**NSF Workshop for a Plant Cyberinfrastructure Center**

To provide feedback on this document, please send your comments to Sue Rhee (mailto:rhee@acoma.stanford.edu).

**Introduction**

Recent progress in plant science has resulted in the development of a wide range of new tools and resources for research and education. These include genome scale information (such as whole genome sequences and annotations, and high resolution genetic and physical maps), extensive phenotypic data sets (derived from gene expression profiling, proteomics, and metabolic data sets) genetic resources (such as sequence-catalogued collections of mutants, recombinant inbred lines, and transgenic collections), functional genomics assets (including reporter-gene expression collections, high-throughput deletion and RNAi sets, protein interaction surveys, and high-throughput protein structure resources) and informatics resources (software tools, algorithms, databases, and web sites). While each of these resources is quite useful individually, their highest value generally lies in their use in combination, and by researchers outside the groups that created the resources.

Optimal use of all of these resources, then, requires combining the information represented among them in innovative ways to achieve a better understanding of fundamental principles in plant biology; and it also requires that individuals from multiple fields and disciplines be able to find, understand and effectively employ these resources in novel ways. Such innovative, synthetic approaches do not yet come 'off-the-shelf' and, instead, present a significant new challenge that can only be met by plant researchers/educators, information scientists, and others working together.

To discuss the means by which to achieve such a synthesis of resources, and to design the appropriate cyberinfrastructure for their best utilization, a workshop on Plant Cyberinfrastructure was held at the National Science Foundation on October 17 and 18, 2005. The participants represented the widest possible range of plant and computational biologists, with experts in plant genomics, development, metabolism, ecology, and evolutionary biology, ecoinformatics and experts in computational modeling, databases, computer infrastructure, software, and mathematics.

List of Participants:

| Name | Institution | Department |
|---|---|---|
| Dan Ashlock | Iowa State | Mathematics |
| Reed Beaman | Yale | Research & Collection |
| Volker Brendel | Iowa State | Bioinformatics |
| Ed Buckler | Cornell University | Life Science/Genomics |
| Clint Chapple | Purdue University | Biochemistry |
| Gloria Coruzzi | NY University | Biology |

| | | |
|---|---|---|
| Cliff Cunningham | Duke | Biology |
| Julie Dickerson | Iowa State | Department of Electrical and Computer Engineering |
| Peggy Lemaux | UC Berkeley | Cooperative Extension Specialist in Plant Biotechnology |
| Sean May | Univ. of Nottingham | Plant Science |
| Pedro Mendez | Virginia Tech | Plant Pathology, Physiology, and Weed Science Department |
| Elliot Meyerowitz | Caltech | Biology |
| Mark Miller | Carnegie Mellon Un. | Robotics Institute |
| Brent Mishler | UC Berkeley | Integrative Biology Department |
| Eric Mjolsness | UC Irvine | ISC-Info & Computer Science |
| Magnus Nordborg | USC | Molecular & Computational Biology |
| James Reichman | UC Santa Barbara | National Center for Ecological Analysis and Synthesis (NCEAS) |
| Sue Rhee | Carnegie Inst. | Plant Biology |
| Johanna Schmitt | Brown University | Ecology and Evolutionary Biology |
| Heiko Schoof | Max-Planck Inst. | Plant Computational Biology |
| Kimmen Sjolander | UC Berkeley | Bioengineering |
| Lincoln Stein | Cold Spring Harbor | Bioinformatics |
| Mike Sussman | U of Wisc.-Madison | Biotechnology Center Biochemistry |
| Fedora Sutton | S. Dakota State | Plant Science and Biology/Microbiology |
| Todd Vision | U of N. Carolina | Biology |
| Cathy Wu | Georgetown | Biochemistry & Molecular Biology and Department of Oncology |

The participants, despite their wide variety of specialties and interests, achieved consensus on a number of issues, among them the need for a plant science center to provide the cyberinfrastructure for future work in plant science and for the developing new field of plant computational modeling, a vision for the mission and goals of such a center, and some suggestions for practices that might be followed to assure that such a center plays a central role in establishing software and interface standards and in outreach.

**Goals for a Plant Cyberinfrastructure Center**

The goals of a plant cyberinfrastructure center should be fourfold:

- To enable new research approaches to fundamental questions in plant biology, which are not currently possible in individual laboratories, by providing the cyberinfrastructure for synthesis and collaboration by interdisciplinary teams
- To train a new generation of scientists to embrace multi-disciplinary approaches to develop a better understanding of plant systems
- To interact with the broader biological science community to set standards for data and meta data description, software and computational interfaces, so that the data created by the community of plant scientists can be used broadly and simply, and so that it will remain useful for long after the specific researchers responsible for it have moved on to other subjects
- To serve as an umbrella under which outreach programs in plant biology will be developed, coordinated and disseminated

**Goal 1: New research approaches**

Many possible models for a center exist, and many were discussed. The workshop participants achieved a strong consensus that a center should foster an environment to address large biological problems, rather than focusing on development of methods or materials for the plant science community (though these may result from a problem-focused approach). The scientific goal of the center, then, would be the solution of a critical set of biological problems that require new interdisciplinary applications of biological, engineering, mathematical, computational and possibly other expertise. By doing this, the center would not only advance plant science, but would also serve as a focal point for integration of data and ideas from other laboratories interested in some aspect of the same large problems.

Rather than focusing on a single biological question or approach, a center should enable collaborative research into a range of questions that could span molecular, cellular, organismal, ecological, and evolutionary levels.

As examples of the scale of problem that a center might approach, a set of problem areas discussed in the workshop included the following list. It is only a sample of possible questions that a center could address and which were brought up at the workshop. It is not meant to be an exclusive list of scientific directions on which a center should focus:

- Regulation of temporal and spatial patterns of gene expression, protein abundance and metabolite accumulation in plants
- Regulation of plant growth and development, cell-cell communication and responses of all relevant systems to the environment, with the ultimate goal of modeling how a plant responds to intrinsic and extrinsic cues
- Elucidation of essential mechanisms that underpin how plants evolve and adapt to diverse environments such as the transition from aquatic to land environments

- Integration of all mechanisms of inheritance in plants, for example at the Mendelian and epigenetic levels and their impact at the individual and populational levels
- Identification and manipulation of the factors that limit agricultural productivity and quality
- Understanding the social and biological factors involved in plant domestication
- Understanding biotic and abiotic interactions in the rhizosphere
- Understanding the mechanism and origin of cell types, tissues, organs, and organ systems
- Understanding the functionality and evolution of plant genomes

The center should be designed and staffed so as to facilitate the integration of existing data to enable investigators to answer important questions. To do so it will need to serve the following functions:

- Bring together people with diverse and complementary expertise to facilitate novel and effective collaborations
- Provide a training environment for postdoctoral scholars that will create a new type of scientist who can deal with both quantitative and qualitative biological datasets
- Facilitate the mining of published data for existing information that would be amenable to subsequent database deposition
- Facilitate the creation and maintenance of common representation schemas for cellular components and interactions
- Enable the community to synthesize existing data to provide a framework for the identification of important gaps in our knowledge of plant biology that should be addressed in the future

A plant cyberinfrastructure center, then, should provide the environment, infrastructure and research platform for researchers to answer central questions in plant biology through the use of community resources such as plant genomes, databases, and through the development of new computational approaches. For example, the environment of the center should make it possible for plant scientists to analyze and synthesize large data sets using computational methods developed and executed in collaboration with the cooperation of mathematicians and computer scientists.

**Physical or virtual center**: The issue of whether a virtual center should be considered was discussed; the consensus is that physical proximity of center members is necessary. A number of different mechanisms to achieve this can be envisioned, however, with mechanisms for a center ranging from geographically distributed to a cluster of institutions that are nearby, to a single location. In cases where a center spans large geographic distances, a key component of any center plan would be an outline of how collaboration across these institutions will be facilitated and true community would be established. One model for this that was discussed was a hub and spoke model, wherein a single physical center serves as the locus for the infrastructure and for training, with a series of laboratories that have necessary experimental equipment and expertise

acting in a formal collaboration that allows scientists from the center to pay long-term visits to the collaborating laboratories, and requires scientists from the labs at the ends of the spokes to visit and interact productively with the physical center at the hub. There may be other viable models for the architecture of the center. Regardless of the model, the group felt strongly that the center should be able to operate beyond the boundary of a single institution and that it should be as inclusive as possible in engaging the larger research community.

**Goal 2: Education and Training**

The proposed center should provide a mechanism for the scientific community to train a new generation of world-class scientists in plant biology – scientists who will be facile with mathematical and engineering approaches, and also with cutting-edge methods and intellectual questions in plant biology, and who will therefore represent a new type of scientist, trained from the start in what are now several different disciplines, and capable of taking new types of approaches to fundamental problems of plant science. A center could serve as an important locus for postdoctoral training and for short or extended visits from faculty at other institutions.

**Postdoctoral and faculty training**: Since some postdoctoral scholars and visiting sabbatical faculty may come from diverse backgrounds (e.g., including those with PhDs outside plant biology or with exclusively computational backgrounds), an ideal cyberinfrastructure center would include the opportunity for specialized training in several core areas including cutting-edge "wet" experimental technologies (e.g., tandem mass spectrometry, DNA chip arrays, advanced optical methods, and so on) and "dry" experimental technologies (e.g., bioinformatics methods for data analysis and interpretation, including cluster analysis for the results of expression data studies, homology detection, protein fold prediction and phylogenetic tree construction; advanced image processing and analysis, etc.). A center should provide both the fundamental computational and wet-bench experimental infrastructure to enable participants to investigate scientific questions effectively. It is not necessary that the experimental technologies, resources and expertise be available within the primary host institution (in fact, the workshop participants expect that not all components will be available in any one institution). This infrastructure might be distributed across the several institutions associated with the center, as described above in the "hub and spoke" example, with extended visits of center postdoctoral fellows to the collaborating institutions for experimental work.

**Mentoring and advising of postdoctoral scholars**. Given that postdoctoral scholars may not be assigned to individual faculty members or principal investigators at the center, and may indeed work at the center and at participating laboratories at different times, the question of appropriate mentoring of postdoctoral scholars should be addressed in the organization of the center. Senior investigators at the host institution (or affiliated labs) who are willing to serve as postdoctoral mentors/advisors should be identified. Alternatively, postdoctoral researchers may be asked to form an advisory committee of

his/her choice, and mechanisms by which the postdoc will interact with the committee as a group or individually could be described.

**Think-tank environment**: An ideal center will also provide a think-tank environment in which center participants (postdoctoral scholars, sabbatical visitors, investigators at the home institution and invited scientists) can discuss big questions in plant biology and brainstorm about solutions to these questions. The center should enable participants to envision and possibly develop new technologies that would catalyze advances in the field. A center should ideally provide an environment that attracts world-class investigators in different fields as visiting speakers or participants.

**Communication, inclusiveness and community building**: The workshop participants envisioned the center as functioning as a hub, connecting groups with specialized expertise. Therefore, effective communication mechanisms across all groups associated with the center should be outlined. A center should develop a true cyberinfrastructure that enables all center participants to access and understand the research and activities of the center community as a whole. Ideally, this cyberinfrastructure will enable the greater scientific community to access the results of the center activities (i.e., expanding the broader impacts of the center).

**Achieving these disparate aims:** To accomplish these several aims, one can envision a variety of mechanisms, listed here as examples.

- A center could provide workshops in "best practices" in different technologies and methodologies (or encourage center participants to attend such workshops provided elsewhere).
- The host institution could identify institutions or labs with the requisite experimental infrastructure and expertise to serve as the "spokes" of the wheel.
- Postdoctoral scholars and sabbatical visitors could be encouraged to make visits (of weeks or months, or longer) to collaborating institutions.
- A center may wish to include regular seminars in plant biology and technology, inviting leaders in the field from other institutions as external speakers.
- Technical staff could be included in the Center to construct and maintain online resources (including database development or interactivity, if relevant) presenting the results of Center activities and research efforts.

**Goal 3: Software and Interface Standards**

Implicit in the scientific goals described above is a series of goals in setting standards for data and metadata description and software and computational interfaces, so that the data created by the community of plant scientists, and the work done at the center, could be used broadly and simply and remain useful for long after the specific researchers responsible for it have moved on to other subjects.

To do this, the center should use and encourage the adoption of understandable, published data interchange standards whenever practical. These standards should not apply only to static data objects but to data processing techniques such as pipelines. Where possible, data, metadata, and data processing protocols should be readable by users and machines.

Active support of these standards in the form of consulting experts associated with the center who are conversant in existing standards, and who take a leadership role in convening standards working groups, should be a part of the center's mission. Reference implementations and standards compliance testing suites should be a part of the center's implementation efforts.

An imposed standard creates friction. A consensus standard is intrinsically outdated. A middle road between imposition and community consensus is thus desirable. The center should proactively support the creation or adoption of draft standards, where practical, which are sufficiently extensible and flexible to permit the national and international community to participate in the development of the data exchange and processing cyberinfrastructure. The center's mission could include support for the implementation of standards including acting as a repository for those standards and the associated software libraries.

Some of the types of data (both raw and analyzed) that should be generated, described, exchanged and analyzed in standard ways include: genomic data, transcriptome data, proteomic data, metabolome data, interactome data, phenotypic data including morphologic, developmental, biochemical and anatomical data, image data of several types such as confocal microscopy or photographs of phenotypes, crystallographic and protein structural data, phylogenetic data, relevant geographical information system data such as climate or ecotype, ontologies, mathematical and computational models - both derivation and implementation, biodiversity data, including specimen and observational data, metadata for data and data pipelines.

Providing basic cyberinfrastructure to make these types of data, models, and data and model descriptors available to the broad plant science community should be a key part of the center's core mission. Some of these data standards either exist or are emerging. The center should interact with these groups to facilitate the development and adaptation of community-wide standards and play a central role in facilitating the dissemination of these standards to the larger research community.

**Community Training and Standard Implementation:** A critical element of cyberinfrastructure is the training of community members in the use of these standards and their associated software and applications. To this end the center should enable the development of novel software or integration of existing software into the center's operational and training framework.

Software supported by the center should have a flexible interface useable, when appropriate, by biologists, and an interface specification that permits the linking of new

analytic and visualization modules with a minimum of additional coding. In addition, software should have an understandable, published Application Programming Interface (API). Whenever possible, disseminated software and software standards associated with the center should be widely available. The use of open source is therefore strongly encouraged.

Interoperability of software tools and databases associated with the center should be a part of the center's core mission. Such interoperability is critical for successful cyberinfrastructure to facilitate discovery and third party use of data and tools. This interoperability mission should encompass the broad international plant science community, not only the center's research mission. The center should seek collegial partnerships with existing biological database and repositories, where practical, in consulting on issues of standards and interoperability. The center may serve as a repository and assume responsibility for curation for data sets within its core mission if it is deemed to fill a missing link to enable existing databases to interoperate. The center should seek to enable the persistence of critical data sets and software tools. Grants under the aegis of the center could address the issue of persistence, possibly with center support. An example of such a role could be to provide a permanent home for project-specific plant databases and data sets for which there is not an immediate solution for permanent archiving. As this is a critical issue beyond the plant community, the center should be proactive in enabling a more permanent and robust mechanism of data persistence with other centers.

**Goal 4: Outreach to the broader community**

The Center should serve as an umbrella under which outreach programs will be developed and coordinated. The primary participants in the Center's outreach program should be those integrally involved in outreach. One of the goals of such a program in addition to the target audiences of the outreach programs is to encourage the culture of all scientists involved to become proactive and sharing in terms of outreach, data sharing, data annotation, and management of new technologies.

Utilizing the developed infrastructure and resources in the proposed Center, these efforts should be consolidated into a well-designed program that would have more impact because it would be an integral part of the Center and not just an afterthought. In addition the Center should provide more long-term infrastructure than individual efforts and therefore provide a foundation upon which future outreach and education efforts will be built. The Center should foster an opportunity for close interactions between outreach and science professionals, and the Center outreach program could provide leverage in outreach by providing opportunities for research into impact and needs assessments of science outreach programs in general.

What efforts might a Center engage in with regard to outreach? Some examples are:

- Organize a well-annotated portal of available tools to facilitate outreach (software, websites, Powerpoint presentations) and sponsor workshops to

introduce other scientists to these resources – to avoid reinventing the outreach wheel and to optimize efforts.

- Sponsor workshops for people involved in doing outreach, including scientists and outreach coordinators. Working groups could be organized to discuss specific types of outreach like the workshop organized by Lenore Reiser that led to the publication in *Genetics* (**166**: 1601-9, 2004) about not reinventing the "Outreach wheel". The Center could sponsor workshops that would offer training and equipment for particular outreach efforts. These workshops could explore the specific problems of developing outreach programs with minority populations, such as American Indians.
- Provide travel funds and release time for faculty at minority-serving institutions and community colleges to participate in the ongoing science efforts at the Center – a type of sabbatical for providing scientific training for individuals who would otherwise not have such opportunities. In this light, one suggestion was that postdoctoral fellows at the center may wish to have some teaching experience, and after suitable introductory training, may agree to teach at minority-serving and community colleges as a way of providing teaching leave for faculty at those institutions.

**Outreach Target Audiences:** Since Center outreach cannot effectively support all potential end-users, the target audiences should be prioritized with respect to the relative amount of impact such efforts would have and the appropriate coverage of underrepresented minorities. The workshop participants suggested that the high impact tier could be:

- Journalists
- Policy-makers at all levels
- K-12 and small/community college teachers

**Other Issues for Consideration**

In addition to the goals and missions of a plant cyberinfrastructure center and the possible mechanisms for realizing the goals as articulated above, participants discussed and provided general suggestions for possible core infrastructure and staffing needs, assessment and international collaboration.

**Core infrastructure and staffing:** The Center could support on-site activities as well as activities that take advantage of unique off-site technical facilities. There should be a core infrastructure and staff at the center to accommodate the postdoctoral fellows, visiting scientists, and project participants. The core staff could largely be grouped into Directorate (executive director, education and training director and outreach coordinator), IT support (IT director, systems and database administrators and bioinformatics analyst) and Administration (business administrator and administrative assistant). The Center should be associated with a University or other research institution, to provide access to libraries, advanced scientific computing, and other facilities. However, the Center should also be administered in a way that ensures independence in pursuit of its scientific mission. To ensure effective pursuit of the core mission, implementing

cyberinfrastructure that facilitates synthesis in plant biology, feedback and creative input of scientific staff and visitors into decision making for the center is essential. A key recommendation is that checks and balances be instituted through two bodies, a Scientific Advisory Committee concerned with strategic planning and a Board of Directors concerned with more immediate goals and projects.

**Assessment:** The center should include a robust assessment component both to determine the needs of the scientific community it serves and to measure the impact of its activities, including training and outreach. One possibility of implementing assessment made at the workshop was to form partnerships with social science departments and/or individuals who study large-scale coordination projects. There may be other models of carrying out assessment.

Measurement of impact could include

1. How many visiting scientists who would not normally have access to the various facilities utilize the center?
2. How much synthesis occurred as reflected in the start of new areas of research or the rapid contributions to current areas of research?
3. Is the general public more scientifically literate? If outreach efforts (via Center journalists, etc) are effective, then the average citizen's science knowledge base should have improved. This can be assessed in various ways.
4. Track the careers of postdoctoral fellows that are trained at the Center to observe the manner in which they develop their own research careers in terms of its multidisciplinary nature and their outreach and training activities.

**International collaboration:** An ideal center should be as inclusive as possible in the participation of the projects by scientists internationally, and should leverage other similar activities being carried out in other countries. Internationally, several limited efforts towards improving cyberinfrastructure, data availability and integration are underway, but there is still a gap to bridge until these bear fruit for many plant biologists. An ambitious center could, in this background, play a vital role in consolidating these efforts and bringing them together to have impact on essential plant biology questions. One recommendation is that the center should be open to international scientists as postdoctoral scholars. In addition, the implemented cyberinfrastructure will require global components, and international coordination (e.g., standards for interoperability) is essential. Other efforts towards consolidating cyberinfrastructure for plant biology need to be considered as well as experimental data and analytical tools. The center should be both open and outreaching. For example, information, tools, data and standards should be open internationally, and the center should also actively participate in international standardization efforts and seek interoperability with international data centers. Frequently, funding is restricted to a specific project. The center should aim to provide a space where international projects can meet to discuss and synchronize their approaches, e.g. by setting up workshops for standards implementation. The center should aim at international leadership in defining strategies for plant cyberinfrastructure in an inclusive fashion, i.e., as stated above, not dictating standards, but also not waiting for broad

consensus. International collaboration will be essential for the broad effort required for successful implementation of standards.

The Center could also play a role in coordinating outreach efforts both in developed and developing countries. A loose group of international scientists now attempts to focus on issues (e.g., regulatory, ethical, scientific) of concern to the general public and end users of the technologies that plant scientists develop. With the leadership of a Center these efforts could be made more coordinated and also allow the expertise of scientists integrally involved in these efforts to be made available to others.

It would seem appropriate for this center, which is expected to become one of the most prestigious in the world, to play a role in enabling investigators in developing countries to benefit from cyberinfrastructure. The challenges in accessing cyberinfrastructure are especially acute in developing countries. The center could play an important role in this area. by inviting scientists from developing countries for short internships in the center, providing a limited helpdesk service where scientists from developing countries can request computational support, and including international participants in the workshops and planning boards.

**Conclusion**

A plant science synthesis center could provide a unique opportunity to bring together in a cohesive resource a robust combination of complex infrastructure, support services, expertise, and focus that is not readily achievable by individual research groups. A center can act as a nucleus to draw together communities that, otherwise, may have few opportunities for real interaction. By drawing scientists from diverse fields such as mathematics and statistics, the center could attract individuals who may otherwise not be attracted to the plant sciences. Acting together with center resources, such communities of scientists can bring novel insights and creative new approaches, taking on the grand challenges presented by modern plant biology.

**Recommendations**

1. There is a strong need to create a plant cyberinfrastructure center to promote the integration of diverse and large-scale genomics and other data to address a few fundamental problems in plant biology using multi-disciplinary approaches.

2. A core mission of the center should be to train a new generation of scientists who can combine multi-disciplinary approaches and utilize data in public repositories maximally.

3. The center should be composed of an entity that is connected to a number of adjunct institutions with different types of scientific expertise and facilities.

4. The core infrastructure and staffing should follow the successful models of NCEAS and NESCent and designate enough core staff to facilitate the activities of the center efficiently.

5. The center should provide an umbrella infrastructure to coordinate disparate outreach activities and train students and faculty in achieving a comfort level in interacting with the public about plant science and its importance.

6. The center should maximally leverage existing databases and standards to promote international integration of data.

**Disclaimer**

The workshop described in this report was supported by the National Science Foundation under Grant number DBI-0550931. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the National Science Foundation.