iPG2P_Progress

Executive Summary

A solution to the genotype-to-phenotype (G2P) problem can be described as an analytic process that allows an investigator to begin with a trait of interest in a species possessing limited genetic resources and progress towards the ability to predict trait scores for known genotypes in given, non-constant environments. To this end, the iPG2P Steering Committee is identifying common, abstract workflows that will support this process and has defined high-level use cases that the working groups should strive to implement. The two prioritized use cases are 1) Carbon metabolism (including C3 and C4) and flowering time and 2) Hypothesis generation through data mining, processing, and visualization. (need to attach summary document)

Working Group Formation

In late July 2009, 16 scientific research community members with interests in phenology, drought stress, photosynthesis, bioinformatics, and machine learning and 8 iPlant faculty/staff members met in Chicago at the “iPG2P Project Kickoff” meeting to establish the specific focal areas for this Grand Challenge and to develop a high-level implementation plan. Prior to this meeting, a survey was conducted among planning group members and meeting participants to gain insight into what cyberinfrastructure areas were of highest interest to the participants and what use cases might be established. At the meeting, a series of plenary sessions with the entire group, smaller breakout sessions, and extensive group discussions were used to define five working groups that would cover the intellectual and technical range of the "genotype-to-phenotype" problem. Leads/co-leads for each working group were recommended by consensus and working group membership was drafted from appropriate meeting participants and non-participating community members.

Working groups are comprised of a lead and co-lead, who are senior investigators in fields pertinent to the group's focus, one or more iPlant Engagement Team Analysts, and five to ten invited members of the plant science, computer science, and biological informatics communities. They are advised and guided by the G2P Scientific Lead, and are supported by the G2P Project Manager and an Administrative Assistant. Groups meet biweekly via teleconference, have access to a shared collaborative web space, and use of group-specific mailing lists.

Working Group Progress

- NextGen Sequencing Pipeline: The goal of the NextGen Sequencing Pipeline working group is to develop tools to permit efficient use of next generation sequencing (NGS) data by members of the plant science research community involved in genotype to phenotype research. Requirements analysis for this working group is the most advanced. A draft statement has been developed describing the first iteration of a NGS discovery environment, which includes a core web application framework with data upload capability, user authentication and collaboration tools, pre-processing and quality control tools, support for a variant detection workflow, and support for a transcript quantitation workflow. Also included is support for command-line access to RESTful services for advanced users. Software development on this iteration is scheduled to start in early Q1 2010. The working group is currently working to assemble a list of prioritized development activities for future releases. The group is also working in collaboration with the Visual Analytics Working Group to understand the types and forms of data that should emerge from NGS workflows to best facilitate visualization. In conjunction with the Data Integration Working Group, NGS is defining what data needs to flow through the pipelines. In addition, several matters of standards and practices are being addressed via formation of sub-working groups: base-calling in polyploid genomes, standard formats for representing variants, standard formats for representing sequencing-based transcriptional data, and logical representation of genomic structural variants.

- Statistical Inference: The Statistical Inference group is working to develop a Discovery Environment that can make advanced computational approaches to statistically link genotype to phenotype more available to the general user and more rapid for the specialist user. A first iteration for this system will be described by the end of Q1 2010. The group has identified and prioritized general classes of statistical genetics methods that should be supported by this platform. These include General Linear Models (GLM), Mixed Models, Machine Learning and Bayesian approaches. General Linear Models, being most pertinent to the widest cross-section of plant biologists, is being addressed first. A test implementation of parallel GLM is being developed by iPlant scientific programmers, which should enable larger, more intensive genetic mapping analyses to be conducted. A prototype of this tool is expected to be complete in Q1 of 2010. In addition, a comprehensive description of GLM-based QTL analysis is being developed for two research computing teams affiliated with iPlant who will develop implementations of this algorithm for GPU and FPGA architectures, with the goal of dramatically decreasing execution time for GLM analyses. The group has also initiated discussions with the Visual Analytics group on how to view and explore the large (2.5E+6 points) multidimensional data sets which are expected to emerge from genetic association studies with the advent of relatively inexpensive whole-genome resequencing, as wel as how to make the results of such analyses more accessible to the general research community. Finally, they are working to develop universal standards for defining and describing genotype/phenotype mapping experiments.

- Modeling Tools: The Modeling Tools working group seeks to develop framework tools to support construction, parameter and confidence estimation, sensitivity analysis, verification testing, and utilization of models. To date, an exemplar modeling workflow has been described and is currently under review by the working group. This workflow includes integration with the activities and products of the NextGen, Statistical Inference, and Visual Analytics working groups. In addition, the group is evaluating model description languages such as SBML and modeling repositories and platforms such as BioModels.net and OpenMI for potential synergy with iPlant-led efforts.
- Visual Analytics: The goal of the Visual Analytics working group is a Discovery Environment capable of displaying diverse types of data from laboratory, field, in silico analyses and simulations, and other sources specific to genotype-to-phenotype research in ways that reveal underlying patterns, lead to novel hypotheses, provide concise syntheses, and support publication, collaborations, and education. In early November 2009, members of the Visual Analytics working group met at TACC to bring together the plant biologists and computer scientists in the group to develop a mutual understanding between them. During this meeting, the working group identified major issues in the G2P analysis and emerged with test cases for a canvas/widget approach to analysis and visualization. Currently, a series of demonstration applications are being developed to showcase this approach. Workflows have been generated to describe visualization needs for a "Maize Gene Analysis" and of an "Interactive Gene Expression and Metabolomics Analysis".
- Data Integration: The Data Integration group seeks to build Discovery Environment software atop existing middle-ware systems that use metadata to achieve situational awareness of available data, logical relationships between different data sets, and tools that enable users to find relevant information even when they are not sure what data may exist. The group is currently analyzing workflows from the various G2P working groups as they are developed. One initial focus has been on identifying data integration needs in the NextGen Sequencing variant detection and transcript abundance workflows. To this end, a survey was created and distributed to reference sequence data providers asking for details on the types and formats of data they offer, as well as inquiring about communications standards and methods for integration service applications. A similar questionnaire will be sent out to reference sources identified in the Visual Analtyics (maize gene analysis and stress biology analysis) and Statistical Inference workflows (GWAS/QTL mapping). The group has also begin to define metadata/provenance/data quality standards for iPlant, in collaboration with Sudha Ram and her graduate students. Finally, the working group is exploring approaches for integrating expression data, molecular pathways, metabolite profiles, and biological networks.

Engagement Team

- The composition and role of the iPG2P Engagement Team and working groups are described. The Engagement Team is the primary source of contact between the five working groups. The Engagement Team has two main roles: to perform requirements analysis and to provide project management and coordination for the working groups.

The iPG2P Engagement Team

The Engagement team is an outward facing part of iPlant that serves as the interface between the Grand Challenge projects and the cyberinfrastucture developers.

Team Composition

Leaders

Matthew Vaughn, Scientific lead

Karla Gendler, Project Manager

Administration

Tina Lee, Special Assistant
Tina Lee is a Special Assistant at the iPlant Collaborative, assisting with organizing iPlant events, such as the GC workshops and team meetings. With degrees in Biology and Anthropology, she knows just enough to be dangerous, and is thus helping produce *The iPlant Leaflet* (e-newsletter). Prior to working at iPlant, Tina was an aide to the Tucson City Council and an environmental consultant specializing in water resources and land use planning for the private and public sectors.

Mary Margaret Sprinkle, Special Assistant
Mary Margaret has degrees in Business Administration. Prior to working for iPlant, she was the project manager for an NSF-funded Science and Technology Center based at UC Berkeley. She is based at the University of Arizona. Mary Margaret provides administrative and budget support for the iPG2P engagement team.

Engagement Team Analysts

Adam Kubach
Adam has a BS in computer science and extensive experience in C++, Python, Java, OpenGL/scientific and geospatial visualization, multi-threaded programming, software architecture and design and SQL.

Zhenyuan (Jerry) Lu
Jerry received MS degrees in both Computer Science and Biochemistry. His scientific background is in scientific visualization, large-scale data integration, numeric simulation and rapid prototype development for biological applications. His computing experience is in Perl, Java; MySQL and Oracle databases.

Bernice Rogowitz
Bernice obtained her PhD in psychology from Columbia University and postdoctoral training in phychophysics at Harvard University. She is a fellow in the Society for Optics, Photonics and Imaging. She is currently based at TACC. Bernice has just joined iPlant and will contribute her expertise in human perception and data visualization in both the iPToL and iPG2P visualization working groups.

Liya Wang
Liya received his PhD in biophysics. His research experience includes algorithm development, protein NMR, image analysis, signal analysis, optimization and machine learning. He has computational expertise in MATLAB, PHP (MySQL), c/C++, JAVA, Python, Perl and R.

Floating Team members
Staff members at TACC and CSHL can be brought in on a shorter term basis as required. TACC members contribute high performance computing and software engineering expertise and CSHL staff contribute expertise in bioinformatics.

Engagement Team Role

The primary purpose of the Engagement Team is to translate the needs of the biologist to help guide the computer scientists in creating the appropriate cyberinfrastructure. This involves several components:

Requirements Analysis

Broadly stated, requirements analysis involves the identification of high level software and infrastructure deliverables that will address research and analysis needs. The deliverables and high-level requirements are documented by the Engagement Team. In collaboration with the core software needs analysis specialists, these are decomposed into logical components, dependencies, user personas, user stories etc. that will inform orderly development of the final software product. The process is bilateral and involves further iteration with the domain experts in the working groups as required.

Project Management and Coordination

Another key responsibility of the Engagement Team is to coordinate activities of the working groups, the Engagement Team, the Core Software group and other iPlant staff. This support consists of meeting and other administrative logistics, budgeting, staffing, planning, etc. The Engagement Team has a project manager who works with all stakeholders to assure that documentary requirements are met, that milestones and time-lines are mapped out, that progress is monitored and that necessary adjustments are made to ensure that the project succeeds.

Prototyping

Another facet of requirements analysis is to test ideas and generate proofs of concept that will facilitate early innovation and inform promising directions for further development. Although the scope of future prototypes will be limited, this engagement method will be employed in some cases to facilitate needs assessment prior to commitment of resources to core software development.

The iPG2P Working Groups

The iPlant Genotype to Phenotype (iPG2P) Grand Challenge Project is currently composed of five working groups with specific development goals: NextGen Sequencing Pipeline, Modeling Tools, Statistical Inference, Visual Analytics, and Data Integration. Each working group is composed of a lead, co-lead, members of the scientific community, and iPlant Engagement Team Analysts (ETAs).

Ultra High Throughput Sequencing Working Group

https://pods.iplantcollaborative.org/wiki/display/ipg2p/NextGen+Sequence+Pipeline

Unable to render {include} The included page could not be found.

Modeling Tools Working Group

https://pods.iplantcollaborative.org/wiki/display/ipg2p/Modeling+Tools

Unable to render {include} The included page could not be found.

Statistical Inference Working Group

https://pods.iplantcollaborative.org/wiki/display/ig2p/Statistical+Inference

Unable to render {include} The included page could not be found.

Visual Analytics Working Group

https://pods.iplantcollaborative.org/wiki/display/ipg2p/Visual+Analytics

Unable to render {include} The included page could not be found.

Data Integration Working Group

https://pods.iplantcollaborative.org/wiki/display/ipg2p/Data+Integration

Unable to render {include} The included page could not be found.

Other Collaboration Projects

The following projects represent other areas of interest seeing attention by the iPG2P group.

BrachyBio!

https://pods.iplantcollaborative.org/wiki/display/ipg2p/BrachyBio

Unable to render {include} The included page could not be found.

High Throughput Image Analysis

https://pods.iplantcollaborative.org/wiki/display/ipg2p/PhytoBisque

Unable to render {include} The included page could not be found.

Core Software and the iPG2P Discovery Environment

Core Software

https://pods.iplantcollaborative.org/wiki/display/coresw/Home

The iPlant core software engineering group is based at the Unversity of Arizona. It serves all grand challenge projects as well as other iPlant cyberinfrastructure development activities. The lead developer in this group is Sonya Lowry.

Interface with the iPG2P engagement team

Management

Planning and high level design issues are communicated directly from Matt Vaughn and Karla Gendler to Sonya Lowry, the lead developer. This is primarily off-line communication.

Requirements Analysis and Development

Each engagement team analyst (ETA) has primary responsibility for at least one working group and secondary responsibility for a second. The ETAs attend working group meetings and work directly with the working group lead and other members to assess the scientific and technical requirements of the working group. The ETAs then communicate these requirements, with appropriate triage and refactoring, to Nicole Hopkins, core software's needs assessment specialist. Acquisition of scientific and computational domain knowledge for the working is primarily the responsibility of the ETA.

Transparency

There is direct communication on detailed needs assessment between all members of the engagement team and Nicole Hopkins, the core software needs analysis specialist. This level of communication is almost entirely documented on the confluence wiki space for core software (http s://pods.iplantcollaborative.org/wiki/display/testdev/Home+-+Core+Software). The confluence wiki helps to track design discussions and development issues associated with the discovery environment. Engagement team members also contribute documents and comments to the core software wiki space. Detailed development issues are tracked internally by the core software group on the JIRA content management system, excerpts of which are also posted on the wiki. Core software group members also post "virtual standup" reports on an ongoing basis on the wiki.

Meetings

There is a bi-weekly design and retrospective meeting held by the core software group and attended by the iPG2P engagement team analysts. This is a platform for discussion of architecture design decisions and detailed reports on development activities.

Near Term Road Map

Gantt Charts

Unable to render {include} The included page could not be found.