

# **Программная система ExpertDiscovery анализа регуляторных районов ДНК.**

**Шипилов Т.И., Хомичева И.В., Витяев Е.Е.**

## **АННОТАЦИЯ**

В последнее время наблюдается лавинообразный рост объема информации о первичных нуклеотидных последовательностях. Для их анализа применяется множество методов поиска и распознавания контекстных сигналов. Количество различных решений одной задачи порождает массу результатов, анализируемых экспертами в буквальном смысле вручную, тогда как их исследование требует интенсивного использования не просто баз данных по нуклеотидным, аминокислотным последовательностям и метаболическим путям, но и знаний, концепций, логических схем, построенных на основе таких баз данных. Таким образом, необходимо применение современных компьютерных технологий, связанных с интеллектуальным анализом данных (Data Mining and Knowledge Discovery). Нами разработана программная система ExpertDiscovery, позволяющая эксперту осуществлять анализ, как самих последовательностей, так и сигналов, обнаруженных на этих последовательностях при помощи различных методов и программ. Находить связи и закономерности во множестве сигналов и объединять их со знаниями экспертов для достижения биологически значимого результата при помощи метода Discovery. В процессе разработки системы был формализован вид гипотезы эксперта, разработаны механизмы её формулирования и проверки. Также был разработан алгоритм автоматического построения и уточнения гипотез на основании заданных экспертом критериев. Система применена для распознавания сайтов связывания транскрипционных факторов, распознавания промоторных районов генов эукариот и построения модели регуляторных районов.

## **Введение**

Знания о регуляторной функции ДНК, РНК и белков имеют важнейшее значение при решении широкого круга задач молекулярной биологии, молекулярной генетики, биотехнологии, медицины. Осуществление крупномасштабных проектов по секвенированию геномов человека, животных, растений, бактерий и вирусов привело к лавинообразному росту объема информации о первичных нуклеотидных последовательностях. Их анализ, обобщение и накопление знаний о структуре и функции генетических молекул становится, в настоящее время одной из наиболее важных проблем современной молекулярной генетики.

Биологические процессы обладают большой специфичностью и сложностью, поэтому их исследование требует интенсивного использования не просто баз данных по нуклеотидным, аминокислотным последовательностям и метаболическим путям, но и знаний, концепций, логических схем, построенных на основе таких баз данных. Технически проблема состоит в формализации и унификации разнородной информации по физико-химическим, структурным, информационным свойствам регуляторных последовательностей генов, экспериментальных данных об их функционировании и выявлении на этой основе с помощью специализированных алгоритмов новых биологических знаний в легко интерпретируемой форме. Для её решения необходимо применение современных компьютерных технологий, связанных с интеллектуальным анализом данных (Data Mining and Knowledge Discovery).

На сегодняшний день не один из известных методов не в состоянии полностью решить эту проблему. В большинстве случаев эксперты-генетики вынуждены анализировать огромное количество информации, часто противоречивой, чтобы добиться биологически значимого результата.

Для анализа нуклеотидных последовательностей регуляторных районов генов применяется множество методов поиска и распознавания контекстных сигналов. Среди них метод весовых матриц, метод Discovery, нейронные сети, различные оптимизационные подходы, основанные на генетических и других эвристических алгоритмах. Количество различных решений одной задачи порождает массу результатов, анализируемых экспертами института в буквальном смысле вручную. Конечно, существуют алгоритмы и программы, облегчающие часть рутинной работы, однако их применение ограничено иногда из-за неудобства интерфейса пользователя, иногда из-за ограниченности их возможностей, иногда из-за отсутствия наглядности, необходимой для анализа результатов. Инструменты позволяющие проверить произвольную гипотезу эксперта отсутствуют, а важность их очень велика, потому что за годы работы эксперты не просто освоили генетические цепочки, с которыми они работают, а в прямом смысле этого слова почувствовали их.

Нами была разработана программная система, позволяющая эксперту осуществлять анализ, как самих последовательностей, так и сигналов, обнаруженных на этих последовательностях при помощи различных методов и программ. Находить связи и закономерности во множестве сигналов и объединять их со знаниями экспертов для достижения биологически значимого результата.

В процессе разработки системы был формализован вид гипотезы эксперта, разработаны механизмы её формулирования и проверки. Также был разработан алгоритм автоматического построения и уточнения гипотез на основании заданных экспертом критериев.

## §1. Определения

Для формализации задачи и вида гипотез экспертов, введём следующие обозначения.

**Определение 1.** Текстовые строки, состоящие из символов A, C, T, G латинского алфавита, будем называть символьными последовательностями (или просто последовательностями) в 4-х буквенном коде.

**Определение 2.** Символы 15-ти буквенного кода: A, C, T, G, M, W, R, Y, S, K, H, V, D, B, N. Соответствующие им символы 4-х буквенного кода и пояснение можно найти в таблице 1.

Таблица 1. Символы 15-ти буквенного кода.

15-ти буквенный символ	Набор 4-х буквенных символов	Пояснение
A	A	Аденин
C	C	Цитозин
T	T	Тимин
G	G	Гуанин
M	A,C	Амино-содержащие
W	A,T	Слабые взаимодействия (две водородные связи)
R	A,G	Пурины (большие остатки)
Y	C,T	Пиримидины (малые остатки)
S	C,G	Сильные взаимодействия (три водородные связи)
K	T,G	Кето-содержащие

H	A,C,T	He G
V	A,C,G	He T
D	A,T,G	He C
B	C,T,G	He A
N	A,C,T,G	Любой

**Определение 3.** Текстовые строки, состоящие из символов 15-ти буквенного кода, будем называть символьными последовательностями (или просто последовательностями) в 15-ти буквенном коде.

Как известно, генетическая информация представляется в виде символьных последовательностей в 4-х или 15-ти буквенном коде. Любая последовательность может быть отнесена к определённому классу, определяющему её назначение.

**Определение 4.** Сигнал – система правил, определяющих свойства участков последовательностей ДНК.

Задача распознавания сигналов заключается в том, чтобы обнаружить сигналы, отличающие последовательности одного класса от всех остальных. При решении этой задачи эксперт формулирует некоторую гипотезу, проверяет её на данных и на основе результатов проверки производит уточнение гипотезы, после чего цикл повторяется.

Уточним вид гипотез экспертов. Для этого введём ещё несколько определений.

**Определение 5.** Элементарный сигнал – неделимый сигнал, который характеризуется именем и местами в последовательности, где он присутствует.

Такие сигналы могут формулироваться экспертом непосредственно, а также загружаться в систему в виде аннотации последовательностей. 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.

**Определение 6.** Гипотезы экспертов формулируются в виде Комплексных Сигналов (КС) определяемых рекурсивно на основе элементарных сигналов и операций над ними:

1. Элементарный сигнал является комплексным сигналом;
2. Результат воздействия на комплексный сигнал операций (подробное определение операций приведено ниже) повторения или принадлежности интервалу является комплексным сигналом;
3. Результат воздействия на два комплексных сигнала операции дистанция между сигналами является комплексным сигналом.

Множество операций может быть расширено. Здесь перечислены лишь важнейшие из них, которые реализованы в программе.

**Определение 7.** Над комплексными сигналами определены следующие операции:

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

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

*Принадлежность сигнала интервалу.* Указывает, что входной сигнал 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 комплексными сигналами может быть измерена различными способами, такими как:

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

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

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

### **§3. Задача обнаружения комплексных сигналов.**

Задача автоматического обнаружения комплексных сигналов является достаточно сложной и в настоящее время не существует готовых методов для её решения. Для решения этой задачи мы использовали оригинальный реляционный (Relational Data Mining) подход к обнаружению знаний [2, 10, 11, 12]. Этот подход и реализующая его система Discovery применялись для решения большого количества практических задач в психофизике, диагностики раковых заболеваний и предсказания курсов акций ценных бумаг. В процессе решения задач было проведено сравнение систем Discovery с другими методами интеллектуального анализа данных (Data Mining and Knowledge Discovery) такими как Neural Network (Brainmaker, California Scientific Software), Linear Discriminant Analysis (SIGAMD, StatDialogue, Moscow), Decision Tree (SIPINA, Lyon, France), Inductive Logic Programming (FOIL, First Order Inductive Logic, York, UK). Система Discovery всегда давал лучшие результаты, чем другие методы и в среднем в 1.5 раза превосходил их. Детали сравнения приведены в статьях и представлены на сайте Scientific Discovery [9]. Там же можно ознакомиться с результатами применения системы Discovery в различных предметных областях.

В основе системы Discovery лежит семантический вероятностный вывод, обладающий целым рядом теоретически важных свойств (см. [2, 11]) и позволяющий, в частности, находить все максимально вероятные и специфичные закономерности в данных. Идея состоит в последовательном уточнении гипотезы таким образом, чтобы на каждом следующем шаге получались гипотезы с большей вероятностью и определённой. При этом осуществляется проверка значимости полученного результата при помощи статистических критериев.

**Определение 8.** 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* (УВ) правила  $\text{Prob}(C_i) = \text{Prob}(G/A_1^i \& \dots \& A_{k_i}^i) = \text{Prob}(G \& A_1^i \& \dots \& A_{k_i}^i) / \text{Prob}(A_1^i \& \dots \& A_{k_i}^i)$ ;
4.  $C_i$  – *Вероятностные Законы* (ВЗ), 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* (СВЗ), 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 [2].

На основе системы Discovery разработана специализированная система ExpertDiscovery обнаружения комплексных сигналов в участках последовательностей ДНК.

#### §4. Система ExpertDiscovery.

Ключевым отличием системы ExpertDiscovery от системы Discovery является класс рассматриваемых гипотез SetKC и процесс их уточнения. Для работы алгоритма требуется определить набор операций SetO, которые будут использоваться для генерации KC, а также критерии, по которым будет производиться отбор KC.

На первом шаге алгоритма рассматривается популяция комплексных сигналов являющихся элементарными. На последующих шагах мы уточняем KC текущей популяции. Для уточнения рассматриваемого KC делается следующее:

- (1) выбирается один из элементарных сигналов T;
- (2) из набора операций SetO берётся одна из операций, и осуществляется замена T на эту операцию примененную к некоторым другим элементарным сигналам;
- (3) у полученного комплексного сигнала проверяются критерии отбора KC и, в случае их выполнения, он записывается в результат;
- (4) иначе проверяются критерии ветвления. В случае их выполнения сигнал переносится в следующую популяцию.
- (5) если ни один из предыдущих критериев не выполнялся, то KC отсеивается.

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

Этот цикл продолжается до тех пор, пока не получится пустая популяция комплексных сигналов. Результатом работы алгоритма является совокупность ResKC комплексных сигналов, полученных на этапе (3) работы алгоритма.

Для проверки комплексного сигнала необходимо две выборки символьных последовательностей. Назовём их условно YES и NO. Выборка YES содержит последовательности, которые принадлежат к классу содержащему сигнал. Выборка NO содержит последовательности других классов или случайно сгенерированные и используется для подсчёта статистических параметров сигнала.

В текущей реализации алгоритма присутствуют следующие критерии для отбора результатов:

- порог условной вероятности KC;

- порог статистической значимости по критерию Фишера;
- порог покрытия позитивной выборки.

Для принятия решения о продолжении ветвления:

- порог условной вероятности КС;
- порог статистической значимости по критерию Фишера;
- 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  $P$  того, что последовательность принадлежит выборке YES при условии, что в ней есть сигнал:

$$P = \frac{a_{11}}{a_{10} + a_{11}},$$

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])
3. покрытие позитивной выборки в % (доля последовательностей позитивной выборки, содержащих сигнал).
4. покрытие негативной выборки в % (доля последовательностей негативной выборки, содержащих сигнал).
5. для операции дистанция оценивается уровень корреляции между аргументами.

## §5. Программная реализация

Программа ExpertDiscovery реализована на языке C++, в среде программирования Microsoft Visual Studio .NET, предназначена для использования в операционных системах Win9x/NT/2k/XP. Интерфейс пользователя построен на базе библиотеки Microsoft Foundation Classes (MFC). Код разработан с использованием архитектуры «документ/вид».

Выбор языка программирования объясняется жёсткими требованиями к производительности программы. Язык C++ предоставляет широкие возможности для написания эффективных программ и при этом поддерживает разработку объектно-ориентированных приложений, что позволяет создавать хорошо структурированный и легко модифицируемый код.

Операционные системы семейства Windows были выбраны потому, что эксперты пользуются именно такими ОС на своих рабочих компьютерах.

Для разработки интерфейса пользователя была выбрана библиотека MFC, потому что её использование по сравнению с Windows API позволяет существенно ускорить и упростить процесс разработки, придать программе большую структурированность.

Логически всю программу можно разделить на две относительно независимых по своим функциям части, а именно:

- *вычислительная часть.* Содержит структуры данных для организации внутреннего представления комплексных сигналов, их проверки, а также классы для сохранения результатов счёта, классы, представляющие символьные последовательности и их выборки. Она реализована в виде библиотеки C++ классов, которая может быть собрана как для ОС Windows, так и для \*nix. Такая реализация позволяет использовать алгоритмический блок программы в других приложениях.
- *интерфейс программы.* Содержит серию классов, обеспечивающих визуальный анализ сформулированного комплексного сигнала на загруженных базах последовательностей, отображение результатов проверки сигнала, диалоги ввода входных данных и комплексных сигналов.

## §6. Описание интерфейса программы

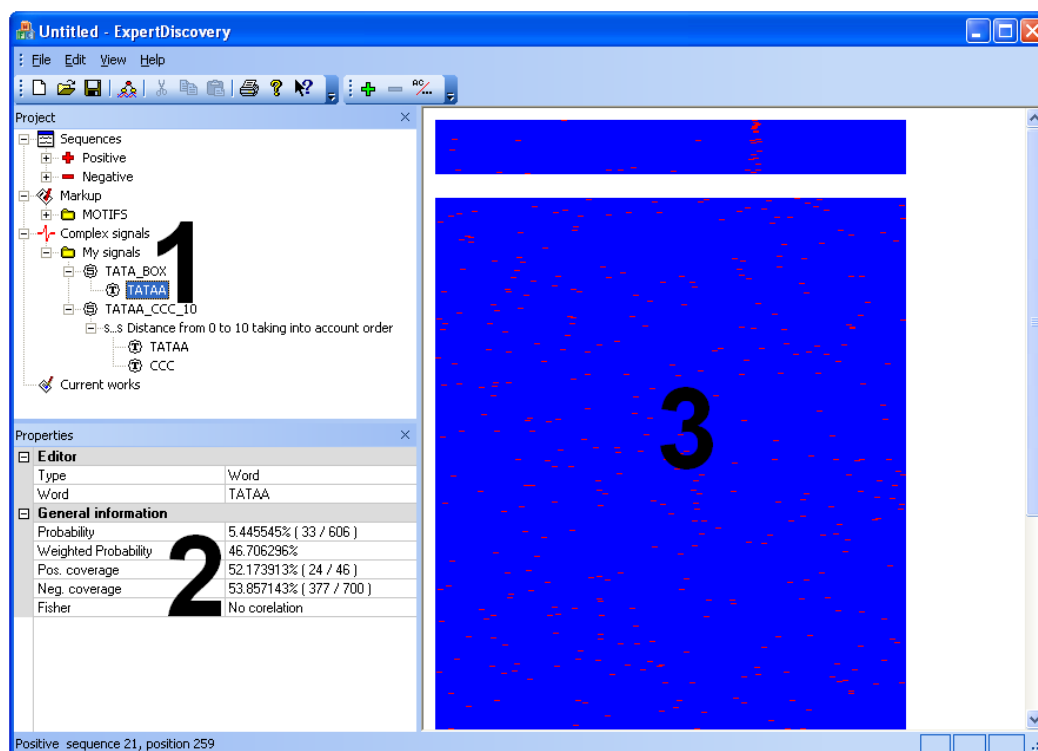


Рис. 1. Общий вид программы ExpertDiscovery

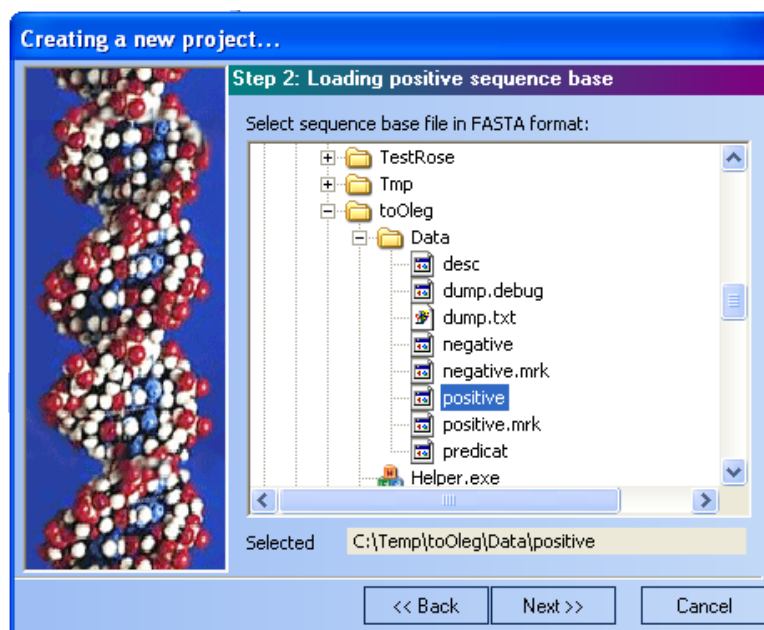
Интерфейс программы построен по принципам архитектуры «документ/вид». В основном окне программы можно выделить следующие элементы:

1. панель проекта;
2. панель свойств текущего элемента;
3. визуализация текущего элемента.

Их внешний вид можно видеть на рисунке 1. Также там присутствуют панели инструментов и меню.

Панель проекта служит для навигации по данным, загруженным в программу. На ней данные упорядочены в иерархическую структуру и разбиты на четыре группы:

- последовательности (Sequences);
- разметка последовательностей (Markup);



- комплексные сигналы (Complex signals);
- текущие работы (Current works).

В ветке последовательностей, в подгруппах Positive и Negative, соответственно, содержатся последовательности позитивной и негативной выборки. В ветке разметки содержится информация об элементарных сигналах, сгруппированная по семействам. В разделе «комплексные сигналы» находятся определённые экспертом или построенные автоматически сигналы. В «текущих работах» находятся выполняющиеся в текущий момент операции по обработке данных.

При выборе любого элемента панели проекта на панели свойств отображаются его характеристики и параметры (рис. 1), последние могут быть отредактированы. В окне визуализации отображается графическое представление текущего элемента. Для комплексного сигнала это распределение по последовательностям. Так на рисунке 1 изображено распределение сигнала TATAA по загруженным в систему последовательностям. Для последовательностей отображено имя и код.

Такой способ отображения позволяет эксперту видеть всю доступную информацию в одном месте и избавляет его от необходимости переключаться между многочисленными окнами.

**Загрузка данных в программу.** Входными данными для программы являются две выборки последовательностей. Дополнительно может быть загружена их разметка.


Для загрузки данных используется команда меню *File->New* или кнопка  панели инструментов. Процесс загрузки производится при помощи мастера. На первой странице производится пояснение последующих шагов. На второй странице требуется указать имя файла в формате FASTA (см. приложение 1) с позитивной выборкой последовательностей (рис. 2). На третьей – имя аналогичного файла с негативной выборкой.

Рис. 2. Выбор файла с позитивной выборкой последовательностей

Сразу после загрузки последовательностей на экране появится мастер для загрузки разметки. Можно отказаться от его услуг, нажав кнопку Cancel. В будущем разметка может быть загружена отдельно при помощи команды контекстного меню всплывающего при щелчке правой кнопкой мыши по группе Markup панели проекта (рис. 3).



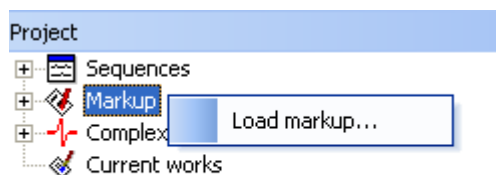


Рис. 3. Контекстное меню загрузки разметки

Мастер загрузки разметки содержит четыре шага. На первом шаге приводится описание последующих шагов. На втором требуется указать файл с разметкой позитивной последовательности (см. формат в приложении 1). На третьем – негативной последовательности. На последнем шаге требуется указать файл с описанием имён внешних сигналов (см. формат в приложении 1), либо установить галочку для автоматической генерации этой информации по разметке.

**Создание и редактирование комплексного сигнала.** Для создания комплексного сигнала необходимо воспользоваться контекстным меню, которое появляется при щелчке правой кнопкой мышки на группе «комплексные сигналы» панели проекта или одной из папок в этой группе. В этом меню требуется выбрать пункт *New signal* (рис. 4). Появится новый сигнал с именем *NewSignalXX*, где XX – число необходимое для обеспечения уникальности имени сигнала.

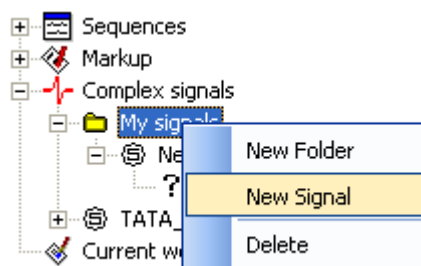



Рис. 4. Контекстное меню для создания нового сигнала

В силу своего определения комплексный сигнал представим в виде дерева, в узлах которого стоят операции, а листья являются терминальными символами Т. Именно эта форма представления была взята за основу его представления в интерфейсе программы. После создания дерево содержит корень с именем сигнала и один неопределённый элемент. После щелчка на неопределённом элементе на панели свойств можно установить тип этого элемента и его параметры. В программе присутствуют следующие типы узлов комплексного сигнала:

- Операция дистанция (Distance). На панели свойств для этой операции можно установить способ измерения дистанции (Distance type), диапазон расстояний, которые могут быть между аргументами операции (Distance from min и Distance to max) и указать, важен ли порядок аргументов (Order).
- Операция повторения (Repetition). Свойствами этой операции являются искомое количество повторений (от Count from Nmin до Count to Nmax), дистанция между этими повторениями (от Distance from min до Distance to max) и способ измерения дистанции (Distance type).
- Операция интервал (Interval). Для этой операции на панели свойств устанавливается диапазон позиций последовательности (от Interval from min до Interval to max), в которых ищется аргумент операции.
- Слово (Word). Соответствует элементарному сигналу являющемуся словом. Имеет один параметр – слово в 15-ти буквенном коде (Word).

- Элемент разметки (Markup Item). Соответствует элементарному сигналу из разметки. Параметрами являются семейство (Family) внешнего сигнала и его имя (Name).

**Автоматическое построение комплексных сигналов.** Для автоматического построения комплексных сигналов используется пункт меню *Edit->Extract signals...* или кнопка  панели инструментов. Для запуска процесса используется мастер, в котором на 1-м шаге устанавливаются параметры отбора результатов и отсева неперспективных гипотез. Часть параметров вынесена на вкладку Advanced. На втором шаге требуется сформулировать предикаты, которые будут использоваться при построении комплексных сигналов (рис. 5).

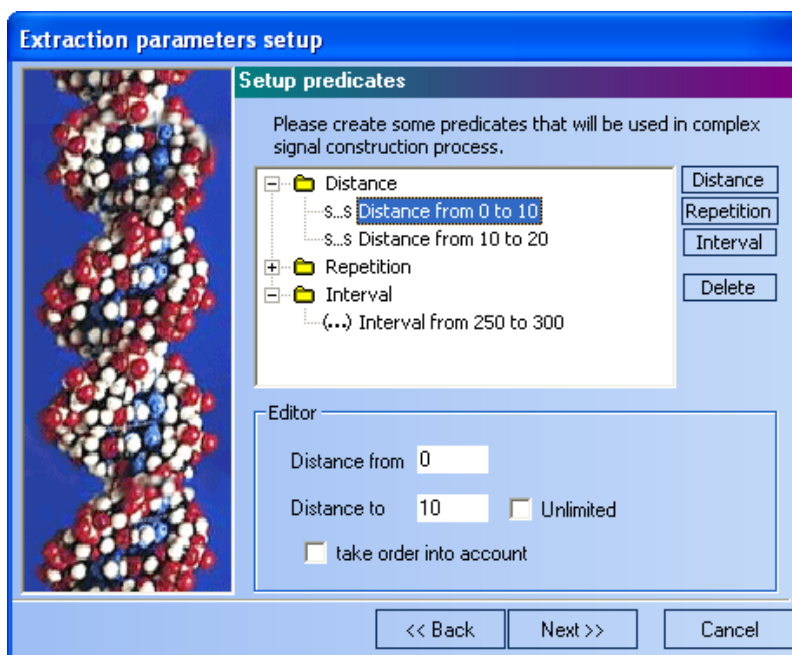


Рис. 5. Окно для спецификации предикатов

Чтобы задать предикат, необходимо создать его, используя одну из кнопок Distance, Repetition или Interval. После чего установить в окне Editor его параметры и перейти к спецификации следующего предиката.

На последней странице мастера требуется указать папку из ветки «комплексные сигналы» проекта, в которую будет помещён результат.

После завершения мастера в списке текущих задач появится новая задача с именем Extracting signals (x%), где x – процент завершения этой задачи. Это выглядит, как показано на рисунке 6.

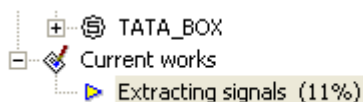


Рис. 6. Отображение созданного процесса в списке текущих работ

## §6. Применение

В процессе исследования была показана высокая эффективность ExpertDiscovery как системы распознавания сайтов связывания транскрипционных факторов (ССТФ), коротких участков ДНК, служащих местом посадки для регуляторных белков [4-7]. Проведено сравнение системы ExpertDiscovery с оптимизированным методом весовых матриц (Position Weight Matrix, PWM), точность которого оценивалась на разной длине сайта. В качестве модельных объектов использовались ССТФ SF1, EGR1, SREBP. Сравнение проводилось в соответствии со стандартной процедурой бутстреп.

В таблице 2 представлена точность распознавания ССТФ SF1, EGR1, SREBP системой ExpertDiscovery и методом PWM, полученная для контрольных объектов.

Таблица 2. Ошибки второго рода для ССТФ SF1, SREBP, EGR1, вычисленные при ошибке 1-го рода 50% для системы ExpertDiscovery и PWM.

ССТФ	ExpertDiscovery	PWM
SF1	5.01E-05	6.87E-05
SREBP	1.97E-04	8.32E-04
EGR1	8.09E-04	4.06E-03

Сравнение показало, что система улавливает закономерности организации нуклеотидных последовательностей и не уступает оптимизированному методу весовых матриц в задаче распознавания ССТФ.

Показана возможность генерации знаний в виде комплексных сигналов, выделяющих подгруппы биологически осмысленных последовательностей, что является принципиальным отличием системы от PWM и других подходов.

Также была показана высокая эффективность системы ExpertDiscovery для распознавания промоторных районов генов эукариот как совокупности ССТФ необходимых для инициации транскрипции гена, и других регуляторных районов, определяющих стадие- и тканеспецифичный характер транскрипции.

Несмотря на то, что многие методы сосредоточены на исследовании индивидуальных ССТФ, очень немногие предназначены для анализа их взаимного расположения и исследования взаимодействий между ними. Эти методы дают только общую оценку, насколько случайно расположение сайтов, но не учитывают всего многообразия особенностей расположения сайтов, и не учитывают другие контекстные характеристики, которые могут быть важны для регуляции транскрипции гена. Известные методы распознавания, как правило, чувствительны только к конкретным характеристикам и, как следствие, дают хорошие результаты распознавания на одной группе районов и низкую точность распознавания на другой. Система ExpertDiscovery использовалась для распознавания промоторов путем интегрирования контекстных характеристик регуляторных элементов и ССТФ, полученных другими методами.

Закономерности (комплексные сигналы), обнаруженные системой, являются биологически интерпретируемыми правилами расположения различных контекстных характеристик регуляторных районов генов [6]. Система ExpertDiscovery осуществляет иерархический анализ регуляторных районов, обнаруживая биологически-целесообразные иерархически-усложняющиеся модели районов от простых композиционных моделей,

состоящих из двух ССТФ, до сложных моделей сигналов, соответствующих комплексной регуляции транскрипции.

## §7. Заключение

В работе впервые был использован интеграционный подход к анализу генетической информации. Существующие методы фокусируются на выявлении отдельных сигналов. Созданная система призвана объединить результаты работы этих методов со знаниями экспертов для получения наилучшего биологического результата. Таким образом, она позволяет проводить исследования на любом уровне от нуклеотидного до геномного и даёт возможность интеграции специфичных методов на каждом уровне.

В результате проведённой работы была создана программа «ExpertDiscovery». Она представляет собой удобный и эффективный инструмент для автоматизации работы экспертов, который обеспечивает удобные средства для визуализации данных, простой механизм ввода, проверки, анализа и уточнения комплексных сигналов. Система прошла тестовую эксплуатацию в Институте Цитологии и Генетики и показала свою эффективность при решении практических задач.

Разрабатываемый оригинальный подход к интеграции данных от различных методов и знаний экспертов показал свою эффективность на практике. Полученные результаты опубликованы в серии статей, например в [4-8, 13-14] и в главе в книге [1].

## Литература

1. 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.
2. Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. // НГУ, Новосибирск, 2006. 293 с.
3. Кендал М., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973. С. 899.
4. 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
5. Khomicheva I.V., Vityaev E.E., Ananko, V.G., Levitsky V.G., Shipilov T.I. 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, 2007, pp. 142-144.
6. Khomicheva I.V., Vityaev E.E., Ananko E.A., Shipilov T.I., Levitsky V.G., ExpertDiscovery application for the hierarchical analysis of the eu-karyotic transcription regulatory regions based on the 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 (in press)

7. 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.
8. 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, pp.173-192.
9. Scientific Discovery Web Site, <http://www.math.nsc.ru/AP/ScientificDiscovery>
10. E. 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
11. Evgenii Vityaev, Boris Kovalerchuk, Empirical Theories Discovery based on the Measurement Theory. *Mind and Machine*, v.14, #4, 551-573, 2004
12. 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.
13. Vityaev E.E., Shipilov T.I., Pozdnyakov M.A., Vishnevsky O.V., Proscura A.L., Orlov Yu.L., Arrigo P. Analysis of gene regulatory sequences by knowledge discovery methods, In: "Bioinformatics of Genome Regulation and Structure II." (eds. by Nikolay Kolchanov and Ralf Hofstaedt), Kluwer-Springer, Inc. 2005
14. Vityaev E.E., Orlov Yu.L., Pozdnyakov M.A., Vishnevsky O.V., Kolchanov N.A., Kovalerchuk B.K. Knowledge Discovery for Promoter Structure Analysis In: Proceedings of the International Conference on Imaging Science, Systems, and Technology Eds.: Hamid R. Arabnia, Youngsong Mun, Las Vegas, Nevada, USA, June 24-27, 2002, CSREA Press, v.1, 122-128