A semantic web approach to integrating heterogeneous yeast genome data

Kei-Hoi Cheung[1,4], Andrew Smith[2], Kevin Y. Yip[2], Michael Seringhaus[3]
Shawn M. Douglas[3], Mark Gerstein[3]

[1]Center for Medical Informatics, [2]Computer Science Department
[3]Molecular Biophysics and Biochemistry Department, [4]Genetics Department
Yale University, New Haven, Connecticut, USA

Email addresses
kei.cheung@yale.edu, andrew.smith@yale.edu, yuklap.yip@yale.edu,
michael.seringhaus@yale.edu, shawn.douglas@yale.edu, mark.gerstein@yale.edu

Introduction

The budding yeast *Saccharomyces cerevisiae* was the first fully sequenced eukaryotic genome [1]. Given its ease of genetic manipulation and the fact that many of its genes are strikingly similar to human genes, the yeast genome has been characterized extensively by a wide range of biological experiments such as DNA microarrays and SAGE [2]. A large quantity of such experimental data have been published and distributed in disparate formats through numerous web-accessible databases including MIPS (http://mips.gsf.de/genre/proj/yeast/index.jsp) and SGD (http://www.yeastgenome.org). While each of these databases serves as a valuable and unique resource that meets some specific research needs, a broader need can be served if the data provided by these resources can be mined or analyzed in an integrated way. For example, the reason that a particular group of yeast genes are found over-expressed (or under-expressed) in a microarray experiment may be explained by integrating such gene expression data with related categories of data such as protein-protein interactions and subcellular localizations. Bioinformatics efforts have been underway to perform large-scale integrative analysis on diverse genomic databases [3]. However, such integrated data analysis has been hampered by a number of factors including the following.

1. It is not uncommon that the same genome object (e.g., gene) may be identified using different schemes in different databases.
2. Although most biological data can now be accessible through a web interface, different data sources may make their data available in different ways, requiring different programmatic interfaces to be used to implement data access methods.
3. To use or interpret the data that have been retrieved, the users need to be aware of the different formats that are used in representing the data. A wide variety of formats ranging from unstructured to structured text files have been used for data representation.
4. Different data models such as the relational model and object oriented model can be used to describe the data. Even the same model is used, different model constructs can be used to describe the same object.

5. Different databases may use different terms (synonyms) to code the same concept or use the same term (homonym) to represent different concepts. This inconsistent use of nomenclature makes cross-database comparison and validation challenging.
6. When collecting data from multiple resources, we should also consider the following issues: (i) how up-to-date the data are, (ii) whether the data are curated or not, and (iii) how stable or reliable the resources are, and (iv) how evolvable the resources are.

To address these problems, we propose to use RDF as a common language for describing different yeast data sources and integrating different types of yeast genome data represented in heterogeneous formats.

Position

1. Translating different data formats into the RDF format

Different types of yeast data may be available in the relational database format as well as a variety of XML formats defined using Xschema or DTD. It is desirable to build in a high level translation language into RDF to faciliate translation of these formats into the common RDF format.

2. Scalable RDF-based database

As we propose to use RDF to describe different yeast data sources as well as to integrate the data from such sources, we need to seek an efficient way to store and query large amounts of RDF-formatted data. RDF/DB (http://www.guha.com/rdfdb/) is a native RDF database that can handle a reasonably large dataset in RDF format. However, it may not be efficient enough to handle or query very large datasets generated using high-throughput genomic/proteomic technologies. There is a need for a scalable RDF-native database.

3. Flexible querying capability

To facilitate data analysis, we need a more powerful and flexible query language for accessing and manipulating RDF data.

References

1. Goffeau A, Barrell B, Bussey H, Davis R, Dujon B, Feldmann H, Galibert F, Hoheisel J, Jacq C, Johnston M, Louis E, Mewes H, Murakami Y, Philippsen P, H HT, SG SO, *Life with 6000 genes.* Science, 1996. **274**(5287): p. 546, 563-7.
2. Kal AJ, van Zonneveld AJ, Benes V, van den Berg M, Koerkamp MG, Albermann K, Strack N, Ruijter JM, Richter A, Dujon B, Ansorge W, Tabak HF, *Dynamics of Gene Expression Revealed by Comparison of Serial Analysis of Gene Expression Transcript Profiles from Yeast Grown on Two Different Carbon Sources.* Mol. Biol. Cell, 1999. **10**(6): p. 1859-1872.

3. Gerstein M, *Integrative database analysis in structural genomics.* Nat Struct Biol, 2000. **7**(Suppl): p. 960-3