The BioRS^{TM} Integration and Retrieval System: An open system for distributed data integration

Andreas Kaps[1], Konstantin Dyshlevoi[2], Klaus Heumann[1], Ralf Jost[1], Ioannis Kontodinas[1], Martin Wolff[1], Jean Hani [1]

[1]Biomax Informatics AG, Lochhamer Str. 9, D-82152 Martinsried, Germany, www.biomax.com

[2]RingRows GmbH, Lochhamer Str. 9, D-82152 Martinsried, Germany

Summary

Modern academic and industrial research in life sciences generates huge amounts of data and information. To extract knowledge from this information space, optimized integration and retrieval software tools are essential. In the last years, a number of academic as well as commercial systems have been developed to solve this problem. However, as scientific projects are distributed at different locations (e.g., subsidiaries of companies, academic partnerships), data exchange and availability must be realized in a way that avoids data replication.

In this article, we describe a global solution for integrating distributed information by applying the BioRS^{TM} Integration and Retrieval System and its inter-BioRS communication capability that goes beyond the standard issue of local data integration. Each site integrates and maintains locally generated data using a local copy of the BioRS software. Applying the inter-BioRS communication, all available BioRS instances can communicate with each other realizing a global network of integrated databanks. All databanks integrated in this network can be accessed from any site without any data replication. This open system allows the addition of new information and sites dynamically. However, access privileges for certain databanks can be maintained on a per user and databank level ensuring data security when required.

1 Introduction

Modern academic and industrial research in life sciences generates huge amounts of data and information. This information, as well as the underlying primary data, is heterogeneous, reflecting the different research subjects and experimental techniques used by various projects. This body of data and knowledge is archived in specialized project specific databank systems. In addition, the broad variety of publicly available databanks must be interconnected with the individual data resources to realize an exhaustive reflection of current knowledge in a dedicated scientific field. These various data sources lack a common representation of information in the field of life sciences. Additionally, each databank available from the Internet provides its own mechanism for searching (e.g., [5]). Entrez ([10]) is a gateway to data sources maintained by NCBI. PubMed allows medical literature to be searched using a simple query interface. Software solutions facilitating integrated searches of heterogeneous databanks in parallel have been developed in the last ten years. Such software includes ATLAS ([9]) and the former academic, now commercially available, SRS system ([2], [3], [4]), which has also been incorporated as

a third-party retrieval system into sequence analysis packages, such as EMBOSS ([6]). Especially for searching flat-file databanks, software solutions are published (e.g., [7]).

In recent years, an increasing number of genomes are being sequenced and annotated. Concurrently, more protein and gene interaction data are being published and expression data are generated in large-scale proportions. Taken together, these lead to an explosion in the amount of data available in the life sciences. Consequently, the integration of these databanks is still a major issue ([11]): "Viewing data integration as simply an 'IT problem' underestimates the novel and serious scientific and management challenges it embodies" ([12]). All the databanks available fulfill different functions and, although they are technically and architecturally similar, no standard solution for their integration has been established, despite attempts to standardize data exchange formats which were made in the early application of bioinformatics (e.g., [8]).

Within projects, both academic and commercial, the involved partners must be able to access and process information in a consistent and transparent way. In particular, partners might be distributed around the world as different research groups or subsidiaries of companies. However, access to up-to-date information is required for realizing effective and efficient science. Data replication between all involved sites is prohibitive because of the various disadvantages including bandwidth problems, inconsistencies between replicas and the enormous amount of administrative work required. Maintenance of databanks must be performed where the data are generated. Publicly available databanks must be updated, ideally incrementally, and incorporated into the productive system in a way that avoids downtime. Especially in a global setting of groups working in different time zones, time outside of working hours for performing tasks such as system update may be reduced.

All available retrieval and integration systems focus on the local integration of databases. In these solutions, all databases have to be copied to a central instance for integration. Such systems lack homogeneous communication and integration of multiple instances of the retrieval software at different sites for accessing integrated databanks remotely and transparently.

Therefore, we developed a communication mechanism that allows import of databanks which are integrated by a remote instance of our BioRS Integration and Retrieval System without the need to duplicate any underlying data source. By applying this strategy, individual research groups still control their data and application space of their projects and resolve their data integration issues locally. However, by importing remote BioRS databanks, a network of communicating retrieval systems can be established. The amount of information available to local users can be increased dramatically with a minimal effort concerning data integration maintenance tasks. Data duplication is avoided completely.

2 The BioRS Integration and Retrieval System

2.1 Introduction

The BioRS Integration and Retrieval System is a data retrieval system that allows the integration of relational and flat-file databanks into a common and homogeneous environment. The databanks to be integrated, both public and proprietary, are organized differently according to storage (flat-file vs. relational) and format (e.g., EMBL, XML, etc.).

The BioRS software allows rapid retrieval of data (e.g., sequence, literature, information about structure) from multiple databanks in parallel. By using convenient Web-based forms, searches can be as simple or complicated as necessary. A subquery option is provided for refining search results. The BioRS system also supports searching with phrases and search-term synonyms. For example, a thesaurus of gene names might contain all corresponding alternative names for each gene. When searching a databank for a specific gene name, entries containing synonyms of the gene name will also be retrieved.

The retrieval function can be incorporated into proprietary programs or scripts allowing transparent access of entries in databanks. By mapping semantically equivalent elements of different databanks to the same BioRS element that is used for retrieval, the same information entities can be found in all integrated databanks in parallel. Cross-references between related information in different databanks ensure convenient access to all available information which can be also searched. Downloading relevant results allows further processing. This enables the system to be integrated into local workflows of processing steps requiring a data access layer to the huge information space beyond.

2.2 Technical Architecture

The BioRS software is based on a client-server architecture and provides convenient Web-based interfaces (HTML/http(s)) for both users searching the databanks and administrators managing the databanks. In addition to the Web interface, a command-line interface is provided for most search and administrative functions allowing incorporation into proprietary scripts or programs. This flexible client-server architecture allows the incremental support of additional upcoming technologies, such as Web services (SOAP).

The retrieval system is supplemented by an update tool that automates the tasks of checking and downloading new versions of databanks via the Internet (e.g., by using FTP).

The BioRS system is built on object-oriented principles. To achieve distributed processing, which is required to handle the huge amount of data, an object-oriented middleware is applied. Every object is constructed according to the Common Object Request Broker Architecture (CORBA(R)[14]) and interfaces are specified in the Interface Definition Language (IDL(TM)[14]), both well-established industry standards.

This architecture ensures that the system is extensible and scaleable and addresses increasing requirements concerning data handling and processing. Index generation as well as query processing is distributed across a flexible number of machines/processors that have been made available for the system. By integrating additional computer power, the performance of the system can be increased dynamically.

2.3 Integrating flat-file databanks

Flat-file databanks consist of one or more plain text files that organize the biological information contained in these files into independent units called entries. Information in each entry is further organized into a set of named fields which contain specific biological information, such as names of organisms or protein sequences. To define a BioRS databank format for a flat-file databank, a format-specification file must be written and compiled by applying the standard

tools *flex* and *bison*. The format-specification file contains the parsing rules and instructions for extracting and representing data from entry fields. In particular, the format-specification file defines XML paths for elements, sub-elements and attributes corresponding to relevant fields in the flat-files. Format-specification files are written in the BioRS Format-Specification Language. Defined elements can be then indexed to increase search performance, i.e., all values occurring in a defined element are indexed to realize a full-text index. The generated index files are stored on the file system where the BioRS server is installed.

The BioRS system suggests a list of global elements, i.e., element names that have been proven to be meaningful and can be used by biologists for formulating a query. The administrator, who integrates databanks, can map these elements from the BioRS system to elements from the corresponding database. This strategy allow semantically identical entities that are named differently in different databanks to be handled in a homogeneous way. The use of global elements reduces the arbitrariness of element naming by creating a unified naming context for elements used in multiple databanks. However, this mapping is not limited to the elements provided by the BioRS system but can be enhanced dynamically by the administrator, allowing optimal integration according to the particular data source. Each user can use personal setting to display all elements of all databanks or to display only the intersection of elements of all selected databanks. This option can be set independently for querying and displaying results.

Since data in flat-file databanks are provided in a variety of formats, the BioRS system creates an internal representation of databanks using a consistent XML-based structure. This internal representation is generated on the fly and is not stored persistently. Thus, original data sources are not copied to a centralized data storage system. All data sources that are integrated are accessed in the original format.

XML is being used with increasing frequency to represent biological information appropriately. XML-based databanks can be integrated very easily into the BioRS system. However, XML is not a human-friendly format for viewing results, such as a query hit. Consequently, the BioRS system allows user-defined views for data formats that are incorporated into the system. More than one view can be defined for a databank format, allowing information to be presented according to the different needs of the users.

2.4 Nested Properties

Information properties in entries of some databanks are organized within a hierarchical structure. The sub-structured properties within a databank are stored as “nested databanks” within the BioRS system. For example, the “Features” field of GenBank entries contains one or more “FtKey” sections. Each “FtKey” section contains one or more “FtQualifier” sections. Entries in the nested databanks are linked to the parent entry. Multiple entries in nested databanks can be derived from a single entry of the parent databank. Therefore, related information that falls directly under the scope of a nested entry can be searched. When the feature key and feature qualifier elements are indexed as part of a nested databank, the BioRS system is able to properly maintain the hierarchical relationship.

2.5 Integrating relational databanks

Not all retrieval systems that are available support the integration of relational databanks. However, by the means of relations, biological or medical networks and dependencies can be modeled more easily in relational databanks than in flat-files. The BioRS system, therefore, supports the integration of relational databank management systems (RDMS).

However, the task of integrating a relational databank into the BioRS system requires that the administrator has a full understanding of the relational databank scheme. The choice of tables and columns to be integrated into the BioRS system depends both on how information will be presented to the user and what kind of relations exist between the chosen tables. As described for flat-file databanks, entities from the underlying databank are mapped to BioRS elements used for querying.

Using the BioRS system, administrators can decide to either access relational databanks to search for information directly by using the database system internal indices or use pre-computed BioRS indices. Both approaches have advantages and disadvantages. When BioRS indices are used, searching speed is improved and cross-references between relational databanks and other integrated databanks can be created. When RDBMS indices are used, access to up-to-date information is ensured. Of course, flat-file and relational databanks can be searched in parallel with a single query.

2.6 Integrating external applications

In addition to presenting search results within the context of the BioRS system, hits can be easily delegated to third-party products for presentation of results in the context of the application that generated the data returned by the query. For example, databanks generated by the Pedant-Pro program for exhaustive sequence analysis can be integrated ([13]). Proteins found using the BioRS system in a Pedant-Pro databank can be viewed within the context of the Pedant-Pro software. The focus of the user is thereby moved from retrieval to the sequence analysis tool allowing further exploration and navigation through annotation provided by the Pedant-Pro system.

3 Global data integration with inter-BioRS communication

3.1 Novelty of the inter-BioRS approach

The functionality of the BioRS system as described in the previous section addresses typical local data integration problems that are solved also by other software products (e.g. SRS). However, in a global environment (e.g., in a scientific project of numerous groups distributed around the world or in a company having numerous subsidiaries realizing one company-wide research group), data integration must be addressed at a global level. By solving the data integration problems only at the local level, the global and integrated aspect of collaborative work is lost.

One solution to this problem could be the replication of relevant information at sites for all partners or colleagues who need access to the information resource or the establishment of a central

data storage system where data from all subgroups is collected. Such a data-replication strategy presents numerous disadvantages. Databanks, which can be huge in size, must be distributed over computer networks having a limited bandwidth. Data replication requires time and leads to inconsistencies between the data source and the replication site. Although data replication can be automated, manual interaction is required to handle error situations adequately (e.g., network problems, downtime, etc.).

A superior strategy is the establishment of local retrieval systems that are connected with each other to form a network of retrieval systems which allows access to all data sources without data replication.

3.2 Technical realization a network of BioRS instances

The BioRS system allows a local BioRS instance to search databanks administered by remote BioRS instances. This feature increases convenience for situations where disk space and/or network bandwidth are limited or when the overall setup of the working environment requires databank sets to be stored at different locations.

To set up such a network of communicating BioRS instances, all groups or sites involved must have a local instance of the BioRS system. All the locally generated databanks are integrated into the local BioRS instance. In addition, administration of integration and update of these databanks is performed locally. Local users can access databanks at remote sites after they have been imported from the remotely installed BioRS instances into the local BioRS system.

Databanks can be imported by a remote BioRS instance only if the providing BioRS instance explicitly publishes this particular databank. This ensures that only databanks that are to be seen by other BioRS instances are available. This security management is performed by all local BioRS administrators.

When importing a remote databank into a local BioRS system, neither the databank nor the databank indices must be copied. Only the databank format specification, which contain all definitions integrating databank entries, is imported.

Databank import requires access to the remote host on which the remote BioRS instance is running. In particular, databanks from remote BioRS instances running on different operating systems can be imported to a local instance, i.e., this import functionality is platform-independent.

The BioRS system handles access to remote and local databanks transparently. Users working with a local instance of the BioRS system do not have to be aware of the location of any given databank. From the user's point of view, local and imported databanks are selected and searched homogeneously. Technically, parts of a query that address a remote databank are forwarded from the local BioRS instance to the remote BioRS instance where the corresponding databank is integrated. Queries are executed in a distributed fashion. The results from all involved instances are collected by the local instance and are presented as a single result set to the user. The data transfer between the communicating BioRS instances is thereby minimized. When a unspecific query results in thousands of hits from a remote databank, only the information about the number of hits and a specified number of results (e.g., 10) are transferred. The network load is decreased and the query performance is increased. With this information, the user has the possibility to perform a sub-query (or sub-queries) to minimize the result set before transferring the data across the network.

By applying this scenario, up-to-date information is always provided globally to all users within this network of BioRS instances and query executing power is distributed over all the computing resources of this BioRS network. With respect to data management for a networked retrieval system, one strategy would be to dedicate a single BioRS instance to integrating and maintaining all publicly available databanks that are of interest. All other instances focus on proprietary databanks. By interconnecting all resources together, users at all sites have immediate access to both public and proprietary databanks.

4 Outlook

Results retrieved by applying the BioRS software can be examined in detail by the user. By improving the querying possibilities with fuzzy searches and by going deeper into semantics by applying computer linguistics approaches, a more association-based retrieval can be realized for the user. The user must be aware of fewer details concerning the content of the databanks and more benefit can be obtained by using such a homogeneous gate into the globally unstructured and heterogeneous life sciences information space.

5 Acknowledgement

The authors like to thank Shannon Frances for carefully reviewing the manuscript.

References

[1] Etzold, T., Argos, P.: Transforming a set of biological flat file libraries to a fast access network. Comput. Appl. Biosci. **9** (1993) 59–64

[2] Etzold, T., Ulyanov, A., Argos, P.: SRS: information retrieval system for molecular biology data banks. Meth. Enzymol. **266** (1996) 114–128

[3] Zdobnov, E., Lopez, R., Apweiler, R., Etzold, T.: The EBI SRS server - recent developments. Bioinformatics **18** (2002) 368–373

[4] Zdobnov, E., Lopez, R., Apweiler, R., Etzold, T.: The EBI SRS server - new features. Bioinformatics **18** (2002) 1149–1150

[5] Hubbard, T., et al.: The Ensembl genome database project. Nucleic Acids Res. **30** (2002) 38–41

[6] Rice, P., Bleasby, A., Longdon, I., Williams, G., Curven, W.: EMBOSS

[7] Ramu, C.: SIR: a simple indexing and retrieval system for biological flat file databases. Bioinformatics **17** (2001) 756–758

[8] George, D., Mewes, H.-W., Kihara, H.: A standardized format for sequence data exchange. Protein Seq. Data Anal. **1** (1987) 27–39

[9] George, D., Barker, W., Mewes, H.-W., Pfeiffer, F., Tsugita, A.: The PIR-International protein sequence database. Nucleic Acids Res. **24** (1996) 17–20

[10] Schuler, G., Epstein, J., Ohkawa, H., Kans, J.: Entrez: molecular biology database and retrieval system. Meth. Enzymol. **266** (1996) 141–162

[11] Stein, L.: Integrating Biological Databases. Nature Reviews Genetics **4** (2003) 337–345

[12] Searls, D.: Data Integration: Challenges for Drug Discovery. Nature Reviews Drug Discovery **4** (2005) 45–58

[13] Riley, L., Schmidt, Th., Wagner, Ch., Mewes, H.-W., Frishman, D.: The PEDANT genome database in 2005. Nucleic Acids Res. **33** (2005) D308–D310

[14] http://www.corba.org