Programming and Computer Software, Vol. 26, No. 3, 2000, pp. 170–176.
Original Russian Text Copyright © 2000 by Kolpakov, Podkolodnyi, Lavryushev, Grigorovich, Ponomarenko, Kolchanov.

Methods for Integration of Heterogeneous Information Resources in Molecular Biology in the Digital Library GeneExpress

F. A. Kolpakov, N. L. Podkolodnyi, S. V. Lavryushev, D. A. Grigorovich, M. P. Ponomarenko, and N. A. Kolchanov

Institute of Cytology and Genetics, Siberian Division, Russian Academy of Sciences,
pr. Akademika Lavrent'eva 10, Novosibirsk, 630090 Russia

Abstract—Difficulties in integrating information resources (IRs) in molecular biology are due to a complex hierarchical and/or network organization of data, to their heterogeneity, complex interrelations, insufficient formalization, and to incompleteness. To overcome these difficulties, a digital library called GeneExpress has been under development in the Institute of Cytology and Genetics of the Siberian Division of Russian Academy of Sciences. This system, which belongs to a new class of information systems, integrates a great number of databases and hundreds of computer programs designed for processing information on the structure and functions of DNA, RNA, and proteins. The foundation of our approach is provided by hypertext integration, integration on the basis of a unified object-oriented environment by mapping the data into a canonical model with the use of specially designed mediators, and semantic data integration. A prototype of an implementation of this approach used in the current version of GeneExpress is described.

1. INTRODUCTION

To enhance the efficiency of studies in molecular biology, a large amount of information on the primary and spatial structure of macromolecules of DNA, RNA, proteins, and their complexes, as well as information about their interaction, must be used.

To systemize and accumulate this information, over 500 specialized databases [1] have been developed, the greater part of which are available on the Internet. Dozens of specialized systems have been developed that provide access and navigation across these databases; some systems also provide graphical representation of the available data.

An important characteristic of data in molecular biology is incompleteness. For example, the primary sequence of a gene or protein may be known, but their functions and details of interaction with other macromolecules may remain unclear. The missing information can be reconstructed by using various computer programs designed for predicting the structure, functions, and details of interaction of macromolecules of DNA, RNA, and proteins [2].

Thus, there is a need for integrating heterogeneous information sources and developing an effective system for searching the required information; performing complex analysis, involving various databases and programs; storing the results obtained in databases; and so on.

Since IRs are distributed and heterogeneous, the integration is very difficult, both conceptually and technically. To provide an effective solution to the problems described above, we have been developing an information system of a new class— the digital library GeneExpress [3– 5].

2. PROBLEMS OF INFORMATION INTEGRATION IN MOLECULAR BIOLOGY

A salient feature of biological systems and their components is the block modular, hierarchical and (or) network structure [6, 7]. For example, organs consist of tissues, tissues consist of various kinds of cells, cells consist of compartments (cytoplasm, nucleus, vacuoles, etc.), and DNA, RNA, and protein macromolecules are distributed across the compartments. These macromolecules interact with each other (they can form complexes, undergo various reactions, move across the compartments, cells, tissues, and organs, etc.), forming a complex network of interactions— the so-called genetic network. Because of this, the data in molecular biology are tightly interrelated, which makes it impossible to decompose them into independent databases. Thus, solving particular problems requires the use of information obtained from several databases in various combinations.

To use information obtained from several databases, the data must be semantically integrated. The integration involves two steps: (1) the identification of biological objects in different databases that can be considered equivalent in the context of a particular problem; (2) the correlation of information about the same biological object obtained from different sources. It must be stressed that many integration criteria are formulated by the user at the moment of problem solving, and these criteria can differ when different problems are solved.

0361-7688/00/2603-0170$25.00 © 2000 MAIK "Nauka/Interperiodica"

The complexity of data used in molecular biology leads to a variety of methods and formats used for description in different databases. The greater part of databases in molecular biology are organized as one or several text files. Every file contains a set of records that describe properties of a set of objects. These records consist of a set of text fields, which may have a complex hierarchical or recursive structure. Additional metainformation and (or) complex semantic analysis may be required to process these data. In addition, it is often necessary to transform data and represent them in a different form, depending on the problem. The transformation can be resource intensive. One example is the calculation of local conformational characteristics of a protein on the basis of atom coordinates; the information on atom coordinates is accumulated in the PDB database.

There exist a variety of experimental methods available for the analysis of molecular biological systems and their components. Any method can measure only a limited set parameters of a macromolecule or a molecular biological system. Often, the data obtained reflect the required properties only indirectly or approximate them with a certain accuracy.

Thus, analyzing a molecular biological object, we almost always face a lack of data and the problem of correlating the data obtained in various experiments. The researcher often has the alternative of obtaining either a more complete description of the object or a more accurate description. This is one of the problems solved by GeneExpress.

3. METHODS FOR INTEGRATING INFORMATION RESOURCES IN MOLECULAR BIOLOGY

Our approach to the integration of heterogeneous IRs involves five stages (Fig. 1):

(1) Hypertext integration. For this purpose, standard Web-technologies and tools for the automatic generation of hyperlinks are used that establish relationships between various information resources and between databases using key fields.

(2) Creation of a unified object-oriented environment based on the mapping of data into a canonical model, with the help of wrappers and mediators.

(3) Knowledge-based semantic data integration that involves a formal and biologically justified test of data integrity, correctness, and consistency.

(4) Data "enrichment" based on automatic processing, purposeful transformation, search for regularities, information retrieval, and construction of child databases.

(5) Constructing a system that can perform queries against several databases, use thesauruses, and a knowledge base for the automatic construction of queries, and generate scenarios of data analysis and processing.

Our approach corresponds to the model of heterogeneous IR integration suggested in [9, 10]. According to this approach, the system architecture includes three layers: user layer, federal layer, and IR access layer. At the access layer, special agents called mediators are used. They transform the information obtained from different databases to a set of objects corresponding to the canonical model, which gives a unified representation of all the data. The federal layer includes tools for the semantic analysis and integration of data, tools for data processing, prediction, generation of child databases, generation of knowledge on the basis of an automatic search for regularities, and so on. The user layer includes agents (mediators) that transform data to meet the user's demands.

4. IMPLEMENTATION OF THE APPROACH PROPOSED

The digital library GeneExpress is designed for collecting experimental data, searching for information, analyzing data, and investigating relationships in the field of the regulation of gene expression. It integrates a large number of databases and hundreds of programs designed for processing information about the structure and functions of DNA, RNA, and proteins; it also integrates other IRs available on the Internet that are important for describing gene expression.

In the current version of GeneExpress, stages 1, 2, and 4 are implemented. Stages 3 and 5 are under development.

4.1. Hypertext Integration of Information Resources

At stage 1, we performed the hypertext integration of databases (Table 1) with programs developed at the Institute of Cytology and Genetics [3– 5] and with other resources (available on the Internet) dealing with the regulation of gene expression. We use the SRS (Sequence Retrieval System) [11] to access those databases. SRS was developed by the European Institute of Bioinformatics and has become a *de facto* standard for access to databases in molecular biology represented as text files. SRS based on the Icarus language for describing the structure and syntax of data that are then used by the SRS to decompose text files into records, fields, and subfields.

The standard CGI interface is used to access WWW servers. The query manager makes it possible to combine queries as logical conditions of any complexity. The query result is obtained as an HTML file, and the SRS includes flexible tools to transform data when they are displayed as an HTML page. It also contains tools for the automatic generation of explicit hyperlinks, inclusion of those references in hypertext documents, the automatic generation of queries to databases, and built-in standard formats for representing molecular genetics data. Due to these advantages of the SRS, it is

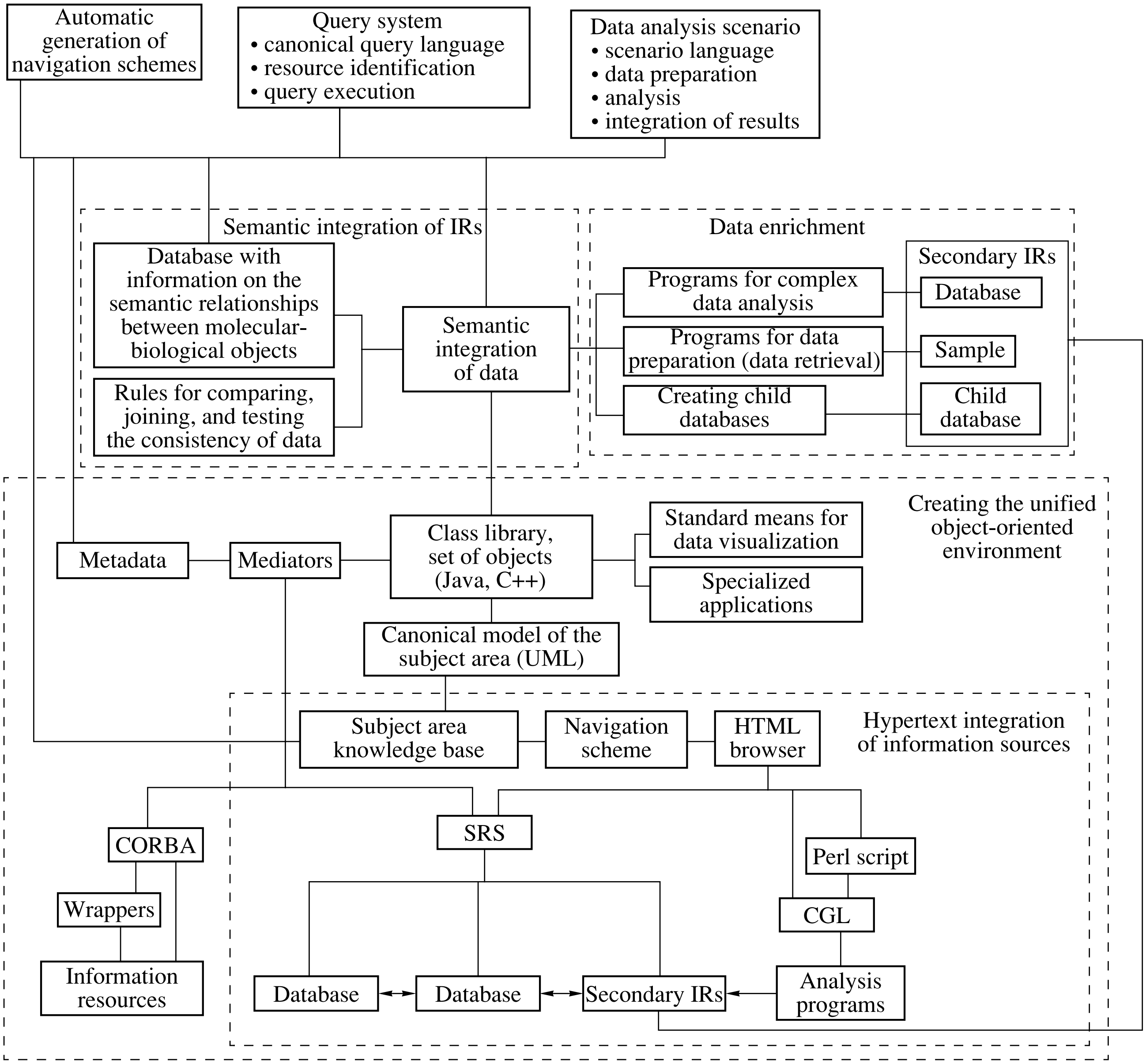
Automatic generation of navigation schemes
Query system
• canonical query language
• resource identification
• query execution
Data analysis scenario
• scenario language
• data preparation
• analysis
• integration of results
Semantic integration of IRs
Database with information on the semantic relationships between molecular-biological objects
Rules for comparing, joining, and testing the consistency of data
Semantic integration of data
Data enrichment
Programs for complex data analysis
Programs for data preparation (data retrieval)
Creating child databases
Secondary IRs
Database
Sample
Child database
Creating the unified object-oriented environment
Metadata
Mediators
Class library, set of objects (Java, C++)
Standard means for data visualization
Specialized applications
Canonical model of the subject area (UML)
Subject area knowledge base
Navigation scheme
HTML browser
Hypertext integration of information sources
CORBA
Wrappers
Information resources
SRS
Perl script
CGL
Database
Database
Secondary IRs
Analysis programs

Fig. 1. Integration stages of molecular biological resources in GeneExpress.

used as one of the basic means of integration at the first stage of GeneExpress.

The CGI interface is used to access programs designed for processing data in molecular biology; Perl scripts are used for processing HTML forms and data input and output; and special programs are used to represent data to be analyzed in the form required by analysis programs.

Figure 2 presents the functional scheme of GeneExpress. Each module includes experimental data represented as a database or a sample, programs for data analysis, results of automatic analysis, and the means used for graphical representation of the data and analysis results. The knowledge of the subject area (regulation of gene expression) underlies the decomposition of the system into functional modules. It includes such notions as the gene, regulatory region, factor, site, organism, organ, tissue, stage of development, and so on. The following types of relationships are used: general– specific, composition, classification, origin, function, regulator of the function, location, participation in the process, temporal relationships, and the like.

4.2. Constructing the Unified Object-Oriented Environment

At the next stage, a unified object-oriented environment is created. A prerequisite for such a type of environment is the development of a unified method for rep-

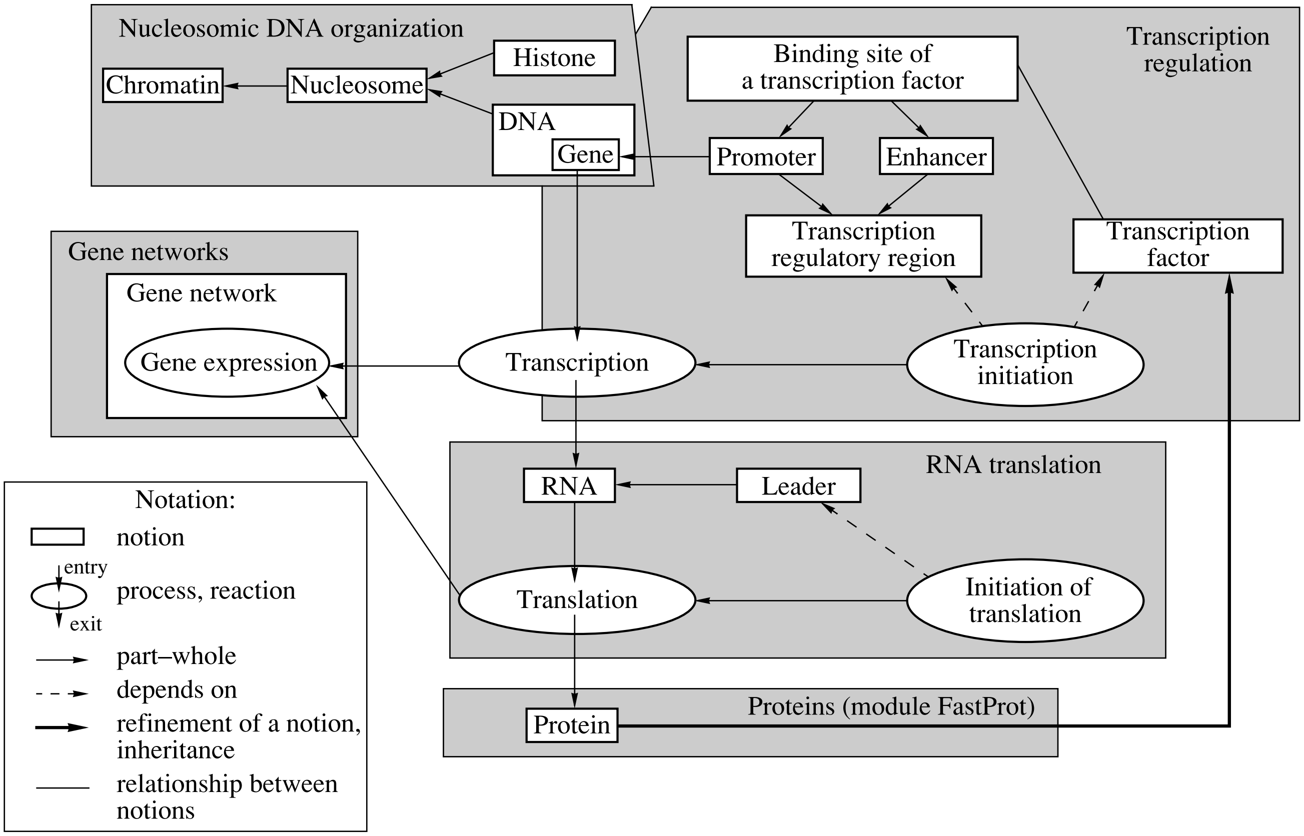

Fig. 2. The correspondence between modules of GeneExpress (gray rectangles) and the concepts of gene expression.

resenting molecular-biological data and the results of their analysis. For this purpose, the canonical model was developed on the basis of the knowledge of the subject area.

The canonical model is described using the graphical notation of the UML language [12] as a set of diagrams of classes; the model was also implemented as a library of Java classes. Thus, every molecular-biological object described by a notion in the knowledge base (e.g., gene, protein, RNA, functional site) is assigned a distinct data type in the UML diagram or a class in Java.

To embed molecular-biological databases available through SRS in the object-oriented environment, we use mediators that take information from the databases and represent it as a set of objects of Java classes of the canonical model. At present, such mediators are developed for the databases TRRD [3], GeneNet [13, 14], and EBML; the last database contains all the established sequences of DNA and RNA.

The canonical model includes tools for graphically representating the basic types of molecular-biological data (a library of classes in Java). This allows for the standard visualization of data obtained from different databases and the results of data analysis. Two types of graphical representation are available:

(i) diagrams (graphs) are used to represent the organization of molecular-biological systems (e.g., gene networks, metabolic paths, paths of signal transduction and so on);

(ii) maps are used to represent data about the structure and functional organization of sequences of gene clusters, genes, transcript regulatory regions, RNA, and proteins.

This library is used in such applications as Java applets in the TRRD Viewer [3], GeneNet Viewer [13], and the graphical interface of data input into GeneNet through the Internet [14].

4.3. Data Enrichment

A salient feature of molecular-biological data is the need for automatic analysis, in particular, of automatic genome annotation; i.e., distinguishing functionally important sections of the genome and predicting its structure and functions. For this purpose, approaches similar to data mining [8] are used. GeneExpress includes a large number of procedures for data processing that can be used for data enrichment, i.e., for step-by-step transformation of the data into a more informative (from the user's standpoint) semantic space. In the process, every subsequent level uses the preceding semantic space as the initial one. The direction of this

Table 1. The main databases containing data on the regulation of gene expression available from GeneExpress

Database	Short description
TRRD	Transcription Regulatory Regions Database. Contains the description of DNA regions responsible for the regulation of gene transcription
GeneNet	Contains a description of graphs of gene networks and elements of these graphs: cells, genes, RNA, proteins, various chemical compounds, as well as a description of interactions between these elements
Selex	Contains sequences of various functional DNA regions determined by special experiments (Selex protocol)
Activity	Data on the activity of the sites involved in the regulation of gene expression, as well as the physicochemical, conformational, and statistical DNA and RNA properties significant for the activity of these sites
Property	Contains information on conformational and physicochemical properties of the double helix of the DNA
Leader	Contains samples of RNA regions (leaders) responsible for initiating the translation
Samples	This database contains samples of regulatory genome sequences of various types

Table 2. Systems of knowledge production and procedural knowledge available from GeneExpress

Knowledge production system	Procedural knowledge (programs) obtained as the result of knowledge production
B-DNA produces knowledge on conformational and physicochemical characteristics of sites*	Features contains 1402 programs for recognizing sites by functionally significant conformational and physicochemical characteristics
Activity produces knowledge on context, conformational and physicochemical characteristics of sites that are important for predicting their activity	Knowledge contains 49 programs for predicting site activity by their nucleotidic sequences
CONSFREQ is designed for generating and using knowledge on context characteristics of sites that are important for their recognition	Matrix contains 567 programs for recognizing sites by the frequency of short "words" (oligoneucleotides); Consensus contains 66 programs for recognizing sites by their evolution–conservative invariants
Leader produces knowledge on the effectiveness of translation for various leading sequences	Leader Knowledge contains specific features of leading sequences that are significant for the effectiveness of the translation

* Site is a functionally significant region of a macromolecule DNA or RNA.

multilevel data enrichment depends on the particular problem.

At the first stage of this process (Fig. 3), information for systems of automatic knowledge production is prepared. For this purpose, the computer system MGL [15] is used that can automatically retrieve information about different parts of genes (e.g., the leading sequences) based on a semantic analysis of the description of the structure and functions of genes (this information is taken from the EBML database). MGL also constructs sequences of binding sites of transcription factors by integrating information obtained from the TRRD and EMBL.

At the next stage, systems of knowledge production (Table 2) analyze the information obtained and find regularities that are important for predicting activity and searching for functional sites. The procedural knowledge obtained at this stage include the procedure (a program or script) for searching or predicting activity of a certain type of functional sites, the description of this procedure purpose, conditions for its application, the format of the input data, constraints imposed on the input data, the format of the output data, and so on. This knowledge makes possible the automatic synthesis of complex scenarios for searching functionally important parts of genome sequences being annotated, and predicting their structure and function.

5. PROSPECTS

At present, tools of object-oriented access to molecular-biological databases [16] and their integration on the basis of the CORBA technology is under development [17]. This project is headed by the European Institute of Bioinformatics. It is supposed that at the first stage of the project the developers of the main databases (EMBL, SwissProt, PIR, MSD, GDB, TRANSFAC, P53, and RHdb [1]) will develop object-oriented representations and CORBA IDL specifications [18] of their databases. Then, these developers are supposed to adopt a common object model of the information stored in their databases. Our approach, in which the common (canonical) model is constructed based on ontological considerations, can prove more promising.

To implement this project, we develop mediators that map the data into an object environment built on the basis of the canonical model. Presently, the following parts are at the development stage: basic object adapters for databases maintained in the Institute of Cytology and Genetics, repositories of interfaces and

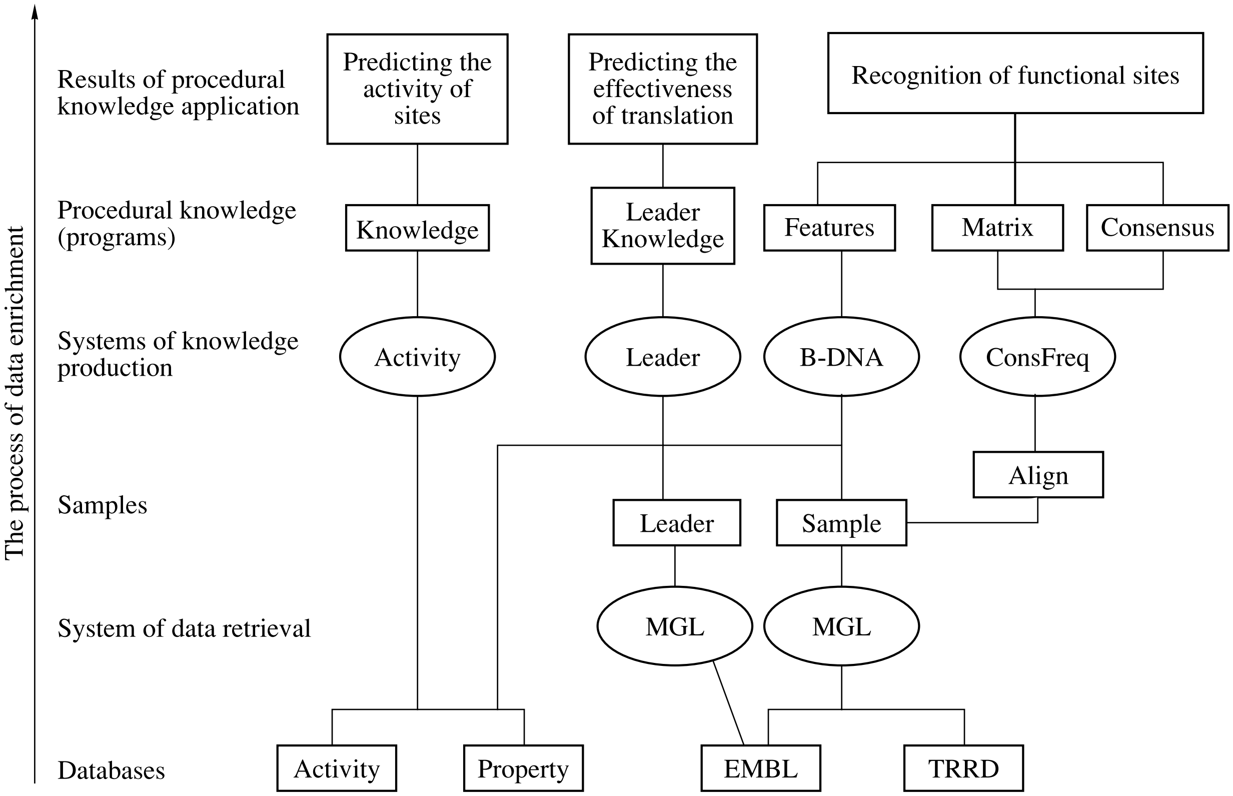
The process of data enrichment
Results of procedural knowledge application
Procedural knowledge (programs)
Systems of knowledge production
Samples
System of data retrieval
Databases
Predicting the activity of sites
Predicting the effectiveness of translation
Recognition of functional sites
Knowledge
Leader Knowledge
Features
Matrix
Consensus
Activity
Leader
B-DNA
ConsFreq
Align
Leader
Sample
MGL
MGL
Activity
Property
EMBL
TRRD

Fig. 3. An example of using the data enrichment technology in GeneExpress.

implementation of objects, services for working with objects (registration, generation and interpretation of objects, calling methods, and so on).

The next version of GeneExpress will focus on the semantic ontology based integration of data. We plan to create a knowledge library that will include, in addition to the subject area knowledge base, the following knowledge bases:

(i) *a base of basic knowledge* that is similar to the subject area knowledge base. However, the concepts described in this knowledge base are common for many areas and are used as the basis and templates for constructing the subject area knowledge base. Here, the concepts of state, event, process, action, function, and so on will be described;

(ii) *a terminology and information knowledge* base containing thesauri and a metadescription of available databases, for example, the database scheme, description of fields, their interpretation in terms of the subject area knowledge base, and so on. This knowledge base will also include a set of rules and metarules for identifying the same objects stored in different databases and integrating information from several objects into a single object;

(iii) *a knowledge base of methods for solving problems* that includes conceptions underlying descriptions of problem solving methods. The metabase that describes methods for accessing various programs, program interfaces, the format and description of the input and output data, and so on;

(iv) *application knowledge* base contains concepts that are used in knowledge modeling when solving concrete problems. The application knowledge base includes the subject area knowledge base, the base of basic knowledge, information knowledge base, and the knowledge base of methods for solving problems.

The knowledge library described above will be used by semantic data integration tools and the query system that can make queries against several databases.

A salient feature of GeneExpress is the integration of a large amount of various information and software sources. When working with these sources, a great variety of methods for using GeneExpress (the so-called navigation schemes) is possible. A navigation scheme is a sequence of user's actions when he deals with GeneExpress and remote information resources. At present, we develop a system for automatic generation of navigation schemes based on the metainformation contained in the knowledge base of methods for solving problems and in the terminology and information knowledge base.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project nos. 99-07-90203, 98-07-

91078, and 98-07-90126. The authors are grateful to L.A. Kalinichenko for helpful discussion of the concepts of integration of heterogeneous information resources, and to M.R. Kogalovsky for useful advice and remarks.

REFERENCES

1. Catalog of Databases in Molecular Biology, http://www.infobiogen.fr/services/dbcat/.
2. Catalog of Programs for Analyzing Data in Molecular Biology, http://www.ebi.ac.uk/biocat/.
3. Kolchanov, N.A., Ponomarenko, M.P., Kel, A.E., Kondrakhin, Yu.V., Frolov, A.S., Kolpakov, F.A., Kel, O.V., Ananko, E.A., Ignatieva, E.V., Podkolodnaya, O.A., Stepanenko, I.L., Merkulova, T.I., Babenko, V.N., Vorobiev, D.G., Lavryushev, S.V., Ponomarenko, Yu.V., Kochetov, A.V., Kolesov, G.B., Podkolodny, N.L., Milanesi, L., Wingender, E., Heinemeyer, T., and Solovyev, V.V., GeneExpress: A Computer System for Description, Analysis, and Recognition of Regulatory Sequences of the Eukaryotic Genome, *ISBM*, 1998, pp. 95–104.
4. Digital Library GeneExpress, http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/.
5. Kolchanov, N.A., Ponomarenko, M.P., Frolov, A.S., Ananko, E.A., Kolpakov, F.A., Ignatieva, E.V., Podkolodnaya, O.A., Goryachkovskaya, T.N., Stepanenko, I.L., Merkulova, T.I., Babenko, V.V., Ponomarenko, Yu.V., Kochetov, A.V., Podkolodny, N.L., Vorobiev, D.V., Lavryushev, S.V., Grigorovich, D.A., Kondrakhin, Yu.V., Milanesi, L., Wingender, E., Solovyev, V.V., and Overton, G.C., Integrated Databases and Computer Systems for Studying Eukaryotic Gene Expression, *Bioinformatics*, 1999, vol. 15, no. 7, pp. 669–686.
6. Ratner, V.A., Biology–Modular Principle of the Organization of Evolution in Molecular Genetics Control Systems, *Genetika*, 1992, vol. 28, no. 3, pp. 5–25.
7. Ratner, V.A., *Molekulyarno geneticheskie sistemy upravleniya* (Molecular Genetics Control Systems), Novosibirsk: Nauka, 1975.
8. *Knowledge Discovery through Data Mining: What Is Knowledge Discovery?* Tandem Computers, 1996.
9. Kalinichenko, L.A., *Metody i sredstva integratsii neodnorodnykh baz dannykh* (Methods and Means for Integration of Heterogeneous Databases), Moscow, Nauka, 1983.
10. Kalinichenko, L.A., Integration of Heterogeneous Semistructured Data Models in the Canonical One, *Trudy 1-oi Vserossiiskoi nauchnoi konferentsii Elektronnye biblioteki: perspectivnye metody i tekhnologii* (Proc. First All-Russian Conf. Digital Libraries: Advanced Methods and Technologies), St. Petersburg, 1999, pp. 3–15.
11. Etzold, T. and Argos, P., SRS—an Indexing and Retrieval Tool for Flat File Data Libraries, *Comput. Appl. Biosci*, 1993, vol. 9, pp. 49–57.
12. UML Specification, *OMG* Documents ad/97-08-02–ad/97-08-09.
13. Kolpakov, F.A., Ananko, E.A., Kolesov, G.B., and Kolchanov, N.A., GeneNet: A Gene Network Database and Its Automated Visualization, *Bioinformatics*, 1998, vol. 14, pp. 529–537.
14. Kolpakov, F.A. and Ananko, E.A., Interactive Data Input into the GeneNet Database, *Bioinformatics*, 1999, vol. 15, pp. 713–714.
15. Kolpakov, F.A. and Babenko, V.N., Computer System MGL—a Tool for Retrieving, Graphical Representation, and Analysis of Regulatory Genome Sequences, *Mol. Biol.*, 1997, vol. 31, no. 4, pp. 647–655.
16. Grant Linking Biological Databases Using CORBA, http://corba.ebi.ac.uk/CORBA_grant/.
17. Common Object Request Broker Architecture. Version 2.3, Object Management Group, *OMG* Documents formal/99-07-01–formal/99-07-28.
18. Kalinichenko, L.A. and Kogalovsky, M.R., OMG Standards: Interface Definition Language in the CORBA Architecture, *SUBD*, 1996, no. 2.