Mediation of heterogeneous information resources in the gene expression regulation domain

[1]***L.A. Kalinichenko,*** [1]***D.O. Briukhov,*** [1]***V.N . Zakharov,*** [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
Keywords: subject mediator, heterogeneous information sources, semantic information integration, ontology, gene expression regulation

Resume

Motivation: Semantic integration of heterogeneous information and procedural sources in bioinformatics is a necessary pre-requisite for efficient research.

Results: Application of the subject mediation approach in bioinformatics is shown for the gene expression regulation domain.

Availability: The results are available on request from the authors.

Introduction

A discriminative feature of molecular-genetic systems is their complex hierarchical and/or network organization. For instance, an organ consists of tissues, a tissue – of various cell types, a cell – out of compartments (i.e., cytoplasm, nucleus, vacuoles, etc.) that contain the macromolecules of DNA, RNA, and proteins. These macromolecules intensively interact with each other (they organize complexes, act in various reactions, move through cell compartments, cells, tissues, and organs, etc.), thus forming a composite net of interactions, namely, the gene network.

While solving concrete problems that are important in practice it is necessary to use a large number of heterogeneous, weakly structured molecular-genetical databases accumulating the results of numerous, complementary, intersecting, and probably contradictory experimental data. Databases on molecular-genetic information store the sequences, structures, 3D descriptions, attributive information, along with program software tools for data analysis, search of regularities, and prediction of different properties of objects, data reorganization, visualization, etc.

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.

Such data usually are semistructured. For their processing, a significant amount of additional metainformation, complex semantic analysis combining various methods may be required. The problem becomes even more complicated because data stored in different sources are obtained for different research entities, with different precision describing real processes in the living organism.

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. These definitions constitute specification of a subject mediator. Due to that, 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.

We develop a mediator for integration of heterogeneous molecular-genetic data in the area of gene expression regulation. The three level mediator architecture consists of federated, local and intermediate layers. The federated layer keeps subject mediator specifications, such as ontological definitions of the subject domain, schema description defining structural (types, classes, attributes) and functional (e.g., facilities for semantic data analysis and predictions, knowledge discovery based on the automatic methods) capabilities of the mediator. The local layer represents canonical specifications of the heterogeneous sources registered at the mediator. The intermediate layer defines a mapping of the source specifications into the specifications of the mediator.

Advantages of the proposed approach include the following:

- Semantic integration of heterogeneous information collections can be reached by taking into account structural, value, semantic, quality data heterogeneity;
- Users should know only subject definitions that contain concepts, structures and methods as defined by the community. Querying the subject definitions, users have integrated access to all information registered at the mediators up to the moment of a query.
- Personalization providing convenient views for specific groups of users can be formed above the subject definitions. This process is independent of the existing collection and their registration.

The mediator structure includes the metainformation base, tools for information sources registration, query interpretation facilities, facilities for collecting the query results and providing them to users.

The mediator for gene expression regulation

The model of the subject domain (gene expression regulation) has been developed. The model has a multilevel structure and includes ontological definition of the related concepts and thesauri, definition of information structuring, types of experiments, data analysis methods, as well as the related models of the respective theory.

The mediator is oriented on a broad class of problems. The intuition behind them can be provided by an example sequence of interrelated queries to the mediator that are intended for preparation of the training samples of regulatory regions, which may be used by recognition programs: to output the set of transcription factor binding sites sequences, which have a definite type of DNA-binding domain, search for transcription factors corresponding to the proteins found, search for transcription factor binding sites; search for the sequences of pre-ordered length including relevant transcription factor binding sites.

Examples of the ontological definitions represented in the metainformation base:

```
Name        "protein"
Definition  "A large molecule composed of one or more chains of
            amino acids in a specific order; the order is determined by
            the base sequence of nucleotides in the gene coding for the
            protein. Proteins are required for the structure, function,
```

```
            and regulation of the bodys cells, tissues, and organs, and
            each protein has unique functions. Examples are hormones,
            enzymes, and antibodies."

Name        "transcription factor"
Definition  "A protein that regulates transcription after nuclear
            translocation by specific binding with DNA or by
            stoichiometric interaction with a protein that can be
            assembled into a sequence-specific DNA-protein complex."
Part-of     "transcription complex"
Subclass-of "regulatory protein"
Subclass-of "protein"
```

Fig.1 shows the fragment of the mediator schema specification in UML notation. These specifications are used to illustrate heterogeneous sources registration and querying of the mediator.

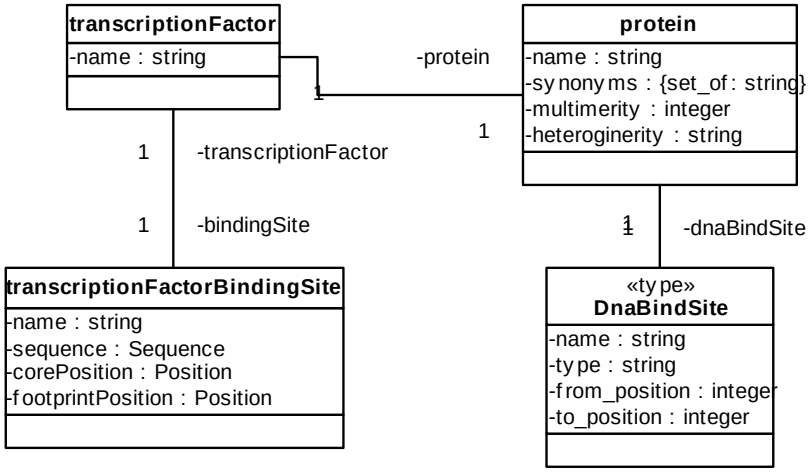

Fig.1 The fragment of mediator schema specification in the UML notation.

Information sources

Initial set of information sources to be registered at the mediator includes:

- The database TRRD developed at the Institute of Cytology and Genetics, unique informational resource that has neither world-wide analogs and that contains information about structural and functional organization of extended transcription regulating regions of eukaryotic genes and their expression. A subset of the TRRD schema using in this paper contains classes *sites, factors* and types *SITES, FACTORS*. (Kolchanov N.A., 2002a)
- The database SWISSPROT contains an information about the structure and functions of genes, about their domain structure, sequences, etc. A subset of the SWISSPROT schema using in this paper contains class *sprotein* and types *SProtein, Description, Feature, Dna_bind*.
- The databases EMBL/GenBank accumulate information about the sequences DNA, RNA, their exon-intron structure, and other functional layout.
- The database Medline/PubMed stores bibliography that is necessary for supporting and verifying the data presented.

Registration of the information sources in the mediator

The process for registration of heterogeneous information sources at the subject mediator is based on the LAV (Local as view) approach. In LAV the registered collections schemas are considered as materialized views above virtual classes of

mediator. In the specific registration method developed (Briukhov D.O. et al, 2001) the materialized views mentioned are designed applying compositional development method (source schema is treated as a specification of requirements and class schemas of the mediator are treated as component specifications). This approach is intended to cope with a dynamic, possibly incomplete set of sources. Sources may change their exported schemas, become unavailable from time to time. To disseminate the information sources, their providers should register them at a respective subject mediator. Such registration can be done concurrently and at any time. Specific methods and tools supporting process of information sources registration have been developed to make mediators scalable with respect to a number of sources involved.

During the registration a local source class is modeled as a set of instances (objects) of the class instance type, and the description of the source in terms of the mediator schema specifies the constraints on the class instances to be admissible for the subject mediator. The process of registration includes the ontological-based reconciliation of the application contexts of the registered sources and that of the mediator, identification of relevant classes of the mediator schema, constructing of the most common reducts for the mediator type specifications and respective types of the sources, constructing of views, specifying source class constraints in terms of the mediator classes.

Due to the space limit we cannot show here how contexts are reconciliated and types of the mediator are represented by the related types of the sources by means of the so-called concretizing reducts (Kalinichenko L.A. et al, 2000). We only show how views are expressed by means of the inverse rules (O. Duschka and M. Genesereth, 1997) expressing classes of the mediator through classes of sources (first rule relates protein to a class in SWISSPROT, two other rules relate *transcriptionFactor* and *transcriptionFactorBindingSite* to the respective classes in TRRD):

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

transcriptionFactor(t/TranscriptionFactor_FACTORS) :- factors(t/TranscriptionFactor_FACTORS)

transcriptionFactorBindingSite(s/TranscriptionFactorBindingSite_SITES) :- sites(s/TranscriptionFactorBindingSite_SITES)

Query rewriting in terms of the sources

We consider here an example of a query to the mediator:

Display the transcription factor binding sites with the definite types of DNA binding domain

In the mediator's canonical model this query is expressed as:

transcriptionFactorBindingSite(s) & transcriptionFactor(t) & protein(p) & s.transcriptionFactor = t & t.protein = p & p.structure.type = "HOMEBOX"

After query rewriting applying the inverse rules above, we get the query:

sites(s/TranscriptionFactorBindingSite_SITES) & factors(t/TranscriptionFactor_FACTORS) & sprotein(p/Protein_SProtein) & s.transcriptionFactor = t & t.protein = p & p.structure.type = "HOMEBOX"

This query is implemented by a subquery SQ1 to TRRD and a subquery SQ2 to SWISSPROT with the remaining postprocessing in the mediator SQ3:

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

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

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

Conclusion

The paper is of the "work in progress" kind. It shows how the subject mediation approach can be applied in bioinformatics. Gene expression regulation domain has been chosen to define an example of the subject mediator. The paper briefly introduces a notion of a subject mediator and explains how it can be defined. Issues of heterogeneous sources registration at the mediator and query rewriting in terms of registered sources are given in more details. The paper shows benefits that can be obtained applying the subject mediation approach.

An approach developed will be used as the tool for integration of information-software resources entering the integrated system on gene expression regulation, GeneExpress, which is being developed at the Institute of Cytology and Genetics of SB RAS (Kolchanov N.A., 2002b). This system integrates heterogeneous program-informational resources (a large bulk of databases, and hundreds of programs). We believe that developed by us technology of mediators is an adequate tool to accomplish the task that faces us.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 01-07-90376, 01-07-90084, 00-07-90337), Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Project No. 65).

References

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, 3**0**, p. 312-317.

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.

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.

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

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.