Проблемы создания предметного посредника для интеграции молекулярно-генетических информационных ресурсов

Л. А. Калиниченко*
ИПИ РАН
leonidk@synth.ipi.ac.ru

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

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

Аннотация

В данной работе рассматриваются вопросы создания конкретного предметного посредника для интеграции различных информационных источников в области регуляции экспрессии генов.

KEYWORDS

интеграция гетерогенных информационных ресурсов, посредники, молекулярная генетика, экспрессия генов, консолидация посредника

1 ВВЕДЕНИЕ

Современные отрасли науки являются сложными инфраструктурами накопления, систематизации, производства и распространения знаний. В научную деятельность в определенной отрасли науки вовлечены многие сотни организаций и групп. Качество исследований зависит от эффективности взаимодействия этих групп, темпа публикации и ассимиляции результатов всем научным сообществом. Масштабы исследовательских процессов, динамика производства новых результатов требуют совершенно новых подходов к организации взаимодействия ученых и организации информации в конкретных отраслях знания. Электронные библиотеки являются одним из наиболее перспективных направлений достижения такой организации. В настоящей работе, говоря об Электронных библиотеках, авторы ставят перед собой задачу выработки практичного подхода к организации информации в рамках конкретных областей научного знания. Основой подхода является тезис о том, что для успеха требуется установление определенного порядка, правил представления и использования информации, более жестких, чем те, что используются в Web. Существенно, что этот порядок должен вырабатываться самим научным сообществом и соответствовать насущным потребностям накопления и развития научного знания. Вместе с тем, при существующих масштабах процессов исследования, устанавливаемые формы и правила должны содействовать публикации результатов по мере их производства так, чтобы они немедленно становились достоянием научного сообщества.

По степени независимости, автономизации и динамики развития информационных коллекций, такие действия сравнимы с теми, которые имеют место на сайтах Web. Однако, для немедленной интеграции новой информации в общую структуру данных и знаний в данной области науки, отражения ее семантики, однозначно понимаемой сообществом, ее согласованной с сообществом структуризации, поддержки понятийной базы требуется определение научным сообществом соответствующих семантических моделей представления информации. На основе таких моделей должны устанавливаться достаточно жесткие рамки, терминология, структура представления информации в конкретных, достаточно узких предметных областях. Эти рамки определяются самим научным сообществом так, чтобы в них можно было вкладывать новые факты и результаты исследований, представляющие интерес для сообщества. Такие рамки предлагается устанавливать подобно тому, как это делается в развитых базах данных, базах знаний, информационно-поисковых системах: должна быть задана система понятий, принятых в рассматриваемой области науки, словари, а также схемы представления соответствующих фактов и знаний. С другой стороны, каждая исследовательская группа, каждый специалист, должны сохранять возможность представления своих результатов в предопределенной форме автономно, одновременно с другими группами и независимо от них, но так, чтобы они интегрировались с уже накопленными в установленной форме данными.

Вопросам создания среды, способной семантически интегрировать многочисленные неоднородные и независимые информационные источники, посвящен проект РФФИ "Создание интегрированных библиотек на основе неоднородных распределённых электронных коллекций научной информации", грант 98-07-91061 [7]. Согласно этому проекту, основным компонентом среды являют-

*Работа поддержана РФФИ 98-07-91061, 99- 07-90203, 98-07-91078

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

ся предметные посредники, образующие промежуточный слой между производителями и потребителями информации. Каждый посредник функционирует в определенной предметной области. Рамки контекста посредника определяются его метаинформацией, задающей систему понятий, рубрик, терминов, типов данных, видов классов фактов и методов анализа и обработки данных, представимых в данном посреднике. Этот уровень представления информации в посреднике называется интероперационным, или федеративным. Произвольные коллекции данных и знаний представимы в посреднике так, чтобы соответствовать указанной метаинформации.

Первой фазой создания конкретного предметного посредника является фаза его консолидации. На этой фазе усилия научного сообщества в определенной области направлены на формирование метаинформации интероперационного (федеративного) уровня. Предполагается, что в этот процесс вовлекаются наиболее авторитетные группы исследователей, имеющие значительный опыт метаописания достигнутых результатов. Создаваемая на этапе консолидации метаинформация считается достаточно консервативной. После согласования научным сообществом, метаинформация федеративного уровня фиксируется на определенный период времени, в течение которого она может лишь расширяться.

После фазы консолидации посредник вступает в операционную фазу. Во время этой фазы произвольные коллекции данных и знаний могут быть зарегистрированы в посреднике в терминах федеративного уровня. Процесс регистрации новых коллекций автономен и выполняется одновременно и независимо силами держателей таких коллекций - в нашем случае, исследовательских групп. Существенно, что среда посредников имеет рекурсивную природу, так что посредник сам может выступать в качестве коллекции, которая может быть зарегистрирована в другом посреднике. Пользователи посредника имеют доступ ко всей накопленной в нем информации на основе метаинформации интероперационного уровня. Метаинформация в посреднике представляется в рамках канонической модели данных Синтез [8].

В настоящее время, различными научными коллективами ведутся работы, способствующие консолидации информации в различных областях. Например, группа Life Sciences Research является консорциумом представителей фармацевтических компаний, академических институтов, фирм - разработчиков информационных технологий, объединяющим специалистов всего мира в рамках OMG. Цели консорциума заключаются в достижении интероперабельности информационных ресурсов в исследовательских работах в области наук о жизни (life sciences). В частности, недавно опубликован запрос предложений в области генной экспрессии и уже подготовлены первые предложения [10]. Целью этого запроса является определение набора интерфейсов, структур данных и сервисов, которые должны позволить более просто обмениваться данными между системами в области генной экспрессии. Эти стандарты (как принято в OMG) должны способствовать разработке общего каркаса, на основе которого более перспективные сервисы в области экспрессии генов могли бы быть созданы. Европейская лаборатория по биоинформатике ведет работы по созданию распределенной объектной среды, используя CORBA технологию [5, 6].

Сложность предметной области, разнообразие решаемых задач и большое число распределенных слабоструктурированных источников данных обуславливает необходимость тщательной проработки технологии формирования метаинформации предметного посредника на этапе консолидации. Эти проблемы рассматриваются в настоящей работе в контексте проблем интеграции молекулярно-генетических информационных ресурсов при исследовании регуляции экспрессии генов, возникающих на этапе консолидации. Выделены этапы и определена последовательность проектирования среды посредника. Проведен анализ особенностей распределенных источников данных в предметной области и определен представительный класс источников данных по проблеме исследования экспрессии генов. Рассмотрен состав метаинформации предметного посредника и технология ее формирования. Консолидируемый посредник предполагается входящим в коллектив посредников в области молекулярной биологии.

Эта работа выполняется объединенными усилиями проектов РФФИ 98-07-91061, 99-07-90203, 98-07-91078, разрабатываемых ИЦиГ СО РАН и ИПИ РАН.

2 ПРОЦЕДУРА КОНСОЛИДАЦИИ БАЗЫ МЕТАИНФОРМАЦИИ ПОСРЕДНИКА

Возможны различные схемы консолидации базы метаинформации посредника:

- отталкиваясь от конкретной области знания, независимо от решаемых задач и существующих источников данных и информационных ресурсов. Этот способ методологически представляется наиболее последовательным. Он позволяет определить модель конкретной области знания в ее терминах и в формализме, принятом в соответствующем научном сообществе. Вместе с тем, этот способ представляется трудно применимым на практике в сложных, динамически быстро развивающихся предметных областях;

- исходя из известных классов задач, которые являются актуальными для соответствующей области знания. Этот способ позволяет описывать модель предметной области ограниченно, в рамках определенного класса (классов) задач, но в терминах и в формализме соответствующей области знания;

- исходя из известных, уже накопленных информационных ресурсов. Этот способ, кажущийся наиболее простым, методически не является строго последовательным: он закрепляет текущее состояние источников, которое является результатом прагматических решений отдельных провайдеров на протяжении некоторого периода времени. Результатом этого подхода может быть создание "глобальной"схемы имеющейся совокупности источников (примерно так, как это делается при интеграции разнородных баз данных). Результат выражается средствами канонической модели данных;

- комбинированные подходы, включающие некоторую композицию рассмотренных схем.

Существенно, что первые две схемы должны включать этап перехода от абстрактной модели к представлению соответствующих понятий, структур, поведения средствами канонической модели данных.

Ввиду сложности и быстрого развития молекулярной биологии как предметной области, в рамках данной работы предполагается использование схемы консолидации базы метаинформации посредника исходя из известных классов задач. Под конкретный класс (классы) задач определяется модель узкой предметной области в молекулярно-биологических терминах. Затем выбираются представительные коллекции и ресурсы. Они могут иметь отображение в построенную модель, иметь частичное отображение, или не иметь такого вообще. Таким образом, происходит регистрация коллекций и проверка модели посредника. Если с очевидностью полезные коллекции не смогут быть должным образом зарегистрированы, то модель посредника необходимо модифицировать.

Рассмотрение последующего класса (классов) задач приводит к образованию нового посредника, либо к расширению существующего. Консолидация посредника, претендующего на некоторый уровень общности, таким образом, рассматривается как процесс объединения частных посредников, определенных для различных классов задач, относящихся к более общему посреднику.

Общая схема посредника, демонстрирующая его рекурсивную природу по отношению к процессам регистрации посредника как коллекции в другом посреднике, задания запросов и формирования ответов в иерархии посредников, показана на рис. 1.

3 АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ И ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ

Отличительной особенностью биологических систем и их компонентов является их блочно-модульная, иерархическая и/или сетевая организация. Например, орган, состоит из тканей, ткань состоит из различных типов клеток, клетка состоит из компартментов (цитоплазма, ядро, вакуоли и т.д.), по которым распределены макромолекулы ДНК, РНК и белков. Эти макромолекулы активно взаимодействуют друг с другом (образуют комплексы, осуществляют различные реакции, перемещаются по клеточным компартментам, клеткам, тканям и органам, и т.д.), формируя сложную сеть взаимодействий - генную сеть, которая описывает качественную модель регуляции экспрессии генов. В настоящее время существуют сотни источников слабоструктурированных данных, которые отражают те или иные аспекты регуляции экспрессии генов [3]. Поэтому, при решении реальных конкретных задач возникает необходимость использовать информацию из различных источников информации в тех или иных комбинациях. Следствием сложности молекулярно биологических данных является разнообразие способов и форматов их описания, используемых в различных базах данных. Часто отдельные данные являются слабо структурированными. Для их обработки может потребоваться значительный объем дополнительной метаинформации и/или их сложный семантический анализ. Кроме того, часто возникает необходимость преобразования данных и представления их в разной форме в зависимости от задачи. Такое преобразование может быть ресурсоемким. В качестве примера можно привести расчет локальных конформационных характеристик белка по координатам атомов, информация о которых накапливается в базе данных PDB [15]. Необходимо отметить, что представленные в различных источниках знания зачастую получены на различных объектах исследования, с разной степенью адекватности описывающие реальные процессы, происходящие в живом организме.

Можно построить иерархию таких экспериментальных модельных объектов, с различной степенью абстрагирования описывающих реальный объект исследования. Сопоставление этих знаний может оказаться нетривиальной проблемой.

Таким образом, сложность интеграции информационных ресурсов по регуляции генной экспрессии объясняется сложной организацией молекулярно- биологических данных, их неоднородностью, высокой степенью связанности, недостаточной формализацией и структурированностью, различным качеством, а также часто их неполнотой.

4 ВЫБОР КЛАССОВ ЗАДАЧ ДЛЯ ФОРМИРОВАНИЯ ПРЕДМЕТНОЙ ОБЛАСТИ ПРЕДЛАГАЕМОГО ПОСРЕДНИКА(ОВ) И ЕГО (ИХ) БАЗЫ МЕТАИНФОРМАЦИИ

Рассмотрим проблемы интеграции информационных ресурсов на конкретном примере, возникающие, при исследовании возможных режимов функционирования сложной нелинейной генной сети, описывающей регуляцию некоторого физиологического процесса (например, дифференцировка и созревание эритроидной клетки, антивирусный ответ, стероидогенез, регуляция клеточного цикла и т.п.), и при поиске оптимальных стратегий управления этим процессом. Это – типичная, практически важная задача, возникающая при поиске новых лекарственных средств, основанных на направленных воздействиях на геном человека с учетом индивидуальных особенностей генотипа. При создании такого рода фармакологических препаратов и индивидуальных схем коррекции патологических состояний, часто используются математические модели, описывающие динамику молекулярно- генетических процессов, на которых отрабатываются всевозможные реакции организма на фармакологическое воздействие. Генная сеть, наряду с другими типами молекулярно-биологических систем (например, метаболических путей, путей сигнальной трансдукции и т.д.) является одной из важнейших моделей, используемых при исследовании физиологических процессов. В построении такой модели можно выделить несколько этапов :

1. построение структурно-функционального описания генной сети;

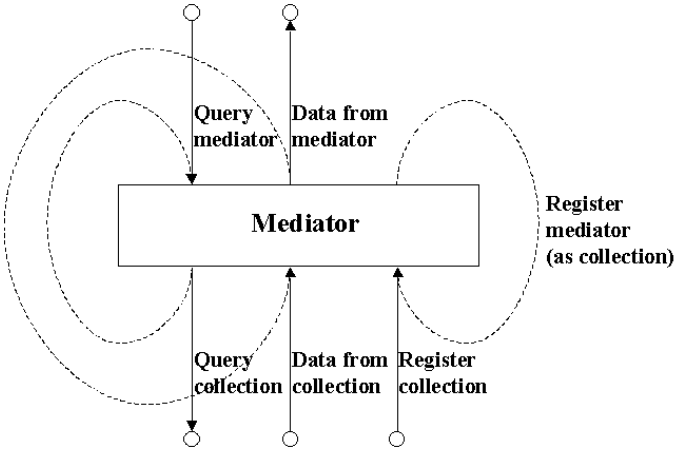

Рис. 1: Рекурсивная структура посредников

2. построение математической модели, описывающей динамику генной сети, например, логические переключательные модели, сети Петри, системы нелинейных дифференциальных уравнений или гибридные модели;

3. идентификация параметров модели;

4. исследование математической модели динамики генной сети.

Как правило, на каждом из перечисленных этапов у исследователя может возникнуть необходимость возврата на предыдущий шаг для уточнения и модификации принятых решений.

Каждый из этих этапов является достаточно сложной задачей, требующей привлечения дополнительной информации для обоснования выбора того или иного варианта решения. В большей степени это касается выполнения первого этапа, на которым происходит анализ и использование гетерогенных источников данных. Поэтому, рассмотрим его более подробно.

При построении структурно-функционального описания генной сети требуется :

1. вычленить ключевые физиологические процессы, регуляция которых выполняется на молекулярно-генетическом уровне;

2. выделить гены, координированное функционирование которых обеспечивает выполнение исследуемого молекулярно-генетического процесса;

3. определить все варианты мРНК и белков, кодируемых этими генами;

4. выявить регуляторные области этих генов;

5. определить локализацию содержащихся в них регуляторных элементов (сайтов связывания регуляторных факторов, энхансеров, сайленсеров и т.д.);

6. выделить небелковые низкомолекулярные соединения, которые участвуют в регуляции экспрессии генов и/или производство которых регулируется белками, которые кодируются этими генами;

7. выявить регуляторные факторы (белки), связывающиеся с этими регуляторными элементами и их влияние на экспрессию генов, включая управление переключением альтернативных сплайсингов мРНК и трансляций, усиление или ослабление экспрессии генов на различных стадиях развития организма, для разных видов, в разных органах, тканях или клетках и т.д.;

8. выявить и описать другие типы элементарных событий - взаимодействий между элементами генной сети на уровнях сплайсинга мРНК, трансляции и т.д.;

9. реконструировать структуру генной сети, описывающей исследуемый молекулярно-генетический процесс.

Результатом данного этапа является построение генной сети, в которой должны быть представлены гены, регуляторные факторы, белки, кодируемые этими генами, небелковые низкомолекулярные соединения, участвующие в функционировании генной сети, различные виды взаимодействий (белок-регуляторный элемент, белок-мРНК, белок-белок, белок-лиганд и т.д.).

Множество задач, решаемых на данном этапе, определяет достаточно замкнутую (в смысле системы понятий и задач) предметную область, которая, в свою очередь, традиционно разделена на подобласти исследования со своими специфическими проблемами (исследование ДНК, РНК, белка, молекулярно-генетических процессов).

Поскольку в целом задача построения генной сети является сложной, предполагается формирование нескольких иерархически организованных частных посредников. Примером такого частного посредника является посредник, соответствующий задачам, решаемым на уровне ДНК.

5 ВЫБОР ПРЕДСТАВИТЕЛЬНЫХ КОЛЛЕКЦИЙ ДЛЯ ФОРМИРОВАНИЯ ПОСРЕДНИКА

В качестве представительных коллекций на этапе консолидации предполагается использовать информацион-

ные коллекции и ресурсы из электронной библиотеки GeneExpress, которая включает большой объем различного вида слабоструктурированных данных, базы знаний, программы и сценарии, осуществляющие обработку и преобразование данных, поиск закономерностей, распознавание сайтов связывания и предсказания их активности и др. [11, 4]. При этом активно используются методы анализа данных и выявления знаний (Data mining & Knowledge discovery) [12].

Все ресурсы системы электронной библиотеки GeneExpress разделены на компоненты в соответствии с естественной иерархической организацией молекулярно генетических систем : (1) уровень ДНК, (2) уровень РНК, (3) уровень белка, (4) уровень генных сетей рис. 2.

5.1 УРОВЕНЬ ДНК

На этом уровне представлены знания об экспериментально определенной структуре и функциях ДНК и способах оценивания и предсказания этих знаний по первичным данным, например, первичным последовательностям.

База данных TRRD (TRANSCRIPTION REGULATORY REGIONS DATABASE), представляющая этот уровень, предназначена для накопления экспериментальной информации по структурно-функциональной организации регуляторных областей эукариотических генов. Кроме описания самих регуляторных областей, база данных TRRD включает: а) описание иерархии всех регуляторных единиц, находящихся в данной регуляторной области (таких как сайты связывания транскрипционных факторов, промоторы, энхансеры, сайленсеры и т.д.); б) информацию об особенностях экспрессии генов, описанных в базе данных; в) информацию о физиологических системах, органах и типах клеток, в которых экспрессируются описываемые гены.

База данных TRRD содержит также описание отдельных функциональных групп генов, важных с медико-биологической или фармакологической точки зрения: гены, которые индуцирует интерферон; гены, специфичные для эритроидной системы; гены липидного метаболизма; гены, которые контролируются глюкокортикоидами; гены, зависимые от клеточного цикла; гены эндокринной системы; гены теплового шока и растительные гены.

На данном уровне представлен программно-информационный ресурс **SITE RECOGNITION**, который содержит знания о конформационных и физико-химических особенностях сайтов связывания транскрипционных факторов, базы знаний по методам распознавания сайтов связывания транскрипционных факторов и программы, реализующие эти методы.

База знаний **SelexDB** содержит описание синтетических ДНК- последовательностей - аналогов сайтов связывания транскрипционных факторов. Для каждого сайта связывания приводится величина его афинности к соответствующему транскрипционному фактору. В базу знаний включены программы распознавания в произвольных последовательностях сайтов связывания транскрипционных факторов, представленных в **SELEX** и построения профилей афинности для соответствующих транскрипционных факторов.

Система **REGSCAN** предназначена для изучения протяженных регуляторных районов последовательностей ДНК эукариотических генов. Позволяет выявлять контекстные, конформационные и физико-химические свойства на основании анализа протяженных районов ДНК.

База знаний **SITE ACTIVITY PREDICTION** предназначена для описания, анализа и предсказания количественных характеристик специфической активности функциональных сайтов ДНК и РНК. Базы экспериментальных данных содержат описание сайтов с количественной величиной их специфической активности, контекст-зависимые конформационные и физико- химические свойства двойной спирали ДНК и функциональных сайтов, значимые для предсказания активности, и т.д. В базу знаний включены программы предсказания активности функциональных сайтов и параметры, необходимые для настройки программ на конкретную последовательность.

База знаний **DNA NUCLEOSOMAL ORGANIZATION** предназначена для хранения информации по контекстным, конформационным и физико-химическим особенностям нуклеосомных сайтов и программ распознавания этих сайтов в произвольных нуклеотидных последовательностях. База знаний включает выборки сайтов связывания нуклеосом, которые используются в качестве обучающих при создании алгоритмов распознавания, а также программы распознавания нуклеосомных сайтов на основе их значимых конформационных и физико-химических характеристик и на основе их контекстных характеристик.

5.2 УРОВЕНЬ РНК

На данном уровене представлены программно-информационые ресурсы, предназначенные для оценки трансляционных свойств мРНК. В частности, последовательности 5’ нетранслируемых районов высоко- и низкоэкспрессирующихся мРНК млекопитающих, одно- и двудольных растений, которые используются как обучающие выборки. База знаний для РНК уровня содержит описание выявленных особенностей мРНК, которые могут быть использованы для разделения высоко и низкоэкспрессирующихся мРНК, а также программы для предсказания трансляционной эффективности мРНК на основе значимых контекстных и структурных характеристик 5’- нетранслируемых районов мРНК. Имеется возможность предсказания вторичной структуры РНК.

5.3 УРОВЕНЬ БЕЛКА

Включает как описание первичной и пространственной структур и сайтов связывания белков, так и программ их распознавания, программ распознавания белковых доменов по свойствам аминокислот, обнаружения и анализа координировано эволюционирующих позиций.

5.4 УРОВЕНЬ ГЕННЫХ СЕТЕЙ

Данный уровень представлен следующими модулями [1] :

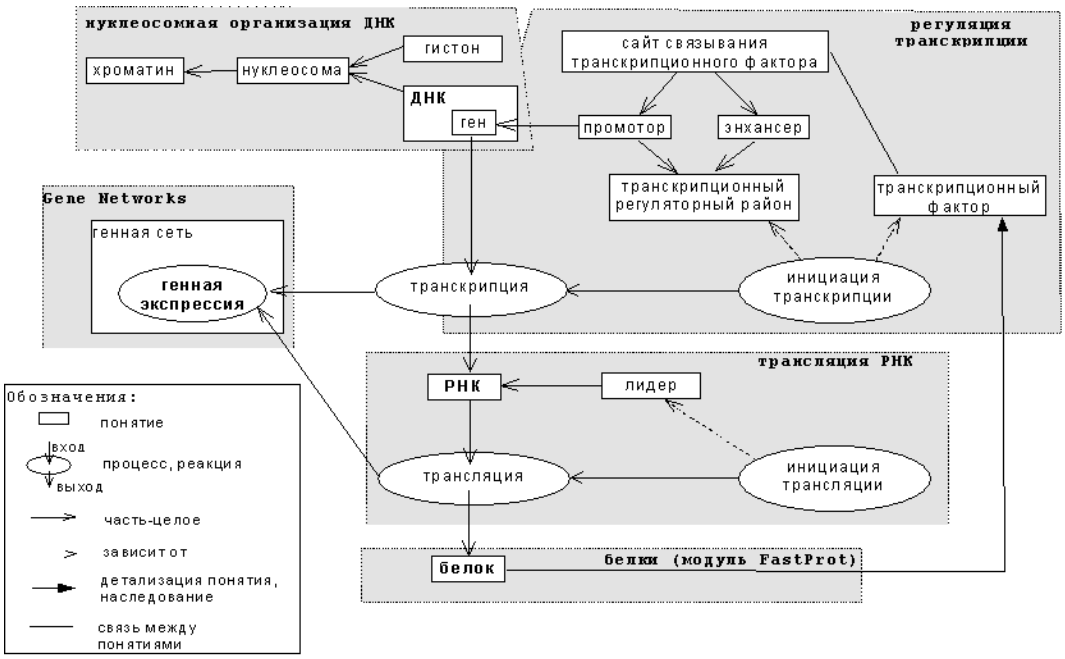

Рис. 2: Уровни представления генных сетей

- База данных **GeneNet,**предназначенная для накопления информации о генных сетях - группах координировано функционирующих генов, обеспечивающих выполнение жизненно важных функций организмов, визуализации генных сетей и моделировании их динамики.

- База знаний **GeneNetModels,** предназначенная для накопления информации о математических моделях генных сетей.

5.5 Другие информационные источники

Кроме этого существуют другие базы данных, которые предполагается использовать как на этапе консолидации посредников, так и на операционном этапе, в частности :

- База данных **Transfac,** содержащая информацию о транскрипционных факторов и методах предсказания сайтов связывания. Объединение этого источника данных с базами данных системы **ГенЭкспресс** (**TRRD**, **Samples**, **Selex**) позволяет строить представительные выборки сайтов связывания экспериментально определенных как для естественных, так и искусственных последовательностей, конструировать алгоритмы их предсказания для автоматического аннотирования геномной последовательности.

- База данных **COMPEL,** включающая описание композиционных регуляторных элементов (кластеров сайтов связывания). Эти данные позволяют учитывать влияния на экспрессию гена одновременного воздействия различных взаимодействующих факторов.

- Базы данных **EMBL/GenBank,** содержащие нуклеотидные последовательности регуляторных и кодирующих частей генов. Это первичные данные и их использование обязательно для любых молекулярно-генетических исследований.

- База данных **GDB,** в которой представлена информация о генах: структура и функция гена, продукты, цитогенетическая локализация гена, последовательности, полиморфизм, мутации, фенотипические проявления и т.д.

- База данных **SWISS-PROT,** содержащая информацию о структуре и функциях белков, их классификации, доменной структуре, последовательностях и т.д.

- База данных **PDB,** содержащая информацию о 3D структуре белков. Механизм взаимодействия белков с другими биомолекулами во многом определяется их 3D структурой. Анализ особенностей пространственной конформации взаимодействующих биомолекул позволяет выявлять сайты связывания и оценивать их активность. База данных **PDB** может служить примером целого класса источников информации о 3D структурах биомолекул. Класс задач, связанных с визуализацией и выявлением пространственных закономерностей, может быть отработан на этой модели.

- База данных **Medline,** содержащая библиографические ссылки. Этот класс источников данных необходим для обоснования представленных данных и, при необходимости, их уточнения и проверки. В конечном счете, большая часть информации получена из научных публикаций.

Таким образом, данный набор источников данных, доступных в сети Интернет, включает различную информацию о регуляции экспрессии генов на различных уровнях (ДНК, РНК, белок, генные сети), которая используется для решения реальных задач в области молекулярной биологии и, вследствии этого, может служить представительным набором источников данных, который позволит в будущем специфицировать предметный посредник.

6 МОДЕЛЬ И ИНСТРУМЕНТАРИЙ ПОДДЕРЖКИ МЕТАИНФОРМАЦИИ ПОСРЕДНИКА

6.1 Информационная модель

Абстрактная модель посредника, выраженная в терминах молекулярной биологии, в конечном счете, должна быть представлена (реализована) средствами информационной модели для формирования базы метаинформации посредника. Информация, необходимая для интерпретации молекулярно- биологических моделей, исключительно разнообразна. Спектр возможных форм представления данных простирается от структурированных (объектных) баз данных, слабоструктурированных баз данных и баз знаний до полностью неструктурированных данных (включая текстовые и визуальные данные). В слабоструктурированных моделях и в моделях баз знаний поведение, заданное схемой или отдельной сущностью, менее детерминировано, чем в структурированных базах данных.

Значительный объем информации должен быть представлен в посреднике на основе частично определенной схемы. Полностью неструктурированные текстовые данные характеризуются словарями, предоставляющими пользователям наборы терминов и их связей, которые являются характерными для текстовых документов предметной области.

Наконец, программные сервисы представляют методы обработки информации, которые вовлекаются в процесс обнаружения информации, вычисления и поиска требуемых данных. Такие сервисы (программные ресурсы) должны быть представлены в базе метаинформации для идентификации и формирования интероперабельных композиций при решении задач.

Модель для однородного представления различных информационных моделей в рамках одной парадигмы называется канонической. Одна и та же модель должна обеспечивать представление неоднородных коллекций данных, знаний и программ для формирования их интероперабельных композиций, а также средства представления онтологических и терминологических моделей предметной области. Многообразие требуемых форм представления информации предопределяет необходимые свойства канонической информационной модели посредника, которая реализуется на основе языка СИНТЕЗ, обладающего разнообразными способностями:

1. Способностью определения произвольных информационных ресурсов в унифицированном однородном представлении, независимом от использованных при реализации ресурсов языков и моделей данных (знаний). Обеспечением средств адекватной спецификации ресурсов для их эквивалентного унифицированного представления и интероперабельного использования.

2. Способностью представления развитых моделей предметных областей для спецификации требований и решения задач на основе баз данных, баз знаний и модельных (алгоритмических) знаний в однородном представлении. Поддержкой разнообразных средств представления знаний о предметной области (метаинформации, понятийно-структурной информации, знаний о законах и правилах поведения объектов предметной области).

3. Способностью модульного описания предметной области, позволяющего рассматривать совокупность информационных ресурсов как мультибазу ресурсов. Средства модульности должны обеспечивать возможность гибкого формирования контекстов решения задач (такие контексты могут быть созданы динамически).

4. Наличием специальных средств для представления слабоструктурированных данных и понятийно-структурной информации о предметной области как базы фреймов, а также для представления самоопределенных объектов.

5. Поддержкой полной системы типов. Тип рассматривается как специфический вид объектов. Обеспечивается возможность структурированного описания объектов предметной области на основе объектно-ориентированной модели представления иерархий типов и классов, множественного наследование спецификации типов. Поддерживается гибридная модель, позволяющая представлять и манипулировать объектами и фреймами при описании сущностей предметной области.

6. Возможностью определения информационных ресурсов как одиночных объектов и (или) классов объектов при обеспечении должного уровня инкапсуляции. Обеспечение континуума типизации: от практически бестиповых структур данных (представимых при помощи фреймов) до строго типизированных объектов

7. Гибкими средствами задания утверждений об информационных ресурсах - - встроенных и программируемых – на основе объектного исчисления. Задание семантических утверждений на уровне атрибутов объектов, классов и совокупности классов.

8. Поддержкой спецификации процессов для описания интерактивных систем и потоков работ.

9. Использованием одного и того же формализма для представления утверждений об информационных ресурсах, функций (предикатов) как логических программ, запросов к базе информационных ресурсов, предикативных спецификаций и условий конкретизации спецификаций информационных систем информационными ресурсами.

В языке СИНТЕЗ сочетаются достоинства объектной и фреймовой модели. Трактовка фреймов как слабоструктурированных данных позволяет формулировать запросы, не обладая полным знанием схемы. С другой стороны, все объектно-ориентированные свойства модели (наличие функций, иерархии типов и классов) также присутствуют. Выражения путей в языке СИНТЕЗ соответствуют базам данных, которые могут включать только структурированные или только слабоструктурированные данные, либо могут включать смешанные данные, включая обозначение перехода из структурированных данных в слабоструктурированные, или наоборот.

Методы отображения различных моделей данных в каноническую модель с целью их интеграции рассмотрены в [9]. Их применение приводит к формированию однородного представления совершенно различных видов используемых данных в одной модели. В соответствии с этим подходом разработаны адаптеры, позволяющие подключать к посреднику SRS и XML коллекции даных [13, 14].

6.2 Структура спецификации предметной области в посреднике

Посредник вводит уровень представления информации, характеризующий содержание зарегистрированных источников данных в интегрированном виде. Каноническая модель обеспечивает возможность задания запросов к такой совокупности источников и вычислять результат. Метаинформация посредника используется одновременно потребителями информации, провайдерами источников данных и самими предметными посредниками. Представление метаинформации в посреднике соответствует конструкциям канонической модели данных, что позволяет связывать различные контексты и представления метаданных неоднородных источников между собой.

База метаинформации в посреднике имеет модульное построение, позволяет задавать межмодульные связи, что необходимо при загрузке метаданных и регистрации коллекций.

Спецификация одной предметной области представляется в посреднике при помощи соответствующей схемы. Схема может включать модули различных видов, определяющие: структуру, онтологию, тезаурус, расширения тезауруса, рубрикатор. Спецификации одной предметной области различных уровней (федеративного, локального) входят в состав одной и той же схемы. При регистрации коллекций определяется конкретная предметная область (области).

Понятие предметной области относительно: спецификации предметных областей более высокого уровня образуются интеграцией схем включаемых предметных областей более низких уровней. Синтаксически иерархия предметных областей устанавливается посредством импорта соответствующих схем.

6.3 Представление онтологических спецификаций и тезауруса

Онтологические описания предметных областей содержат спецификации онтологических понятий, отношений между ними и ограничений, налагаемых на такие определения понятий. Для моделирования контекста предметной области достаточно возможностей языка СИНТЕЗ, включая конструирование онтологической модели, модели тезауруса и классификатора.

Существует отображение метаинформации в каноническую модель для известных моделей представления онтологий, таких как Ontolingua, OKBC, различных дескриптивных логик.

Каноническая модель онтологий и тезаурусов включает в качестве основных понятия категории, концепта (унифицированного для онтологических и лексических определений), их свойств, отношений и связанных с ними утверждений.

Различные информационные коллекции формируются в различных предметных областях со специальной терминологией и трактовкой структур объектов. Для описания этих особенностей вводится онтологический контекст как совокупность онтологических определений, гарантирующих правильность интерпретации концептов в конкретной предметной области.

Любое имя может быть введено в метаинформации как лексический концепт (лексическая единица). Предполагается наличие определений всех имен на естественном языке. Более формальные онтологические определения, относящиеся к именам, также допускаются.

Онтологические концепты представляют собой сущности представления знаний, отражающие характеристики класса подобных объектов реального мира. Их структурные и логические свойства могут быть выражены в терминах абстрактных типов данных. Каждому концепту может соответствовать также спецификация класса. Экстенсионал этого класса содержит объекты базы метаинформации (другие концепты, классы, элементы спецификации схемы), которые семантически связаны с данным концептом.

Тезаурусы представляются как коллекции лексических единиц и связей между ними. Модель определения лексического концепта рассматривается как подмножество онтологической модели концепта. В целом, используемая модель согласована с требованиями стандарта к многоязычным тезаурусам.

Для категоризации предметной области вводится иерархия классов. Каждый класс определяет некоторую категорию (рубрику) предметной области. Экземпляры этого класса представляют конкретные артефакты предметной области, включая:

- лексические единицы тезауруса;
- спецификации онтологических концептов;
- спецификации типов в различных модулях схемы.

6.4 Средства поддержки базы метаинформации посредника

В схеме предметной области содержится один онтологический модуль федеративного уровня. Для конкретной коллекции, содержащей онтологические спецификации, создаётся свой онтологический модуль локального уровня, принадлежащий предметной области, в которой регистрируется коллекция. При этом, в этой же предметной

области для каждой коллекции создается также модуль связующего уровня, в котором определяются отображения между онтологическими понятиями федеративного и локального уровней.

В рамках предметной области создаётся единственный (возможно, многоязычный) тезаурус федеративного уровня. Также в предметной области на федеративном уровне может быть определён модуль, содержащий словарь дополнительной лексики, привнесенной конкретными коллекциями. Тезаурусы коллекций хранятся на локальном уровне.

Рубрики конкретной предметной области определяются в спецификации модуля федеративного уровня, принадлежащего данной предметной области. Такой модуль для предметной области является единственным. Рубрикатор конкретной коллекции (если он задан) хранится в модуле локального уровня, содержащемся в предметной области, в которой зарегистрирована коллекция

Можно выделить следующие особенности межмодульных связей метаобъектов в базе метаинформации. Элементы структурных спецификаций становятся экземплярами рубрики или класса, соответствующего концепту. Если собственный тип класса является метаобъектом типа концептов, то такой класс содержит в качестве экземпляров элементы структурных спецификаций, по смыслу относящиеся к понятию тезауруса или онтологии, описываемому этим метаобъектом.

Типы и классы могут становиться как экземплярами классов, соответствующих концептам, так и экземплярами рубрик. Связи онтологической релевантности устанавливаются заданием в качестве экземпляра класса концепта, соответствующего элемента структурной спецификации в схеме.

Спецификации концептов содержат элементы, необходимые для описания списка дескрипторов, то есть лексических единиц тезауруса, определяющих данное понятие, позитивных, ассоциативных связей и родо-видовых связей с другими концептами. Эти связи устанавливаются на множестве понятий тезауруса и онтологии. В онтологических понятиях, кроме того, можно специфицировать их внутреннюю структуру средствами спецификации типов.

Концепты связаны с классами для хранения множеств элементов структурных спецификаций, соответствующих семантически данному понятию. Для таких классов собственным типом является спецификация типа понятия.

Рубрики являются классами, содержащими типы, классы и фреймы спецификации схемы, а также онтологические концепты. Рубрика может иметь вербальное определение, терминологический портрет из понятий тезауруса. Рубрики объединятся в иерархию с помощью отношения класс/подкласс. С классом рубрики в качестве собственного типа данного класса связан концепт. Именно в нём определяется вербальное описание и список дескрипторов как терминологический портрет рубрики. Более детально вопросы манипулирования онтологическими и терминологическими определениями в базе метаинформации рассмотрены в [16]. Средства поддержки репозитория метаинформации реализованы на языке Java над Oracle 8.

7 ОЖИДАЕМЫЕ ПРЕИМУЩЕСТВА АРХИТЕКТУРЫ, ОСНОВАННОЙ НА ИДЕЕ ПРЕДМЕТНЫХ ПОСРЕДНИКОВ

Основные преимущества архитектуры предметных посредников по сравнению с другими архитектурными решениями информационной интеграции заключаются в следующем:

- предметные посредники обеспечивают достижение семантической интеграции неоднородных информационных источников. При этом принимается во внимание структурная разнородность, разнородность значений, семантическая разнородность, возможность изменения локальных структур источников по мере развития и уточнения, различие в качестве данных (например, в точности);

- потребители информации должны знать только определения предметной области, включающие понятия, термины, структуры, методы, определенные сообществом в данной предметной области;

- поставщики информации могут распространять информацию для интеграции в посреднике независимо друг от друга и в любое время. Для распространения они должны регистрировать свои источники информации в предметном посреднике (посредниках). При регистрации источников не только их данные, но и их контекст должен быть отображен в контекст посредника. Потребители не вовлекаются в процесс распространения информации. Таким образом, достигается масштабируемость посредника по числу зарегистрированных источников. Это число может быть произвольным, ввиду того, что провайдеры регистрируют коллекции одновременно и независимо от других. Существенно, что регистрироваться могут не первичные источники, а обогащаемые и аннотируемые при преобразовании к уровню посредника данные;

- контексты, модели данных и языки, используемые платформы реализации в различных информационных источниках являются совершенно независимыми от посредника и его консолидированного определения метаинформации;

- формулируя запрос в терминах определений предметного посредника, потребители информации получают интегрированный доступ ко всей информации, зарегистрированной в посреднике к моменту запроса;

- посредники имеют рекурсивную природу: каждый посредник может быть зарегистрирован в другом посреднике. Таким образом, можно формировать и семантически интегрировать многоуровневые различные предметные области, определяя посредники более высоких уровней;

- персонализация информации для конкретных групп потребителей, связанных общностью интересов, формируется над определениями предметной области в посреднике. Этот процесс не зависит от существующих источников данных и состояния их регистрации в посреднике.

8 ОСОБЕННОСТИ ПРОЦЕССА РЕГИСТРАЦИИ КОЛЛЕКЦИЙ НА ОПЕРАЦИОННОМ ЭТАПЕ

Хотя данная статья посвящена этапу консолидации посредника, уместно кратко охарактеризовать процедуру регистрации в посреднике, тем более, что представительные коллекции подлежат регистрации на этапе консолидации.

Прежде всего, осуществляется выбор предметной области из набора предметных областей посредника, в которой будет производиться регистрация. На локальном уровне посредника создаются модули для загрузки онтологических спецификаций, рубрикатора, тезауруса, а также схемы (структурной спецификации локального уровня). В модулях образуется ссылка на схему выбранной предметной области.

Загружаются спецификации онтологических понятий, связи между элементами схемы и онтологическими понятиями, классы рубрикатора, связи между элементами схем и классами рубрик, загружаются локальные тезаурусы коллекций.

Осуществляется интеграция локального тезауруса (его лексики и связей) в тезаурусе посредника. При этом, если требуется, может быть обогащен словарь дополнительной лексики тезауруса федеративного уровня. При регистрации коллекции в базе метаинформации сохраняется также статистическая информация о вхождении терминов локального словаря в данную коллекцию.

Устанавливаются связи между онтологическими понятиями локального модуля коллекции и понятиями тезауруса и онтологии. Производится слабая онтологическая интеграция, при которой создаются межуровневые позитивные связи и родо-видовые связи понятий.

Если требуется, производится сильная интеграция, при которой элементы структурных спецификаций локального уровня становятся экземплярами классов, соответствующих понятиям онтологии федеративного уровня. Идентификация релевантных элементов структурных спецификаций является частью этого процесса.

Относительно структурных спецификаций (схем) предполагается, что преобразование спецификаций локальной модели данных в каноническую к моменту регистрации уже выполнено. Загружаются спецификации типов, классов и других элементов схемы канонической модели в созданный модуль структурной спецификации. Если спецификация коллекции является модульной, то ее модульная структура сохраняется (ее модули становятся подмодулями созданного модуля спецификации структуры локального уровня). Допускается также загрузка структурных спецификаций, представленных в других форматах (например, CDIF). Имена элементов структур соответствуют именам в коллекции.

Далее, типы и классы локальной схемы представляются как композиции типов и классов федеративного уровня. Эта важнейшая процедура отображения локальных коллекций в виртуальные коллекции федеративного уровня выполняется аналогично композиционному проектированию систем [2]. При этом создается модуль связующего уровня, содержащий конкретизирующие и композиционные типы и взгляды.

9 ЗАКЛЮЧЕНИЕ

В данной статье проанализирован подход к определению посредника неоднородных информационных коллекций в конкретной области молекулярной биологии - области регуляции экспрессии генов, а точнее ее моделирование при помощи генных сетей. Показано, что такой посредник целесообразно формировать как композицию более простых. Определены представительные информационные коллекции, необходимые для этапа консолидации посредника. Охарактеризованы информационная модель и инструментарий, необходимые для реализации этапа консолидации посредника.

Список литературы

[1] Ананько Е.А., Лихошвай В.А., Колпаков Ф.А., Подколодный Н.Л., Ратушный А.В., Игнатьева Е.В., Подколодная О.А., Колчанов Н.А. Электронная библиотека GeneNet: описание и моделирование генных сетей животных и растений.//Труды Второй Всероссийской конференции по Электронным Библиотекам. Протвино, Сентябрь 2000.

[2] Briukhov D., Kalinichenko L. Component-based information systems development tool supporting the SYNTHESIS design method. Springer LNCS, *Proceedings of the East European Symposium on "Advances in Databases and Information Systems"*, September 1998, Poland

[3] Burks C. Molecular Biology Database List//NAR, 1999, Vol.27, No.1., P.1-9.

[4] ГенЭкспресс. http://wwwmgs.bionet.nsc.ru/mgs/systems /geneexpress/

[5] CORBA at European Bioinformatics Institute (EBI). http://corba.ebi.ac.uk/

[6] Coupaye T. Wrapping SRS with CORBA: from Textual Data to Distributed Objects. Bioinformatics Journal, Chris Sander, Gary Stormo, Eds., Oxford University Press, 15(4), 1999.

[7] Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N., Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections.//Proceedings of the Second All Russian Conference on Digital Libraries. Protvino, September 2000.

[8] Калиниченко Л.А., СИНТЕЗ – язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов, ИПИ РАН, Москва, 1993.

[9] Kalinichenko L.A. Method for data models integration in the common paradigm. In *Proceedings of the First East European Workshop 'Advances in Databases and Information Systems'*, St. Petersburg, September 1997.

[10] Life Sciences Research, Gene Expression, LSR RFP-7, Request For Proposal, OMG Document: lifesci/2000-03-09. ftp://ftp.omg.org/pub/docs/lifesci/00-03-09.pdf

[11] 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.

[12] Колчанов Н.А., Подколодный Н.Л., Пономаренко М.П., Витяев Е.Е., Иванисенко В.А. Анализ данных и продукция знаний в системе GeneExpress - электронной библиотеке по структуре и функции ДНК, РНК и белков.// Труды Второй Всероссийской конференции по Электронным Библиотекам. Протвино, Сентябрь 2000.

[13] Котляров Ю. В., Подколодный Н.Л. Подключение баз молекулярно- генетических данных к посреднику среды создания интегрированных электронных библиотек.// Труды Второй Всероссийской конференции по Электронным Библиотекам. Протвино, Сентябрь 2000.

[14] Осипов М.А., Калиниченко Л.А. Интеграция XML-коллекций данных в посреднике неоднородных коллекций электронных библиотек.// Труды Второй Всероссийской конференции по Электронным Библиотекам. Протвино, Сентябрь 2000.

[15] База данных PDB. http://www.rcsb.org/pdb/

[16] Skvortsov N.A., Kalinichenko L.A. An Approach to Ontological Modeling and Establishing Intercontext Correlation in the Semistructured Environment.// Proceedings of the Second All Russian Conference on Digital Libraries. Protvino, September 2000.