Анализ данных и продукция знаний в системе GeneExpress – электронной библиотеке по структуре и функции ДНК, РНК и белков

Н.А. Колчанов
ИЦиГ, Новосибирск, Россия
kol@bionet.nsc.ru

Н.Л. Подколодный
ИВМиМГ, Новосибирск, Россия
pnl@omzg.sscc.ru

М.П. Пономаренко
ИЦиГ, Новосибирск, Россия
pon@bionet.nsc.ru

Е.Е. Витяев
ИМ, Новосибирск, Россия
Vityaev@math.nsc.ru

В.А. Иванисенко
ИЦиГ, Новосибирск, Россия
salix@bionet.nsc.ru

Реферат

В данной работе описаны методы анализа данных, средства поиска, обнаружения, извлечения знаний и обогащения знаний, накапливаемых в электронной библиотеке ГенЭкспресс с целью исследования структурно-функциональных особенностей ДНК, РНК и белков. Применяемые подходы являются инструментальными, т.е. позволяют генерировать процедуры для решения не только отдельных задач, но и целых классов задач.

1 ВВЕДЕНИЕ

Знания о регуляторной функции ДНК, РНК и белков имеют важнейшее значение при решении широкого круга задач молекулярной биологии, молекулярной генетики, биотехнологии, медицины. Такого рода данные и знания накапливаются в Электронной библиотеке GeneExpress, разрабатываемой в ИЦиГ СО РАН [1,2].

Система GeneExpress предназначена для сбора экспериментальных данных, навигации и поиска информации, анализа данных и исследования зависимостей в области регуляции генной экспрессии. Она интегрирует большое количество распределенных баз данных, баз знаний по структурно-функциональным особенностям ДНК, РНК, белков и фундаментальных молекулярно-генетических процессов, в которых эти объекты задействованы, сотни программ для обработки этой информации, а так же другие доступные через Internet информационные ресурсы (ИР), важные для описания экспрессии генов.

Информационные и программные модули системы ГенЭкспресс могут быть разделены на несколько групп, в зависимости от выполняемой ими функции.

1) Базы данных, содержащие информацию о различных аспектах (особенностях) структурно-функциональной организации ДНК, РНК и белков, значимые для регуляции экспрессии генов.

2) Системы продукции знаний, позволяющие анализировать информацию, представленную в системе ГенЭкспресс с целью выявления особенностей структурно-функциональной организации генетических макромолекул, значимых для их функционирования, уровня специфической активности, а также для их распознавания и классификации.

2 ПРОДУКЦИЯ ЗНАНИЙ ПО СТРУКТУРНО-ФУНКЦИОНАЛЬНОЙ ОРГАНИЗАЦИИ РЕГУЛЯТОРНЫХ ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Извлечение знаний в базах данных (Knowledge Discovery in Databases and Data Mining KDD&DM) – интенсивно развиваемое направление. Оно является многоступенчатым интерактивным процессом, включающим в себя: создание «выборки»; «очищение» данных и их предобработку; выбор способа представления данных (визуализация, проекция, feature extraction и т.д.); выделение априорных знаний (background

© Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

knowledge, domain theory); выбор задачи извлечения знаний; выбор алгоритма; интерпретация результатов; использование извлеченного знания [3].

Применительно к электронной библиотеке GeneExpress методы и алгоритмы извлечения знаний должны осуществлять автоматическую обработку информации из баз данных (выборок) и давать закономерности, значимые для предсказания активности или поиска функциональных сайтов. Полученные в ходе этого процесса процедурные знания включают процедуру (программу, скрипт) для поиска или предсказания активности определенного типа функциональных сайтов: описание назначения процедуры, условия ее применения, формат входных данных, ограничения на входные данные, формат выходных данных и т.д. Эти знания позволяют автоматически синтезировать сложные сценарии для поиска различных функционально значимых участков аннотируемых последовательностей генома и предсказывать их структуру и функцию.

Основные блоки продукции знаний системы ГенЭкспресс перечислены ниже:

1. B-DNAvideo. Модуль предназначен для выявления контекстно-зависимых конформационных и физико-химических характеристик ДНК- сайтов, значимых для их функционирования, и построения методов их распознавания на основе этих характеристик.

2. Activity. Модуль предназначен для выявления контекстно-зависимых конформационных и физико-химических характеристик ДНК- сайтов, определяющих количественный уровень их специфической активности и построения методов для предсказания активности сайтов на основе их нуклеотидных последовательностей.

3. Cons-Freq. Модуль предназначен для выявления значимых контекстных характеристик сайтов и построения методов их распознавания на основе различных подходов, таких как консенсусы, весовые матрицы и т.д.

4. Nucleosome. Модуль предназначен для анализа ДНК-сайтов формирования нуклеосом, выявления их значимых контектсных и конформационных характеристик особенностей и построения методов их распознавания в нуклеотидных последовательностях геномной ДНК эукариот.

5. Система Leader_mRNA, предназначенная для анализа нуклеотидных последовательностей 5'-нетранслируемых последовательностей мРНК, выявления контектсных и структурных характеристик этих последовательностей, значимых для определения уровня их трансляционной активности, а также для создания методов предсказания трансляционной активности эукариотических мРНК по их нуклеотидным последовательностям.

Значимые результаты анализа регуляторных геномных последовательностей (РГП), полученные с помощью описанных выше программных модулей, накапливаются в соответствующих базах знаний. Каждая из этих баз знаний содержит формализованное описание значимых результатов анализа, а также программы распознавания РГП или оценки уровня их специфической активности, построенные на основе этих значимых характеристик.

Несмотря на то, что РГП, выполняющие различные функции, характеризуются принципиально различающейся структурно-функциональной организацией, для всех них является общей особенностью наличие двух типов характеристик, значимых для их функционирования и выполнения специфических функций и количественного уровня активности этих последовательностей облигатных и факультативных.

Облигатные характеристики обеспечивают базальный уровень активности РГП данного типа. Они одинаковы для всех вариантов РГП данного типа по их расположению и количеству. Облигатные характеристики абсолютно необходимы для выполнения специфических функций РГП при взаимодействии с регуляторными макромолекулами и регуляторными супрамолекулярными комплексами, такими как РНК-полимераза, сплайсома, рибосома и т.д.

Факультативные характеристики обеспечивают модуляцию уровня активности РГП по отношению к базальному уровню, определяемому облигатными характеристиками. Факультативные характеристики отличаются в пределах конкретной РГП количеством и взаимным расположением. Именно уникальный набор факультативных характеристик, присутствующих в определенной РГП, а также их взаимное расположение и определяют специфический уровень активности конкретной РГП.

Ассоциации факультативных характеристик, взаимодействующих друг с другом, и с облигатными характеристиками функциональных сайтов, определяют их специфику, то есть то, какую функцию выполняет РГП, а также величину ее специфической активности.

3 ТЕОРИЯ ПОЛЕЗНОСТИ ДЛЯ ПРИНЯТИЯ РЕШЕНИЙ.

В математической основой многих алгоритмов, использующихся в перечисленных выше модулях продукции знаний для анализа данных и поиска закономерностей является теория полезности для принятия решений, представленная, например, в работах Фишберна. Этот подход показал свою эффективность для анализа информации о регуляторных геномных последовательностях и выявления контекстных, конформационных и физико-химических характеристик РГП, значимых для их функционирования [11].

К числу достоинств этого метода относятся следующие:

- Возможность автоматического анализа данных при выявлении значимых структурно-функциональных характеристик РГП и построении методов их распознавания (автоматическая продукция знаний).
- Устойчивость получаемых результатов и высокая степень их воспроизводимости на контрольных выборках данных.
- Наличие статистической меры значимости выявляемых характеристик, которая называется полезностью и вычисляется в рамках теории полезности для принятия решений.
- Интерпретируемость получаемых результатов.

4 МАТЕМАТИЧЕСКИЕ МОДЕЛИ ПРОДУКЦИИ ЗНАНИЙ. ОБНАРУЖЕНИЕ ЗАКОНОМЕРНОСТЕЙ В ЛОГИКЕ ПЕРВОГО ПОРЯДКА

В общем виде информация об РГП представляется как совокупность реляционных таблиц и, следовательно, может быть представлена совокупностью отношений в языке логики первого порядка. В частности, с помощью 5-х таблиц можно, в достаточно общем виде, представлять информацию о локализации сигналов, содержащихся в любой группе РГП:

Таблица 1: *<Суперкласс РГП><Класс РГП>*

Используя данную таблицу можно задать иерархическую классификацию РПГ в виде отношения типа kind_of.

Таблица 2: *<Класс РГП><Имя РГП> <Характеристики РПГ>*

Данная таблица определяет принадлежность конкретного РПГ тому или иному классу. По сути, имя класса РПГ является только одной из характеристик РПГ, которые могут использоваться для формирования гипотез. В таблице предусмотрена возможность включения дополнительных характеристик РПГ. Если классификация РПГ не задана, то имеется возможность формального построения ее по характеристикам РПГ.

Таблица 3: *<Суперкласс сигнала ><Класс сигнала>*

В данной таблице задается иерархическая классификация сигналов.

Таблица 4: *<Класс сигнала><Имя сигнала> <Характеристики сигнала, независимые от РГП и позиций сигнала>*

В данной таблице задается принадлежность конкретного сигнала к тому или иному классу сигналов. Также предоставляется возможность задать дополнительные характеристики сигнала, которые не зависят от того, в каком РПГ и в какой позиции этот сигнал встречается.

Таблица 5: *<Имя РГП><Имя сигнала><Позиция сигнала><Характеристики особенностей сигнала для данной позиции и РГП>*

Эта таблица описывает конкретную последовательность. Число наблюдаемых сигналов и их позиция меняется в зависимости от последовательности.

Таким образом, например, в случае такого Суперкласса РГП, как промоторы эукариот в качестве класса РПГ может выступать группа промоторов генов, экспрессирующихся в определенной ткани или органе. Имя РГП - название конкретного промотора в базе данных TRRD из соответствующего класса, а в качестве сигнала могут рассматриваться сайты связывания транскрипционных факторов, либо другие контекстные, конформационные или структурные особенности, значимые для функционирования промоторов. Позиция сигнала обычно определяется относительно старта транскрипции гена. Безусловно, данный набор таблиц может быть дополнен информацией о процессах, в которых участвуют данные РПГ, регуляторных факторах и т.д.

Другим основным типом данных, используемым в базах данных, является числовое представление признака. Объекты в этом случае представляются наборами значений признаков. В [3,7] предлагается использовать *Теорию Измерений* для представления этого типа данных в языке логики первого порядка и тем самым в реляционном виде. В *Теории Измерений* показано, что числовые значения величин определяются отношениями. Следуя *Теории Измерений*, в [3,7] показано, как наиболее известные способы представления данных – таблицы объект-признак, матрицы упорядочений и близости, множественные и парные сравнения, могут быть представлены в языке первого порядка.

Методы KDD&DM работающие в языке логики первого порядка называются Реляционными DM методами [3]. Система Discovery, используемая нами для выявления комплексов (ассоциаций) признаков, значимых для функционирования РГП и автоматической продукции знаний по структурно-функциональной организации и распознаванию РГП, является реляционным DM методом.

Как показано в [3] реляционные DM методы позволяют снять практически все ограничения стандартных DM методов:

- сформулировать в языке первого порядка практически любое знание о предметной области (Background Knowledge) и использовать его для обучения;
- расширить понятие Data Type, за счет практически неограниченной выразительной возможности языка первого порядка;
- использовать *Теорию Измерений* для представления разнообразных (и необычных) величин в языке первого порядка (отношений: предпочтения, частичного порядка, решеток и т.д.; шкал: наименований, порядка, лог-линейных и т.д.; структур: древовидных, сетей, графов и т.д.; сме-

си всех этих величин, что особенно важно для систем связанных баз данных.);

- ввести понятие Rule Type как типа гипотез, которые могут проверяться в базах данных.

Поскольку возможности языка первого порядка в формировании гипотез также практически неограниченны, то необходимо определять тип высказываний проверяемых на данных. В качестве Rule Type могут быть сформулированы практически все типы гипотез проверяемые различными DM методами. Например, классы кусочно-линейных правил или m-of-n правил, используемых нейронными сетями; классы правил для любого типа деревьев; практически любой тип логических решающих правил; правила проверяемые в Inductive Logic Programming; булевы функции и т.д. В общем случае, разработана система все более точных типов гипотез (система вложенных Rule Types) позволяющая реализовать стратегию все более точного и детального анализа теории предметной области [10].

К реляционным DM методам относятся также методы Inductive Logic Programming (ILP), работающие в языке первого порядка. Но в отличие от этих методов система Discovery может обнаруживать вероятностные закономерности в языке первого порядка и работать с данными с высоким уровнем шумов, какими являются, например, финансовые данные [3].

Было проведено сравнение системы Discovery другими методами на примере двух задач [8]: диагностика раковых заболеваний [9] и предсказаний курсов акций ценных бумаг [3]. Сравнение производилось со следующими методами: 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). Во всех случаях система Discovery показала лучший результат. В среднем результаты системы Discovery превосходили результаты других методов в 1.5 раза. Более подробно результаты сравнения представлены на WebSite “Scientific Discovery” в разделе “comparison”: http://jnt.novosoft.ru/~vityaev.

Приведем примеры представления различных Data Types в языке первого порядка и примеры Rule Types.

Возьмем матрицы упорядочений: (r_{ij}), i = 1,…,m; j = 1,…,n; где r_{ij} - оценка i-го объекта по j-му признаку. Такие матрицы могут выражать упорядочение k объектов n экспертами или упорядочение k объектов n ранговыми признаками. Такие матрицы обрабатываются методами многомерного шкалирования или методами ранжирования. Поставим в соответствие каждому признаку j отношение P_j определенное следующим образом:

$$P_j(a_{i1},a_{i2}) \Leftrightarrow r_{i1j} \le r_{i2j} . \quad (1)$$

Получим совокупность бинарных отношений $\{P_1,\ldots,P_n\}$. Пусть $A = \{a_1,\ldots,a_m\}$ - множество объектов, на которых получена матрица упорядочений. Тогда представлением матрицы упорядочений на множестве A будет модель $pr^V = \langle A; P_1,\ldots,P_n\rangle$ (как она определяется в математической логике).

Рассмотрим матрицы близости. Матрицей близости для множества объектов $A = \{a_1,\ldots,a_m\}$ называется матрица (r_{ij}), i,j = 1,…,m; r_{ij} - числовые оценки меры близости (сходства, различия) в порядковой шкале (имеет смысл только сравнение величин $r_{i1j1} < r_{i2j2}$). Такие матрицы возникают в различных областях при сравнении или оценке экспертом двух объектов в некотором отношении. Матрицы близости обрабатываются методами многомерного неметрического шкалирования. Определим на множестве A отношение:

$$P(a_{i1},a_{i2},a_{i3},a_{i4}) \Leftrightarrow r_{i1i2} < r_{i3i4}. \quad (2)$$

Так как это отношение определено на всем множестве A, то представлением матрицы близости является модель $prV = \langle A;P\rangle$.

Представим структурные величины, например, графы. Если A = {a1,…,am} – множество вершин графа, а X = {<ai,aj>} – множество его ребер, то граф определяется как модель <A, P>,

где отношение $P(ai,aj) \Leftrightarrow \langle ai,aj\rangle \in X$.

Рассмотрим пример Rule Type. Сформулируем свойство монотонности

$$\forall a,b(F(a)\&F(b)\&(a \le_1 b)^{\varepsilon 1}\&\ldots\&(a \le_m b)^{\varepsilon m} \Rightarrow (a \le_0 b)^{\varepsilon 0}) \quad (3)$$

где $\le_1,\ldots, \le_m, \le_0$ – отношения порядка (возможно, определенные для различных признаков);

$\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_m \in \{0,1\}$ – значения истинности отношений, $(a \le b)^1 = (a \le b)$, $(a \le b)^0 = (a > b)$. Формула F(a) определяет некоторый участок (область) в обобщенном признаковом пространстве относительно переменной по объектам **a**. Свойство монотонности довольно часто встречается и достаточно полезно, однако, уже оно выходит за рамки правил, обнаруживаемых существующими методами. Далее будут приведены более сложные примеры типов правил.

В [4] показано, что система Discovery способна обнаружить любые закономерности в языке первого порядка, имеющие максимальные оценки условной вероятности. Поэтому система Discovery способна решать задачу извлечения знаний из GeneExpress.

Система Discovery генерирует гипотезы в виде некоторого параметрического семейства формул типа:

$$A_1\&\ldots\&A_n \Rightarrow A_0 \quad (4)$$

где $A_0,A_1,\ldots,A_n$ - логические выражения (включающие логические связки AND, OR, NOT, скобки и произвольные арифметические выражения с параметрами). Параметрами могут быть номера признаков, интервалы изменения признаков, выделенные значения признаков, параметры, модифицирующие признак (подвергающие его различным преобразованиям) и т.д. Система позволяет реализовать перебор гипотез с помощью определенной стратегии, представляющей собой семантиче-

ский вероятностный вывод [4]. Уточнения гипотез осуществляются путем добавления новых условий в посылку, либо применением подстановок.

Описанные подходы были апробированы при решении различных задач. Например, для обнаружения закономерностей описания различных типов промоторов эукариот по вырожденным олигонуклеотидным мотивам проверялись следующие гипотезы:

$\forall a \exists p_1,p_2,\ldots,p_i$ $((Pos(p_1) < Pos(p_2))\&(Pos(p_2) < Pos(p_3))\&\ldots\&(Pos(p_{i-1})<Pos(p_i))\&(Sign(p_1)=s_1)\&(Sign(p_2) = s_2)\&\ldots\& (Sign(p_i) = s_i) \Rightarrow (Class(a) = cl_j)$), (5)

где a – последовательность олигонуклеотидов; $p_1,p_2,\ldots,p_i$ – номера олигонуклеотида; $Pos(p_j)$ - номер позиции олигонуклеотида p_i в последовательности a, j = 1,…,i; $Sign(p_j)$ – знак олигонуклеотида; $s_1,\ldots,s_i \in \{+,-\}$ знак означает, что олигонуклеотид расположен в прямом (+) или обратном (-) порядке; Class(a) – номер класса последовательности a.

В результате работы программы обнаружено большое число закономерностей и, в частности, обнаружена закономерность :

$\forall a \exists p_1,p_2,\ldots,p_{10}$ $((Pos(p_1) < Pos(p_2))\&(Pos(p_2) < Pos(p_3))\&\ldots\&(Pos(p_9)<Pos(p_{10}))\&(Sign(p_1)=s_1)\&(Sign(p_2) = s_2)\&\ldots\&(Sign(p_i)=s_i)\Rightarrow (Class(a) = clj)$) (6)

$p_1=9$, $p_2=27$, $p_3=3$, $p_4=4$, $p_5=2$, $p_6=1$, $p_7=3$, $p_8=2$, $p_9=11$, $p_{10}=6$;

$s_1= +$, $s_2= -$, $s_3= -$, $s_4= -$, $s_5= -$, $s_6= -$, $s_7= +$, $s_8= +$, $s_9= +$, $s_{10}= +$; Clj = 1;

Это означает, что промоторы, в которых олигонуклеотиды расположены в последовательности $9 < 27 < 3 < 4 < 2 < 1 < 3 < 2 < 11 < 6$ и определенным образом ориентированы (прямо или обратно), относятся к классу 1.

Другая задача, на которой апробировались данные подходы, связана с распознаванием функциональных групп промоторов, в частности, промоторов генов, участвующих в регуляции липидного метаболизма, процесса дифференцировки эритроидных генов и интерферон - регулирующих генов, описанных в базе данных TRRD.

Кроме экспериментально выявленных регуляторных сайтов, в промоторах, представленных в базе данных TRRD, для их описания использовались потенциальные сайты, выявленные методом весовых матриц.

Первичные последовательности промоторов выбирались из базы EMBL. Использовались весовые матрицы, определенные в базе данных Transfac. В качестве класса "нет" использовались случайные выборки последовательностей, полученные с помощью системы Samples, входящей в GeneExpress.

5 ПРОДУКЦИЯ ЗНАНИЙ ПО СТРУКТУРНО-ФУНКЦИОНАЛЬНОЙ ОРГАНИЗАЦИИ БЕЛКОВ.

Третичные структуры белков задаются совокупностью данных по координатам атомов молекулы, их типу, сетью ковалентных связей между атомами, а также типу аминокислотных остатков и других молекул, входящих в состав белка. На основе этих данных могут быть рассчитаны многие структурные и физико-химические характеристики белков, необходимые для решения задач по их структурно-функциональной организации.

Одним из важных классов задач структурно-функциональной организации белков является предсказание связывания белков с определенным типом молекул. При этом большой интерес для исследователей могут представлять взаимодействия белков с самым разнообразным типом молекул включая другие белки, пептиды, ДНК, РНК, сахара, ионы металлов и т.д. Данный класс задач включает в себя а) задачи докинга, или, другими словами, поиск участков связывания двух молекул на основе анализа третичных структур каждой из взаимодействующих молекул, б) задачи предсказания потенциальных сайтов связывания на основе анализа общих особенностей таких сайтов лишь одной взаимодействующей молекулы. Для решения задач докинга необходимо представление поверхности молекул в виде набора характеристик, позволяющих проводить оценки комплиментарности при сопоставлении различных фрагментов поверхности одной молекулы фрагментам другой молекулы. Для осуществления связывания имеют важное значение геометрическое соответствие участков поверхностей подобно ключ-замок, водородные связи между определенными группами атомов, а также электростатические и гидрофобные взаимодействия контактных областей. Кроме этого, определенное влияние на связывание могут оказывать эффекты индуцированной комплиментарности контактных областей, т.е. локальное изменение рельефа поверхности молекулы, вызванное контактом с другой молекулой. Эффекты индуцированной комплиментарности могут быть особенно существенными в случаях, когда в контактную область входят гибкие петлевые участки белка. Набор решающих правил, использующих эти характеристики и многие другие может быть получен при анализе пространственных структур комплексов белков с лигандами.

Задачи второго типа направлены на поиск потенциальных сайтов связывания не с конкретной молекулой, а с обобщенным образом молекул заданного класса. Например, в случае изучения антигенной структуры белка интерес представляют сайты, которые являются наиболее вероятными эпитопами, по отношению к антителам. В данном случае, задача сводится к определению общих закономерностей пространственной организации сайтов связывания белков с определенным классом молекул-лигандов.

Рассмотрим для примера простейшую схему решения задачи второго типа с помощью системы Discovery. Пусть будет сформирована выборка пространственных

структур сайтов связывания белков с неким классом молекул-лигандов. Пусть для некоторой пространственной структуры сайта из этой выборки String элементы этой пространственной структуры составляют множество A(String) = $\{a_1,a_2,a_3,a_4,\ldots\}$. Зададим связи пространственной структуры двуместными предикатами $P_1(a,b)$, $P_2(a,b)$, $P_3(a,b)$,…, где первый предикат будет задавать валентные связи, второй – водородные, третий – солевые мостики и т.д. В пространственной структуре кроме предикатов должны быть определены еще расстояния между элементами ρ(a,b). Автоматический поиск таких сайтов в пространственных структурах других белков может быть тогда осуществлен следующим типом правил:

$\forall String \exists a_1,a_2,a_3,a_4[P_1(a_{i1},a_{i2})^{\varepsilon 11}\&\ldots\&P_1(a_{ik-1},a_{in})^{\varepsilon 1n}$
$\&P_2(a_{j1},a_{j2})^{\varepsilon 21}\&\ldots\&P_2(a_{jl-1},a_{jl})^{\varepsilon 2l}\&P_3(a_{k1},a_{k2})^{\varepsilon 31}\&$
$\&P_3(a_{km-1},a_{km})^{\varepsilon 3l}\&(par_1 < \rho(a_{t1},a_{t2}) < par_2)\&\ldots\&(par_s < \rho(a_{tu},a_{tu+1}) < par2s) \Rightarrow Activity(String)]$

Следует заметить, что анализ решающих правил открывает перспективы в решении другого класса важных задач, связанных с изучением закономерностей, лежащих в основе избирательного распознавания белками молекул лигандов. Например, для белков существует проблема выделения из совокупности признаков сайтов облигатных и факультативных признаков. Факультативные признаки определяют специфичность связывания внутри класса молекул лигандов, а облигатные между классами. Очевидно, что правила докинга опираются как на облигатные, так и на факультативные признаки, а правила решающие задачу второго типа только на облигатные. Таким образом, ответ на данный вопрос для молекул-лигандов заданного класса может быть получен при сопоставлении этих правил.

Для анализа структурно-функциональной организации белков разрабатывается особый модуль системы GeneExpress - **FASTProt (F**unction**, A**ctivity and **ST**ructure of **Prot**eins) [12].

5.1 Базы данных системы FASTPROT.

(1) База данных EnPDB по пространственным структурам ДНК, РНК и белков, созданная на основе базы данных PDB. При ее создании был выбран формат представления данных, обеспечивающий их наиболее полное индексирование средствами SRS, и интернет-интеграцию с другими базами данных по молекулярной биологии (TRRD, SwissProt, EMBL и др.). Также были введены новые поля для описания структурно-функциональной организации белков, ранее отсутствовавшие в PDB.

(2) База данных по фаговому дисплею (ASPDB). Эта база данных создается на основе экспериментальных данных по технологии «фагового дисплея», позволяющей выявлять аминокислотные последовательности ДНК-связывающих доменов транскрипционных факторов, взаимодействующих с различными ДНК-сайтами.

(3) Компиляции амонокислотных последовательностей ДНК-связывающих доменов природных белков.

(4) Другие базы данных, содержащие информацию по структурно-функциональной организации белков.

5.2 Программы анализа данных и продукции знаний системы FASTPROT.

(1) Программы построения дочерних баз данных, которые предназначены для накопления результатов анализа пространственных структур белков, требующих для своего получения больших вычислительных затрат. К их числу относится база данных PDBSite, содержащая информацию об особенностях пространственной организации биологически активных сайтов в белках. Кроме аминокислот, непосредственно входящих в сайты, в PDBSite содержатся характеристики окружающих их аминокислот. Пространственное расположение аминокислот представлено координатами центров масс аминокислот и Cα-атомов. Вычисляются взаимные расстояния между аминокислотами, входящими в активные центры, а также значения физико-химических свойств сайтов (гидрофобность, заряженность, доступность аминокислот для растворителя и т.д.).

(2) Программы распознавания активных сайтов в пространственной структуре белков. Конформация сайта и его функциональная активность во многом определяется его окружением в третичной структуре, что может быть важным для распознавания сайтов. Разрабатываются методы поиска сайтов на основе трехмерного сходства между шаблонами сайтов, описанными в базе данных PDBSite и участками пространственной структуры белков, представленных в EnPDB. Однако, вариабельность аминокислот в сайтах не всегда позволяет напрямую использовать координаты атомов в качестве шаблонов для поиска сайтов в третичной структуре белков. Поэтому разрабатываются программы для поиска сайтов в третичной структуре белков на основе совокупности методов распознавания образов и классификации, нейронных сетей, дискриминантного анализа.

(3)Программы автоматического конструирования методов распознавания структурно-функциональных доменов (СФД) белков основе анализа их аминокислотных последовательностей. СФД - участки аминокислотной последовательности длиной до несколько десятков аминокислот, которые имеют определённую первичную, вторичную и трехмерную структуру и несут определённую функцию. Примерами таких СФД могут быть сайты фосфорилирования, сайты гликозилирования, сайты протеолиза, сайты связывания металлов и лигандов, сайты, обеспечивающие различные типы ферментативной активности, ДНК-связывающие домены, сиг-

нальные пептиды, трансмембранные сегменты, антигенные детерминанты и т.д. Для распознавания СФД используется совокупность методов: весовые матрицы, дискриминантный анализ, нейронные сети, регулярные выражения, скрытые Марковские модели и т.д. Они аналогичны тем, которые применяются нами для построения методов распознавания сайтов в нуклеотидных последовательностях, описанным выше.

(4) Программы предсказания активности функциональных сайтов белков по их аминокислотным последовательностям. Они предназначены для количественного анализа взаимосвязи между структурой и активностью белков; выявления облигатных остатков, абсолютно необходимых для активности сайтов; выявления групп факультативных остатков, обеспечивающих модуляцию уровня активности сайтов относительно их базального уровня, построения методов для предсказания количественного уровня активности сайтов по их аминокислотным последовательностям.

6 ЗАКЛЮЧЕНИЕ.

Описанные в данной работе методы анализа данных и средства поиска, обнаружения и извлечения знаний активно используются для исследования структурно-функциональных особенностей ДНК, РНК и белков. Эти методы позволяют осуществлять все этапы итерационного процесса KDD&DM по анализу и обогащению знаний, накапливаемых в электронной библиотеке ГенЭкспресс. Более того, применяемые подходы являются инструментальными, т.е. позволяют генерировать процедуры для решения не только отдельных задач, но и целых классов задач. Важнейшим направлением дальнейшей работы будет приложения изложенных выше подходов к поиску структурно-функциональных закономерностей в генных сетях, закономерностей динамики генных сетей, построению согласованных оценок скорости элементарных реакций в определенных условиях по косвенным данным и т.д.

В частности, анализ закономерностей переходов модели генной сети из одного состояния (патология) в другое (норма) могут оказаться весьма важным для поиска эффективного лечения т.е. соответсвующих управляющих воздействий на организм, например, пищевых добавок [13].

Список литературы

[1] *Колпаков Ф.А., Подколодный Н.Л., Лаврюшев С.В., Григорович Д.А., Пономаренко М.П., Колчанов Н.А.* Методы интеграции неоднородных информационных ресурсов по регуляции генной экспрессии в электронной библиотеке GeneExpress. // Программирование. 2000.N3.

[2] Колчанов Н.А., Лаврюшев С.В., Григорович Д.А., Пономаренко М.П., Фролов А.С., Подколодный Н.Л., Колпаков Ф.А., Пономаренко Ю.В., Кочетов А.В., Ананько Е.А., Подколодная О.А., Игнатьева Е.В. (1999) ГЕНЭКСПРЕСС: электронная библиотека по структурам и функциям ДНК, РНК и белков. Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Санкт-Петербург, 19-22 окт. 1999 г, 161-169.

[3] Kovalerchuk B., Vityaev E. Data Mining in finance: Advances in Relational and Hybrid Methods. Kluwer Academic Publishers, 2000, p.308

[4] Витяев Е.Е. Семантический подход к созданию баз знаний. Семантический вероятностный вывод наилучших для предсказания ПРОЛОГ-программ по вероятностной модели данных. // Логика и семантическое программирование (Выч., сист. 146), Новосибирск, 1992, с.19-49.

[5] Krantz D.H., Luce R.D., Suppes P., Tversky A. Foundations of measurement. Vol. 1,2,3 - NY, London: Acad. press, 1971, 1989, 1990.

[6] *Kolchanov N.A., Lim H.A. Computer analysis of Genetic Macromolecules. World Scientific, 1994, p.556.*

[7] Витяев Е.Е. Обнаружение закономерностей (методологоя, метод, программная система SINTEZ). 1. Методология // Методологические щроблемы науки (Вычислительные системы, 138), Новосибирск, 1991, с. 26-60.

[8] Kovalerchuk B, Vityaev E. Comparison of relational methods, attribute-based methods and hybrid methods, *Proc. 14th IEEE International Symposium on Intelligent Control /Intelligent Systems and Semiotics ISIC/ISAS'99, September 15-17*, 1999, Cambridge, Massachusetts, USA.

[9] B. Kovalerchuk, E. Vityaev, J. Ruiz. Consistent knowledge discovery in medical diagnosis. Special issue of the journal: IEEE Engineering in Medicine and Biology Magazine: “Medical Data Mining”, July/August 2000.

[10] Витяев Е.Е., Москвитин А.А. Введение в теорию открытий. Программная система DISCOVERY. // Логические методы в информатике (Вычислительные системы, вып. 148), Новосибирск, 1993, с.117-163.

[11] Фишберн П.С. Теория полезности для принятия решений. М., Наука, 1978. - 352с.

[12] Иванисенко В.А., Григорович Д.А., Афонников Д.А., Куропатов Д.А., Валуев В.П., Колчанов Н.А. Информационная система FrameProt по пространственным структурам ДНК, РНК и белков в составе GeneExpress // Первая Всероссийская научная конференция Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. 19-21 октября 1999г., Санкт-Петербург. С.175-186.

[13] Ананько Е.А., Лихошвай В.А., Колпаков Ф.А., Подколодный Н.Л., Ратушный А.В., Игнатьева Е.В., Подколодная О.А., Степаненко И.Л., Колчанов Н.А. Электронная библиотека GeneNet: описание и моделирование генных сетей животных и растений // // Вторая Всероссийская конференция по электронным библиотекам. 26-28 сентября 2000 года, Протвино. (Настоящий выпуск).