

# Web-Based Genomic Information Integration with Gene Ontology

Kai Xu<sup>1</sup>

IMAGEN group, National ICT Australia, Sydney, Australia,  
kai.xu@nicta.com.au

**Abstract.** Despite the dramatic growth of online genomic data, their understanding is still in early stage. To have meaningful interpretation, it requires the integration of various types of data that are relevant, such as gene sequence, protein interaction and metabolic pathway. *Gene Ontology*, an ontology proposed by molecular biologist community, is a possible tool to help address some of the difficulties in such integration. In this paper, we exam the formality of Gene Ontology, and study the possibilities and potential problems in applying Gene Ontology to both structured (such as database) and semi-structured (such as the publications in the literature) data integration for online genomic information.

## 1 Introduction

The amount of genomic information available online increases dramatically in the last a few year. For instance, the number of DNA base pair in GenBank ([www.ncbi.nlm.nih.gov/genbank/](http://www.ncbi.nlm.nih.gov/genbank/)) increase from 680,338 in 1982 to 44,575,745,176 in 2004. Despite the abundance of data, the understanding of these data lags far behind the collection. A key question that molecular biologists try to understand is the gene regulation mechanism, i.e., why the variation in gene sequence can lead to diseases such as cancer. The underlying principle of the answer to this question is the “central dogma of molecular biology”: the genetic information stored in DNA sequence is passed through RNA to protein, which eventually performs the encoded regulation function by its interaction with its environment. The central dogma implies that any type of genomic data alone, such as the DNA sequence data, is not sufficient for meaningful interpretation of its genetic function; this can only be achieved through the integration of relevant genomic data such as DNA sequence, protein interaction and metabolic pathway, which are all publicly available online.

Data integration has been a challenging problem studied by computer scientist for years due to the prevalence of heterogeneity. The integration of genomic data has additional difficulties such as the requirement of biological expertise. Merging various types of genomic data poses a even greater challenge due to the complexity introduced by the variety of data and their context. Some attempts have been made recently, but few of them are successful [1]. The *Gene Ontology* [2], an recent collaboration within the molecular biologist community, tries to alleviate the semantic heterogeneity in genomic data representation by providing a

shared vocabulary. Though still in its early stage, we believe the Gene Ontology can play an important role in genomic information integration besides facilitating the communication between molecular biologists. An compelling property of the Gene Ontology is that by building an ontology the genomic expertise is captured in the concept definitions and the relationships among them in a more machine-friendly form. In this paper, we discuss the possibilities and challenges of employing the Gene Ontology to address some issues in integrating online genomic information, including both structured data (such as database) and semi-structured data (such as the publications in the literature). This study is not meant to provide a complete solution but rather a pilot study of the feasibility of applying Gene Ontology to data integration.

## 2 Background

### 2.1 Gene Ontology

The goal of the Gene Ontology is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing [2]. It is maintained by the “Gene Ontology Consortium” ([www.geneontology.org](http://www.geneontology.org)). The Gene Ontology includes not only terms and relations among them, but also the associations between the terms and gene products in online databases.

### 2.2 Web-Based Biological Information Integration

There are two type of approaches commonly used for web-based biological information integration, centralized (or warehouse) integration and distributed (or mediator-based) integration. Centralized integration duplicates data from multiple sources and stores them in a data warehouse. All queries are then executed locally rather than in the actual sources. An example of such system is GUS (Genomics Unified Schema) [3]. The centralized integration approach relies less on the network to access the data, and using materialized warehouses also allows for an improved efficiency of query optimization. This approach however has an important and costly drawback that it must regularly check throughout the underlying sources for new or updated data and then reflect those modifications on the local data copy.

Unlike the centralized/warehouse approach, none of the data is cached locally in a distributed/mediator-based integration. Instead a query is reformulated by the mediator at runtime into queries on the local schema of the underlying data sources. Example of such system is the DiscoveryLink [4]. The two main approaches for establishing the mapping between each source schema and the global schema are global-as-view (GAV) and local-as-view (LAV) [5]. The LAV is considered to be much more appropriate for large scale ad-hoc integration because of the low impact changes to the information sources have on the system maintenance, while GAV is preferred when the set of sources being integrated is known and stable.

## 3 Genomic Information Integration with Gene Ontology

### 3.1 The Formality of Gene Ontology

A commonly quoted definition for ontology is “a formal, explicit specification of a shared conceptualization” [6]. Therefore, an ontology should have:

1. A vocabulary of terms that refer to the things of interest in a given domain;
2. Some specification of meaning for the terms, (ideally) grounded in some form of logic.

However, as stated by the creators of the Gene Ontology [7], they have consciously chosen to begin at the most basic level, by creating and agreeing on shared semantic concepts; that is, by defining the words that are required to describe particular domains of biology. They are aware that this is an incomplete solution, but believe that it is a necessary first step. The argument is that these common concepts are immediately useful and can be used ultimately as foundation to describe the domain of biology more fully. In this sense, Gene Ontology is still an incomplete ontology. It has a defined vocabulary (requirement 1 previously), but is lack of formal specification (requirement 2). This can affect the application of Gene Ontology to data integration and is discussed in detail in Section 3.2 and 3.3.

### 3.2 Structured-Data Integration

For structured-data, we focus on online databases. Generally, there are two types of heterogeneity in database integration: the structural and semantic heterogeneity. Three frameworks are available when using ontology to address the semantic heterogeneity: single ontology model, multiple ontology model and hybrid ontology model [8]. The hybrid ontology model has a high-level vocabulary that is shared by the ontologies of participating databases. The semantic mapping between databases is done by the following transformation:

$$\textit{local ontology 1} \rightarrow \textit{shared vocabulary} \rightarrow \textit{local ontology 2}$$

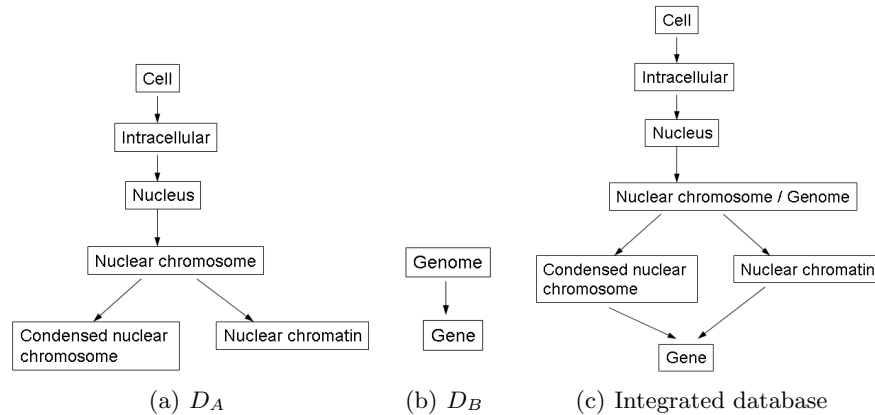
This eliminates the mapping between every pair of local ontologies. Instead, only the mappings between the shared vocabulary and the local ontologies are required. The Gene Ontology fits well into the hybrid ontology model because:

- It is a high-level vocabulary that is well-defined
- It is general enough for extension and thus easy for databases to adopt.

In fact the semantic heterogeneity is much less severe for databases that choose to annotate their data with Gene Ontology because essentially they are following the same ontology by doing this (the single ontology model). For other genomic databases that have their own ontology, the Gene Ontology can be served as the global shared vocabulary to help build ontology mappings between these

databases and with databases already integrated. In this sense, the Gene Ontology provides a nice tool for semantic integration among genomic databases.

Though devised mainly for semantic heterogeneity, ontology can also contribute to resolve structural heterogeneity due to the fact that ontology and database schema are closely related [9]. One of the similarities is the considerable overlap in expressivity, which includes objects, properties, aggregation, generalization, set-valued properties, and constraints. For example, entities in an ER model correspond to concepts or classes in ontologies, and attributes and relations in an ER model correspond to relations or properties in most ontology languages. For both, there is a vocabulary of terms with natural language definitions. Such definitions are in separate data dictionaries for database schema, and are inline comments in ontologies. Arguably, there is little or no obvious essential difference between a language used for building database schema and one for building ontologies. The similarity between ontology and schema can be used for resolving structural heterogeneity in database integration. Here we use an example to illustrate deriving the global schema of the integrated system from the mapping between the database ontologies. The two databases in this example are referred as  $D_A$  and  $D_B$ . Their ontologies are shown in Figure 1(a) and 1(b) respectively.  $D_A$  follows Gene Ontology, whereas  $D_B$  does not. For simplicity



**Fig. 1.** Global schema derivation based on ontology mapping

we assume their ER models have the same structure as the ontology; thus the global schema can be represented as the integration of two database ontologies. With term definitions, it is possible to build a mapping between the ontologies of  $D_A$  and  $D_B$ . It is easy to see that "nuclear chromosome" and "genome" both refer to the entire DNA sequence, therefore they are semantically equivalent. When two terms are referring to the same concept, then one global term can represent both in the integrated ontology, i.e., the global term definition subsumes the two local term definitions. In this example "nuclear chromosome" and

”genome” will be represented by one term (for instance ”nuclear chromosome / genome”) in the integrated ontology. If two terms refer to two concepts in a specialization relation, then such relation should be kept in the integrated ontology. In this example, ”condensed nuclear chromosome” (a highly compacted molecule of DNA and associated proteins resulting in a cytologically distinct structure that remains in the nucleus) and ”nuclear chromatin” (the ordered and organized complex of DNA and protein that forms the chromosome in the nucleus) are two specializations of ”nuclear chromosome”. Such relation should be kept between these two terms and the new term replacing ”nuclear chromosome” in the integrated ontology. This general rule also applies to other relations in the ontology, such as the ”part-of” relation between ”genome” and ”gene” in database  $D_B$  ontology. Therefore the integrated ontology is as the one shown in Figure 1(c). Based on our assumption, the global schema can be derived from the integrated ontology in a straightforward manner.

### 3.3 Semi-structured Data Integration

Besides the structured genomic data stored in various online databases, there are also large amount of semi-structured data, which mainly includes the publications in the literature. The knowledge in such publications can be made structural (more understandable by machine) by extracting and annotating them. However, such task requires significant effort and usually requires biological expertise. The major hurdle for any algorithm to perform such task is its inability to understand the semantics of previous research and based on them derive new knowledge from publications.

Theoretically it is possible to use ontology to capture the semantics of publications, but is almost impossible in practice. For structured data, the semantics of a database can be captured by mapping its ontology to some known one. It is much more difficult to build such mapping for semi-structured data. First, semi-structured data hardly have its ontology available at all. Second, the number of such ontologies can be prohibitive because every paper could have its own ontology. Therefore, it is too early to discuss using Gene Ontology to map publication semantics. However, we think it is feasible to use Gene Ontology as a prior knowledge so the algorithm can understand the literature in a deeper semantic level. For example, given the knowledge that gene MCM2 is associated with molecular function ”chromatin binding”, an algorithm can ”guess” that the paper is relevant to ”chromatin binding” when it finds gene MCM2 in it.

Such semantic reasoning may not be able to extract meaningful (to human) knowledge from publication yet, but it can improve the effectiveness of existing machine learning / data mining algorithms for genomic knowledge discovery. For instance, the algorithm may be able to distinguish the type of data used in a paper, whether it is alphanumeric, DNA sequence, or interaction network, without understanding every specific data type. Such capability is also important when integrating online services. The possible better understanding of the service functionalities, which is usually semi-structured data, can lead to better query execution planning for integrated systems that adopt distributed/mediator-based

approach. This also provides possible solutions for the data redundancy and inconsistency. By understanding the data semantics, the algorithm can identify duplications in databases. Such information with extra knowledge regarding data quality, it is also possible for algorithm to recognize correct data from contradictory copies. All these semantic reasonings heavily rely on the formal representation of ontology, which is still lack in the current Gene Ontology. Without such formal specification, it is difficult for algorithm to follow the knowledge in Gene Ontology and use them to derive new knowledge from online resources.

## 4 Conclusions

In this paper, we study the feasibility of using Gene Ontology to facilitate online genomic information integration. Our findings confirm that Gene Ontology is a valuable tool to resolve semantic and sometimes structural heterogeneity in database integration. The potential of applying Gene Ontology to tasks in semi-structured data integration, such as automatic literature knowledge discovery and data inconsistency resolving, is limited by its lack of formal specification. With the development of the Gene Ontology, we believe it will play an vital role in genomic information integration and understanding.

## References

1. Hernandez, T., Kambhampati, S.: Integration of biological sources: current systems and challenges ahead. *ACM SIGMOD Record* **33** (2004) 51–60
2. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. *Nature Genetics* **25** (2000) 25–29
3. Davidson, S.B., Crabtree, J., Brunk, B.P., Schug, J., Tannen, V., Overton, G.C., C. J. Stoeckert, J.: K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. *IBM Systems Journal* **40** (2001) 512–531
4. Haas, L.M., Kodali, P., Rice, J.E., Schwarz, P.M., Swope, W.C.: Integrating life sciences data-with a little garlic. In: *Proceedings of the 1st IEEE International Symposium on Bioinformatics and Biomedical Engineering*. (2000) 5–12
5. Lenzerini, M.: Data integration: a theoretical perspective. In: *Symposium on Principles of Database Systems*. (2002) 233 – 246
6. Gruber, T.: A translation approach to portable ontology specifications. *Knowledge Acquisition* **5** (1993) 199–220
7. The Gene Ontology Consortium: Creating the gene ontology resource: Design and implementation. *Genome Research* **11** (2001) 1425–1433
8. Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hubner, S.: Ontology-based integration of information - a survey of existing approaches. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01) Workshop: Ontologies and Information Sharing*. (2001) 108–117
9. Uschold, M., Gruninger, M.: Ontologies and semantics for seamless connectivity. *SIGMOD Record* **33(4)** (2004) 58–64