Semantic Digitization of Experimental Data in Biological Sciences

Saurabh Raghuvanshi
Department of Plant Molecular Biology
University of Delhi South Campus
New Delhi, India
Email: saurabh@genomeindia.org

Abstract— A major bulk of published experimental data, referred to as 'Gold Standard' data, is available in a format that cannot be easily accessed by computers unless effectively curated. Most curation techniques bank on mining the text for information. Here we propose and demonstrate the efficacy of curating the experimental data itself. The data models facilitate digitization of the every aspect of the information associated with the experimental data. The models utilize several universally accepted ontologies as well as in-house developed alphanumeric notations for digitizing different aspect of the data. The data models have sufficient flexibility to address the extensive variability in experimental data. They have a very generic nature and can be used to curate and digitize experimental data from any organism. The digitized data is easily stored in a relational database management system and can thus be rapidly searched and integrated. These models have been successfully used to digitize data from over 20,000 experiments spanning over 500 research articles on rice biology. The entire dataset is available as a database entitled 'Manually Curated Database of Rice Proteins' at www.genomeindia.org/biocuration.

Keywords—Digitization, Ontologies, Rice, Gold standard data

I. INTRODUCTION

A 'Systems' level understanding of any organism requires integration and analysis of multi-dimensional experimental data [1]. Due to its very nature the underlying experimental data is extremely diverse and complex. Thus, integration of such data is never straightforward. These are several issues that impede seamless integration of experimental data. The foremost is the fact that there is no standard data representation format, especially for the 'Gold standard' or published experimental data [1]. Experimental data coming out of different techniques is presented in very different formats and thus cannot be easily correlated and integrated. Moreover, the description of the experimental conditions, biological material etc. is extremely variable in terms of details and thus integration of such poorly explained data might not be very useful. Further, in general, the published experimental is presented in pictorial format either as an image of graph and is thus not amenable to computerized search, let alone seamless integration. Several of the above mentioned issues could be partially resolved by extensive manual curation which is very slow and labor intensive and cannot keep up with the huge amount of experimental data that gets published every year.

II. DESCRIPTION OF THE DATA PRESENTATION FORMATS

Here we address these issue by proposing formats or data-models for digitization and semantic representation of the experimental data in biological sciences. These formats have a very generic nature and are flexible enough to be used for digitization of diverse experimental data. In essence, these models attempt to represent every aspect of the data in terms of an alpha-numeric notation which can be an ontology term or custom notation. As depicted in Fig. 1, every experimental data is composed of several data units, which are usually represented by a single 'bar' of a bar-chart or a 'band' of a gel profile. Each of these data-points is associated with two types of information. The first is the actual value of the data-unit (height of bar) while the other is an array of information such as Gene id., plant type, tissue, growth condition etc. Depending on the complexity of the experiment each data-unit may be associated with upto 10-15 different types of information. As per the data-models each of these information types is represented by either a custom alpha-numerical notation or an ontology term. Thus, every data unit is now represented as a collection or group of alpha-numeric terms. This collection can be easily visualized as an array or equation as depicted in Fig. 1. Every term of this array is actually stored in a specific relational database table. Consequently, since experimental data consists of several data-units, there will be as many groups of alpha-numerical notations, which can be easily stored in an relational database table.

III. ALPHANUMERIC NOTATIONS FOR DATA DIGITIZATION

The data models extensively utilize several ontologies such as plant ontology, environment ontology, trait ontology as well as gene ontology terms to digitize various aspects of the information associated with every data-unit [2][3]. As is obvious, plant ontology terms are used to describe plant part/tissue or development stage whereas environment ontology term is used to digitize the growth conditions or any other treatment (chemical or physical) that may have been administered. Similarly, trait ontology terms define any trait (molecular, biochemical or phenotypic) that has been studied in the experiment. Besides these ontologies other notations have also been used extensively. Some of these include GenBank accessions or locus identifiers, taxon db ids as well as several in-house developed notations to represent

information such as promoter type, mutant type, transgenic line etc. Several of the ontology terms need to be qualified in order to precisely digitize the information. For example, in order to represent plant developmental stage, the age of the plant (in days) need to be added. Similarly, concentration of duration of treatment is appended to environment ontology terms.

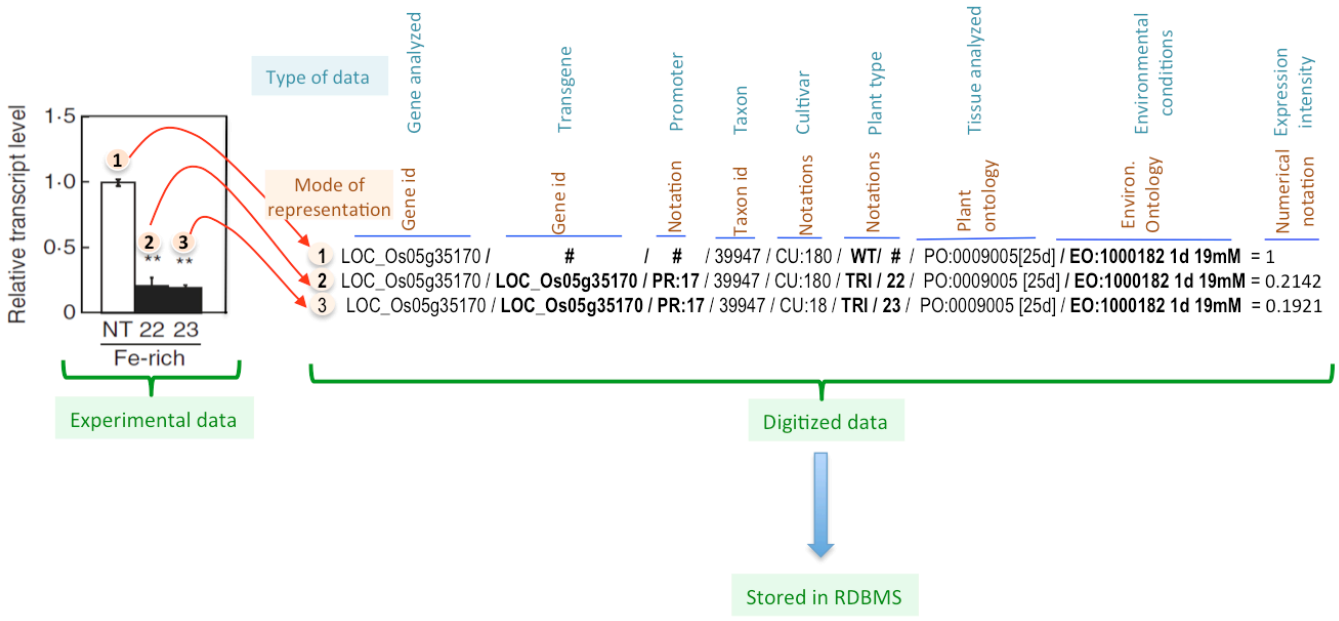

Figure 1. The figure depicts the basic fundamental of data model for digitization of experimental data. Usually experimental data is represented in published research articles as a graph or image. In the present example, digitization of experimental data from a gene expression graph is shown. Every bar of the graph is considered as a 'data-unit'. Each data-unit is associated with several types of information/data such as gene type, transgene, plant type etc. As per the digitization model, each of these data types can be represented as an alphanumeric notation. These notations might be an ontology term or any custom notation. Thus, every data-unit can be represented by a structured array of such notations, while the whole experiment can be ultimately represented as a collection many such arrays. Since the notations are alphanumeric, they can be easily stored in a relational database table.

IV. Implementation of the Data Models

Based on the aforementioned principles, manual data curation and digitization portals were created. These portals were used to digitize experimental data from over 500 published research articles on rice biology. The entire data set has been organized as a database entitled 'Manually Curated Database of Rice Proteins' (MCDRP) [4]. The database contains digitized experimental data pertaining to over 2300 rice proteins that has been curated from over 20,000 different experiments. Altogether, these experiments have over 80,000 data units for which the information has been digitized by manual curation. The digitization of the data has been done by utilizing >600 plant ontology, >350 environment ontology, >800 trait ontology and >350 gene ontology terms. The database is updated twice a year.

Since every aspect of information associated with data-unit of an experiment has been digitized as alpha-numeric notations, can be retrieved from several different aspects. For example, in the browse page one can easily retrieve all the genes that have been studied in a particular plant part or developmental stage. Similarly, all genes that have been associated or studied for a particular trait can be easily accessed.

V. Issues that need to be addressed

The process of digitization is primarily dependent on the depth of various ontologies. During the process of digitization, it was realized that there is an extensive amount of untapped variability in all aspects of data that is not completely described by the current ontologies. Consequently, a significant number of terms used in the curation endeavor had been coined anew (www.genomeindia.org/biocuration).

While several new terms had to be coined since there was no equivalent term, many were formulated since the existing terms had a different perspective. Thus, there is a requirement for consistent efforts to enrich and modify the existing ontologies by defining more and more terms. Further, several new ontologies might also be required such as an ontology describing different methodologies. This is important since the

interpretation of the data can only be done in light of the precise information about the experimental methodology.
It was also observed, that many times that description of the experimental data in published literature is not very precise. This has major consequences in terms of data integration. Thus, usage of precise ontology terms to describe different aspect of the experimental data must be encouraged.

VI. Future

The data digitization formats for the experimental data in biological sciences enable a very precise digitization of the data. This makes the experimental data (Gold standard data) amenable to computerized search and integration. Thus, consistent curation efforts must be done to digitize the already published experimental data. Further, while currently, the data models are being used for digitization of published experimental data, the main aim is to develop data curation and exchange formats that can be implemented in every lab itself even before publication. Thus, at the time of publication all the experimental data is in a pre-defined digitized format than can be easily integrated in any database or public repository.

VII. Discussion

Seamless availability and semantic integration of experimental data is essential to comprehend the complex behavior of biological systems. The data digitization fundamentals briefly explained in this article facilitate digitization of almost every aspect of the experimental data and thus should prove instrumental in achieving a higher understanding of any biological system, if implemented universally. The basic fundamental is very generic in nature and can be applied to data from any biological system. In essence, the data models facilitate digitization of every data-unit of the experimental data in terms of an organized array of alpha-numeric notations. The structure of the array is important since same notation can be used in different positions to mean differently. For example, a plant ontology term can be used to describe the tissue wherein expression of a particular gene has been studied. It can also be used to qualify a gene ontology term to associate a particular 'molecular activity' in a specialized tissue or developmental stage. Thus, same notation can be used to signify different aspect of the information. This gives flexibility as well as universality to the concept.
One of the basic aims of this endeavor is to develop data digitization, archiving and exchange formats that can be used at a very early stage (pre-publication) to organize the experimental data. The idea is to implement these data models as lab data management portals/softwares. This will greatly facilitate rapid and seamless access of the experimental data since there would be no or very minimal need for post-publication data curation [5]. Consequently, curators can address the meta-analysis of such data instead of basic archiving.
We regard our study as a step of a much wider area of research, since experimental data has several dimensions of interpretations. So far we have only addressed the initial digitization and archiving of the data. Models for correlating the digitized data from different studies need to be worked out. Nevertheless, it was possible to demonstrate an effective digitization of experimental data in biological sciences.

Acknowledgment

The author acknowledgs financial support from Department of Biotechnology, Government of India.

References

[1] Rhee, S.Y. and Mutwil, M, "Towards revealing the functions of all genes in plants", Trends Plant Sci., 19: 212–221, 2014.
[2] Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., Clark, K., Teytelman, L., Cartinhour, S., Stein, L., and McCouch, S, "Gramene: Development and integration of trait and gene ontologies for rice", Comp. Funct. Genomics 3: 132–136, 2002.
[3] Gene Ontology Consortium: going forward. Nucleic Acids Res., 43: D1049–56, 2014.
[4] Gour, P., Garg, P., Jain, R., Joseph, S. V., Tyagi, A.K., and Raghuvanshi, S, "Manually curated database of rice proteins". Nucleic Acids Res., 42, 2014.
[5] Baumgartner, W.A., Cohen, K.B., Fox, L.M., Acquaah-mensah, G., and Hunter, L, "Manual curation is not sufficient for annotation of genomic databases", 23: 41–48, 2007.