ПРОБЛЕМЫ СОЗДАНИЯ ПРЕДМЕТНОГО ПОСРЕДНИКА ДЛЯ РЕГУЛЯЦИИ ЭКСПРЕССИИ ГЕНОВ

Л.А. Калиниченко[1], Д.О. Брюхов[1], В.Н. Захаров[1], О.А. Подколодная[2], Н.Л. Подколодный[2,3]
[1]Институт Проблем Информатики РАН, 117333, Москва, ул. Вавилова 44-2
e-mail: {leonidk,brd}@synth.ipi.ac.ru
[2]Институт цитологии и генетики СО РАН, 630090, Новосибирск-90, пр. Лаврентьева, 10
[3]Институт вычислительной математики и математической геофизики СО РАН, 630090, Новосибирск-90, пр. Лаврентьева, 6
e-mail: pnl@bionet.nsc.ru

PROBLEMS OF SUBJECT MEDIATOR DEVELOPMENT FOR GENE EXPRESSION REGULATION DOMAIN

[1]L.A. Kalinichenko, [1]D.O. Briukhov, [1]V.N. Zakharov, [2]O.A. Podkolodnaja, [2,3]N.L. Podkolodny
[1]Institute for Problems of Informatics RAS, Moscow, Russia
e-mail: {leonidk,brd}@synth.ipi.ac.ru
[2]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[3]Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia
e-mail: pnl@bionet.nsc.ru

For efficient organization of research in the domain of bioinformatics it is required to organize properly the relevant information in specific research areas. One of the important outcomes of such organization would be provision of access to and querying of a large number of distributed information sources including various data on the primary and spatial structure of DNA and RNA macromolecules, proteins and their complexes as well as data on peculiarities of their interactions with each other.

To provide for semantic integration of nonsystematic population of autonomous information sources kept by different information providers into a well-structured information collection it is required to create the global unified representation of the existing information sources and services. To reach that it is proposed to form a special middleware consisting of the *subject mediators*. For each subject mediator, the application domain model is to be defined by the experts in the field. This model may include specifications of data structures, terminologies (thesauri), concepts (ontologies), methods applicable to data, processes (workflows), characteristic for the domain. The mediators provide a uniform query interface to the multiple data and procedure service sources,

thereby freeing the users from having to locate the relevant sources, query each one in isolation, and combine manually the information from them.

In the paper we discuss an approach for development of the mediator for integration of heterogeneous molecular-genetic data in the gene expression regulation domain.

1. Введение

Для организации эффективных исследований в области биоинформатики необходимо обеспечить доступ и выполнение сложных запросов к большому числу распределенных информационных ресурсов, включающих разнообразную информацию о первичной и пространственной структуре макромолекул ДНК, РНК, белков и их комплексов, а также об особенностях их взаимодействия друг с другом.

Эти данные обычно являются слабо структурированными. Для их обработки может потребоваться значительный объем дополнительной метаинформации, сложный семантический анализ или обработка данных, выполненная разными методами. Проблема осложняется тем, что представленные в различных источниках знания зачастую получены на различных объектах исследования, с разной степенью точности описывающих реальные процессы, происходящие в живом организме.

Для обеспечения семантической интеграции многочисленных неоднородных и независимых информационных ресурсов требуется организовать глобальное унифицированное представление включаемых информационных ресурсов и предоставляемых услуг. Предлагается создать специальный промежуточный слой, состоящий из предметных посредников [3], обеспечивающих унифицированный интерфейс запросов к многочисленным источникам данных и освобождающих пользователя от необходимости находить подходящую коллекцию, осуществлять в ней поиск требуемой информации и вручную объединять информацию, полученную из различных коллекций. Каждый посредник функционирует в определенной предметной области.

Нами развивается посредник[1] для интеграции гетерогенных молекулярно-генетических данных в области регуляции экспрессии генов. Посредник состоит из 3 уровней: федеративного, локального и промежуточного. На федеративном уровне задаются онтологические понятия данной предметной области, схема, описывающая структуру (типы, классы, атрибуты) и функциональность посредника (средства семантического анализа и интеграции данных, средства обработки данных, предсказания, генерации

[1] Работа частично поддержана РФФИ (гранты 01-07-90376, 01-07-90084, 00-07-90337), Министерством промышленности, науки и технологий Российской Федерации (№ 43.073.1.1.1501), СО РАН (Интеграционный проект № 65).

дочерних баз данных и генерации знаний на основе автоматического поиска закономерностей и т.д.). На локальном уровне представлены спецификации информационных ресурсов. На промежуточном уровне задаются соответствия между спецификациями посредника и информационных ресурсов.

Основные преимущества такого подхода заключаются в обеспечении:

- развитых средств семантической интеграции неоднородных информационных источников. При этом принимается во внимание структурная разнородность, разнородность значений, семантическая разнородность, различие в качестве данных (например, в точности) и т.д.;
- независимости интерфейса пользователя от существующих источников данных. Пользователи должны знать только определения предметной области, включающие понятия, термины, структуры, методы, определенные сообществом в данной предметной области.

Архитектура посредника включает базу метаинформации, средства регистрации коллекций [1], средства выполнения запросов, получения результатов и представление их пользователю.

2. Описание посредника для регуляции экспрессии генов

Описание посредника включает описание онтологических понятий для области регуляции экспрессии генов и описание федеративной схемы. Онтологическая модель предметной области (регуляции транскрипции генов эукариот) включает онтологическое определение системы понятий, тезаурусов и словарей.

Примеры онтологических понятий, представленные в библиотеке онтологий:

```
Name "nucleotide"
Definition "A subunit of DNA or RNA consisting of a nitrogenous base
(adenine, guanine, thymine, or cytosine in DNA; adenine, guanine,
uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule
(deoxyribose in DNA and ribose in RNA)."
Part-of "DNA"
Part-of "RNA"

Name "nucleus"
Definition "The cellular organelle in eukaryotes that
contains the genetic material."
Part-of "cell"
Subclass-of "organelle"
```

Схема посредника включает описание источников данных, типов экспериментов и порождаемых ими типов данных, методов анализа данных и построения теории предметной области и модели.

На Рис.1 представлен фрагмент спецификации схемы посредника для области регуляции экспрессии генов, содержащий спецификации ти-

пов, их атрибутов, ассоциаций между ними и ограничений, налагаемых на такие типы:

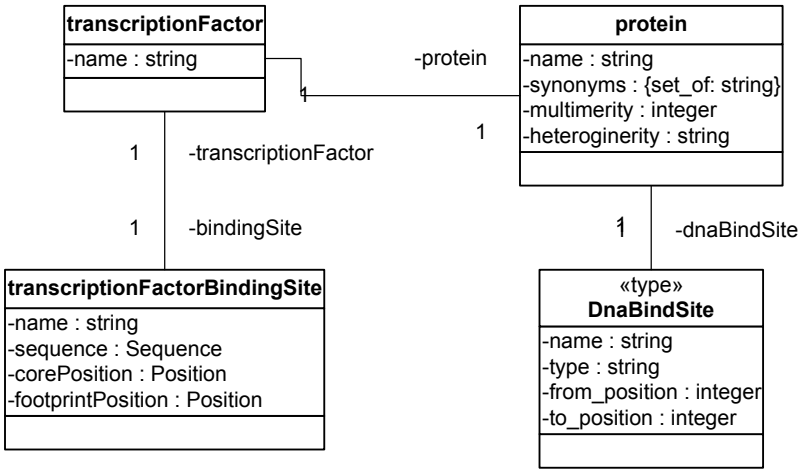

Рис.1 Фрагмент UML спецификации посредника

3. Описание информационных ресурсов

В качестве информационных ресурсов, представленных в среде посредника используются следующие Интернет доступные базы данных:

- База данных TRRD [4,5], развиваемая в ИЦиГ СО РАН, является уникальным, не имеющим в мире аналогов, информационным ресурсом, содержащим информацию о структурно-функциональной организации протяженных транскрипционных регуляторных областей генов эукариот и экспрессии этих генов.
- База данных SWISSPROT, содержащая информацию о структуре и функциях белков, их классификации, доменной структуре, последовательности и т.д.
- База данных EMBL/GenBank, содержащая информацию о последовательностях ДНК, РНК, их экзон-интронной структуре и другой функциональной разметке.
- База данных Medline, содержащая библиографические данные, необходимые для обоснования и уточнения представленных данных.

На Рис.2 представлен фрагмент спецификации базы данных TRRD:

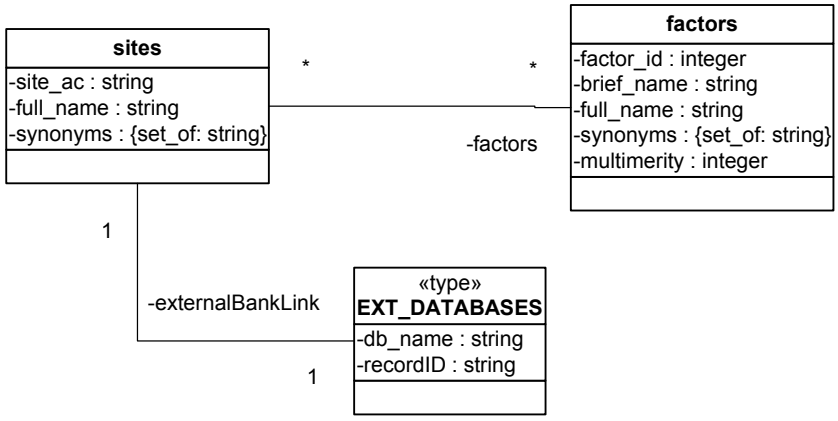

Рис.2 Фрагмент UML спецификации базы TRRD

На Рис.3 представлен фрагмент спецификации базы данных SWIS-SPROT на UML.

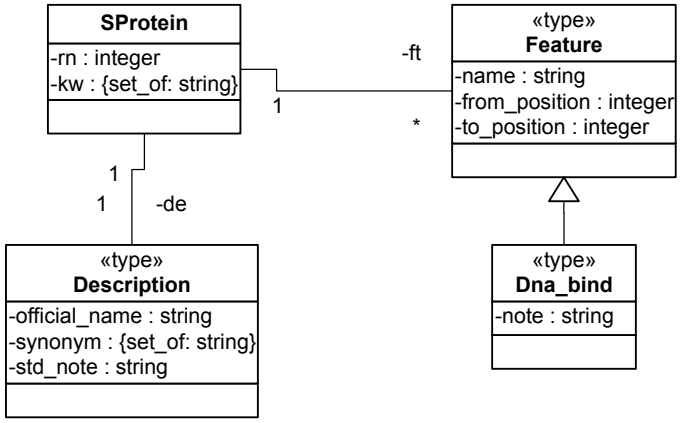

Рис.3 Фрагмент UML спецификации базы SWISSPROT

4. Регистрация информационных ресурсов в посреднике

Процесс регистрации неоднородных информационных коллекций в предметном посреднике базируется на подходе Local As View (LAV). Этот подход рассматривает схемы регистрируемых коллекций как материализованные взгляды над виртуальными классами посредника. Это позволяет работать с динамическим, возможно неполным набором коллекций. Коллекции могут менять свои схемы, могут становиться временно недоступными. Для обеспечения доступа к информационным коллекциям, их поставщики должны зарегистрировать их в соответствующем предметном посреднике. Такая регистрация может быть осуществлена в любое время и независимо от других коллекций. Предлагаемые программные средства обеспечивают масштабируемость предметных посредников по отношению к количеству коллекций.

В процессе регистрации классы локальных коллекций моделируются как множество экземпляров (объектов) типа экземпляра классов, а их описание в терминах федеративной схемы посредника определяет ограничения, которым должны удовлетворять экземпляры данного класса. Получение определения класса локальной коллекции как взгляда над федеративным уровнем означает, что не требуется добавлять локальный класс к федеративному уровню при регистрации. Описанное представление обеспечивает масштабируемость архитектуры посредника: представление одного класса не зависит от других классов, зарегистрированных на локальном уровне.

Процесс регистрации [1] включает в себя: согласование прикладных контекстов регистрируемой информационной коллекции и данного посредника на основе онтологического подхода, идентификацию релевантных классов федеративной схемы, конструирование наиболее общих редуктов, и конструирование взглядов, задающих ограничения в терминах федеративных классов.

Общая схема процесса регистрации информационных коллекций в посреднике выглядит следующим образом:

1. идентификация релевантных классов федеративной схемы

Для каждого класса локальной коллекции формируется список возможно релевантных классов федеративной схемы посредника, на основании интеграции онтологических понятий, соответствующих данным классам. Т.е. если между онтологическими понятиями, соответствующими классу коллекции и классу посредника, установлена позитивная связь или связь понятие-подпонятие, то данный класс посредника добавляется в список возможно релевантных классов. Аналогично формируются список релевантных типов для типов экземпляров классов коллекций, список релевантных атрибутов (функций) для каждого атрибута (функции) типа.

2. конструирование наиболее общих редуктов

Для типа экземпляра каждого найденного релевантного класса посредника выполняется конструирование наиболее общего редукта между ним и типом экземпляров заданного класса коллекции. В наиболее общий редукт попадают как общие атрибуты, так и атрибуты, соответствующие атрибутам типа посредника, которые могут быть получены из атрибутов типа коллекции. Для этого осуществляется поиск и разрешение структурных конфликтов между спецификациями коллекции и схемы посредника.

Пример конкретизирующего редукта для типа посредника *Protein* и SWISSPROT *SProtein*:

```
{R_Protein_SProtein;
  in: reduct;
  metaslot
    of: Protein;
    taking: {name, synonyms, keywords, dnaBindSite};
    c_reduct: CR_Protein_SProtein
  end
}
```

```
{CR_Protein_SProtein;
  in: c_reduct;
  metaslot
    of: SProtein;
    taking: {de, kw, ft};
    reduct: R_Protein_SProtein
  end;
  simulating: {
    R_Protein_Protein.name ~ get_name,
    R_Protein_Protein.synonyms ~ get_synonyms,
    R_Protein_Protein.keyWords ~ R_Protein_Protein.kw
    R_Protein_Protein.dnaBindSite ~ get_dnaBindSite,
  }
  get_name: {in: function;
    params: {+ext/CR_Protein_SProtein, -returns/string};
    predicative: {ex p/SProtein ((p/CR_Protein_SProtein = ext)
      & returns = p.de.official_name)}}  get_synonyms: {in: function;
    params: {+ext/CR_Protein_SProtein, -returns/string};
    predicative: {ex p/SProtein ((p/CR_Protein_SProtein = ext)
      & returns = p.de.synonym)}}  get_dnaBindSite: {in: function;
    params: {+ext/CR_Protein_SProtein, -returns/DNABindSite};
    predicative: {ex p/SProtein ((p/CR_Protein_SProtein = ext)
      & ex d/Dna_bind (in(p.ft, d)
      & returns = d/CR_DnaBindSite_Dna_bind))}}}
```

3. конструирование частичных взглядов

Для каждого релевантного класса конструируется частичный взгляд, задающий ограничения в терминах классов посредника, которые должны удовлетворяться значениями соответствующих наиболее общих редуктов типов экземпляров класса коллекции.

Пример спецификации взгляда для класса SWISSPROT *sprotein* над классом посредника *protein*:

```
{v_sprotein_protein;
  in: class;
  class_section: {
    lav: invariant, {subseteq (v_sprotein_protein(p),
      protein(p/R_Protein_SProtein))}
  };
  instance_section: CR_Protein_SProtein
}
```

4. композиция частичных взглядов

Для данного класса коллекции конструируется взгляд как композиция частичных взглядов над его релевантными классами посредника. При этом конструируется композиция наиболее общих редуктов типов коллекции, полученных для типов экземпляров всех релевантных классов посредника. Полученный взгляд является материализованным взглядом, задающим формулу выражения класса коллекции через классы посредника.

Поскольку классу SWISSPROT *sprotein* релевантен только один класс посредника, полученный взгляд *v_sprotein_protein* является финальным.

Для данного класса формула вычисления локальных классов через федеративные имеет вид:

```
sprotein(p/CR_Protein_SProtein)⊆protein(p/R_Protein_SProtein)
```

Инверсное правило [2], задающее как классы посредника выражаются через классы коллекции, для данной формулы выглядит следующим образом:

```
protein(p/Protein_SProtein) :- sprotein(p/Protein_SProtein)
```

5. Выполнение запроса к посреднику

В качестве тестовой задачи рассматривается проблема составления обучающих выборок регуляторных районов для использования программами распознавания.

Пример запроса пользователя к посреднику:

Выдать список сайтов связывания транскрипционных факторов, имеющих определенный тип ДНК связывающего домена.

Данный запрос на языке посредника выглядит следующим образом:

```
Q: transcriptionFactorBindingSite(s) & protein(p) &
s.transcriptionFactor.protein = p & p.dnaBindSite.type = "HOMEBOX"
```

Перепишем запрос, добавляя классы, которые участвуют в ассоциации (в данном примере s.transcriptionFactor.protein = p заменяем на transcriptionFactor(t) & s.transcriptionFactor = t & t.protein = p):

```
Q': transcriptionFactorBindingSite(s) & transcriptionFactor(t) & pro-
tein(p) & s.transcriptionFactor = t & t.protein = p & p.structure.type
= "HOMEBOX"
```

Применяем инверсные правила:

```
RQ1: FACTORS(t/TranscriptionFactor_FACTORS) &
SITES(s/TranscriptionFactorBindingSite_SITES) & spro-
tein(p/Protein_SProtein) & s.transcriptionFactor = t & t.protein = p &
p.structure.type = "HOMEBOX"
```

Посредник разбивает данный запрос на подзапросы:

1) к базе TRRD

```
SQ1(s,t):- FACTORS(t/TranscriptionFactor_FACTORS) &
SITES(s/TranscriptionFactorBindingSite_SITES) & s.transcriptionFactor =
t
```

2) к базе SWISSPROT

```
SQ2(p):- sprotein(p/Protein_SProtein) & p.structure.type = "HOMEBOX"
```

3) объединение результата первых двух запросов в посреднике

```
SQ3(s,t,p) :- SQ1(s,t) & SQ2(p) & t.protein = p
```

В процессе выполнения запроса используется информация из двух различных баз данных: SWISSPROT, TRRD.

6. Заключение

В статье рассмотрен пример предметного посредника для регуляции экспрессии генов.

Разрабатываемые технологии отрабатываются на программно-информационных ресурсах в области регуляции экспрессии генов, разрабатываемых в ИЦиГ СО РАН (http://wwwmgs.bionet.nsc.ru/mgs/gnw/).

Литература

[1] Briukhov D.O., Kalinichenko L.A., Skvortsov N.A. Information sources registration at a subject mediator as compositional development Proceedings of the Fifth East European Symposium on Advances in Databases and Information Systems (ADBIS'01), Springer, LNCS, 2001.

[2] O. Duschka and M. Genesereth. Answering Queries Using Recursive Views. In Principles Of Database Systems (PODS), 1997.

[3] 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. Proc. of the Second Russian National Conference on "Digital Libraries: Advanced Methods and Technologies, Digital Collections, Sep. 26-28, 2000, Protvino.

[4] Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*.-2002a, 30, p. 312-317.

[5] Kolchanov N.A., Podkolodny N.L., Ananko E.A., etc. Integrated system on gene expression regulation GeneExpress – 2002// Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).-2002b.