ГЕНЭКСПРЕСС: электронная библиотека по структурам и функциям ДНК, РНК и белков

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.
Институт Цитологии и Генетики СО РАН
630090, Новосибирск, пр. акад. Лаврентьева, 10

1 Введение

Бурное развитие молекулярной биологии и генетики в последние десятилетия привело к появлению огромного количества экспериментальных данных по структуре и функции ДНК, РНК и белков. Решение задач молекулярной биологии и генетики, биотехнологии, генетической и белковой инженерии требует использования разнообразной информации о первичной и пространственной структуре этих макромолекул, распределенной по большому количеству баз данных. Для упорядочения и накопления этой информации создано не менее 500 специализированных баз данных, большинство из которых доступны по Интернет [1].

Для анализа информации, накопленной в этих базах данных разработано огромное количество программ, а так же десятки систем, обеспечивающих доступ к базам данных, навигацию по ним, и графическое представление имеющихся данных. Распределенность информации по базам данных и ее представление в различных форматах делают проблему Интернет-интеграции весьма сложной, как в концептуальном, так и техническом аспектах. Сложной и нерешенной в настоящее время является проблема автоматической продукции молекулярно-биологических и молекулярно-генетических знаний на основе компьютерного анализа информации, накапливаемой в базах данных. Любые интегрированные информационные ресурсы по молекулярной биологии и генетике должны обеспечивать широкий набор средств для работы пользователей, в первую очередь — эффективную систему для выполнения сложных запросов и поиска информации по большому количеству распределенных ресурсов; возможность осуществления сложных сценариев анализа, требующих использования большого количества различных баз данных и программ; возможность хранения значимых результатов анализа в соответствующих базах знаний и т. д. Очень важным является создание средств для эффективной навигации по интегрированным Интернет-ресурсам. Для решения этих задач нами разрабатывается электронная библиотека ГенЭкспресс по пространственным структурам и функциям ДНК, РНК и белков [10, 3] описание которой дается в настоящей статье.

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

2 Ресурсы, интегрированные в рамках электронной библиотеки ГенЭкспресс

В рамках ГенЭкспресс интегрировано большое количество баз данных, баз знаний и сотни программ для обработки информации по структуре и функции ДНК, РНК и белков (рис. 1).

2.1 Базы данных

ГенЭкспресс включает в себя большое количество внутренних информационных ресурсов — баз данных по структуре, функции и эволюции ДНК, РНК и белков, разработанных в Лаборатории теоретической генетики ИЦиГ СО РАН.

2.1.1 Базы данных по структуре и функции ДНК и РНК

База данных Transcription Regulatory Regions Database (TRRD) содержит информацию о строении и функционировании районов геномной ДНК, обеспечивающих регуляцию транскрипции генов эукариот [9]. Данные вносятся в базу на основании аннотирования научных статей. TRRD содержит информацию о следующих регуляторных элементах: 1) сайтах связывания транскрипционных факторов, 2) композиционных элементах; 3) промоторах, 4) энхансерах и сайленсерах; 5) транскрипционных регуляторных районах. Один вход в базу данных соответствует описанию регуляторных районов одного гена. В настоящее время в базе содержится описание 689 генов, 984 регуляторных районов (промоторов, энхансеров и сайленсеров), и 3335 сайтов связывания. Эта информация получена на основании реферирования более 2311 научных статей.

SAMPLES, база данных регуляторных геномных последовательностей. Она содержит информацию о выборках последовательностей сайтов связывания транскрипционных факторов и функциональных районов других типов (сайты связывания нуклеосом, 5' нетранслируемые районы мрнк эукариот, промоторы генов эукариот и т. д.). База данных формируется на основе баз данных TRANSFAC, TRRD и EMBL, а также литературных данных.

База данных по активности функциональных сайтов ДНК и РНК. Она содержит информацию о нуклеотидных последовательностях сайтов различных типов с ука-

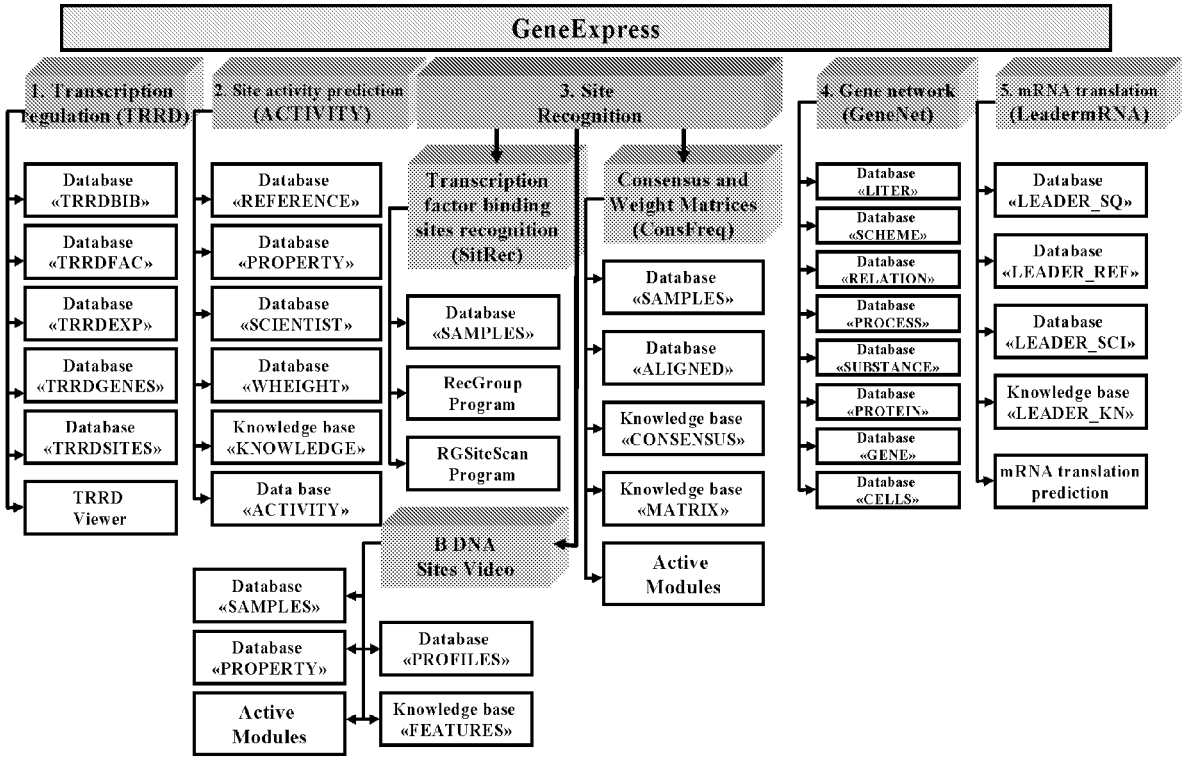

Рис. 1: Основные ресурсы, представленные в электронной библиотеке GeneExpress

занием количественных величин их специфической активности. База данных содержит описание более 700 экспериментов на основе 225 научных статей. Один документ базы описывает один эксперимент. База содержит данные о промоторах различных генов, сайтах связывания транскрипционных факторов, 5'-мрнк районах и многих других функциональных сайтов в ДНК и РНК про- и эукариотических геномов.

База данных GENENET. Она содержит информацию о молекулярных механизмах функционирования генных сетей — координировано функционирующие групп генов, обеспечивающих выполнение жизненно важных процессов организмов. Genenet содержит информацию о генных сетях, регулирующих противовирусный ответ, липидный метаболизм, дифференцировку эритроцитов, азотфиксацию у бобовых, регуляцию функции генов теплового шока, накопление запасающих белков в процессе созревания семян и т. д. [11].

База данных PROPERTY. Эта база данных содержит информацию о контекстно-зависимых конформационных и физико-химических свойствах двойной спирали ДНК. В текущей версии она содержит описание 38 таких свойств.

База данных по спонтанным мутациям содержит информацию о более чем 8000 нуклеотидных замен, произошедших в геномах позвоночных.

2.1.2 Базы данных по структуре и функции белков

Основной среди них является база данных EnPDB, содержащая информацию о протранственной структуре белков. Она создана в Лаборатории теоретической генетики ИЦиГ СО РАН на основе базы данных PDB. Описание EnPDB и других информационных ресурсов по структуре и функции белков, входящих в состав ГенЭкспресс приведено в этом же выпуске докладов [4].

Внешние информационные ресурсы электронной библиотеки ГенЭкспресс. К ним относятся доступные по Интернет базы данных по молекулярной биологии и генетике такие как EMBL, EPD, EpoDB, MEDLINE, TRANSFAC, SWISSPROT, COMPEL и т. д.

2.2 Системы продукции знаний и базы знаний

Создание баз знаний по молекулярной биологии и генетике имеет особенно важное значение для успешного анализа огромных объемов данных, получаемых в ходе экспериментального исследования ГЕНОМа ЧЕЛОВЕКА, важнейшей задачей которого является расшифровка нуклеотидных последовательностей геномной ДНК с ее последующим компьютерным анализом. Целью этого анализа является предсказание неизвестных генов, регуляторных районов и функциональных сайтов и т. д.

В рамках ГенЭкспресс разрабатываются технологии автоматической и интерактивной продукции знаний о структуре и функции ДНК и РНК, в том числе — генерации кодов программ для распознавания. Наиболее важ-

ным видом знаний являются программы для распознавания регуляторных районов, контролирующих функцию генов и предсказания их свойств. Знания записываются в соответствующие базы знаний. В ГенЭкспресс входят такие системы продукции знаний, как B- DNA VIDEO, ACTIVITY И CONSFREQ, LEADER MRNA и некоторые другие. B- DNA VIDEO обеспечивает продукцию знаний о конформационных и физико-химических характеристиках ДНК-сайтов, значимых для их функционирования и распознавания. ACTIVITY обеспечивает продукцию знаний о контекстных, конформационных и физико-химических особенностях ДНК и РНК-сайтов, значимых для предсказания их активности. CONSFREQ предназначена для продукции и использования знаний о контекстных характеристиках сайтов, значимых для их распознавания. Полученные знания хранятся в одноименных базах знаний, являющихся составными частями электронной библиотеки ГенЭкспресс. Детальное описание указанных выше систем продукции знаний можно найти в специальных выпусках журнала Bioinformatics за 1999 год [5, 8, 10, 12, 14–17].

В качестве примера рассмотрим базу знаний B-DNA-FEATURES. На рис. 2 приведена карточка, соответствующая функциональному сайту HNF1. Имя сайта (HNF1) указывается в поле NM. Поле DR содержит ссылку на базу данных SAMPLES, содержащую последовательности этого сайта. Поле PV содержит имя наиболее значимого sдля распознавания этого сайта физико-химического свойства ДНК, выявленного системой продукции знаний — температуры плавления ("Melting temperature"). Поле DP содержит ссылку на базу данных PROPERTY, в которой содержится описание зависимости величины параметра "Melting temperature" от динуклеотидного контекста. Согласно информации, записанной в этой карточке, HNF1 сайт в участке $[-21; 4]$ статистически значимо отличается по средней величине температыры плавления ДНК ("Melting temperature" property) от случайных последовательностей. Полезность этой характеристики для распознавания сайта HNF1 равна U=0.867 (поле UT). Средняя величина этого температуры плавления "Melting temperature" в пределах участка $[-21; +4]$ реальных сайтов равна $70.68 \pm 3.84°C$ (поле ST). Для случайных последовательностей эта величина составляет $73.55 \pm 4.61°C$ (поле NT). Значимость различий между сайтами и случайными последовательностями по указанному свойству, определнная по критерию $\chi 2$ составляет $\alpha < 0.005$. Это иллюстрируется соответствующим графиком, ссылка на который находится в поле WW GALERY. Кроме того, имеется поле WW PROGRAM, содержащая ссылку на программу для распознавания сайта HNF1 в произвольной нуклеотидной последовательности, построенную на основе значимого свойства "Melting temperature". В карточке содержится также с-код этой программы.

В текущей версии ГенЭкспресс содержатся следующие базы знаний:

KNOWLEDGE — 49 знаний-программ для предсказания активности функциональных сайтов по их нуклеотидным последовательностям;

ALIGN — 45 знаний-выборок выравненных последовательностей сайтов связывания транскрипционных факторов;

FEATURES — 1402 знаний-программ распознавания сайтов по функционально важным конформационным и физико-химическим свойствам ДНК сайтов;

MATRIX — 567 знаний-программ распознавания сайтов по частотам коротких "слов"-олигонуклеотидов;

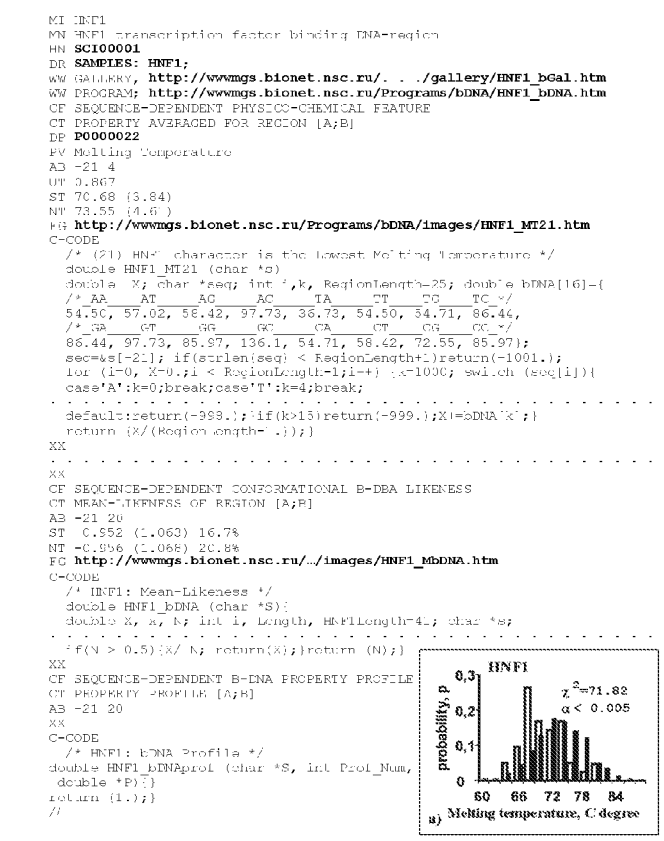

Рис. 2: База знаний FEATURES (карточка)

CONSENSUS — 66 знаний-программ распознавания сайтов по эволюционно-консервативным инвариантам этих сайтов.

Место Баз Знаний в структуре электронной библиотеки GeneExpress схематически показано на рис. 3.

Посредством Баз Знаний осуществляется взаимосвязь между экспериментальными данными и компьютерными программами, предназначенными для анализа геномной ДНК. В Базах знаний помимо текстов программ документируются закономерности, выявленные при анализе экспериментальных данных. Взаимосвязи между компьютерными программами и экспериментальными данными, осуществляемые посредством Баз Знаний, обеспечивают следующие возможности для решения молекулярно-генетических задач.

Результаты работы каждой программы, хранящейся в базе знаний, могут быть объяснены пользователю (рис. 3, стрелки вниз) путем представления тех экспериментальных данных, на основе которых была осуществлена генерации соответствущей программы и тех закономерностей, коорые были выявлены при анализе экспериментальных данных.

Из сотен компьютерных программ, документированных в Базе Знаний пользователь с помощью стандартных информационно-поисковых средств может быстро найти именно те программы, которые необходимы ему для конкретного исследования.

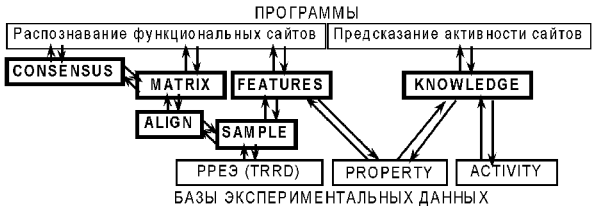

Рис. 3: Базы Знаний по молекулярной биологии и генетике в электронной библиотеке GeneExpress [1]. Стрелки: вверх, процесс анализа данных; вниз, процесс интерпретации результатов.

Например, пользователь, который желает распознать сайт «TATA-бокс» в исследуемой последовательности ДНК, может (рис. 4): войти в базу знаний FEATURES, содержащую информацию о методах распознавания сайтов на основе конформационных и физико-химических свойств ДНК; найти карточку, соответствующую TATA-боксу и вызвать из поля WW PROGRAM этой карточки соответсвующую программу распознавания (рис. 4 а). На рис. 4 б представлен пользовательский интерфейс вызванной программы распознавания, а на рис. 4 в — результат работы этой программы при анализе нуклеотидной последовательности.

3 Технологии интеграции разнородных биологических данных, используемые в GeneExpress

3.1 SRS — интеграция баз данных на основе перекрестных гиперссылок

Большинство баз данных по молекулярной биологии и генетике являются набором из нескольких текстовых файлов. Основной системой для работы с такими базами данных является система SRS — Sequence Retrieval System [7]. Основой SRS является объектно-ориентированный язык Icarus на котором описываются структура и синтаксис данных. Средства индексирования данных позволяют осуществлять быстрый поиск информации. SRS имеет специальный язык запросов в виде логических условий любой сложности, который обеспечивает поиск и доступ к базам данных через Internet. Результат запроса выдается в виде HTML файла, причем SRS имеет гибкие средства преобразования данных при отображении их виде HTML документа и встроенные стандартные форматы представления молекулярно-генетических данных.

С учетом этих положительных качеств SRS используется как одно из основных средств интеграции в системе ГенЭкспресс. Для того, чтобы можно было собрать воедино информацию о заданном молекулярно-биологическом объекте, описанном в системе GeneExpress, многие из баз данных интегрированы друг с другом на уровне перекрестных ссылок. Эти ссылки поддерживаются системой SRS, в виде гипертекстовых ссылок, содержащими SRS запрос, чтобы получить требуемый вход из другой базы данных. Таким образом, пользователь получает возможность легкой навигации через Internet по связанным между собой входами (см. пример на рис. 5) из разных баз данных.

Однако, интеграция на основе SRS не обеспечивает возможности решения всех задач, возникающих у молекулярного биолога. SRS лишена наглядности, не обеспечивает интеграция баз данных на основе семантического анализа; не позволяет составлять выборки нуклеотиднх последовательностей в автоматическом режиме, в SRS отсутствет возможность графического представления данных. С учетом этого нами разрабатывается новый подход к интеграции биологических баз данных на основе специализированног объектно-ориентированног языка запросов MGQL.

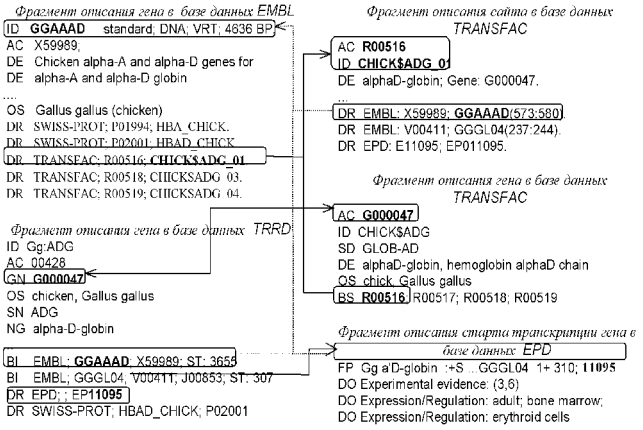

Рис. 5: Пример интеграции баз данных TRRD, EMBL, TRANSFAC и EPD на уровне перекрестных ссылок.

3.2 MGCL, MGQL — интеграция баз данных на основе объектно-ориентированного представления содержащейся в них информации

В данной разделе рассмотрим этот подход на примере интеграции 6 взаимосвязанных баз данных по регуляции генной экспрессии: EMBL, TRRD, TRANSFAC, COMPEL, EPD, SWISS-PROT и GeneNet (таблица 1), входящих в состав электронной библиотеки GeneExpress.

3.2.1 MGCL (Molecular genetics class library) — объектно-ориентированное представление информации из баз данных

В рамках предлагаемого нами подхода объектно-ориентированное представление текстовой информации из баз данных состоит из 5 этапов:

1. Выделение основных предметно-ориентированных понятий, описываемых в каждой базе данных.

2. Сопоставление каждому понятию отдельного типа.

3. Определение минимально необходимого набора атрибутов для каждого типа, так чтобы было возможно адекватно представить всю информацию для соответствующего этому типу понятия во всех используемых базах данных.

4. Упорядочивание типов в виде некоторой иерархии, что позволяет существенно упростить описания многих типов, определив общие для них атрибуты и методы в некотором базовом (родительском) типе.

5. Написание драйвера для каждой базы данных, осуществляющего перевод информации в текстовом виде из базы данных в набор связанных между собой объектов.

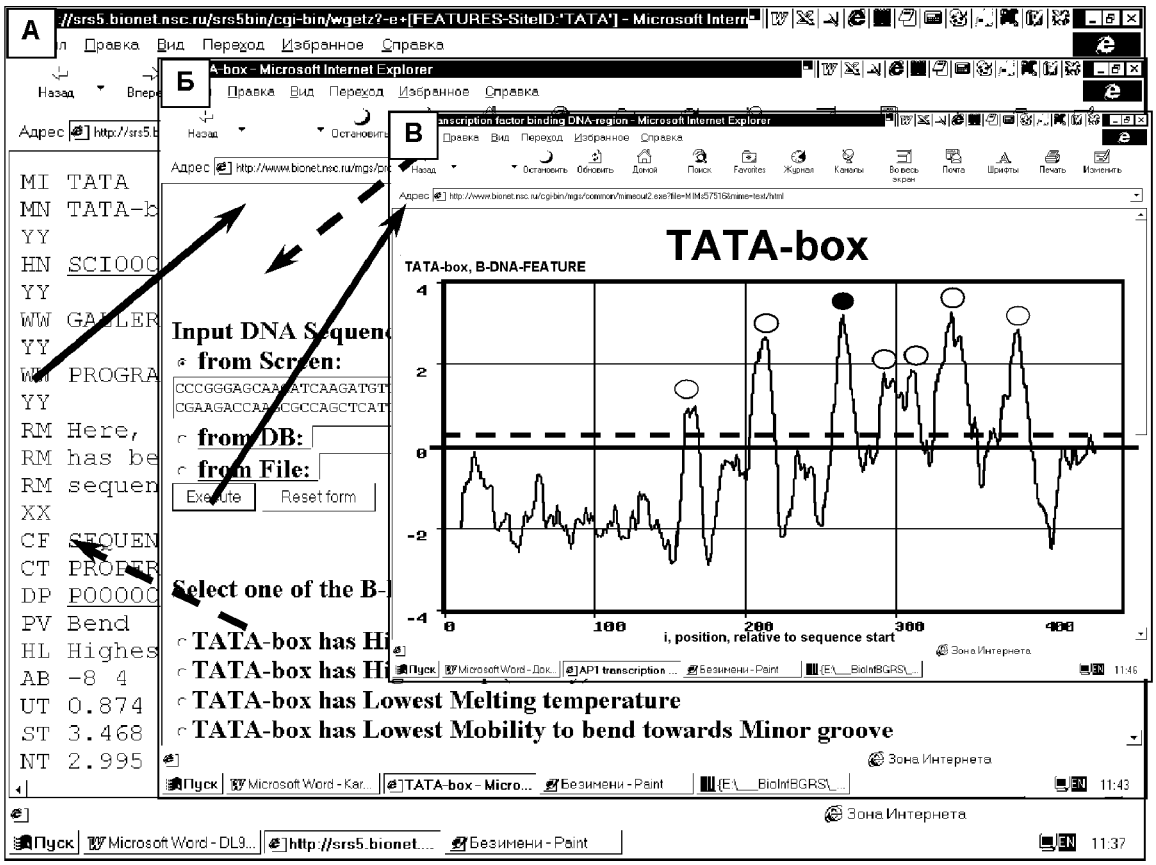

Рис. 4: База знаний FEATURES по конформационным и физико-химическим свойствам сайтов ДНК: а) документ с описанием свойств TATA бокса; б) программа распознавания TATA боксов по этим свойствам; в) результат этой программы для промотора гена CYC1 изо-1-цитохрома С дрожжей (EMBL: X03472). Кружки: темный, известный TATA бокс; жирные, документированные TATA-подобные боксы; тонкие — ошибки II рода (перепредсказание). Стрелки: непрерывные, предсказание; разрывные, интерпретация результата. Пунктир, 95 %-граница сходства с TATA боксом.

Рассмотрим каждый из этих этапов в применении к базам данных EMBL, TRRD, TRANSFAC, COMPEL, EPD, SWISS-PROT и GeneNet. В таблице 1 приведен список основных молекулярно-биологических понятий, описываемых в каждой из этих базах данных. Каждому выделенному понятию сопоставлен свой тип (таблица 2). Вопрос построения минимально необходимого набора атрибутов для каждого типа в рамках данной статьи не рассматривается.

После этого, выделенные типы упорядочены в виде иерархии, представленной на рис. 6. И наконец, для каждой базы данных был создан специальный драйвер, переводящий текстовую информацию из этой базы данных в набор объектов соответствующих типов.

Данный подход реализован в виде библиотеки классов на языках C^{++} и Java. Единственное принципиальное различие в их реализации, это то, что класс Component в версии на Java имеет метод getImage, который представляет в графическом виде данный Component. Производные классы замещают этот метод, чтобы адекватно визуализировать соответствующий объект. Таким образом, данный подход так же позволяет создать средства для графического представления информации из баз данных. Были созданы TRRD и GeneNet viewer, идет работа над объединенным TRRD, TRANSFAC, EMBL viewer.

Для семантического анализа информации построен набор правил и метаправил. Для управления поиском и анализом информации, а так же для ее графического представления нами разрабатывается предметно-ориентированный язык.

На основе этой библиотеки классов разработаны TRRD и GeneNet viewer, которые осуществляют объектно-ориентированное представление информации из соответствующих баз данных в графическом виде, а так же обеспечивают навигацию по этим базам данных.

3.2.2 MGQL — специализированный объектно-ориентированный язык запросов

Для доступа, анализа и графического представления информации из баз данных EMBL, TRRD, TRANSFAC, EPD, SWISS-PROT, PDB и GeneNet нами разработан специализированный объектно-ориентированный язык высокого уровня (Molecular Genetics Query Language). В отличии от обычных языков запросов, например SQL, он ориентирован на работу с информацией из этих баз данных в виде набора объектов, а так же содержит специальные функции для семантического анализа этой ин-

Таблица 1.

Список и краткое описание баз данных, использованных в работе, а так же список основных молекулярно-биологических понятий, информация о которых в них представлена.

База данных	Полное название	Краткое описание	Основные молекулярно-биологические понятия
EMBL	European Molecular Biology Laboratory Nucleotide Sequences Library	содержит нуклеотидные последовательности генов, а так же данные о их структурно-функциональной организации	ген, последовательность, алфавит, структура гена, сайт, структурный сайт
TRRD (Kolchanov et al., 1999)	Transcription Regulatory Regions Database	содержит иерархическое описание транскрипционных регуляторных районов генов	ген, сайт, набор сайтов, структура гена, сайт связывания транскрипционного фактора, комозиционный элемент
TRANSFAC (Heinmaier et. al., 1998)	TRANScription FACtors	данные об особенностях регуляции транскрипции данного гена	ген, сайт, набор сайтов
EPD	Eukaryotic Promoters Database	содержит позицию и характеристику старта инициации транскрипции для данного гена	ген, сайт
SWISS-PROT		содержит аминокислотную последовательность белка, кодируемого этим генам, а так же структурно-функциональную разметку этого белка	ген, белок, сайт, набор сайтов, последовательность, алфавит
GeneNet (Kolpakov et al., 1998)		содержит информацию о взаимодействиях этого гена с другими белками и генами в рамках генной сети	клетка, компартмент, ген, РНК, белок, вещество, состояние, взаимодействие

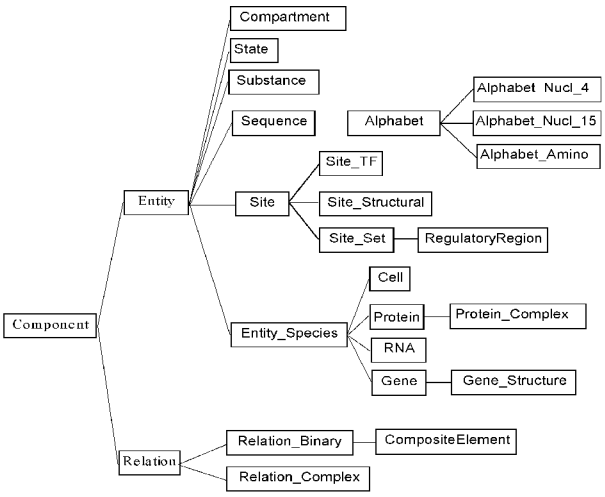

Рис. 6: Иерархия классов, соответствующих основным типам данных по регуляции экспрессии генов.

формации и ее графического представления.

Данный язык основан на описанной выше библиотеке классов, поэтому основные типы

3.2.3 Семантический анализ

Для семантического анализа информации построен набор правил и метаправил, которые можно разделить на 2 группы: 1) описание взаимоотношений между различными биологическим объектами; 2) описание каким образом информация об объектах разнесена по базам данным и каким образом она может быть взаимодополнена и взаимопроверена. Несколько примеров семантического анализа данных приведены на рис. 7.

```
Пример интеграции баз данных TRRD и TRANSFAC
string SRS_adr = "http://srs.ebi.ac.uk/srs5bin/cgi-bin/wgetz";
Database dbTFGENE = Database ( "TFGENE",  SRS_adr );
Database dbTRRD = Database ( "TRRD" );
// создает объект Gene на основе информации о
// заданном гене из базы данных TRANSFAC
Gene gTRANSFAC = new Gene( dbTFGENE, "G000042" );
// создает объект Gene на основе информации о
//  заданном гене из базы данных TRRD
Gene gTRRD = new Gene( dbTRRD, "G000042" );
// возвращает набор сайтов, общих для TRRD и TRANSFAC.
// Для сравнения сайтов используется правило CompareSites
Site_Set  sitesCommon = GetSitesCommon(gTRRD, gTRANSFAC,
CompareSites);
// возвращает набор сайтов, описаных только в TRRD
Site_Set  sitesTRRD = GetSitesDifferent( gTRRD, gTRANSFAC,
CompareSites);
// возвращает набор сайтов, описаных  только в TRANSFAC
Site_Set sitesTRRD =GetSitesDifferent( gTRANSFAC, gTRRD,
CompareSites);
//представляет полученные наборы сайтов в графическом виде
sitesCommon.View("Common sites");
sitesTRRD.View("TRRD sites");
sitesTRANSFAC.View("TRANSFAC sites");
```

Рис. 7: Семантическое объединение информации о сайтах связывания транскрипционных факторов из баз данных TRRD и TRANSFAC.

Рассмотрим один из примеров использования семантического анализа при интеграции баз данных TRRD и TRANSFAC. Обе базы содержат информацию о том, какие сайты связывания транскрипционных факторов содержатся в заданном гене, причем некоторые сайты описываются в обоих базах данных, а другие могут быть

Таблица 2.

Сопоставление типов каждому понятию.

Биологическое понятие	Тип	Комментарий (определение)
Компартмент	Compartment	
Клетка	Cell	Отдельный тип клеток или клеточная линия
Белок	Protein	Отдельный белок
Белковый комплекс	Protein_Complex	Белковый комплекс, состоящий из двух и более полипептидных цепей
Вещество	Substance	Низкомолекулярное химическое вещество
Состояние	State	Состояние генной сети или индуктор общего типа
Ген	Gene	Ген как компонент генной сети, не рассматривая его структуру
Структура гена	Gene_Structure	Ген вместе с описанием его структурно-функциональной организации
РНК	RNA	РНК различных типов
Сайт	Site	Функциональный участок последовательности
Структурный сайт	Site_Structural	Структурный сайт гена
Сайт связывания транскрипционного фактора	Site_TF	
Набор сайтов	Site_Set	Функциональный участок последовательности, состоящий из нескольких сайтов
Регуляторный район	Regulatory_Region	Регуляторный район гена
Последовательность	Sequence	Последовательность ДНК, РНК или белка
Алфавит	Alphabet	Алфавит для представления билогических макромолекул (ДНК, РНК или белка)
Взаимодействие	Relation	Взаимодействие между несколькими компонентами генной сети в самом общем смысле
Бинарное взаимодействие	Relation_Binary	Взаимодействие между двуми компонентами
Комплексное взаимодействие	Relation_Complex	Взаимодействие между тремя и более компонентами генной сети
Композиционный элемент	CompsiteElement	Композиционный элемент, т. е. ...

описаны только в одной из них. Наша задача объединить эту информацию из обоих баз данных. Для этого, сначала находятся сайты, которые описаны в обоих базах данных, затем находятся которые описаны только в TRRD, и затем только те, которые описаны в TRANSFAC. Объединив эти три набора мы получим полную информацию о сайтах связывания транскрипционных факторов, которые находятся в заданном гене.

Основной проблемой при этом является сравнение сайтов, какой сайт из базы данных TRRD соответствует сайту из базы данных TRANSFAC, поскольку (сайты не связаны друг с другом) описание одного и того же сайта в этих 2 базах могут существенно различаться. Во-первых, могут использоваться различные синонимичные названия транскрипционных факторов. Эта проблема решается путем построения таблицы синонимичых названий транскрипционных факторов. Во-вторых, позиции сайта в последовательности и сама последовательность сайта могут несколько различаться, поскольку определение длины флангов сайта подчас носит несколько субъективный характер, и соответственно различные эксперты могут использовать фланги различной длины. Эта проблема решается следующим образом, если название транскрипционного фактора (с учетом синонимичных названий) совпадают, и последовательности сайтов перекрываются, то мы рассматриваем этот сайт как один и тот же. Эта процедура сравнения реализуется функцией CompareSites. При необходимости мы можем определить другое правило сравнения сайтов и использовать его.

И в заключении отметим, что вся эта процедура объединения информации о заданном гене из 2 разных баз данных может быть выполнена одной функцией Complement (GENE gene1, GENE gene2).

4 Перспективы развития системы ГенЭкспресс

(NSAG — автоматическая генерация навигационных схем для электронной библиотеки ГенЭкспресс)

Характерной особенностью системы ГенЭкспресс является интеграция большого количества разнообразных информационных и программных ресурсов. При работе с ними возможно огромное разнообразие навигационных схем, то есть способов работы с системой ГенЭкспресс при получении необходимой информации. Конкретная навигационная схема представляет из себя последовательность действий пользователя с ресурсами Генэкспресс, записанную в виде HTML -страниц. Нами создается система автоматической генерации подобных навигационных схем. Реализация данной задачи предполагает перевод сервера электронной библиотеки ГенЭкспресс на новый технологический уровень. Для этого необходимо создание так называемых технических баз данных, содержащих:

- дескрипторы программ, входящих в состав ГенЭкспресс: совокупность информации о входных и выходных форматах данных, ссылки на запросы к справочной информации, типы связей с внешними модулями;
- дескрипторы внешних ресурсов (имеют аналогичный смысл);
- списки запросов к справочной информации;
- разработка формата конфигурационных файлов для представления навигационной схемы при ее включении в ГенЭкспресс.

После разработки таких средств мы сможем автоматически генерировать интегрированные системы не только на базе ресурсов, содержащихся на сервере, но и использовать также ресурсы мировой сети Internet. Схема интеграции ресурсов электорнной библиотеки ГенЭкспресс приведена на рис. 8. В настоящее время закончена интеграция ресурсов на первом уровне. Заканчивается интеграция ресурсов на втором уровне и начата рабта над третьим уровнем интеграции.

Библиография

[1] URL: http://www.infobiogen.fr/dbcat/

[2] URL: http://wwwmgs.bionet.nsc.ru/

[3] URL: http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/

[4] Иванисенко В.А., Григорович Д.А. и др. *Информационная система FramеProt по пространственным структурам ДНК, РНК и белков в составе GeneExpress.* // Доклады конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Санкт-Петербург, 1999.

[5] Babenko V.N., Kosarev P.S., Vishnevsky O.V. at al. *"A computer tool for investigating extended regulatory regions of genomic DNA sequences".* // Bioinformatics, V. 15, № 7/8, 1999 (in press).

[6] Bernstein F.C. et al. // J. Mol. Biol., 1977, V. 112, P. 535 (version of Nov. 1996).

[7] Etzold T., Argos P. // CABIOS, 1993, V. 9, 49–57.

[8] Kochetov A.V., Ponomarenko M.P., Frolov A.S. at al. *"Prediction of eukaryotic mRNA translational properties".* // Bioinformatics, V. 15, № 7/8, 1999 (in press).

[9] Kolchanov N.A., et al. *"Transcription Regulatory Regions Database (TRRD): its status in 1999".* // Nucleic Acids Res., 1999, 27, 303–306.

[10] Kolchanov N.A., Ponomarenko M.P., Frolov A.S. at al. *"Integrated databases and computer systems for studying the eukaryotic gene expression".* // Bioinformatics, V. 15, № 9, 1999 (in press).

[11] Kolpakov F.A. at al. *"GeneNet: a gene network database and its automated visualization".* // Bioinformatics, V. 14, No. 6, 1998, 529–537.

[12] Kolpakov F.A. and Ananko E.A. *"A graphic interface for interactive data input into the GeneNet database through the Internet".* // Bioinformatics, V. 15, № 7/8, 1999 (in press).

[13] Kolpakov F.A., Babenko V.N. *Computer system MGL: a tool for sample generation, graphical representation, and analysis of genomic regulatory sequences,* Mol. Biol., 31, 1997, 647–655.

[14] Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V. at al. *"Nucleosomal DNA property database".* // Bioinformatics, V. 15, № 9, 1999 (in press).

[15] Ponomarenko M.P., Ponomarenko J.V., Podkolodny N.L. at al. *"Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins".* // Bioinformatics, V. 15, № 9, 1999 (in press).

[16] Ponomarenko J.V., Ponomarenko M.P., Frolov A.S. at al. *"Conformational and physico-chemical DNA features specific for transcription factor binding sites".* // Bioinformatics, V. 15, № 7/8, 1999 (in press).

[17] Sayle R.A. and Milner-White E.J. // Trends in Biochem. Sci., 1995, V. 20, P. 374.

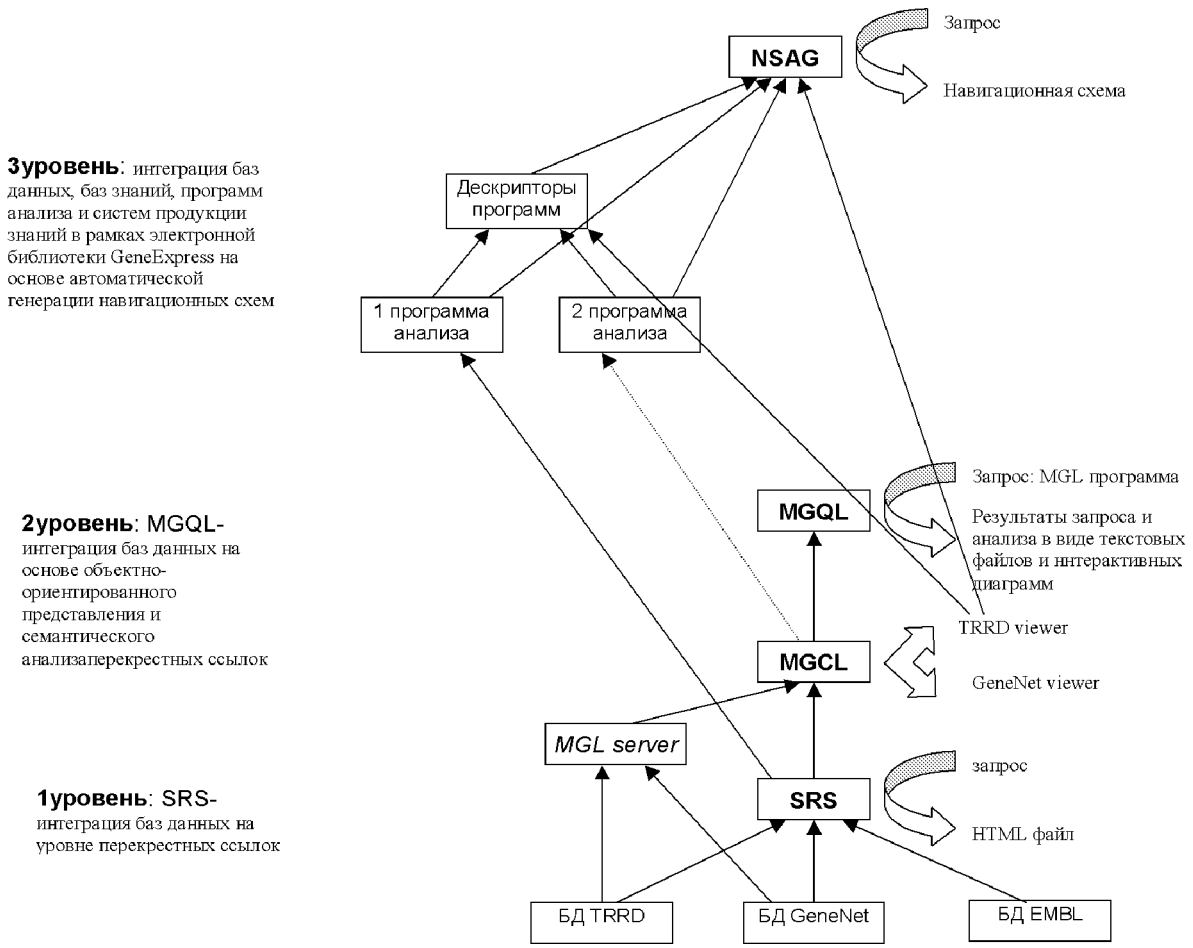

Рис. 8: Общая схема интеграции баз данных и программного обеспечения в рамках электронной библиотеки GeneExpress.