METHOD FOR INTEGRATION OF DATABASES WITH COMMON SUBJECT DOMAINS

Pisarev A.*, Blagov M., Samsonova M.

Department of Computational Biology, Center for Advanced Studies, St. Petersburg State Polytechnical University, 29 Polytechnicheskaya ul., St. Petersburg, 195251, Russia
* Corresponding author: e-mail: pisarev@spbca.ru

Keywords: database integration, natural language interface, conceptual schema, multiagent systems

Summary

Motivation: The integration of biological data from the heterogeneous data sources is one of the central problems of bioinformatics.

Results: We present a novel approach to the integrated retrieval of molecular biology data, based on application of the technology of multiagent systems and design of an adaptive natural language interface. Our approach allows to integrate any information resources which have a common subject domain. The architecture of the system ensures its portability across software/hardware platforms, high adaptivity to functional extensions and modifications, as well as the optimal distribution of query load between several database mirrors.

Availability: http://urchin.spbca.ru/NLP/NLP.htm

Introduction

One of the most widespread systems of unified access to molecular biology data is SRS (Sequence Retrieval System). (Etzold et al., 1996). In the 90s the language SYNTHESIS was created to develop the heterogeneous interoperable environments of information resources (Kalinichenko et al., 2000). The semantic integration of information resources is provided with the use of broker mechanisms and ontological specifications. As an example of implementation of this approach an objective broker for integration of information resources in the area of molecular genetics is currently being developed in the Institute of Cytology and Genetics of the SB RAS [http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/]. A similar approach to the integration of molecular biology databases is applied in the European Bioinformatics Institute (Coupay, 1999). The essential limitation of the approaches mentioned above is that they require using and supporting the actuality of definitions in the specialized languages if the model of subject domain is changed. In this paper we present a new method for integration of databases with a common subject domain. The key components of this method are the conceptual schema of knowledge domain and domain oriented dictionaries; processor of natural language (NL) queries to a database; multiagent architectute to integrate the results of information retrieval from different databases. This method was applied to integrate the information about expression of segmentation genes in fruit fly Drosophila.

Material and Methods

Databases. The genetic network which controls segmentation is one of a few genetic networks fully characterized at genetic and functional level (Ingham, 1988). The initial determination of segments is the consequence of expression of 16 genes. Spatio-temporal expression of segmentation genes is studied in details. The expression of most of segmentation genes starts at embryo stage 4 when the blastoderm is syncytial and not divided into cells and continues at embryonic stage 5. At this time the segments are determined and invagination of membranes and cell cellularization happens. Besides expression at the time of segment determination most segmentation genes are active at later developmental times, and some of them are stably expressed throughout the life of
the fly. Segmentation genes control muscle formation, neurogenesis and other developmental processes (Campos-Ortega, Hartenstein, 1985).

Currently the information about expression of segmentation genes is stored in several databases, namely FlyEx (http://urchin.spbcaas.ru/flyex) and FlyEx mirror at University of New York at Stony Brook (http://flyex.ams.sunysb.edu/flyex), Mooshka (http://urchin.spbcaas.ru/Mooshka), FlyBase ((http://flybase.bio.indiana.edu/) and In situ Database (http://www.fruitfly.org/cgi-bin/ex/insitu.pl). FlyEx and Mooshka store the information about expression of segmentation genes at the time of segment determination. FlyBase contains the information about time and place of expression of each segmentation gene through the life of the fly. In situ database, which is being developed now, stores the images of expression patterns of segmentation genes at all points of development. The determination of segments is the subject of intensive research over the last two decades and the integration of data on segmentation gene expression from the heterogeneous data sources mentioned above is an essential part of the work of researchers in both industry and academia. As the integration-by-navigation is a tedious process, we have developed the system for automatic integration of this information.

**System architecture.** Our goal was the development of technology for integration of databases, which will support the use of traditional Web browsers for information retrieval; continuous work, when new databases are added or old one are removed; failure-resistance, if the malfunction of hardware or software components happens; the optimal distribution of queries to increase the performance of the system. To meet this requirements we have developed the multiagent system, which components are user interface agents, coordinating agent, NLP agents, DB agents and JAS agents, which serve to visualize data in different formats (Fig.). The coordinating agent receives a request from a user interface agent and distributes this request among NLP agents. A NLP agent translates a query in NL in SQL and sends it to the DB agent, which in turn requests the database and transfers the result of the query back to the NLP agent. The later formats the result and sends it to the coordinating agent. Coordinating agent, if necessary, fuses results selected from different databases and transmits the resultant information to the user interface agent.

There exist many standard protocols of agent’s interaction. Most widely used are ACL (KQML/ KIF) [FIPA], CORBA [ODMG], Java RMI (Gavrilova, Horoshevsky, 2000). Till recently these standards imply direct connection between agents, what was a major hindrance to the integration of information resources with FireWall and Proxy servers. We have solved this problem by developing agents interacting via the HTTP protocol and implementing all agents in Java.
The distinctive feature of the approach used in this work is the development of the algorithm of self-organization, which ensures dynamic reconfiguration of the system and optimal distribution of queries with regard to its real load.

**NLP agent.** We use the language understanding technology based on semantic approach. This technology interprets the grammatical and lexical units of any NL into concepts of a knowledge domain (Samsonova *et al.*, 2003). These concepts are introduced in conceptual schema. The conceptual schema of the information on expression of segmentation genes in fruit fly *Drosophila* is an oriented graph, which nodes are concepts of the knowledge domain, and edges define relations between the concepts. This schema serves as a connecting link between the text of a query and a database schema. It helps the natural language processor to interpret higher or lower level concepts and synonyms, as well as equivocal and jargon terms. Moreover, making the specifications of domain knowledge explicit, the conceptual schema guides a user to learn the meaning of each term.

The procedure of processing of NL queries transforms different combinations of word forms to a limited set of terms of the logical level, which are used for a generation of the SQL queries to a database. Firstly, synonyms and high level concepts are substituted by terms of logical level, which are mapped on the database objects. At the step called ‘Search in the Dictionary’ an initial chain of semantic components is constructed. The step ‘Semantic analysis’ converts the initial chain of semantic components into a semantic network, which formally represents a query. At the ‘SQL query generation and optimization’ step the semantic network is converted into the SQL query to the database. The ‘Advanced processing of queries’ step performs the subject domain specific processing of queries, e.g., displays data in different views (as a table, graph or image).

**Implementation**

**Information retrieval.** To formulate and execute queries to the system the HTML form ‘Natural Language Interface’ is to be filled by a user. The text of a query is entered in the text field QUERY; the checkboxes allow to define databases used to retrieve information. The list QUERY EXAMPLES contains a set of predefined standard queries for convenience. By pressing the button SEND QUERY a query will be executed and after a while a result of the query will appear in a new browser window. In the upper part of this window the query in natural language is displayed, in which words used to retrieve the information from databases are shown in red. Below the query the result of retrieval of information from each database is displayed as a table. The SQL query, automatically generated by the system, is presented below the result. SQL query can be edited and returned to the server by pressing the button SEND QUERY.

Selection of the link SWITCH TO RUSSIAN calls the Russian version of the query form. The queries in Russian are submitted and executed similarly to queries in English.

**The capabilities of the NL processor.** To formulate a NL query a user can use any concept described in the conceptual scheme, type the words in any word form, formulate a query as a whole phrase or as a list of keywords, use synonyms or even laboratory jargon, combine selection criteria in a query using logical operators AND, OR, NOT. The query ‘How many...?’ allows to count rows satisfying any criterion (e.g., ‘How many embryos are scanned for expression of bcd and belong to late temporal classes?’); the query ‘Display pattern...’ returns a pattern of segmentation gene expression.

The following combinations of semantic constructions are supported: *larger than, greater than, more than, >, >=, <, less than, smaller than, from n to m, n - m.*
Conclusions

Our approach allows to integrate any information resources (published in the Internet, as well as stored locally) which have a common subject domain. Its benefits are in possibility to formulate arbitrary queries in various languages (in English and in Russian, currently), the optimal transformation of queries from natural language to SQL, as well as in opportunity to present information visually as hyperschemata. Other advantages are the simplicity in access to information and integration of new databases, adaptivity with respect to changes in a knowledge domain and user’s views, increase of the robustness of the system as well as the optimization of distribution of queries load between several database mirrors.

Acknowledgements

The support of this work by EC grant IST-1999-11009 and GAP award RMO-1323 is gratefully acknowledged.

References


