Plant Systems Cyberinfrastructure 10 Year Strategic Plan

Connections: Cyberinfrastructure for Plant Systems Science Over the Coming Decade

Plant Science Research Network

Plant Systems Cyberinfrastructure 10 Year Strategic Plan

Executive Summary

The Plant Science Research Network (PSRN) is a consortium of 12 professional societies with interests in plant science. Its goals are to advance interdisciplinary training and research in plant systems research, including broadening participation and translation of discoveries to practice.

A subcommittee of the PSRN has been developing a 10 year strategic plan for cyberinfrastructure needs for plant systems research spanning all levels of scale from molecules to ecosystems. A visioning workshop was convened by the PSRN on October 18-20, 2016 in Rockville, Maryland. The workshop included participants from the plant and information sciences, from all domains of plant sciences, from academia, industry, non-profits, and government, and from all levels of seniority from students to senior investigators. The workshop drew on inputs from a prior planning workshop, a survey of 52 leaders in the plant and information sciences, and numerous informal inputs from domain conferences during 2016. A draft executive summary was developed from this workshop. We are now seeking input from the wider plant science community on this plan. In particular are there key opportunities and needs that are not fully addressed in this document? Note that a full length strategic plan will be developed from this summary once we are satisfied that the scope of the plan has been adequately determined. Thus most aspects of the plan are addressed at a very high level in the current summary.

To view the summary and post comments, please go to plantae.org, log in (you will need to register - its free), and navigate to the Big Data and Cyberinfrastructure group. The direct link is http://bit.ly/2D6T6To. Please post comments on this group's Discussion or else use the Files menu to upload a comment document. Comments will be most useful if received by February 28, 2017.

Brett Tyler
Chair, PSRN Cyberinfrastructure committee

Plant Science Research Network Steering Committee 2017

Scientific Representatives

Kathy Barton
Association of Independent Plant Institutes

David Baltensberger
Alliance of Crop, Soil and Environmental Science Societies

James Birchler
Genetics Society of America

Todd Dawson
Ecological Society of America

Michael Donoghue
Botanical Society of America

Rebecca Grumet
American Society for Horticultural Science

Barry Pogson
Global Plant Council

Chelsea Specht
American Society of Plant Taxonomists

David Stern
Chair, American Society of Plant Biologists

Brett Tyler
American Phytopathological Society

Andrea Weeks
Council on Undergraduate Research

Eleanor Wurtzel
Phytochemical Society of North America

Connections: Cyberinfrastructure for Plant Systems Science Over the Coming Decade

This document summarizes the vision for plant systems cyberinfrastructure developed from a visioning workshop convened by the Plant Science Research Network on October 18-20, 2016 in Rockville, Maryland. The workshop included participants from the plant and information sciences, from all domains of plant sciences, from academia, industry, non-profits, and government, and from all levels of seniority from students to senior investigators. The workshop drew on inputs from a prior planning workshop, a survey of 52 leaders in the plant and information sciences, and numerous informal inputs from domain conferences during 2016.

Cyberinfrastructure is here defined as an infrastructure ecosystem that is required to convert large amounts of data into knowledge and applications through computational analytics, modeling and simulations. The six overlapping elements of this ecosystem are:

1. Data: reference data sets, metadata (including provenance), data standards, ontologies, interoperability, sharing, access;
2. Tools: data integration, analytics, artificial intelligence, algorithms, software, APIs;
3. Platforms: computing resources, hardware, the cloud, operating systems, middleware, data storage, databases, high-speed connectivity;
4. Communication: collaboration between plant and information scientists (mathematicians, statisticians, computer scientists) and among communities of plant systems scientists;
5. Training: training of current and next generation plant scientists in data literacy and transdisciplinary collaboration, including good data and software practices; familiarization of information scientists with plant science;
6. Public engagement: tools and activities to engage stakeholders, end-users and the general public in plant systems science, including K-16 teachers and students; encompasses translation to use, education and citizen science.

Opportunities

Investments in all six elements of cyberinfrastructure will enable the advance of fundamental knowledge and the achievement of major translational impacts in the areas of natural and managed ecosystems, including: preservation of terrestrial and aquatic ecosystems; intensified but sustainable agriculture, aquaculture, and forestry; and innovations in food, energy and water systems, medicine and biomaterials.

The investments will enable an increasingly sophisticated cross-scale understanding of plant systems spanning molecules, cells, organisms, communities (including microbiomes), ecosystems, and the environment, including increasingly accurate and generalizable predictive models. A strong systems-level understanding of plants and plant communities, and how their components operate, will enable design-driven engineering of plant systems via synthetic biology, including genome editing.

Rapid advances in big data, computing and advanced analytics generally and in the broader life sciences will accelerate the development of plant systems cyberinfrastructure. One example is the recent blossoming of GPU-based deep learning algorithms in the data mining community.

These advances will create major new opportunities for innovation, commercial development and workforce training in areas such as sustainable food and energy security, green chemistry, bio-based manufacturing, and plant products-based healthcare.

Challenges

Considerable investments have been made in generating large scale data sets and software, especially in genomics, but increasingly at other scales of plant systems. However, large amounts of value are being lost because of poor interoperability among data sets, tool sets and databases. This results from heterogeneous development and use of data standards and ontologies, heterogeneous development and use of standard application programming interfaces (APIs), and a heterogeneous culture of open data and software sharing and access among research communities.

Some cyberinfrastructure needs are specific to plant systems research, such as reference data sets, databas-

es to accommodate domain-specific data structures, domain-specific ontologies and semantic concepts, domain-specific models and simulations, and tools and platforms needed for cross-scale integration up to the ecosystem level. Addressing these needs will require close collaboration among plant and information scientists, and new models of training that blend plant and information science.

Ongoing maintenance and updating of data sets, tools, and platforms, including their accessibility, is extremely challenging; existing funding models and professional incentives greatly value innovation over sustaining and strengthening existing resources, and expanding their accessibility. As a result, much value is lost from past investments.

Recommendations

Effective **connections** among data sets, tool sets, platforms including databases, plant and information science researchers and educators, and the public, are required to unlock the potential of plant systems science. Training, funding models and professional incentives are needed that support the development and maintenance of those connections. As much as possible, these connections should extend worldwide.

Funding models are envisaged to include, but not be limited to, federal, state and local government investments, private enterprise, public/private partnerships, non-profit organizations, philanthropy, subscription services, crowd-funding, open sourcing, or combinations thereof, as appropriate.

Professional incentives refer to criteria for professional advancement for scientists in academia, government and industry that influence the amount of time and effort scientists invest in the various dimensions of their professional work. They include the kinds of credit received for individual versus collaborative work, for publications (individual, collaborative, “highly-cited”, open access, etc), other kinds of products (data sets, software, tool sets, ontologies, etc), for various kinds of research funding received, and for synergistic activities such as community building, creation of standards, data curation, benchmarking of tools, and writing of reports like this one. Such incentives arise and change over time through complex interactions among research institutions, funding agencies, journals, scientific societies, and scientists themselves.

Connecting data

Due to investments over the last 10-15 years and rapid advances in technologies, increasingly numerous genomics data sets (including genomes, microbiomes, transcriptomes, epigenomes) have accumulated. Large data sets at other levels of scale (e.g. metabolomes, proteomes, cell and organism images, other phenomes, biodiversity data, and related environmental data) are also rapidly accumulating. Novel technologies including nanopore sequencers, high throughput robotics, sensors, robotic vision, in-field instruments, LIDAR, and wireless data loggers have increasing potential to capture molecular, genomic, organismal, and environmental attributes in real time at an enormous scale.

The ability to integrate the above data types across scales, across time and space, and across species and populations, together with advanced analytics and modeling, will transform our ability to understand plants and plant communities at a systems level. Well documented libraries of interoperable plant “parts” will enable effective application of the tools of synthetic biology and genome editing to plant systems. Acquisition, integration and analysis of large scale data in real time also has the potential to support real time decision making by researchers, end-users, and policy makers.

However, fully realizing these opportunities will require major improvements in data standards and interoperability at all levels of scale. Substantial improvements will be required in the development and use of data standards, ontologies, and detailed metadata (including data provenance), more vigorous enforcement of sharing policies, and development of professional incentives to promote data sharing either directly or via the semantic web & APIs. Standardized data sets are also important for comparing new tools and algorithms. Open data sharing, together with careful adherence to data standards and fully comprehensive documentation of data provenance and collection methods will also be essential to maximizing reproducibility, which is a prerequisite for effective data analysis and interpretation. Preservation of biological materials connected to data sets via seed banks, herbaria, museums and culture collections will also be an important component of reproducibility.

Seamless interoperability between major data repositories, including CyVerse, KnowledgeBase, EMBO EBI, NCBI, and the DOE JGI will also yield substantial synergies. Small

standalone project-specific and taxon-specific databases that are challenging to sustain should be integrated into or federated with major repositories.

In the areas of metabolomics, proteomics, phenomics (including imaging, LIDAR, robotics, etc.), ecoinformatics, biodiversity, and related environmental data, there is a need for comprehensive, high quality reference data sets. Currently these are either poorly accessible or substantially lacking. Furthermore, data sharing is impeded by lack of sustainable, shared, open repositories, lack of data standards, and in some cases a poor culture of data sharing. iDigBio and BISQUE are promising developments for storing and sharing of image data.

Recommendations

1. Funding (including new models), professional incentives, cultural changes, and training are needed to promote open sharing and interoperability of data by research communities, including development and use of data standards, ontologies, and fully comprehensive metadata.

2. Investments in high quality reference data sets, data standards, and shared, open repositories are especially needed in the emerging areas such as plant metabolomics, proteomics, phenomics (including imaging, LIDAR etc), biodiversity, ecoinformatics, and related environmental data.

Connecting tools and services

Driven by the rapid growth in biological systems data (especially genomics), and advances in computer science, statistics and mathematics, a plethora of tools for analyzing and integrating these data has emerged in all areas of life science research, including biomedical science, model systems, and plant science. Some of these have been supported by focused federal investments while others have emerged from smaller projects, from the private sector, and through open source projects. Tools also include growing libraries of models and simulations of various sub-systems of plants including biochemical, regulatory, and ecological networks.

This vast ecosystem of tools provides a rich resource for exploring and mining plant systems data. This opportunity however depends on hardened software tools that have been broadly validated, carefully documented (including user tutorials and algorithmic underpinnings), and that take inputs and produce outputs in standardized formats, so that the tools can work seamlessly with each other and with databases. Many tool sets have these qualities, especially when they have been developed as part of large centralized projects such as iPlant/CyVerse and KnowledgeBase. However, the quality of numerous stand-alone tools that have been developed by the research community is much more heterogeneous; thus, the potential contributions of those tools are much more limited. While there will always be a need for the development and evaluation of draft tools that implement new algorithms, a stronger culture of good software practices and standardized benchmarking is required to ensure that the most valuable of these new tools can quickly transition to mainstream use. Libraries of models and simulations will also benefit from a focus on complete documentation, interoperability and usability. Centralized repositories for interoperable tools that are well documented and validated will facilitate reliable, high quality data analyses that are reproducible across research groups.

Tool sets have been growing to serve the metabolomics, proteomics, phenomics, and ecoinformatics communities. The development of tools in these areas has however been hampered by lack or poor usage of common data standards, complicated by the highly heterogeneous nature of the data in these fields. In some cases such as metabolomics, a number of tool sets have been developed by the private sector, but these can only take proprietary data formats as inputs, limiting the ability to integrate them into a wider ecosystem of tools.

The tools of the semantic web, including the Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML), in principle provide strong potential to promote the connectivity of data and software tools. However, major challenges to this approach persist, including the rapidly evolving nature of knowledge, and the overhead and specialized skills required to accurately represent data and tools in semantic forms.

Recommendations

1. Funding (including new models, such as subscriptions), professional incentives, cultural changes, and training are needed to promote the development of tools, models, and services that have been broadly validated, carefully documented (including user tutorials and algorithmic underpinnings), and that take inputs and produce outputs in standardized formats.

2. Centralized resources and repositories such as CyVerse and Knowledgebase will continue to play a critical role in creating, recruiting, and supporting well-integrated tools and services, and should be supported to do so.

3. Focused investments in the use of semantic web tools to address specific questions in plant science are needed to create informative use cases that advance the adoption of these tools into the mainstream of plant systems science.

Connecting platforms

Common to all areas of data science, the availability and wide implementation of open source operating systems such as Linux, together with standards for high speed networking and data exchange, have provided the basis for wide-scale interconnectivity of data and software platforms. An ongoing challenge, however, is the necessity for rigorous cybersecurity, which slows connectivity. Another challenge is the exponential growth in the sizes of data sets that must be transmitted over networks.

Connecting databases

A diverse ecosystem of data repositories has grown up around specific sets of needs, some of them general to life science, others specific to plant science. These include NCBI, EBI, iPlant/CyVerse, DOE's Knowledgebase, TAIR/Phoenix, Araport, GOBII, FungiDB, DOE JGI, Planteome (Ontologies), Bio-Analytic Resource (BAR), Plant Reactome, Plant Metabolic Network, PhytoPathDB, DataONE, and many others. This diverse ecosystem of resources has enabled data to be represented and served in ways that are usefully customized to particular communities or missions. On the other hand, the different ways that the data are represented in each database, often exacerbated by the use of different data standards, has resulted in the fragmentation of the aggregate data resources, producing major barriers to data integration. In many cases, the isolation of the databases has been exacerbated by competition for limited resources; this could be mitigated by placing a higher funding priority on cross-database integration. The creation of web services on top of some repositories has improved automated access to and integration of the data, however web services can only be implemented where common data standards are in use. Furthermore, latency has emerged as a major challenge, with integrated services being limited by the speed of the slowest web service; an opportunity may exist to use cloud hosting to mitigate the latency challenge.

Federation of databases also provides an opportunity to address the major challenges created by the ongoing diversification of plant systems data. The highly heterogeneous data produced by genomics, metabolomics, proteomics, phenomics, and ecoinformatics, including associated environmental and provenance data cannot be housed in any single relational database. By their nature, relational databases are highly inflexible to changing data structures. A federation strategy allows new customized databases to be built and integrated as new data types emerge.

Connecting to the cloud

Commercial cloud platforms are rapidly growing in flexibility, ease of use, and affordability. These platforms enable hardware-based platforms to be partially or completely replaced by cloud platforms. Cloud platforms also provide opportunities to co-locate computing resources with data sets and tool sets, irrespective of the provenance of those sets. Thus, significant opportunities exist around connecting and integrating hardware- and cloud-based systems. On the other hand, challenges exist in reconciling finite term research funding with the need to continually pay for commercial cloud access. Hardware purchased with a grant can continue to be used at minimal cost after the grant ends, but cloud access cannot be maintained under current funding models. Innovative funding models will be needed if use of cloud resources is to be facilitated and incentivized.

Recommendations

1. Strong funding, professional incentives, and cultural changes are needed to promote the interoperability and federation of major existing plant data repositories, as well as emerging repositories for plant metabolomics, proteomics, phenomics, image, ecoinformatics and environmental data.

2. New funding models including public/private partnerships, and investments in informative use cases are needed to advance the integration of conventional public hardware resources with commercial cloud resources, including better realization of the potential of web services.

Connecting across disciplines through communication

Rapid advances in big data science, high performance computing and advanced analytics are being driven by the explosion in massive data sets in the commercial, social and

healthcare domains, and broadly in life sciences research. Communication and collaboration between plant biologists and information scientists (including computer scientists, statisticians, mathematicians, and engineers) thus will be essential to accessing the latest breakthroughs in advanced analytics and modeling. As detailed in the next section, various levels of cross-disciplinary training, as well as training in the skills and culture of transdisciplinary collaboration, will be needed to unlock this potential. To fully realize the potential for integrating plant systems understanding across scales, communication and collaboration will also need to be facilitated among different communities of plant scientists, including geneticists, genome scientists, molecular biologists, chemists, cell biologists, botanists, physiologists, crop and soil scientists, crop breeders, agricultural and biosystems engineers, and ecologists. Collaborations of robotics engineers with plant and information scientists will also be required to accelerate data collection, especially in areas such as phenomics and environmental data science. Establishment of a strong culture of transdisciplinary collaboration through training and professional incentives will be needed to fully realize the potential for plant systems research.

Recommendation

1. Funding, professional incentives, cultural changes, and training are needed to promote transdisciplinary collaboration between plant and information scientists, and among different communities of plant scientists.

Connecting through training

To be effective in a data-rich plant systems research environment, current and next generations of plant scientists will, as a minimum, require a basic level of data literacy that will enable them to understand and utilize the data sets and tool sets that are available to them. Some plant scientists may acquire much deeper competency in information science. Plant scientists with data literacy will be essential to transforming data analyses into new knowledge about plant systems. Training of some information scientists in the basics of plant science will also facilitate communication and collaboration. Although some sub-disciplines of plant science (e.g. quantitative genetics, agricultural engineering) combine some of the requisite skill sets, lack of sufficient numbers of data-competent plant scientists is a major limitation currently.

Training in the skills and culture of transdisciplinary collaboration also will be needed to fully realize the potential for plant systems research.

Training from the earliest career stages (together with professional incentives) will be key to building an enduring culture of strong data standards, strong software standards, and of open sharing of data and software. Programs such as Software Carpentry and Data Carpentry will likely play an important role in building the requisite culture.

New models for undergraduate and graduate training are needed that integrate plant science, data science, transdisciplinary skills, and a culture of open sharing. Connecting educators who are experimenting with and building such models into communities of practice will facilitate the emergence of best practices in this area.

Recommendations

1. Funding, professional incentives, and cultural changes are needed to support the development of new models for undergraduate and graduate training that integrate plant science, data science, transdisciplinary skills and a culture of open data and software sharing.

2. Funding, professional incentives, and cultural changes are needed to support the cross-disciplinary familiarization of current plant and information scientists with each other's disciplines, and for supporting staff as well as faculty career models at the intersection of plant and information sciences.

3. Support is needed for the development of communities of practice that are experimenting with and building new models for cross-disciplinary education and training.

Connecting with the broader community, including, farmers, the general public, and K-16 teachers and students

In the end, plant science exists to serve the broader community. The knowledge and tools produced by plant systems research should be readily accessible by farmers, environmental planners, teachers, students, and the general public. Continued public support for plant science depends on ensuring that plant science is addressing public needs for food, energy, water, and a clean and attractive environment.

Cyberinfrastructure has an important role to play in engaging with the public. For example, smart phone apps could enable a user to photograph a plant or plant community and obtain immediate accurate information about the plant or community and its role in the biosphere. Related

software could provide useful services to farmers, gardeners, and field ecologists. Farmers could benefit from readily accessible information about crops and cropping systems. Similar cyberinfrastructure could also be adapted to provide resources for teachers and students at the K-12 and undergraduate level, to strengthen the teaching of plant science.

Opportunities also exist for creating cyberinfrastructure that enables teachers, students and the public to contribute to the advancement of plant science though citizen science. Progress is already being made in this area in the field of ecology, such as documentation of bird migrations and phenological observations of plants and animals.

Recommendations

1. Funding and professional incentives are needed to support cyberinfrastructure and training that facilitates the integration of plant systems science into K-12 and undergraduate education.
2. Funding and professional incentives are needed to support the creation of cyberinfrastructure that enables teachers, students and the public to contribute to the advancement of plant science though citizen science.
3. Funding and professional incentives are needed to support cyberinfrastructure that make advances in plant science available and accessible to diverse user groups including farmers, gardeners, and the broader public.

Authors and contributors

Brett Tyler
Center for Genome Research and Biocomputing, Oregon State University

Chair, Plant Science Research Network Cyberinfrastructure Committee

Workshop participants

Plant Science Research Network Cyberinfrastructure Strategic Retreat

October 18-20, 2016

Volker Brendel
University of Indiana

Liliana Florea
Johns Hopkins University

Rodrigo Guitterez
Pontifical Catholic University of Chile

Alex Harkess
Donald Danforth Plant Science Center

Eva Huala
Phoenix Bioinformatics, Inc.

Pankaj Jaiswal
Oregon State University

Eric Lyons
University of Arizona; CyVerse

David LeBauer
University of Illinois, Urbana-Champaign

Lukas Mueller
Boyce Thompson Institute

Molly Megraw
Oregon State University

Allison Miller
Saint Louis University

Mark Miller
San Diego Supercomputing Center

Lijun Ma
University of Massachussetts

Katherine Mejia Guerra
USDA-ARS; Cornell University

James Ostell
National Center for Biotechnology Information

Rebecca Panko
New Jersey Institute of Technology

Nicholas Provart
University of Toronto

Kelly R. Robbins
USDA-ARS; Cornell University

Mark Smith
Amazon Web Services

Daniel Standage
University of California, Davis

Doug Soltis
University of Florida

Lloyd Sumner
University of Missouri, Columbia

John Towns
University of Illinois, Urbana-Champaign

Alex Thomasson
Texas A&M University

Andrew Thornhill
University of California, Berkeley

Matthew Vaughn
Texas Advanced Computing Center; CyVerse

Jacob Washburn
University of Missouri, Columbia

Ya Yang
University of Minnesota

Brett Tyler
Oregon State University

Alain Wouters
Whole Systems

Additional authors and contributors

Ed Buckler
USDA-ARS; Cornell University

Fumiaki Katagiri
University of Minnesota

Carolyn Lawrence-Dill
Iowa State University

Nirav Merchant
University of Arizona; CyVerse

Brent Mishler
University of California, Berkeley

Chris Pires
University of Missouri, Columbia

Steven Salzburg
University of Maryland

Marcela Karey Tello-Ruiz
Cold Spring Harbor Laboratory

Jason Williams
Cold Spring Harbor Laboratory; CyVerse

This work was supported in part by an NSF RCN grant to the Boyce Thompson Institute #IOS-1514765.