МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

---

Кафедра систем информатики
(название кафедры)

Юрий Юрьевич Васькин
(И., О., фамилия студента – автора работы)

# РЕАЛИЗАЦИЯ СРЕДСТВ ИЕРАРХИЧЕСКОГО АНАЛИЗА РЕГУЛЯТОРНЫХ РАЙОНОВ ГЕНОВ ДЛЯ ИНТЕГРИРОВАННОЙ СИСТЕМЫ EXPERT DISCOVERY И UGENE

(полное название темы магистерской диссертации)

**МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ**
по направлению высшего профессионального образования

230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

ФАКУЛЬТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Тема диссертации утверждена распоряжением по НГУ № 9 от «11» января 2012 г.
Тема диссертации скорректирована распоряжением по НГУ № 183 от «14» мая 2013 г.

Руководитель

Витяев Е.Е.
(фамилия, и., о.)
д.ф.-м.н., в.н.с.
ИМ СО РАН
(уч .степень., должность)

Новосибирск, 2013г.

МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

---

Кафедра систем информатики
(название кафедры)

УТВЕРЖДАЮ
Зав. кафедрой Лаврентьев Михаил Михайлович
(фамилия, И., О.)
…….
(подпись, дата)

**ЗАДАНИЕ**

**на магистерскую диссертацию**

Студент Васькин Юрий Юрьевич
(фамилия, имя, отчество)

факультет ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Направление подготовки 230100.68 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

Магистерская программа: Технология разработки программных систем

Тема: Реализация средств иерархического анализа регуляторных районов генов для интегрированной системы Expert Discovery и UGENE.

Цели работы: улучшение базового функционала интегрированной системы Expert Discovery и UGENE, реализация средств распознавания промоторных областей и демонстрация работоспособности системы на биологических данных.

Структурные части работы: обзор литературы по теме регуляции экспрессии генов эукариот, обзор существующих методов и инструментов анализа регуляторных областей, реализация средств полного анализа регуляторных областей, проведение анализа разработанными средствами.

Руководитель
Витяев Е.Е.
(фамилия, и., о.)
д.ф.-м.н., в.н.с.
ИМ СО РАН
(уч .степень., должность)
…….
(подпись, дата)

## Содержание

## ВВЕДЕНИЕ

Дипломная работа выполнялась на кафедре систем информатики факультета информационных технологий Новосибирского государственного университета, в институте математики СО РАН, а также в компании “Унипро” в рамках проекта по созданию программного комплекса для работы молекулярных биологов - UGENE.

Анализ регуляторных последовательностей генов и поиск структурно-функциональных закономерностей представляет актуальную проблему биологии, которая в настоящее время далека от окончательного решения. Во многом это обусловлено сложностью строения регуляторных областей и многообразием механизмов регуляции экспрессии генов. Существует необходимость анализировать разнородные данные по физико-химическим, структурным, информационным свойствам регуляторных последовательностей генов, а также по экспериментальным данным об их функционировании.

Возможность гибкой регуляции экспрессии генов эукариот обусловлена наличием довольно обширных регуляторных областей, имеющих блочно-иерархическую структуру [8].

**Целью данной работы является разработка программной системы, на основе Data mining метода Discovery [28], для проведения качественного анализа регуляторных областей генов эукариот на уровне контекстных характеристик нуклеотидных последовательностей.**

Разработанный ранее метод [1, 4, 5, 6, 38, 39] интеллектуального анализа регуляторных областей основан на интеграции взаимодополняющих инструментов: система ExpertDiscovery – мощный инструмент для иерархического анализа регуляторных районов генов и мультиплатформенный биоинформационный пакет UGENE [7], объединяющий в себе большое число алгоритмов для работы с генетической информацией [1].

Система ExpertDiscovery интегрирована в программный пакет UGENE в виде модуля, как основной результат бакалаврской дипломной работы. В рамках магистерской работы были проведены испытания интегрированной системы, благодаря которым был обнаружен недостающий функционал программы. Обновление системы даст пользователю возможность в полной мере применять метод иерархического анализа регуляторных областей. Таким образом, в работе можно выделить две основные задачи:

1. Реализация недостающего функционала системы

2. Применение обновленной системы для демонстрации результатов

Актуальность работы заключается в предоставлении широкому кругу пользователей кросс-платформенной реализации уникальной системы по поиску и распознаванию комплексных сигналов, а также возможности комбинирования различных алгоритмов с целью получения биологически значимых результатов.

Первая глава работы посвящена обзору предметной области, во второй главе описывается метод и требования к новой системе, в третьей главе представлены программные и биологические результаты.

# ГЛАВА 1

# Обзор предметной области

## 1.1. Регуляция генов

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. На начальных этапах транскрипции РНК-полимераза связывается с геном в его промотороном регионе. У эукариот РНК-полимераза – это лишь часть сложного белкового комплекса, который включает в себя другие белки, которые связываются с особыми участками регуляторных районов гена. Высокая интенсивность транскрипции обеспечивается как наличием РНК-полимеразы, так и связанных с ней регулирующих белков. Примерная позиция различных регуляторных элементов (GC box, CAAT-box, TATA-box, энхансеров), относительная старта транскрипции, показана на рисунке 1. Стоит отметить, что расстояние между любыми регуляторными элементами может варьироваться. Регуляторные элементы, связанные с молекулой ДНК и РНК-полимеразой активируют процесс транскрипции гена, в результате которого генерируется информационная РНК. Далее из полученной молекулы вырезаются интроны, после чего инициируется процесс трансляции. В результате трансляции производится белок, который, приняв нужную форму, будет способен выполнять свою функцию [40].

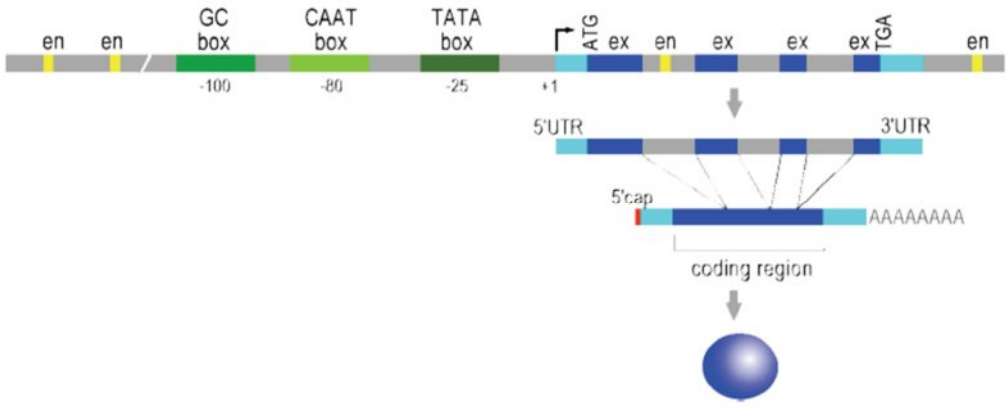


**Рисунок 1.** Структура гена эукариот. Цис-регуляторные элементы (зеленым), энхансеры (желтым), экзоны (синим).

Интенсивность 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) [9]. Регуляция транскрипции генов осуществляется при участии большого количества регуляторных белков: транскрипционных факторов (ТФ), коактиваторов, корепрессоров, медиаторов [10, 38]. 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.

Регуляторные области генов эукариот имеют блочно-иерархическую структуру [8]. Первому уровню иерархии регуляторных областей генов соответствуют ССТФ - короткие последовательности ДНК (10-20 нуклеотидов), являющиеся местом посадки транскрипционных факторов [11].

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 [10]

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 [10]. 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 пар оснований). Энхансеры и сайленсеры могут находиться как на 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 1000 нуклеотидов [12].

Самый высший уровень иерархии строения регуляторных областей генов соответствует системе интегральной регуляции транскрипции [8], 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) [13].

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 1500 транскрипционных факторов [14]. 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 [8]. Разнообразие строения регуляторных районов выражается еще и в том, что, как отмечалось выше, и протяженность регуляторных единиц, и их локализация варьируется 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 [15], метод SITECON [16], SiteGA [17] и ряд других. Все известные подходы имеют определенные недостатки, поскольку не учитывают взаиморасположение сайтов и характеризуются определенными уровнями перепредсказания (или недопредсказания) [8].

Задача, соответствующая одному из уровней иерархии строения регуляторных районов генов, состоит в обнаружении закономерностей расположения ССТФ [7]. Однако, поскольку регуляторные районы генов содержат уникальную 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 вариабельностью и протяженностью регуляторных районов, а также возможностью присутствия большого количества элементарных сигналов:

- ССТФ. Транскрипционные факторы, связанные с соответствующими сайтами, могут образовывать больше группы. Дезактивация хотя бы одного элемента такой группы может приводить к нарушениям экспрессии [10]
- Конформационные. В зависимости от клеточных условий и нуклеотидного состава участков ДНК, молекула может принимать различные формы, что может оказать влияние на экспрессию [41].
- Физико-химические свойства: повышенная гибкость ДНК, легкоплавкость и т.д. [42]

С информационной точки зрения, задача анализа регуляторных районов генов эукариот состоит в иерархическом анализе генетической информации.

### 1.2. Алгоритм работы ExpertDiscovery

Генетическую информацию можно представить в 4-х буквенном алфавите, состоящем из символов A, C, G,T. Также, для записи генетического кода широко применяется 15-ти буквенный код IUPAC [29]: A, C, T, G, M, W, R, Y, S, K, H, V, D, B, N. Перевод такого кода в 4-х буквенный можно осуществить по таблице 1.

| 15-ти буквенный символ | Набор 4-х буквенных символов | Пояснение |
|---|---|---|
| A | A | **A**denosine |
| C | C | **C**ytosine |
| T | T | **T**hymine |
| G | G | **G**uanine |
| M | A или C | A**m**ino groupe |
| W | A или T | **W**eak interaction |
| R | A или G | Pu**r**ine |
| Y | T или C | P**y**rimidine |

| | | |
|---|---|---|
| S | G или C | **S**trong interaction |
| K | G или T | **K**etone |
| H | A или C или T | Not G. **H** comes after G |
| V | G или C или A | Not T. **V** comes after U |
| D | G или A или T | Not C. **D** comes after C |
| B | G или T или C | Not A. **B** comes after A |
| N | A или G или C или T | A**n**y |

**Таблица 1.** Символы 15-ти буквенного кода IUPAC.

Сигнал – система правил, определяющих свойства участков последовательностей ДНК. Элементарный сигнал – неделимый сигнал, который характеризуются 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.

КС может быть представлен схематически в виде дерева:

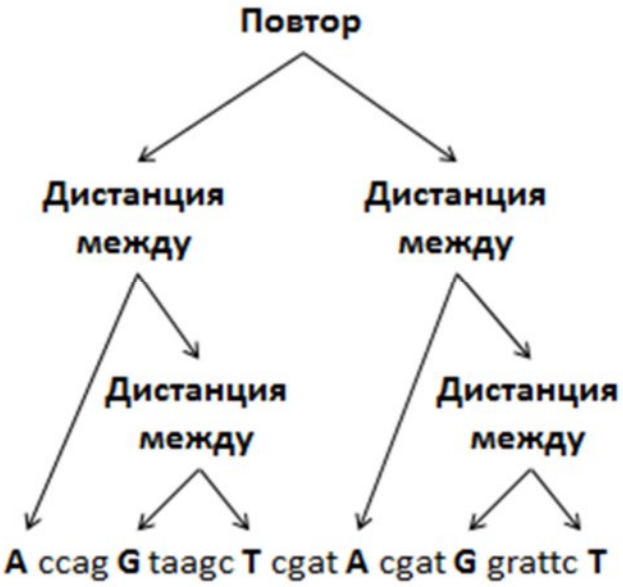


**Рисунок 2.** Схематическое представление КС

В данном случае (Рис. 2), 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.

Над КС определены следующие операции:

Дистанция между сигналами. На вход подаются два КС s1 и s2, 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 s1, и на расстоянии от min до max символов от него найден сигнал s2. В случае, если порядок не имеет значения, сначала может быть найден s2, а потом s1. Параметры min и max задаются экспертом.

Два элементарных сигнала, связанные предикатом «Дистанция» могут соответствовать композиционному элементу – паре сайтов связывания транскрипционных факторов [42].

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

Несколько повторяющихся копий ССТФ – характерная особенность промоторов генов ответа на тепловой шок [42].

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.

С помощью этого предиката можно обнаруживать консервативные в некотором регионе сайты связывания.

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 происходит 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)

a. Если он выполнился, то данный КС записывается в результирующее множество ResCS.

b. Иначе проверяется критерий ветвлений (см. далее). В случае его выполнения сигнал переносится в следующую популяцию.

c. Если ни один из предыдущих критериев не выполнен, то КС отсеивается.

Далее рассматривается следующий КС текущей популяции, когда сигналы в текущей популяции заканчиваются, алгоритм переходит к следующей популяции. Цикл продолжается, пока популяция не опустеет. Результаты работы алгоритма – множество полученных КС ResCS. Стоит 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 [30] 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 [31].
- Порог 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 [3]). 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[1].

### 1.3 Система UGENE

Целью проекта UGENE является качественная интеграция различных алгоритмов анализа генетических данных в единой рабочей среде [7]. Среди таких алгоритмов: поиск шаблонов, локальное выравнивание (Smith-Waterman [32]), HMMER [33], поиск сайтов рестрикции, выравнивание на геном (Bowtie [34], UGENE genome aligner), филогенетический анализ, множественное выравнивание (MUSCLE [35], KAlign [36]) и т.д.

Процесс обработки биологических данных часто является многостадийным. Для организации конвейерной обработки данных реализованы оригинальные конструкторы - для вычислительных схем (Workflow Designer, [7]) и для комплексных запросов (Query Designer). Вычислительная схема в Workflow Designer состоит из вычислительных блоков, которые предоставляются по умолчанию, или тех, которые определяет пользователь. Каждый такой блок реализует некоторый вычислительный алгоритм, который может быть оптимизирован разработчиками. Стандартный набор Workflow Designer входят следующие блоки: чтения данных, записи данных, блоки фильтрации и группировки, алгоритмические блоки.

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 распространённых 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. Имеется возможность взаимодействия с удаленными базами данных (NCBI, Genbank, PDB и др.).

UGENE реализована на Qt4 [2], 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 является достаточно мощным средством иерархического анализа регуляторных областей. Генерируя разметки разными методами, мы позволяем системе осуществлять распознавание на высоких уровнях иерархии, что не могут 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 программный пакет, оформив в виде отдельного модуля, который бы повторял и расширял возможности ExpertDiscovery. Интеграция была выполнена в рамках бакалаврской дипломной работы.

## ГЛАВА 2

### Методы и требования

Система ExpertDiscovery автоматически строит комплексные сигналы по указанным параметрам на основе заданных элементарных сигналов. Подавая различные элементарные сигналы, можно проводить анализ на разных уровнях иерархии строения регуляторных областей. Во встроенной версии системы элементарными сигналами могут быть сигналы, представленные в таблице 1.

| Элементарные сигнал | Описание |
|---|---|
| Нуклеотиды, контекстные сигналы, любые слова в расширенном коде IUPAC [18] | Дополнительно к разметке нуклеотидами, пользователь может загрузить произвольные контекстные сигналы, обнаруженные, например, с помощью программ поиска “по маске”. |
| Потенциальные ССТФ, распознаваемые 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 [16] | Возможно распознавание 44 сайтов связывания генов эукариот, для которых в UGENE имеются модели консервативных, конформационных, либо физико-химических свойств ДНК, выявленных 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. |

**Таблица 1.** Элементарные сигналы ExpertDiscovery.

Был реализован недостающий модуль загрузки мотивов, найденных инструментами поиска мотивов в формате tab-delimited. Также был разработан функционал по

добавлению комплексного сигнала в разметку для анализа на более высоких уровнях иерархии и построения комплексных сигналов, когда элементарными сигналами являются комплексные сигналы.

Основной цикл работы программы ExpertDiscovery выглядит следующим образом (Рисунок 3):

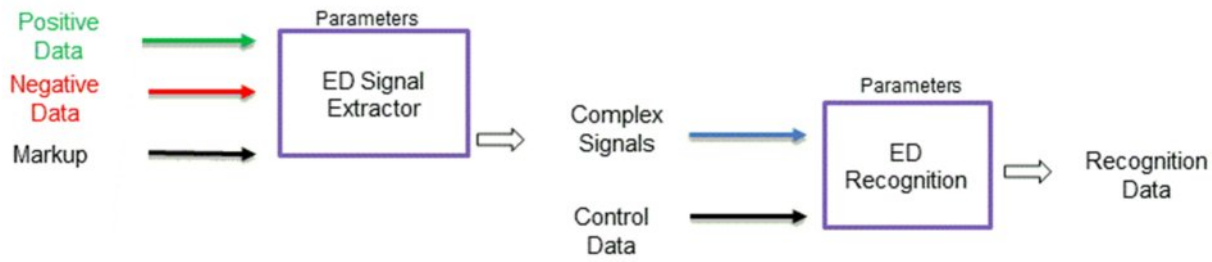


**Рисунок 3.** Основная итерация работы пользователя системы ExpertDiscovery.

В системе ExpertDiscovery логически можно выделить две части первая часть отвечает за построение КС (ED Signal Extractor), а 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.

В ходе испытания описанного выше метода был выявлен ряд мелких и крупных недостатков системы. При этом были найдены, как технические неисправности, так и принципиальные недостатки пользовательского интерфейса. Самым главным недостатком являлось то, что система не позволяла производить автоматический и детальный анализ качества распознавания. Статистик, которыми обладали комплексные сигналы, было

недостаточно для оценки качества текущей модели регуляторной области и метода в целом.. В качестве тренировочных данных использовались последовательности промоторов длины 500 нуклеотидов, специфичные для человеческой печени, а в качестве элементарных сигналов использовались мотивы длины 6 и 8, найденные программой YMF [19]. Испытания проводились с целью получения комплексных сигналов и дальнейшего распознавания промоторов в человеческом геноме. Дальнейшее развитие системы велось по следующим направлениям:

- Загрузка последовательностей только в формате последовательностей. Необходимо сделать возможным загрузку последовательностей в формате множественного выравнивания.
- Для повышения удобства пользователя сделать «Мастер загрузки последовательностей и разметок», содержащий текстовые описания.
- Добавить предикаты для построения комплексных сигналов по умолчанию. Были добавлены предикаты «Дистанция» с перекрывающимися интервалами.
- Загрузка разметок контрольных последовательностей для контроля качества распознавания.
- Добавление выбранного комплексного сигнала в разметку.
- Дополнить отчет распознавания, указать параметры построения комплексных сигналов и файлы с исходными данными.
- Построить график попадания сигналов на каждую последовательность.
- Улучшение окна выбора порога. Отображение зависимостей ошибок первого и второго рода от score.
- Оптимизированное распознавание комплексных сигналов на длинных последовательностях с помощью процедуры скользящего окна.
- Автоматизация процедуры разбиения входных данных на тренировочные и контрольные.
- Экспорт аннотаций комплексных сигналов и последовательностей в файлы.

Помимо этого, были реализованы функции сохранения имен моделей SITECON в UGENE Workflow Designer для автоматической генерации разметок. Была проделана работа по ручному тестированию системы и исправлен ряд критических и мелких программных ошибок.

После того, как указанные недостатки были устранены, программа стала пригодна для получения значимых биологических результатов. Алгоритм распознавания промоторов (или сайтов) таков:

1. Разбиение исходной выборки на тренировочную и контрольную. (Автоматическая генерация негативной выборки).
2. Разметка всех выборок элементарными сигналами. Для этого доступны схемы в Workflow Designer или сторонние инструменты.
3. Загрузка последовательностей и разметок в систему.
4. Построение комплексных сигналов.
5. Анализ комплексных сигналов, подбор порога для распознавания.
6. Генерация отчета или распознавание на длинных последовательностях.

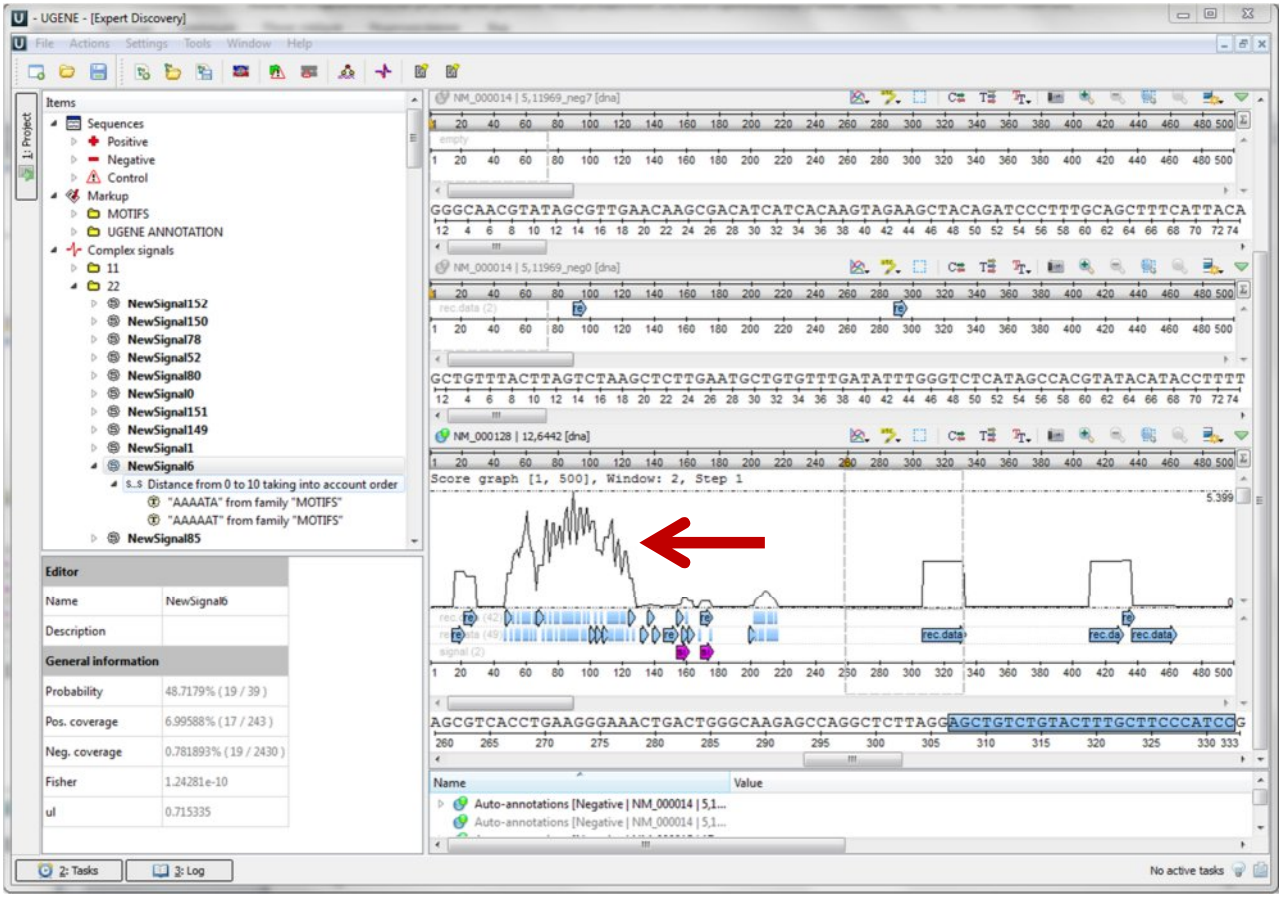


**Рисунок 4.** Окно ExpertDiscovery с выбранным комплексным сигналом

На рисунке 4 продемонстрирован один из комплексных сигналов, который был построен в процессе семантического вероятностного вывода, а также профиль попадания комплексных сигналов на одну из последовательностей. По профилю видно большое

скопление сигналов в интервале [-460, -360] от старта транскрипции, что свидетельствует о наличии ССТФ в этой области.

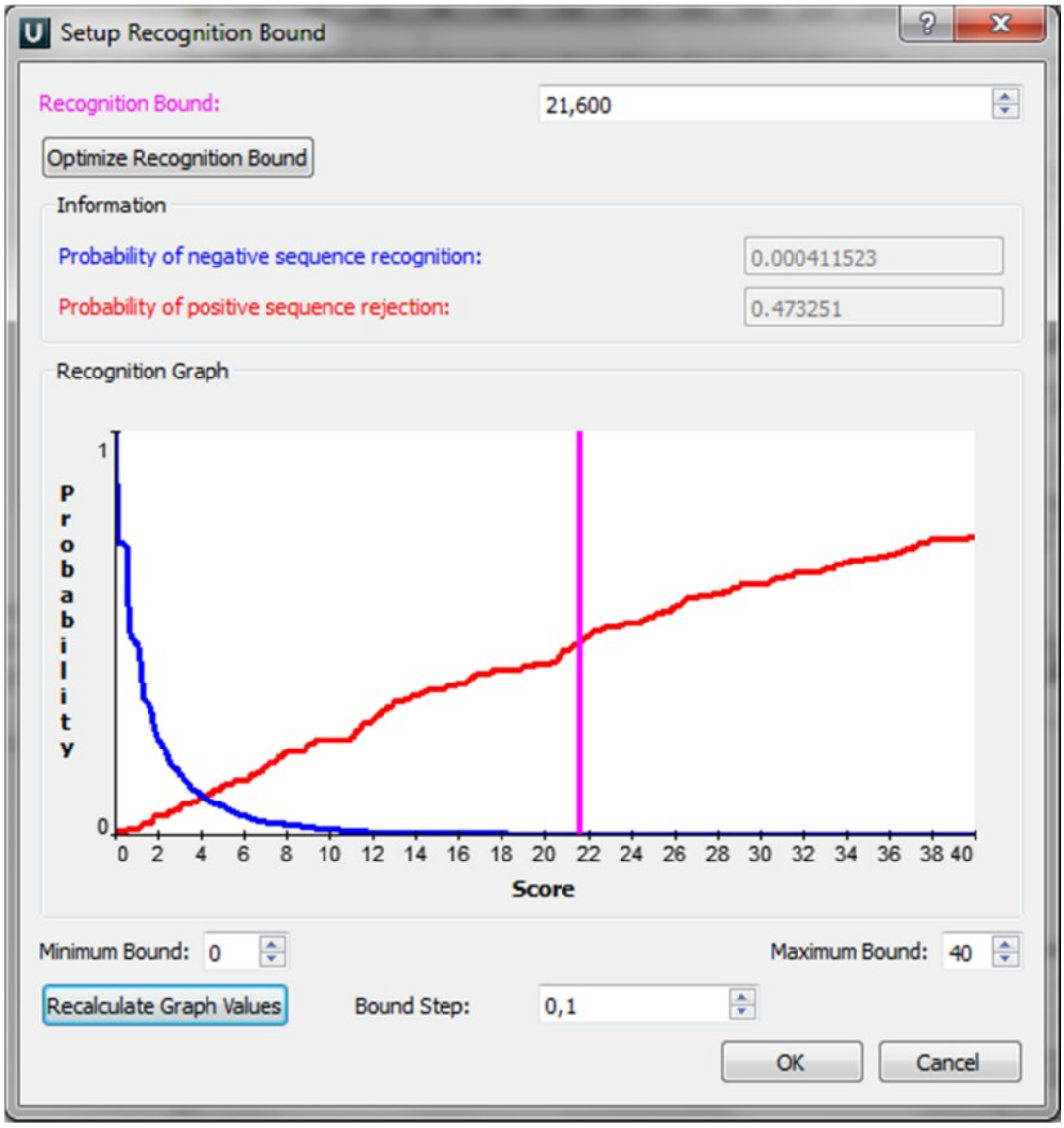


**Рисунок 5.** Окно выбора порога распознавания

На рисунке 5. показано окно выбора порога распознавания. Чем выше порог, тем меньше ошибка второго рода, однако тем больше ошибка первого рода, и наоборот. По графикам зависимостей ошибок от значения порога, исследователь может определить интересующее его значение.

Лучшие практики мирового уровня [16, 17] диктуют форму демонстрации качества результатов распознавания. Используя график в окне выбора порога распознавания можно

определить уровни ошибок, вычисленных на основе обучаюзих данных.Для получения информации о качестве распознавания контрольных данных требуется открыть полный отчет о распознавании. Однако, определенный метод распознавания хорош только в том случае, если он систематически лучше, чем другой. Такую систематику можно продемонстрировать с помощью графиков ROC-кривых и процедуры кросс-валидации. Эту процедуру можно делать вручную, используя несколько итераций алгоритма, как это было продемонстрированов ряде исследований [1, 5, 6]. Но подобные манипуляции требуют значительных временных затрат и, в силу многоступенчатости процесса, повышают вероятность ошибки. Поэтому эта процедура была автоматизирована. Автоматизация потребовала значительной реорганизации кода ExpertDiscovery, так как архитектура программы не была рассчитана на автоматическое итеративное использование системы.

В результате, пользователь может включить процедуру кросс-валидации на этапе построения комплексных сигналов. Для этого нужно загрузить все исходные данные ( последовательности и разметки), указать параметры, а система автоматически построит сигналы и сгенерирует файл, содержащий данные для ROC-кривых. Далее пользователь может построить график по этим данным, чтобы оценить качество распознавания и выбрать порог. Варьируя параметры, можно осуществить несколько способов кросс-валидации: со случайным исключением указанной части выборки, с последовательным исключением k указанных последовательностей.

## ГЛАВА 3

## Применение

Для исследования сложной структуры регуляторных областей генов и для проверки качества работы обновленной системы ExpertDiscovery, использовалась выборка 303 последовательностей человеческих промоторов, специфичных для печени. Выборка генов человека, специфичных для печени, была сформирована с использованием данных о тканеспецифичности, взятых из TiGER[20]. WEB-интерфейс к базе данных TiGER позволяет по указанному типу ткани получить идентификаторы генов, специфичных для этого типа. Последовательности самих промоторов ([-500, -1] от старта транскрипции) были выделены по этим идентификаторам из базы данных UCSC Genome Browser [21]. После выделения последовательностей промоторов, из выборки были удалены дубликаты. Негативная выборка генерировалась автоматически: 40 негативных последовательностей на каждую позитивную, с сохранением частот встречаемости нуклеотидов.

Для дополнительного контроля была использована выборка тканеспецифичных промоторов человеческого мозга: 243 таких последовательности были получены аналогичным методом. Ожидалось, что эти последовательности не должны распознаваться с помощью моделей, полученных для промоторов, специфичных для печени.

Для разметки последовательностей элементарными сигналами был использован алгоритм SITECON, реализованный в UGENE Workflow Designer. Уровень ошибки второго рода был повышен для $10^{-2}$, так как система ExpertDiscovery способна отфильтровать незначимые сигналы за счет используемых статистик. Были использованы модели сайтов SITEON, полученные создателями метода [16], а затем встроенные в UGENE. Среди доступных моделей были выбраны те, которые соответствуют сайтам специфично экспрессирущимся в тканях человеческой печени: CEBP, GATA1, GATA2, GATA3, TATA-box, HNF1, HNF3, HNF4, COUP [43].

Также в качестве стороннего инструмента разметки был использован Weeder motif finder[22], как одно из лучших средств поиска мотивов[23]. Были использованы мотивы длины 6 и 8 нуклеотидов.

Для сравнения результатов распознавания промоторов разными методами использовалась процедура кросс-валидации, 10 итераций со случайным исключением 20% исходной выборки в качестве контрольных данных.

Системы SITECON и weeder motif finder не имеют функции распознавания промоторов. ExpertDiscovery расширяет их функционал, как и функционал любой другой системы, генерирующей элементарные сигналы. Для распознавания промоторов методами SITECON и WEEDER использовались комплексные сигналы одинарного уровня иерархии (не включающие предикатов). Эти сигналы автоматически фильтровались с помощью статистик ExpertDiscovery: покрытие позитивной выборки, порог условной вероятности, критерии Фишера и Юла. Если бы эта фильтрация не проводилась, то качество моделей комплексных сигналов, построенных на основании исходных разметок, было бы хуже, соответственно и качество распознавания было бы хуже.

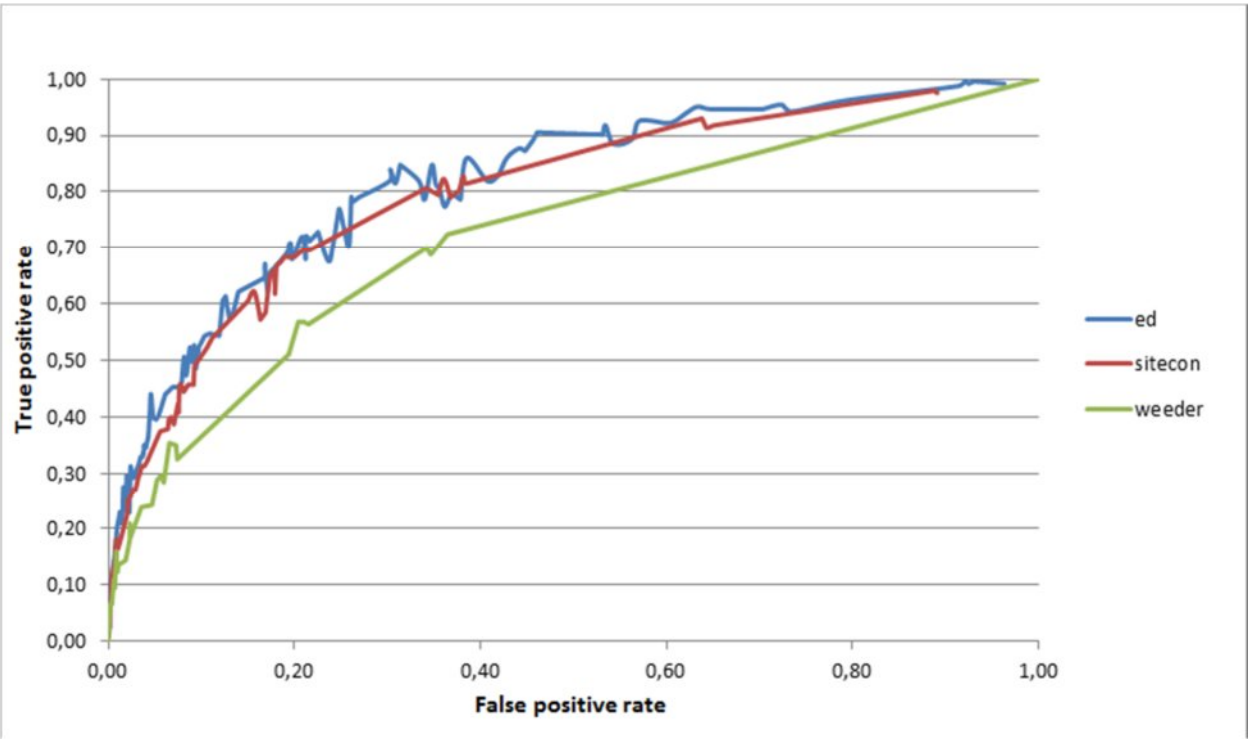


**Рисунок 6.** ROC-кривые распознавания тремя методами

На рисунке 6 показана кривая ошибок распознавания методами ExpertDiscovery, SITECON и WEEDER. Модели сигналов SITECON и WEEDER содержат только значимые элементарные сигналы, без иерархических комплексных сигналов. Модель сигналов ExpertDiscovery содержит комплексные сигналы, построенные на основании

элементарных сигналов и предикатов «Дистанция». Таким образом, модель ExpertDiscovery содержит значимые сигналы трех типов: SITECON, WEEDER и иерархические комплексные сигналы, что объясняет улучшение качества распознавания. Стоит отметить, что при применении всех трех полученных моделей сигналов к выборке промоторов, специфичных для печени, все три метода показали одинаково плохой результат, как и ожидалось, близкий к случайной классификации. Значительной разницы между методами в этом случае не наблюдается.

Среди комплексных сигналов ExpertDiscovery были найдены сигналы, соответствующие биологически значимым регулирующим элементам. Например, пары GATA-HNF или GATA-GATA соответствуют известным композитным элементам, описанным в статьях [24] и базе данных TRANSCompel.

## Заключение

В результате выполненной работы система ExpertDiscovery, встроенная в UGENE, была доведена до уровня, достаточного для получения биологических результатов, принимаемых мировым сообществом [37]. Был выявлен ряд недостатков системы, как в алгоритмической части, так и в части пользовательского интерфейса. Выявленные недостатки были устранены, и обновленная система была применена к реальным данным для демонстрации качества ее работы.

Таким образом, пользователю предоставляется мощное средство анализа регуляторных областей, способное комбинировать результаты работы различных методов распознавания.

Система ExpertDiscovery включена в пакет UGENE, который обладает обширной аудиторией пользователей и распространяется бесплатно.

Промежуточные результаты работы были опубликованы в ряде статей [25-27] и демонстрировались на нескольких конференциях:

- L Международная научная студенческая конференция «Студент и научно-технический прогресс» (Диплом III степени); Васькин Ю.Ю. Анализ регуляторных районов генов. Интегрированная система Expert Discovery и UGENE. p.202
- Международная конференция «Современные проблемы математики, информатики, биоинформатики»; Витяев Е.Е., Васькин Ю.Ю., Хомичева И.В. Анализ последовательностей регуляторных районов генов реляционной системой Expert Dsicovery, встроенной в пакет UGENE. p. 52
- Международная конференция «Постгеномные методы в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика». Vaskin Y.Y., Vityaev E.E, Khomicheva I.V. UGENE and ExpertDiscovery: integrated system for analysis of genes regulatory regions. p.214;
- VIII Международная конференция по биоинформатике, системной биологии, регуляции и структуре геномов. Vaskin Y.Y., Vityaev E.E., Khomicheva I.V. Data Mining tool for analysis of regulatory regions of gene: Integration of ExpertDiscovery and UGENE. P. 324

**Список литературы**


1. Хомичева И. В., Витяев Е.Е., Шипилов Т.И. Анализ регуляторных районов ДНК программной системой ExpertDiscovery. Вестник НГУ, серия: Информационные технологии, 2010
2. Бланшет Ж., Саммерфилд М. Qt 4: программирование GUI на C++. Москва, 2007.
3. Кендал М., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973.
4. Витяев Е.Е., ОрловЮ.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
5. Khomicheva I.V., Vityaev E.E., Ananko E.A., Shipilov T.I., Levitsky V.G. ExpertDiscovery system application for the hierarchical analysis of eukaryotic transcription regulatory regions based on DNA codes of transcription. Intelligent Data Analysis, Special issue on "Machine learning and bioinformatics" eds. Nikolai Kolchanov, Evgenii Vityaev. v.12(5), IOS Press, 2008
6. Nikolay A. Kolchanov, Mikhail A. Pozdnyakov, Yury L. Orlov, Oleg V. Vishnevsky, Nikolay L. Podkolodny, Eugenii E. Vityaev and Boris Kovalerchuk Computer System “Gene Discovery” for Promoter Structure Analysis In: Artificial Intelligence and Heuristic Methods in Bioinformatics, Eds: P. Frasconi and R. Shamir, IOS Press, 2003
7. Okonechnikov K., Golosova O., Fursov M., the UGENE team, Unipro UGENE: a unified bioinformatics toolkit, Bioinformatics 2012 28: 1166-1167
8. Kolchanov N.A., Merkulova TI, Ignatieva EV, Ananko EA, Oshchepkov DY, Levitsky VG, Vasiliev GV, Klimova NV, Merkulov VM, Charles Hodgman T.Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes.Brief Bioinform. 2007 Jul;8(4):266-74.
9. Kel A.E., Kolchanov N.A., Kel O.V., Romashenko A.G., Ananko E.A., Ignateva E.V., Merkulova T.I., Podkolodnaya O.A., Stepenko I.L., Kochetov A.V., Kolpakov F.A., Podkolodniy N.L., Naumochkin A.A. 1997. TRRD – eukaryotic gene regulatory regions database. Mol. Biol., t. 31, p. 626-636
10. Lemon B., Tjian R. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 2000. V.14, №20. P. 2551-2569.

11. Nikolov D.B., Burley, S.K. (1997) RNA polymerase II transcription initiation: A structural view. Proc. Natl. Acad. Sci. USA, 94, 15-22.
12. Caley M., Smale S.T. Transcriptional Regulation in Eukaryotes. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000, 640 p.
13. Trifonov E.N. (1997) Genetic level of DNA sequences is determined by superposition of many codes. Mol. Biol. (Mosk) 31, 759-767.
14. 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;10(3):R29.
15. Stormo G.D. DNA binding sites: representation and discovery. Bioinformatics 2000, 16:16-23.
16. 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.
17. 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 Dec 19;8(1):481
18. Cornish-Bowden A. (1985) Enzyme kinetics in Comprehensive Biotechnology (ed. M. Moo-Young), vol. 1, pp. 521–538, Pergamon, Oxford
19. Sinha, S. and Tompa, M. A Statistical Method for Finding Transcription Factor Binding Sites, Eighth International Conference on Intelligent Systems for Molecular Biology, San Diego, CA, August 2000, 344-354.
20. Liu, X., Yu, X., Zack, D.J., Zhu, H., Qian, J. (2008) TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics. 9:271.
21. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.
22. Pavesi, G., Zambelli, F., Pesole, G.WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. BMC Bioinformatics 2007, 8:46
23. Das M., Dai H. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007 Nov 1;8 Suppl 7:S21.
24. Trainor C.D., Omichinski J.G., Vandergon T.L., Gronenborn A.M., Clore G.M. A palindromic regulatory site within vertebran GATA-1 promoters requires both zinc

fingers of the GATA-1 DNA-binfing domain for high-affinity interaction. Mol. Cell. Biol., 1996, vol. 16, no 5 2238-2247

25. Vaskin Y.Y., Khomicheva I.V., Ignatyeva ,E.V. Vityaev E.E. “ExpertDiscovery and UGENE integrated system for intelligente analysis of regulatory regions of genes”. In Silico Biol. 2011-2012;11(3-4):97-108. doi: 10.3233/ISB-2012-0448.
26. Васькин Ю.Ю., Хомичева И.В., Игнатьева Е.В., Витяев Е.Е. “Анализ последовательностей регуляторных районов генов реляционной системой ExpertDiscovery, встроенной в пакет UGENE”. “Вестник НГУ”, Том 10, Выпуск 1, Новосибирск, 2011
27. Khomicheva I., Vityaev E., Vaskin Y., and Shipilov T. “Analysis and Prediction of Regulatory Regions of Eukaryotic Genes by integrated UGENE and ExpertDiscovery Systems”. Special issue: ECML/PKDD 2011 (5th Workshop on Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions), Athens, 2011
28. Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. // НГУ, Новосибирск, 2006.
29. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and Symbols for Nucleic Acids, Polynucleotides and their Constituents. Recommendations 1970. Arch. Biochem. Biophys. 145, 425-436 (1971); Biochem. J. 120, 449-454 (1970); Biochemistry 9, 4022-4027 (1970); Biochim. Biophys. Acta 247, 1-12 (1971); Eur. J. Biochem. 15, 203-208 (1970), corrected 25, 1 (1972); Hoppe-Seyler's Z. Physiol. Chem. (in German) 351, 1055-1063 (1970); J. Biol. Chem. 245, 5171-5176 (1970); Mol. Biol. (in Russian) 6,167-I 74 (1972); Pure Appl. Chem. 40, 277-290 (1974); also pp. 116-121 in Biochemical nomenclature and related documents (1978), the Biochemical Society, London.
30. Fisher, R. A. (1922). "On the interpretation of $\chi 2$ from contingency tables, and the calculation of P". Journal of the Royal Statistical Society 85 (1): 87–94. doi:10.2307/2340521
31. Yule U. On the Association of Attributes in Statistics // Philosophical Transactions of the Royal Society of London, Ser. A, Vol. 194, 1900, P. 257–319.
32. Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology *147*, 195–197.

33. Eddy, S.R. (2008). A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. PLoS Comput Biol *4*, e1000069.
34. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.
35. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. *32*, 1792–1797.
36. Lassmann, T., and Sonnhammer, E.L. (2005). Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics *6*, 298.
37. Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition *30*, 1145–1159.
38. Vityaev E.E., Orlov Yu.L., Vishnevsky O.V., Pozdnyakov M.A., Kolchanov N.A. (2002) Computer system "Gene Discovery" for promoter structure analysis // (Bioinformation Systems e.V.) In Silico Biology 2(3), 233-247.
39. Kolchanov N.A., Pozdnyakov M.A., Orlov Yu.L., Vishnevsky O.V., Podkolodny N.L., Vityaev E.E., Kovalerchuk B. (2003) Computer System “Gene Discovery” for Promoter Structure Analysis. // In: Artificial Intelligence and Heuristic Methods in Bioinformatics (Eds: P. Frasconi and R. Shamir), IOS Press (ISBN 1-58603-294-1), p. 173-192.
40. Deyhols M., Harrington M. Open Genetics – Winter 2012. http://hdl.handle.net/10402/era.24984. p. 156-169
41. Baikalov, I., Grzeskowiak, K., Yanagi, K., Quintana, J., and Dickerson, R.E. (1993). The crystal structure of the trigonal decamer C-G-A-T-C-G-6meA-T-C-G: a B-DNA helix with 10.6 base-pairs per turn. J. Mol. Biol. *231*, 768–784.
42. Орлов, Ю.Л. (2004). Анализ регуляторных геномных последовательностей с помощью компьютерных методов оценок сложности генетических текстов.
43. DeLaForest, A., Nagaoka, M., Si-Tayeb, K., Noto, F.K., Konopka, G., Battle, M.A., and Duncan, S.A. (2011). HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. Development *138*, 4143–4153.