View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Cold Spring Harbor Laboratory Institutional Repository

Comprehensive Data Infrastructure for Plant Bioinformatics

Chris Jordan and Dan Stanzione
Texas Advanced Computing Center
The University of Texas at Austin
Austin, Texas, United States
ctjordan@tacc.utexas.edu, dan@tacc.utexas.edu

Doreen Ware, Jerry Lu, and Christos Noutsos
Cold Spring Harbor Laboratory
Cold Spring Harbor, New York, United States
ware@cshl.com, cnoutsos@cshl.com, luj@cshl.com

***Abstract*—The iPlant Collaborative is a 5-year, National Science Foundation-funded effort to develop cyberinfrastructure to address a series of grand challenges in plant science. The second of these grand challenges is the Genotype-to-Phenotype project, which seeks to provide tools, in the form of a web-based Discovery Environment, for understanding the developmental process from DNA to a full-grown plant. Addressing this challenge requires the integration of multiple data types that may be stored in multiple formats, with varying levels of standardization. Providing for reproducibility requires that detailed information documenting the experimental provenance of data, and the computational transformations applied to data once it is brought into the iPlant environment. Handling the large quantities of data involved in high-throughput sequencing and other experimental sources of bioinformatics data requires a robust infrastructure for storing and reusing large data objects. We describe the currently planned workflows to be developed for the Genotype-to-Phenotype discovery environment, the data types and formats that must be imported and manipulated within the environment, and we describe the data model that has been developed to express and exchange data within the Discovery Environment, along with the provenance model defined for capturing experimental source and digital transformation descriptions. Capabilities for interaction with reference databases are addressed, focusing not just on the ability to retrieve data from such data sources, but on the ability to use the iPlant Discovery Environment to further populate these important resources. Future activities and the challenges they will present to the data infrastructure of the iPlant Collaborative are also described.**

Keywords-component; bioinformatics; gateways; provenance; data; standards; metadata;

I. Introduction to iPlant and the Grand Challenge Process

The iPlant project was created as part of the (U.S.) National Science Foundation's Plant Science Cyberinfrastructure Collaborative (PSCIC) program. PSCIC looked for a novel approach to construct community driven CI to address the grand challenges of plant biology. While questions of how best to utilize CI exist throughout the life sciences, the PSCIC program was limited in scope to plants, for several reasons. First, plant biology is central to many of the scientific challenges facing society today and the future, including food production and food security, biofuels, health and pharmaceuticals. In a world with increasingly strained water, land, and fossil fuel (for fertilizer) resources, faced with the spread of the western diet, rising populations, increased occurrence of drought, and significant potential climate change, our ability to understand the nature of plants under stress, and breed for greater productivity, is a key to sustainability. Second, a tremendous investment has been made in collecting vast stores of biological data about plant species, and a comprehensive CI is necessary to allow the research community to unlock the knowledge hidden in this data. Biology is rapidly changing into the prototypical data driven science, where the traditional practices of biologists are being transformed, and in many places replaced, by large scale data analysis. Solving the data, CI, and ultimately scientific challenges associated with plant biology will not only offer vast potential benefits to society through advances in agriculture, but also serve as a model for modern biology that can be translated to human disease and other critical problems in life sciences.

The iPlant project was created to leverage the enormous existing investments in biological data collection and bioinformatics tools, and create the virtual organization necessary to allow the plant biology community to progress on grand challenge questions. The iPlant design for cyberinfrastructure rests on the principles expounded for CI at the NSF workshop "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" [4] . This workshop described cyberinfrastructure thus: "Cyberinfrastructure is the set of organizational practices, technical infrastructure and social norms that collectively provide for the smooth operation of research and education work at a distance. All three are objects of design and engineering; a cyberinfrastructure will fail if any one is ignored." Therefore, iPlant is not only a technical product, but also a virtual organization. One of the unusual aspects of iPlant is that, as a Collaborative, iPlant was created not to address any particular scientific question, but rather to build CI to address grand challenge questions designated by the community after the project began.

As a result, the initial challenge for iPlant as a virtual organization was to organize the plant science community to define and prioritize the grand challenges. Over the first year of the iPlant project, this was the dominant task. Starting with a large kickoff conference, iPlant hosted a series of workshop meetings throughout

2008 and through 2009 where self-forming groups throughout the community discussed grand challenge questions that could be addressed through advances in CI. These meetings led to the creation of half a dozen white paper proposals that were provided to iPlant's board of directors for consideration. In April of 2009, the iPlant board announced that the project would proceed in developing CI around two grand challenge questions: constructing the phylogenetic tree for all green plant species (IPTOL, the iPlant Tree of Life), and creating a CI to support investigators in understanding the relationship between genotype information and the expressed phenotypes in plants (IPG2P).

Once these projects were decided, iPlant put together sets of working groups in each grand challenge, consisting of project staff and researchers in the community, to define requirements and design the CI required. The goals of the iPlant CI would be to leverage both the existing base of bioinformatics tools, and the physical resources of existing CIs (e.g. the TeraGrid). The role of iPlant, then, would be to create discovery environments, portals in which researchers can collaborate and make use of the large-scale data and CI resources available. The iPlant infrastructure provides software and interface layers to unify these tools and provide transparent access to these resources, through consistent interfaces, use of standards, and, critically, integration of disparate data sets and types.

II. Genotype to Phenotype Grand Challenge Organization and Methods

While the Tree of Life grand challenge is relatively limited in the variety, if not the size, of the data types to be processed, the Genotype-to-Phenotype grand challenge could be characterized as literally unlimited: potentially almost any type of data generated by plant biology experiments or simulations could be relevant to one or another aspect of the overall problem. In order to approach the problem of understanding the relationship of genotype to phenotype, the iPlant Collaborative created a number of working groups, consisting of both technical and scientific experts, to define various workflows expressing various aspects of the overall challenge. This close collaboration of experts in both CI and a specific scientific domain is one of the more innovative and successful characteristics of the iPlant effort.

The current working groups involved in the iPG2P grand challenge are Ultra-High-Throughput Sequencing, Visual Analytics, Statistical Analysis, Modeling, and Data Integration. The workflows are being developed roughly in the order listed, with the expectation being that both infrastructure and science components can be reused over the course of the project. The Data Integration working group is charged with supporting needs across all the workflows, and as such the working group members are also members of the other working groups and must track the development of workflows for prospective data integration needs.

The Discovery Environment(DE) in which the various workflows are being implemented is a web-based application which takes advantage of modern web application APIs including the Google Web Toolkit along with best-practice development methods including RESTful interaction[5] and an Agile development model[3]. Since the plant biology community has already expended significant effort in developing web-based and other applications that provide components of the required functionality for many of the workflows, the infrastructure will wherever possible make use of such web applications through web service calls, and the overall environment is designed to be open and extensible, either by adding functionality directly or through interaction with external networked applications. While this approach minimizes the overall development effort, it greatly increases the importance of a flexible and robust data integration and data management infrastructure, which presents particular challenges within the realm of bioinformatics.

III. Scale and Diversity of Bioinformatics Data

Plant Bioinformatics is an increasingly data-intensive discipline, driven most obviously by the rise of inexpensive, ultra-high-throughput or "NextGen" sequencing, but also more generally by the increasing number and sophistication of methods and models for analyzing, comparing, and simulating biological processes. One of the distinguishing characteristics of the iPlant Collaborative in general and the Genotype-to-Phenotype grand challenge in particular is the integration of multiple methods for working with a full array of data types in isolation or combination, as opposed to the more common array of bioinformatics applications which address one aspect of the challenge, often for one model organism, i.e. Arabidopsis. In order to support this incredible diversity of data types and analysis methods, a robust data infrastructure is required, including integration of multiple data types and formats, generation and preservation of complex provenance and other metadata, storage and retrieval of extremely large datasets and integration with reference databases across the Internet. All of these various functionalities must be well integrated and extensible, allowing for a true "Discovery Environment" in which plant researchers can focus on the scientific tasks at hand. The following sections describe the various components of this infrastructure and place them within their context as parts of a global-scale CI for plant bioinformatics.

IV. Data Integration Challenges in the Discovery Environment

Initial efforts in the G2P grand challenge have focused on the Ultra-High Throughput Sequence workflow, with relatively few and simpler data types (an important distinction must be made in discussing data for bioinformatics, between data *types*, here used to refer to a category of data such as gene sequences or protein structures, and data *formats*, used to refer to a specific instantiation of a data type such as GFF3). However, data integration efforts

must be cognizant of workflows to be developed by all the working groups, and the abstract data model must be capable of representing all the types of data to be utilized in these workflows, including various multi-dimensional grids, time series, and networks, in addition to chemical structure and pathway data. Another central goal for the DE is extensibility, with open interfaces eventually to be provided allowing for integration of data types and formats not initially supported by the DE, and potentially types of data not even created at the time of the development effort. These requirements argue for a data model that is extraordinarily flexible and capable of handling the structure and semantics of data in an extensible fashion.

V. Description of Data Model

The core assumptions of the iPlant data model are that all, or most, data types can be represented in one of a very few basic forms, most of which are reducible to one or more N-dimensional grids, and that the semantic description of data should be separated from the numerical or other contents of the data. Sequences are the 1-dimensional case of the N-dimensional grid, tabular data is just the 2-dimensional case, and time-series are N+1 cases where N is the dimensionality of the dataset, and networks can also be expressed in a tabular form relating to a separate set of semantic tokens representing, for example, chemical components in pathways, or atomic components in molecular structures (similar to the RDF-triple format for expressing networks of relationships). This model is very similar to that used in the XGAP tool[9] which also addresses many of the components of the Genotype-to-Phenotype challenge, with the significant difference that XGAP attempts to establish a standard file format for storing diverse data types, while the iPlant G2P data model will simply define conversions from existing standard file formats. This was a conscious decision not to further increase the diversity of file formats, as well as to avoid the metastasis and potential incompatibilities that inevitably accompany standardized file formats; early in the process of assessment of bioinformatics data formats we were forced to confront the internal variation within supposedly standard file formats, with variations even at the level of the individual researcher or research group in how file formats may be written and used. A conscious decision was made early in the project, for a variety of reasons, including adoption and technical simplicity, that iPlant efforts would avoid creation of new standards where possible and strive for extension of existing standards where appropriate. Therefore, the data model for the iPlant G2P workflows is used only within the DE, and translation tools can be written with a minimum of complexity to transform data to and from standard file formats for storage and exchange. This allows for maximum interoperability with existing widely-used formats and does not impose the process constraints of a file format.

This basic data model has numerous advantages, the most obvious being great flexibility. In addition, however, the separation of semantic content from numerical content means that mathematical and visualization operations can be written in a generic fashion, without regard to the semantic content of the data they operate on, and the semantics of bioinformatics workflows can be implemented as a separate set of higher-level operators which make use of these lower-level operations. This also allows for easier use and reuse of existing web services and software packages, since these software and services will often either require simple numerical input or will have other specific assumptions about the data to be provided as input.

It is important to note that while in the ideal case the DE would use this comprehensive data model for all workflows and data types, and certain applications such as the Visual Analytics workflows will likely require it due to the complexity and flexibility of the end products to be developed, it is not a requirements that all components or all workflows within the DE utilize this data model; in fact, the initial workflow developed for the G2P Grand Challenge, for processing ultra-high-throughput sequence data, does not utilize this data model, instead supporting only a couple of commonly-used file formats for sequencing data. This was done for the sake of speed in development, as the defined workflow is relatively simple and makes use of several existing tools with specific requirements for formats that are relatively widely used. It is expected that as more complicated workflows are developed, and particularly as the need for more flexibility in the definition and redefinition of workflows grows, the use of this abstract data model will be a requirement from a practical standpoint. However, the goal of openness and flexibility is served overall by the lack of any firm requirements as to how data is handled within the DE.

VI. iPlant Provenance Model

The preservation of comprehensive provenance information is a central component of the goals for all workflows in the iPlant DE. In this context, we define provenance to include both descriptive information regarding the source of data brought into the environment, including experimental description along the lines of the lines of the requirements spelled out by the MIAME and related MIBBI standards[10][2], and information recording the transformations, combinations, and other operations performed on data within the DE. Capturing, encoding, and preserving this comprehensive provenance information will allow for end products of workflows in the DE to be reproduced in their entirety, whether by rerunning a workflow using the same data from reference databases or from within the iPlant data stores, or by reproducing experiments and the processing done on resulting experimental data. This reproducibility is a critical component of the scientific method and a capability of increasing importance as reliance on digitally generated and processed data increases in bioinformatics and other scientific fields. Capturing metadata describing all processing steps also enables workflows and workflow variations to be saved and shared between researchers or groups of researchers, thus allowing for exploration of new

forms of processing and sharing of results when effective techniques or sets of techniques are found. To enable this, a comprehensive provenance metadata model has been developed which will support capture of experimental process descriptions, references to related documentation, and machine-readable descriptions of the full set of operations performed within a DE.

After evaluation of the available standards and extensive discussion of techniques for conceptualizing and capturing provenance, provenance has been split into two broad categories: first, the experimental and other descriptive information necessary to characterize the sources of data being brought into the DE, and second, the complete set of actions and actors within the DE used to produce various derived and combined products from these input data sets. The conceptual model for provenance metadata, particularly for recording actions within the DE, is based on the work of Dr. Sudha Ram in characterizing provenance as the set of "7 Ws": "what", "when", "where", "how", "who", "which" and "why". [8] Efforts are currently underway to develop an efficient means of XML-based expression for both the experimental background metadata and the ongoing tracking of provenance as users submit data and process it through the various workflows. Tracking operations will eventually be associated with all DE components and the provenance "trail" will continually grow from the point where data is loaded into the environment. This will provide the full set information required to reproduce results and to share workflows between users.

VII. Data Management Infrastructure

There are two central aspects of data management for the iPlant DE: integration with external reference data sources and integration with large-scale storage infrastructure for short- and long-term retention of bioinformatics data of relevance. Since open interaction with and leveraging of existing community efforts is a core principle of the iPlant project, integration with the wide variety of well-established reference data sources, particularly in the area of gene sequence and expression data, is an absolute requirement. This integration must be bidirectional, that is we must not only be able to retrieve reference data from genome data sources such as Genbank[1], but where appropriate users of the DE should be able to utilize the infrastructure to directly submit their own experimental data to such reference data bases. As such, an early activity of the Data Integration working group was to survey many of the reference data bases including Gramene, TAIR, and MaizeGDB, to determine the necessary submission formats and preferred interaction styles for these sites, and to form the basis of a generic model for extracting and submitting sequence and other data types. Through the survey, we were able to determine that though there is great diversity in data formats, there are a few formats such as GFF3 and FASTA that are supported by all the major reference data sources. Since these formats must also be supported for interaction with file-based data submitted by users of the DE, compatibility with the submission requirements is not an issue, and we believe that the thorough provenance model we support will be a significant advantage both for determining appropriate targets for submission of data and for providing all necessary information to these sites in an automated fashion. This will encourage submission of data to the reference databases by making the process trivial for the users of the DE, and we hope that increased levels of submission will be a positive side effect of our efforts.

As previously noted, the iPlant Collaborative is a virtual organization, with both human and technical resources spread out across multiple institutions; storage hardware in particular is primarily located at the University of Arizona and at the Texas Advanced Computing Center, as well as commercial cloud storage resources in multiple datacenters. Additionally, the set of available resources is expected to evolve over time, as the number of users of the infrastructure and the corresponding scope of necessary storage grows. From the perspective of interaction between the DE and the available storage resources, however, it is desirable to have a single logical address space that can be used to store and retrieve data from any iPlant storage resource without explicitly identifying the source or destination system. For purposes of both data integrity and performance, it is necessary to have data replicated across resources, some of which may be appropriate for archival purposes while others are oriented towards high-performance data movement. Being able to maintain consistency amongst multiple replicas and move data at very high speeds are crucial given the scale of data to be collected and reused in the DE; initial provisioning allows for 50TB of storage over the next year, and it is expected that this will continue to grow over time. Many individual files will be tens or even hundreds of gigabytes in size, particularly in the UHTS workflow. The required capabilities are provided for the iPlant infrastructure through use of the iRODS "rule-oriented data system" software[6][7]. The iRODS software utilizes a backend Postgres database to provide a virtual namespace across an arbitrary set of resources, and includes features to automate replication and resource selection for both storage and retrieval. It also allows for multiple administrative domains, or "zones", to be federated into a single system with a single set of credentials used to authenticate and perform access control operations across all included resources in the various zones. This allows the DE development team to focus on simple data storage and retrieval operations, while the storage layer can be separately administered and data location and replication features can be managed transparently without requiring user- or developer-level configuration of the storage. Arbitrary metadata can also be stored along with data objects, and objects can be searched for based on this metadata. This ability to handle metadata at the storage layer is another important capability for the data infrastructure of the iPlant DE. Finally, the iRODS data transport mechanisms support common high-performance network transfer features including large window sizes and threaded transfers, allowing the underlying data management system to take advantage of high-speed education and

research networks to provide the necessary performance in moving gigabytes or even terabytes of data between systems inside and outside the network of iPlant systems.

VIII. Future Activities

Two important workflows that will be developed within the next year and will present significant challenges to all aspects of the iPlant data infrastructure involve modeling of plant growth and interaction with environmental conditions, and high-throughput phenotyping through the use of image recognition techniques. In addition to the challenges of assessment and reuse of appropriate algorithms or sets of algorithms for use as components within the workflows, there will be significant issues related to data formats, standards, and data management. High-throughput phenotyping in particular will involve access to and management of large numbers of images, which may be widely distributed across the globe; identification of relevant images and image sources will present a challenge, as will the management of data storage; the use of image caching will likely evolve over time as the necessary quantities of data for various operations becomes more clear, and the addition of a new data type with a much different kind of provenance information will require extensions to the existing provenance metadata standard. Similar issues will need to be addressed with regard to modeling of plant development, with the additional challenge that initial implementations of the modeling workflows will likely support the use of significant variation in input parameters and the need to perform analysis of variation in output and comparison to empirical data using large numbers of output data sets.

An additional development of importance to users of the DE will be integration of the data storage and transport mechanisms with high-performance computing resources such as the Ranger and Lonestar systems at the Texas Advanced Computing Center. This will allow for the most compute- and data-intensive tasks to be offloaded to these powerful resources with minimal delay, and for results to be returned to the user without having to interact directly with the systems in question, which are often difficult to use in comparison to the graphical web interface and may require special access arrangements.

IX. Conclusion

As should be clear, the CI requirements for data in the iPlant DE are diverse and consistently challenging. At every layer of the software and hardware stack there are issues of diversity, complexity, and scale, requiring a combination of solutions, each of which provides a flexible layer of abstraction along with the ability to provide high performance and robustness. Wherever possible existing standards and tools are being used, sometimes through direct integration into the software stack and sometimes through the use of web services to "call out" from the DE. In some areas such as provenance, we are developing solutions we believe represent a level of comprehensiveness that is unprecedented in a scientific computing environment and will enable true reproducibility of results within the digital realm. In the realm of integrating resources, we will provide a web-based interface that will provide access to resources at the largest scale of both storage and computation. The flexibility and use of standards within the core infrastructure enables the development of an open environment which we hope will be not just embraced but extended by the community of plant biologists and bioinformaticians, eventually providing a platform for data-centric exploration of the problems of plant science, now and in the future.

Acknowledgment

The authors would like to thank the National Science Foundation for supporting the iPlant Collaborative.

References

[1] Benson, Dennis et al. "Genbank". Nucleic Acids Research, Vol 27 No 1, 1999

[2] Brazma, Alvis et al. "Minimum Information About a Microarray Experiment – toward standards for microarray data". Nature Genetics, December 2001.

[3] Edmonds, E. A. "A process for the development of software for non-technical users as an adaptive system". General Systems XIX: 215–218, 1974

[4] Edwards, Paul N., Steven J. Jackson, Geoffrey C. Bowker, and Cory P. Knobel. "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures". Ann Arbor Workshop Report, 2007.

[5] Fielding, Roy. "Architectural Styles and the Design of Network-based Software Architectures". Chapter 5, Doctoral Dissertation. Retrieved June 30, 2010 from: http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

[6] Rajasekar, Arcot et al. "A Prototype Rule-Based Distributed Data Management System". Proceedings of HPDC Workshop on Distributed Data Management, Paris, France, 2006

[7] Rajasekar, Arcot et al. "iRODS Primer: Integrated Rule-Oriented Data System". Synthesis Lectures on Information Concepts, Retrieval, and Services, 2010, Vol. 2, No. 1, Pages 1-143

[8] Ram, Sudha and Liu, Jun. "Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling". Lecture Notes in Computer Science. Berlin: Springer 2008

[9] Swertz, Morris et al. "XGAP: a uniform and extensible data model and software platform for genotype and phenotype experiments". Genome Biology 2010, 11:R27

[10] Taylor, Chris et al. "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project". Nature Biotechnology 26, 889 - 896