Dynamic information fusion for genome annotation

Heiko Müller, Peter Rieger, Katja Tham, Johann-Christoph Freytag

Institut für Informatik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
{hmueller, rieger, tham, freytag}@dbis.informatik.hu-berlin.de

Abstract: The need for demand-driven and scaleable integration of heterogeneous data sources is inherent to the process of genome annotation. In this paper we describe the Gene-EYe architecture for data integration and analysis, presenting a platform for dynamic fusion of genome data. The architecture allows uniform access to a large amount of heterogeneous data sources and provides domain specific analysis tools. As genome data is known to be erroneous the ability for data cleansing is a natural part of the annotation process described. Gene-EYe also supports the task of keeping derived annotation data up to date in case of changes in the original data sources.

1 Introduction

Dynamic information fusion is an integral part of genome annotation!

While the genomes of many organisms have been sequenced the next step is to transform the raw data into knowledge. Genome annotation is the process of assigning meaning to sequence data by identifying regions of interest and determine biological function for them. Genome annotation is highly explorative due to incomplete or fuzzy domain knowledge. Often different methods exist to derive a certain annotation. The process therefore cannot be specified completely in advance. It is composed of iteration cycles, which involve data integration and analysis. An example for genome annotation is gene prediction, i.e., given a DNA sequence identify the coding regions (genes). There exist several different methods for gene prediction requiring different data sources. Examples are aligning the sequence with other DNA or protein data [MS+02]. None of these methods is know to be 100% complete and correct. The user decides which of the methods to use, evaluates the results, and decides further steps based on these evaluations.

The process of genome annotation resembles a process of dynamic information fusion. Compared to classical data warehouse approaches the data is not integrated within a global schema for the execution of analytical queries. Instead, we repetitively perform information fusion without the existence of a fixed global schema. Starting with an initial set of sequences and annotations we integrate and analyze them to derive new infor-

mation. The derived data in turn is used within further integration and analysis steps including additional data sources. This is repeated until annotations can be derived based on sufficient evidences using the analysis results. The choice of the next analysis step often depends on the outcome of the previous steps and the quality of their results. We call the iterative process of genome annotation *vertical dynamic information fusion*. This is to be distinguished from the execution of parallel annotation tasks resulting in different annotation data set, which we call *horizontal dynamic information fusion*.

The Gene-EYe architecture presents a platform for dynamic information fusion for genome annotation. It supports (i) the iterative nature of the annotation process, (ii) uniform data access using a standard data model, (iii) selection of sources based on data quality information, (iv) inclusion of data cleansing operations, and (v) ensuring correctness and timeliness of the derived annotation data. Gene-EYe follows a data warehouse approach in physically materializing each of the data sources under the control of a relational database management system. For genome annotation integration and transformation of these data sets is performed utilizing a set of domain specific functions available within Gene-EYe.

The paper is structured as follows: In Section 2 we describe the Gene-EYe architecture and define the dynamic information fusion process for genome annotation. Section 3 gives an overview about how we handle data quality and data cleansing to achieve highly reliable and up-to-date genome annotations. Section 4 outlines the related work. We conclude in Section 5 and give a brief outlook on future work.

2 Gene-EYe: A platform for dynamic information fusion in genome annotation

The Gene-EYe architecture (Figure 1) consists of three layers and is accompanied by a meta data repository supporting data transformation and integration, documentation, and assurance of high quality genome annotations. It resembles a systematic approach by solving orthogonal problems independently, i.e., syntactic data integration (DATA), semantic data integration (CONTENT), and data analysis (KNOWLEDGE). The layers also reflect the main activities of information fusion, i.e., data access, integration, and analysis [CSS99]. Gene-EYe supports the two different processes of physical data integration (ETL) and genome annotation.

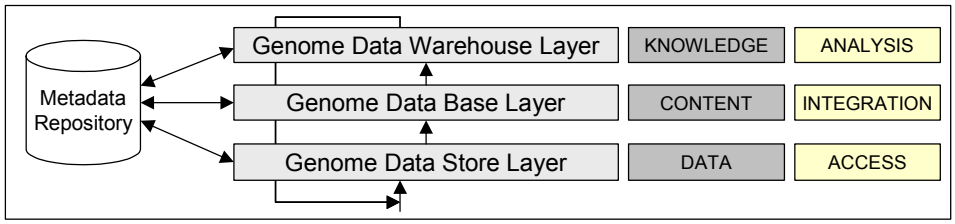

Figure 1: The Gene-EYe architecture

The Genome Data Store Layer (GDS) allows physical integration of heterogeneous data sources within a single relational database. This solves the problem of syntactic heterogeneities, i.e., different data models and access methods. The Genome Data Base Layer (GDB) allows semantic integration among different data sources. It allows combination of data coming from different data sources into data stored in relations that represent "biologically meaningful" objects. While the GDS leaves the source in a schema derived from the original flat file format the GDB performs integration into user-defined schemas. The Genome Data Warehouse Layer (GDW) allows the user to analyze and to process biological entities as represented in the second layer by various tools and programs. Several common algorithms for genome annotation have been implemented within Gene-EYe as User Defined Functions of the database management system.

2.1 The ETL process

Meta data plays a key role for initiating, controlling, and documenting the ETL process and provides a simple mechanism for adding new data sources. We currently focus solely on data sources available in a flat file format. This is a common practice for the sources within the domain of life science. For each data source there are three mayor aspects that have to be described: (i) the characterization of the data provision mechanism, (ii) the assignment of local resources to a data source, and (iii) the characterization of the data source itself, i.e., the file format and the relational mapping of the source. One of the main goals for the extensive management of meta data is to support the most important activities during the ETL process. The supported activities are (i) creating DDL statements from format description for the integration of the source in the data store layer, (ii) generating parser components for preparing the data for loading, (iii) generating administration scripts, and (iv) detecting and reporting errors to allow the application of appropriate data cleansing methods.

2.2 The process of vertical dynamic information fusion

The process of vertical dynamic information fusion (Figure 2) is a sequence of passes through each of the architectural layers. Within each iteration cycle we select the sources of interest (GDS), integrate them (GDB), perform analysis on the resulting data set (GDW), and materialize the results for evaluation and further processing. Each of the cycles is called an annotation step. Initially the system contains a set of annotation sources $Q = \{q_1, \ldots, q_n\}$ and operations for data analysis and transformation (called actions) $A = \{a_1, \ldots, a_m\}$. Within each annotation step we select a subset of the sources $Q^* \subseteq Q$ and an action $a_i \in A$. Each action specifies the required schema for the input data as well as the schema of the results. Using relational schema mappings we map the data in Q^* to a temporary data set q_{temp} following the input schema of a_i. The results q_{result} are materialized as intermediate sources $q_{intermediate}$ and added to Q for further utilization. The process therefore creates a sequence of intermediate data sets on the way to the final results q_{final}.

The resulting annotation process is a sequence of annotation steps, each being composed of a set of data sources Q_i^*, an action a_i, and a relational mapping m_i. This information is

named the annotation lineage L = (<Q_1*, a_1, m_1>, …, <Q_n*, a_n, m_n>). We mainly utilize this information for re-annotation to keep the derived data up-to-data (see next section).

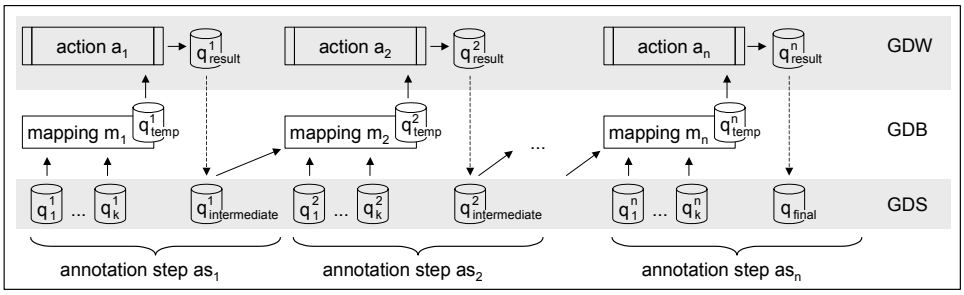

Figure 2: The process of vertical dynamic information fusion for genome annotation in Gene-EYe

3 Achieving high quality for derived data

The existence of errors and inconsistencies in genome data is a well-known fact as production of errors is inherent to the process of genome data production [MNF03]. We therefore have to handle sources or hotspots of poor data quality and perform data cleansing while integrating and analyzing the data.

After transforming and loading the sources into our relational database we specify and check integrity constraints on the data. These constraints reflect domain knowledge about genome biology. Values or tuples violating constraints are flagged. The amount of constraint violations is used as a measure of accuracy of the data source. This information is used when selecting data sources, i.e., the user is enabled to discard sources or tuples within the information fusion process. In [M03] we show the possibility of correcting annotation errors. This also requires fusion of genome data and is therefore performed as an annotation process in Gene-EYe.

When data in external sources is updated these changes might affect derived annotation data. Using the process lineage L we are able to detect the annotations depending on the data sources updated to avoid stale data. The process lineage then enables re-execution of the annotation process. In collecting sufficient metadata about the cleansing process we can also speed-up the necessary cleansing processes. Up to the point where the updated source is used first we can automatically redo the annotation process. Throughout re-execution we are able to check for differences between the new and old derived data. If there are no differences we can proceed automatically. Statistics on update frequencies of data sources can be used in deciding on materializing intermediate results to avoid the necessity for re-execution on future source updates. This makes sense if the sources involved in annotation up to this point do not change frequently. While re-executing, the user might also decide at one point due to the new data to change the annotation process in going new ways or skipping some process steps leading to a new annotation process.

4 Related Work

The area of integration and analysis of genome data has been explored for many years. The existing approaches either focus solely on data integration or constitute proprietary systems driven by particular biological problems, utilizing only a small subset of the available data sources and analysis tools. Data integration approaches have been reviewed and described in [LR03]. Recently, general and extensible genome annotation systems have been developed [MG+03][HR+03]. These systems allow uniform and integrated access to a multitude of data sources and analysis tools. They do not explicitly support the explorative character of genome annotation and have only very limited support for data cleansing. They are also more restrictive in the type of data sources available. Existing genome annotation systems generally lack support for updating derived data in case of data changes in the sources. This shortfall has lead to a large amount of outdated data and the initiation of several re-annotation projects [OK02].

5 Conclusion and Outlook

Dynamic information fusion is an integral part of genome annotation. We presented a high-level view on the Gene-EYe architecture which supports the two main task of physical data integration and genome data annotation with high accuracy and timeliness. The architecture is currently implemented using the IBM DB2™ relational database management system. In future work we plan to accelerate re-annotation by reducing the number of operations that are re-executed. We also intent on adding the ability for user assistance, i.e., propose the user alternatives for the next step within the annotation process based on completed annotation processes and the history of the current process.

Literature

[CSS99] Conrad, S., Saake, G., Sattler, K.: Informationsfusion – Herausforderungen an die Datenbanktechnologie. BTW'99, Freiburg, Germany, 1999

[HR+03] Hoon, S., Ratnapu, K.K., Chia, J., Kumarasamy, B., Juguang, X., Clamp, M., Stabenau, A., Potter, S., Clarke, L., Stupka, E.: Biopipe: A Flexible Framework for Protocol-Based Bioinformatics Analysis, Genome Research, Vol. 12, 2003

[LR03] Leser, U., Rieger, P.: Integration molekularbiologischer Daten, Datenbankspektrum Ausgabe 6, 2003

[MG+03] Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., Pühler, A.: GenDB – an open source genome annotation system for prokaryote genomes, Nucl. Acid Res., Vol. 31, No. 8, 2003

[M03] Müller, H.: Semantic Data Cleansing in Genome Databases, Proc. of the VLDB 2003 PhD Workshop, Berlin, Germany, September 12-13, 2003

[MNF03] Müller, H., Naumann, F., Freytag, J.-C.: Data Quality in Genome Databases, Proceedings of the Conference on Information Quality (IQ 03), Boston, October 2003

[MS+02] Mathé, C., Sagot, M.-F., Schiex, T., Rouzé, P.: Current methods of gene prediction, their strength and weaknesses, Nucleic Acid Research, Vol. 30, No. 19, 2002

[OK02] C.A. Ouzounis, P.D. Karp, *The past, present and future of genome-wide re-annotation*, Genome Biology, Vol. 3, No. 2, 2002