



Planting the Seeds of Cyberinfrastructure: New Projects Nurture Plant Bioinformatics

June 13, 2008

By Vivien Marx

This article has been updated from a previous version to clarify the organizational structure of the Generation Challenge Programme.

While plant genomics often seems to take a backseat to projects focused on human health, several large-scale initiatives hope to change that by developing “cyberinfrastructures” specifically designed to enable research in crop science and plant genomics.

Most recently, the Generation Challenge Programme Platform, a global crop-research consortium, published a [paper](#) describing its progress developing semantic standards and an informatics workbench for crop science. The GCP is an independent research consortium headquartered legally with CIMMYT, the International Maize and Wheat Improvement Center.

The report, which appears in a recent issue of the *International Journal of Plant Genomics*, follows the launch of the Plant Science Cyberinfrastructure Collaborative, or iPlant, which the National Science Foundation funded in January with \$50 million [[BioInform 02-01-08](#)]. In that initiative, researchers from the University of Arizona, Cold Spring Harbor Laboratory, Arizona State University, the University of North Carolina at Wilmington, and Purdue University are establishing a cyberinfrastructure to collaborate on plant biology research.

These projects join PlaNet, a European Commission-funded network of European plant genome-data centers that was set up in 2001 to make it easier to systematically study genomic plant information. The network is developing new ways to facilitate data exchange and database integration via the BioMoby framework. Over 160 web services are integrated into the PlaNet platform so far.

Generation Challenge Programme

Four years ago, the GCP team began developing a bioinformatics platform with a focus on crop research, comparative biology, and genomics, and an eye toward bringing new research to bear on situations affecting resource-poor farmers, such as drought and pests.

Elements of the project include developing domain models, ontologies, and data formats that permit interoperability, and web services to share and integrate data. The project also plans to offer access to high-performance computing centers to support high-throughput data analysis as well as middleware to integrate databases and tools into a unified workbench.

“To my knowledge, and I stand to be corrected, attempting to achieve interoperability between 22 global partners working in

Richard Bruskiwich, a researcher at the Crop Research Informatics Laboratory of the International Rice Research Institute, which is a joint venture of IRRI and CIMMYT, explained in an e-mail to *BioInform* that the platform is still being rolled out. His work focuses on managing rice functional genomics data and conventional crop breeding data, and integrating this data into IRRI’s research. He is also the principal investigator on the GCP domain model project.

As the scientists explained in their paper, the platform is set to be a user-friendly but extensible “workbench” that provides interoperability and data access across all GCP partner sites and, at a later date, across the global plant crop research community.

advanced global crop research has not been directly attempted before to this level of scope complexity of research.”

“To my knowledge, and I stand to be corrected, attempting to achieve interoperability between 22 global partners working in advanced global crop research has not been directly attempted before to this level of scope complexity of research,” Bruskiwich said.

He noted that the National Cancer Institute’s Cancer Biomedical Informatics Grid network is an example of GCP’s vision, but said that there has not been a similar effort applied to plant research so far.

“I think the closest we’ve come to meeting the challenge is that using a domain model-driven architecture seems a unifying approach, since a platform-independent model, such as the GCP domain model attempts to be, may be translated into several ‘platform-specific implementations’ and tie together truly different bedfellows into one system,” said Bruskiwich.

Starting with a Napkin

Bruskiwich explained that the initial idea for GCP “was born on the back of a table napkin” at a 2001 meeting of the Consultative Group for International Agricultural Research, or CGIAR, an association of donors in the area of agriculture and resource management that targets farmers and consumers in the developing world.

CGIAR had been seeking ways to tackle “big” research problems such as water scarcity, and to find opportunities to integrate advances in genomics research, he said.

The IRRI was invited to join the CGIAR project, along with the International Plant Genetic Resources Institute (now called Bioversity) to shape the ideas that eventually grew to become the Generation Challenge Programme. The CIMMYT, along with IRRI and Bioversity, collaborated to get GCP off the ground in late 2001.

Jean Marcel Ribaut is the program director of the GCP, and presides over the GCP from CIMMYT. Theo van Hintum, senior scientist at the Centre for Genetic Resources in Wageningen, the Netherlands, is the current sub-program leader for crop informatics at GCP and Graham McLaren, biometrician and IRRI’s head of the Crop Research Informatics Laboratory, is taking over in the next few months.

The GCP draws together the agricultural research institutes of CGIAR with other research institutes in developed and developing countries.

The IJPG study pointed out that an integrated platform of molecular biology and bioinformatics is “central to GCP activities.”

Funding for the GCP has come from various sources, such as the World Bank, the European Union, and the UK’s Department for International Development. Bruskiwich said that the group has also had interactions with industry, such as Pioneer, Monsanto, and Bayer Crop Sciences, as well as smaller stakeholders. He did not elaborate.

The developers of the GCP crop information platform have to contend with “a large, globally distributed consortium with diverse research, requiring a diversity of tools, large data sets and diverse data types, many legacy informatics systems and tools, and the need for global data integration,” he said.

Designing the Principles

As Bruskiwich explained, in order to create “next-generation” solutions that will enable global collaboration in agricultural research, GCP identified several key design principles, including the specification of a common domain model and shared ontology standards for data integration and analysis; the development of a model-driven software architecture with common software; standards and libraries to wrap existing and new databases and tools for interoperability; and the use of web services to cross-link databases and tools.

The domain model that was developed is “partly based on existing domain expertise of the bioinformaticians and informatics developers of the team, partly by borrowing ideas and data models from other similar international efforts,” Bruskiwich said. These other efforts include CGIAR’s International Crop Information System, the Generic Model Organism Database project, and the Functional Genomics Experiment model, or FUGO, he said.

This model is being refined on an as-needed basis based on its practical application to wrap specific databases and connect specific target third party tools, he added.

Progress in integrating third-party tools has not been as fast as the team would have liked, Bruskiwich said, but he noted that the project has so far integrated the Apollo genome browser, the National Center for Genome Resources’ Comparative Map and Trait Viewer, the Institute for Genome Research’s TMeV gene expression software, the

University of Manchester's MAXD gene expression database and tools, and Cytoscape.

In addition, the platform supports several web service protocols, including SoapLab, BioMoby, TAPIR, and the Virtual Plant Information Network/Simple Semantic Web Access Protocol.

As the researchers explained in the paper, the domain model has generic core model interfaces from which scientific model interfaces are derived. "The decision of what to put in the domain model as object classes, and what semantics to delegate to ontology dictionaries, is largely a bioinformatics judgment call," Bruskiewich said. "The boundaries of the two are sometimes fuzzy; the domain is very large, [which makes this] not an easy task."

The scientists debated the use of Unified Modeling Language versus Ontology Web Language and picked UML. Next was to find a common UML representation and link it to software engineering. As Bruskiewich explained, the idea was to initially focus mainly on Java language implementations of the GCP platform.

Finding a UML tool that also generated Java interfaces and classes was not easy in 2005. "As it happens, industry standards for UML tools were very problematic at the time," he said. For a while, the team used a tool he did not specify that was initially available free of charge for academic users but had to be abandoned when the tool was commercialized.

By 2006, the developers settled on the Open Source Eclipse Modeling Framework, he said.

Interoperating Around the World

The GCP platform is "well on the way" toward interoperating with various data formats and web-service protocols, including BioMoby, BioCASE, GPC, and others, Bruskiewich said.

The project is also drawing from several public ontology standards for genomics and crop data such as the Gene Ontology, the Plant Ontology, IPGRI/Bioversity passport descriptors, and the Ontology for Biomedical Investigations.

"We don't reinvent an ontology if a widely developed community ontology for a given domain already exists," he said. "What we are doing, though, is ensuring that our GCP platform handles a wide range of ontolog[ies] in a consistent fashion."

The basic vision is to have "plug-and-play standards that can be applied to wrap any database relevant to crop research and to enable any target software tool to access those databases," he said.

"This is an extremely ambitious vision and a big 'challenge' — there are some rainy days here that I question whether or not it is truly feasible, or whether or not we are the best team to tackle such an ambitious enterprise — but there it is," he said.

The team plans to showcase the next phase of its work at the annual GCP review meeting in Bangkok in September.

"I think the basic vision of a model-driven ... and web services-driven architecture remains essentially sound, but it is hard to say whether or not we are adequately resourced to realize the full vision," he said, adding that other collaborative efforts such as caBIG face many of the same "design and implementation challenges and frustrations."

Another Project Sprouts

Bruskiewich said that he and his colleagues are following with interest the NSF-funded iPlant Collaborative, which, too, "will need standards and technology comparable to what we've been struggling to develop, on a much smaller budget, over the past four years."

The iPlant project "is the new kid on the block" and "similar in spirit" to GCP's goals, he said. "There certainly is some overlap between iPlant and GCP, but it is not 100 percent in that GCP is a bit more downstream of basic plant research in *Arabidopsis*."

The iPlant framework software development team is led by Lincoln Stein, a researcher at Cold Spring Harbor Laboratory who has contributed heavily to many of the resources and technology used in the GCP, such as the Generic Model Organism Database project, the Gramene comparative cereals database, and BioMoby.

The crop informatics team at GCP is collaborating with US-based partners such as Gramene, the Plant Ontology Consortium, and USDA Germplasm Resources Information Network. "We are hopeful of the potential for synergy with US-based efforts," Bruskiewich said.

In Europe, GCP is working with Bioversity, CIRAD, the Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Wageningen University, the Taxonomic Database Working Group, the EURISCO

genetic resources network, and other partners.

"You can tell who the pioneers are by the arrows in their back," Bruskiewich said. "I feel that we have a lot of arrows sticking out, but we are optimistic that we are making steady progress on the implementation of the vision of the GCP platform."

Matthew Vaughn, a CSHL researcher on the iPlant team, explained in an e-mail to *BioInform* that the goal of that project is to "provide a comprehensive, integrative cyberinfrastructure" to the Plant Sciences community that "is sufficiently forward-facing to facilitate addressing and perhaps solving major-scale 'Grand Challenges' in plant sciences, provide greater degrees of access to next-generation biological tools and data sets to scientists working at smaller institutions, and engage the lay public about biological topics and computational thinking."

The infrastructure in the making will include so-called Discovery Environments, which are software platforms intended to solve the "grand challenges" of interdisciplinary research, such as querying heterogeneous data sets.

Vaughn said that the Discovery Environments are currently at the prototype stage.

"We await direction from the community on what they perceive the Grand Challenges in Plant Biology [to be], and from there, we will try to develop a cyberinfrastructure that facilitates progress towards solving them," he said.

For now, the team is accepting proposals from the plant sciences community, which will lead to a series of workshops. "These planning workshops will form the basis of Grand Challenge teams, who will collaborate with iPlant to build portions of the cyberinfrastructure pertinent to their topics of interest," said Vaughn.

As for the way iPlant and the Generation Challenge Programme relate, Vaughn said, "You could think of GCP as analogous to one of the Grand Challenge teams that will work with iPlant Collaborative. In this case, their Grand Challenge is to provide next-generation agricultural tools for developing world farmers."

Vaughn added that he would not be surprised to see a Grand Challenge proposal submitted "that's in a vein similar to Generation Challenge Program's goals."

© Copyright 2008 GenomeWeb Daily News. All rights Reserved.