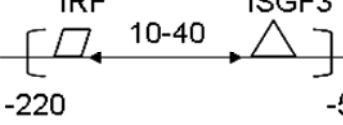
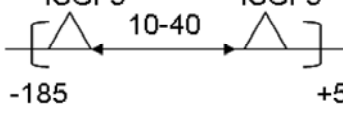

Данный анализ проводился в случае, когда элементарными сигналами для построения комплексных являлись потенциальные ССТФ, обнаруженные методом PWM [Khomicheva *et al.*, 2007a,б].

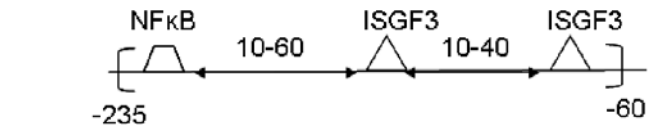

В качестве группы коэкспрессирующихся генов были выбраны гены, экспрессия которых индуцируется в ответ на интерфероны (стимуляторы иммунной системы). Выборка регуляторных районов интерферон-индуцируемых генов эукариот была составлена на основе информации, взятой из базы данных IIG-TRRD [Ананько и др., 1997]. Промоторные последовательности 74 генов были экстрагированы из контигов хромосом человека и фазированы [-500;+200] относительно старта транскрипции. Негативные обучающие данные были составлены из промоторных районов 2140 генов первой хромосомы человека и также фазированы относительно старта транскрипции.

Закономерности, обнаруженные системой, могут быть проинтерпретированы как иерархически вложенные модели сигналов, начиная с простой модели, состоящей из двух ССТФ, удаленных друг от друга на варьирующее расстояние. Если расстояние между ССТФ варьирует в пределах от 10 до 40 нуклеотидов, то такая пара ССТФ образует композиционный элемент, или, в общем случае, статистически значимую пару ССТФ (табл. 3, первые три строки). ExpertDiscovery отбирает наиболее значимые, удовлетворяющие критериям ветвления, простые модели ССТФ и усложняет их в случае, если новые модели обладают большим значением условной вероятности и меньшим уровнем значимости по критерию Фишера (табл. 3, последние три строки). Наше исследование показало, что комбинации не только из двух, трех, но и более упорядоченных ССТФ, являются статистически значимыми, вероятно, соответствуя транскрипционному регуляторному модулю.

Система обнаружила порядка 200 закономерностей. Число закономерностей зависит от выходных параметров, заданных пользователем: условная вероятность (больше 50%) и уровень значимости по критерию Фишера (меньше 0.05).

**Таблица 3.** Иерархически усложняющиеся комплексные сигналы, обнаруженные программой ExpertDiscovery, и соответствующие регуляторным модулям интерферон-индуцируемых генов.

Усложняющиеся модули ССТФ	Биологическая функция
	Синергизм действия факторов, усиление индукции [Leblanc <i>et al.</i> , 1990]
	Прологация индукции [Lew D.J. <i>et al.</i> , 1991]
	Стабилизация комплекса ДНК/белок, усиление стимулирующего эффекта [Li <i>et al.</i> , 1998]
	Кумулятивный эффект [Mirkovitch <i>et al.</i> , 1992]

	Кумулятивный эффект
	Кумулятивный эффект

Приведенные в таблице 3 закономерности имеют биологическую значимость и известны в научной литературе. Так, пара факторов NFκB и IRF влияют на транскрипцию гена значительно сильнее, если их сайты связывания расположены на расстоянии от 10 до 40 пар оснований и образуют так называемый "композиционный элемент" (табл. 3, первая строка). Этот эффект известен как синергизм действия факторов. Если в промоторной области гена присутствуют сайты связывания двух транскрипционных факторов, IRF и ISGF3, то это способствует более длительной повышенной экспрессии гена в ответ на интерферон, поскольку ISGF3 действует только в первые полчаса индукции, а IRF начинает работать после, и его действие длится до 12 часов (продолжения индукции, табл. 3, вторая строка). Если фактор ISGF3 связывается с двумя сайтами, расположенными на небольшом расстоянии друг от друга, это приводит к стабилизации комплекса ДНК и транскрипционных факторов и усиливает стимуляцию транскрипции (табл. 3, третья строка). Одновременное присутствие в промоторном районе сайтов связывания нескольких транскрипционных факторов, работающих в иммунных клетках (IRF, STAT1, ISGF3, NFκB, CEBP, OCT) и активирующихся различными путями передачи сигналов, приводит к кумулятивному эффекту этих факторов на транскрипцию (табл. 3, строки 4-6).

## 6. Заключение

В работе был адаптирован реляционный подход Discovery к задаче анализа регуляторных районов генов эукариот. Прежде всего, был формализован вид гипотез эксперта, комплексный сигнал, создана библиотека элементарных сигналов, разработан универсальный формат разметки последовательностей элементарными сигналами; определены операции над комплексными сигналами. Также был разработан алгоритм автоматического построения, уточнения и проверки гипотез на основании заданных экспертом критериев.

На этапе практического внедрения системы ExpertDiscovery показана применимость системы для анализа контекстной структуры регуляторных районов генов на разных уровнях структурно-функциональной иерархии.

Во-первых, система обнаруживает закономерности контекстной организации последовательностей ССТФ. В отличие от метода весовых матриц система ExpertDiscovery обладает необходимой гибкостью для обнаружения зависимостей между нуклеотидами, достаточно удаленными друг от друга, в общем случае. Комплексные сигналы покрывают биологически осмысленные подгруппы последовательностей. Сравнение оценок точности распознавания системы ExpertDiscovery и PWM показало, что система ExpertDiscovery имеет точность, сходную, или превосходящую метод PWM.

Во-вторых, используя анализ расположения ССТФ, система ExpertDiscovery осуществляет иерархический анализ регуляторных районов, обнаруживая биологически-целесообразные иерархически-усложняющиеся модели районов от простых композиционных моделей, со-

стоящих из двух ССТФ, до сложных моделей сигналов, соответствующих комплексной регуляции транскрипции. Данные модели обладают прогностической силой и позволяют обнаруживать потенциальные регуляторные районы анализируемой группы генов. Учет построенных моделей может обеспечить планирование наиболее экономичного эксперимента.

Общий характер системы выгодно отличает её от других программ распознавания. Так, закономерности, обнаруживаемые программой PromoterExplorer, упомянутой во введении, представляют собой частный случай комплексных сигналов, т.к. данная программа использует линейные комбинации различных характеристик промоторных районов ДНК, и не учитывает закономерные взаимосвязи между ними.

### **Благодарности**

Авторы статьи выражают благодарность Левицкому Виктору Георгиевичу за любезно предоставленную библиотеку весовых матриц, и Кондрахину Юрию Васильевичу за составление контрольных выборок.

Работа поддержана грантом РФФИ 08-07-00272-а; интеграционными проектами СО РАН № 47, 115, 119, Госконтрактом с ФАО № П721, а также работа выполнена при финансовой поддержке Совета по грантам Президента РФ и государственной поддержке ведущих научных школ (проект НШ-335.2008.1, НШ-2447.2008.4)

### **Библиографический список**

1. Ананько Е.А., Бажан С.И., Белова О.Е., Кель А.Э. Механизмы регуляции транскрипции интерферон-индуцируемых генов: Описание в информационной системе IIG-TRRD Молекулярная биология (1997) V.31. P. 701-713
2. Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. // НГУ, Новосибирск, 2006. 293 с.
3. Витяев Е.Е., Орлов Ю.Л., Хомичёва И.В., Шипилов Т.И. Методы извлечения знаний и логического анализа регуляторных геномных последовательностей // Системная компьютерная биология / отв. ред. Н.А. Колчанов, С.С. Гончаров, В.А. Лихошвай, В.А. Иванисенко. Рос. Акад. Наук, Сиб. отд-ние. Новосибирск: Изд-во СО РАН, 2008, стр. 126-136.
4. Кендал М., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973. С. 899.
5. Arnone M.I., Davidson E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10), 1851-1864.
6. Barash Y., Elidan G., Friedman F., Kaplan T. (2003) Modeling dependencies in protein-DNA binding sites, *RECOMB*, 28–37.
7. Benos P.V., Bulyk M.L., Stormo G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* 30, 4442-4451.
8. Blanchette M., Bataille A.R., Chen X., Poitras C., Laganier J., Lefebvre C., Deloix G., Giguere V., Ferretti V., Bergeron D., Coulombe B., Robert F. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* 2006 May;16(5):656-68.
9. Dynan W.S. Modularity in promoters and enhancers. *Cell*, 1989, 58(1), 1-4.

10. Efron B., Gong G. (1983) A leisurely look at the bootstrap the jackknife and resampling. *American Statistician*. 37, 36-48.
11. Kel-Margoulis O.V., Kel A.E., Reuter I., Deineko I.V., Wingender E. (2002) Transcompel: a database on composite regulatory elements in eukaryotic genes, *Nucl. Acids Res.* 30, 332-334.
12. Khomicheva I.V., Vityaev E.E., Ananko E.A., Levitsky V.G., Shipilov T.I. (2007a) Hierarchical analysis of the eukaryotic transcription regulatory regions based on the DNA codes of transcription. Proceedings of the 3-rd Moscow conference on computational molecular biology. Moscow, Russia, July 27-31, p. 142-144
13. Khomicheva I.V., E.E. Vityaev, E.A. Ananko, V.G. Levitsky, T.I. Shipilov (2007b) Hierarchical analysis of the eukaryotic transcription regulatory regions based on the DNA codes of transcription. Proceedings of the 3-rd Moscow conference on computational molecular biology. Moscow, Russia, July 27-31, , p. 142-144
14. Khomicheva I., Demin A., Vityaev E. (2007b) Transcription Factor Binding Site Discovery by the Probabilistic Rules. PKDD Proceedings: Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, Andrzej Skowron (Eds.), Knowledge Discovery in Databases: PKDD 2007. 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Warsaw, Poland, September 17-21, 2007. Proceedings. Lecture Notes in Artificial Intelligence 4702, Springer 2007, ISBN: 978-3-540-74975-2. 104-109.
15. Khomicheva I.V., Vityaev E.E., Shipilov T.I. Discovery of the transcription factor binding sites in the aligned and unaligned DNA sequences. Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008, 22-28 June, Novosibirsk, Russia), ICG, Novosibirsk, 2008, p. 116.
16. 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 // Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS2006, 16-22 July, Novosibirsk, Russia), ICG, Novosibirsk, 2006, v.1, pp.77-80
17. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002) Transcription Regulatory Regions Database, (TRRD): its status in 2002. *Nucleic Acid Res.* 30, 312-317.
18. 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, p.308.
19. Kovalerchuk B., Vityaev E., Symbolic Methodology for Numeric Data Mining. *Intelligent Data Analysis*. Special issue on "Philosophies and Methodologies for Knowledge Discovery and Intelligent Data Analysis" eds. Keith Rennolls, Evgenii Vityaev. v.12(2), IOS Press, 2008, pp. 165-188
20. Leblanc J.F., Cohen L., Rodrigues M., Hiscott J. (1990) Synergism between distinct enhancer domains in viral induction of TI the human beta interferon gene. *Mol.Cell.Biol.* 10(8), 3987-3993.
21. Levitsky V.G., Katokhin A.V., Podkolodnaya O.A., Furman D.P., Kolchanov N.A. (2005) NPRD: Nucleosome Positioning Region Database, *Nucl. Acids. Res.*, 33, 67-70.



22. Lew D.J., Decker T., Strehlow I., Darnell J.E. (1991) Overlapping elements in the guanine-binding protein gene promoter TI mediate transcriptional induction by alpha and gamma interferons. *Mol. Cell. Biol.* 11(1), 182-191.
23. Li X., Leung S., Burns C., Stark G.R. (1998) Cooperative binding of Stat1-2 heterodimers and ISGF3 to tandem DNA elements, *Biochimie* 80, 703-710.
24. Man T.K., Stormo G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay, *Nucleic Acids Res.* 29, 2471-2478.
25. Mirkovitch J., Decker T., Darnell J.E., Jr (1992) Interferon induction of gene transcription analyzed by in vivo footprinting. *Mol. Cell. Biol.* 12(1), 1-9.
26. Nikolov D.B., Burley, S.K. (1997) RNA polymerase II transcription initiation: A structural view. *Proc. Natl. Acad. Sci. USA*, 94, 15-22.
27. Orlov Y.L. and Potapov V.N. (2004) Complexity: an internet resource for analysis of DNA sequence complexity, *Nucleic Acids Res.* 32 (Web Server issue), 628-633.
28. Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucleic Acids Res.* 32(Web Server issue), 208-212.
29. Schneider, T. and Stephens, R. (1990) Sequence logos: A new way to display consensus sequences. *Nucleic Acid Research*, v. 18, 20, p. 6097-6100.
30. Scientific Discovery Web Site, <http://www.math.nsc.ru/AP/ScientificDiscovery>
31. Stormo G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*. 16, 16-23.
32. Trifonov E.N. (1997) Genetic level of DNA sequences is determined by superposition of many codes. *Mol. Biol. (Mosk)* 31, 759-767.
33. Udalova I.A., Mott R., Field D., Kwiatkowski D. (2002) Quantitative prediction of NF-kB DNA-protein interactions. *Proc. Natl. Acad. Sci. USA* 99, 8167-8172.
34. Ulyanov A., Stormo G. (1995) Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions. *Nucl. Acids Res.* 23, 1434-1440.
35. Vishnevsky O.V. and Kolchanov N.A. (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoter, *Nucleic Acids Res.*, 33, 417-422
36. Vityaev E. The logic of prediction. In: *Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia)*, edited by S.S. Goncharov, R. Downey, H. Ono, World Scientific, Singapore, 2006, pp.263-276
37. Vityaev E., Kovalerchuk B., *Empirical Theories Discovery based on the Measurement Theory. Mind and Machine*, v.14, #4, 551-573, 2004
38. Vityaev, B.Y. Kovalerchuk, *Relational Methodology for Data Mining and Knowledge Discovery. Intelligent Data Analysis*. Special issue on "Philosophies and Methodologies for Knowledge Discovery and Intelligent Data Analysis" eds. Keith Rennolls, Evgenii Vityaev. v.12(2), IOS Press, 2008, pp. 189-210
39. Xie X., Wu S., Lam K.-M., Yan H., PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm, *Bioinformatics* 22 (2006), 2722-2728

## **Program system ExpertDiscovery for DNA regulatory regions analysis.**

### **Annotation**

The appearance of advanced experimental technologies in such fields of modern biology as genomics, transcriptomics, proteomics, cell biology, nanobioengineering, etc. resulted in exponential growth of experimental data, that need to be analyzed and mined. The new methods of intelligent data analysis are challenged to solve the task of integration of primary raw experimental data, that are poorly consistent and structured, contain gaps, and separately can't reconstruct completely the biologic system or process. We developed the integrated data mining method ExpertDiscovery, discovering the complex regularities of eukaryotic DNA regulatory regions organization. As the elementary signals to build the complex signals the system takes the different DNA characteristics, obtained, for instance, by another data mining tools. Using the regularities, discovered on the levels of research, the system allows to construct the hierarchical model of regulatory regions of specific group of genes.

### **Ключевые слова**

Complex signal, relational data mining, integrated system, hierarchical analysis, regulatory regions of genes, recognition, accuracy comparison