MT_20100217

iPG2P Modeling Tools Minutes
February 17, 2010; 8am to 1:00pm CST
Kansas City, KS

Present: Steve Welch, Jeff White, Chris Myers, Melanie Correll, Ann Stapleton, Christos Noutsos, Matt Vaughn, Karla Gendler

The meeting was convened at 8am CST.

Agenda/Presentations:

Time	Session/Topics for Discussion	Discussants	Action Items/Major Decisions
8 AM - 8:15 AM	Recap of Day 1	Steve	
8:15 - 9:40	Identify and Match Datasets to Use Cases	All	Wish list/Action Items: • A way to manage data that is in the collection process (tools that make it easy to collect metadata/provenance that puts them in to tabular data model that Jordan had identified) • White to provide Jordan example data as a test case for Universal Semanic Model
9:40 - 10:00	Datasets Action Items		
10:00 - 10:20	**Coffee-Snack Break**		
10:20 - 12 Noon	Prioritize EOT Aspects & Action Items	Steve	Looking for a variety of models to target high-schoolers, possibly earlier, that would involve running models in an OpenGrid type environment. Users would see the outputs and be able to interact with the models much like it was running on their machine. Educational material would be provided. The models should address issues that are of interest: climate change, plant processes, and local food growth.
Noon - 1 PM	**Adjourn & Boxed Lunches (to eat in or take with you)**		

Notes/Summary

Recap of Day 1
Welch started the meeting by recapping yesterday's talks and decisions. The group talked about kinds of things they wanted to see; an initial release focusing on parameter estimation and sensitivity analysis; links to visualization programs; specific details of algorithms used and exports that they'd like to see from the system.

Identify and Match Datasets to Use Cases
In the Steering Committee, the group has talked through use cases relating to C3/C4 modeling and another relating to phenology specifically looking at interactions between photosynthesis and its impact on phenology. An idea is to maybe use the C3 model that exists and connect that to some of phenology models and perhaps provide photosynthesis input to that. In relation to these use cases, it would be good to talk about specific datasets that might be useful. Chris Jordan has been thinking about general data structures, looking at some generic abstractions. When the data and metadata are separated, there are two general types of data: 1) tabular grid of things and 2) data that has metadata with it but grid of points in multi-dimensional space with some things have geometry associate with them. The group needs to start thinking about things to see if the model is generic enough to capture what is needed and if not, then the group needs a basis for refining it.

There are two types of metadata: experimental or descriptive metadata and a semantic description of what the data is. Descriptive is experimental data and/or computational provenance within the domain of the system; semantic describes what it is (i.e. tabular data) and is used for at a higher level of control. One principle to follow would be that a user doesn't lose any of where the data came from (for credit and publications purposes). iPlant does not want to destroy, discard, and/or modify any of the metadata coming in. With modeling, there is probably unique data that needs to be captured and the group should define what that is. It is safe to assume that iPlant will store all of the basic information. Welch stated that the group has two things to talk about: 1) specific data sets that we know are out there relating to use cases and 2) data items relating to metadata.

Myers then described the C3/C4 modeling use case, explaining that it comes out of a specific project to demonstrate the utility in introducing C4 photosynthesis into rice. There is an adjunct project, led by Tom Brutnell, that is generating omics data on gene expression, proteomics in developing leaves and grass etc. The work was first done with maize and now segueing to rice with work also starting on sorghum (EMS mutagenesis). On one hand, the project is trying to make sense of all the system level data in looking for candidate genes involved in C4

differentiation and on the other, is looking at modeling to see how could you push onto a C3 model to make it more C4 like without destroying the basic function.

Welch added that in the FIBR project, they were looking for reasons, at the genetic level, why empirical phenology models are as skillful as they are in terms of predicitive capabilities. They are looking if Arabidopsis can work in a wheat context and vice-versa. White added that the immediate goal is to improve the wheat model and then would attack it.

Datasets

In regards to datasets, Myers is hoping to get his manuscript out so that the data will become available to the public. Stapleton suggested Mark Stitiz's Arabidopsis data. Myers wondered how ready the model is off the shelf and Welch suggested that the group can look at it and decide. Vaughn asked how do you hook up the photosynthesis model physically. Both models are rate driven models. At the very simplest level, you would pick an output of the model and use it as a "sugar signal" or proxy. In the phenology model, you would then use that variable to drive it. Myers added that there are specific data sets that are being generated and maybe an idea would be to look at genetic variation in Arabidopsis.

Welch then asked the group to think about what are the right things to be done with data from a modeling perspective (keeping track of provenance) and if there are tools that would be helpful to help manage data when it came out of the lab. He also asked what is needed for collaboration and for upfront data capture. With data capture, the data is coming from underlying data resources. Stapleton suggested picking one data set for testing purposes and collecting the information that is needed.

For major classes of ecophysiological models, there are standard formats developed. These could be used to get an idea and identify categories of information that we've talked about. Myers said that there is a need for data for the models but the group also should identify what needs to come out of the models. Welch suggested that light and temperature are in common across models, with enzyme levels and planting dates differing. Myers added that SBML is very good for capturing data but is not good at describing it. His group has had to think through a set of abstractions that describe the data. Vaughn said that a controlled vocabulary is needed. The suggestion was to have White send Chris Jordan his data file and ask him what would he do. White also presented some of his visualization work related to the NAM work and this can be found here .

Welch summarized that a need that the group sees at this point relative to modeling will be ways to manage data that is in the collection process (tools that make it easy to collect metadata/provenance that puts them in to tabular data model that Jordan had identified). Jordan and the group will want to look at White's data scheme as a test case to identify issues when it comes be worked with in terms of encapsulation.

Prioritize EOT Aspects & Action Items
Vaughn began discussion by talking about points that were brought up over dinner last night. DNA Subway is an excellent example of an educational based DE and it would be nice to identify similar overlap and synergy with modeling. Some ideas would include taking the NAM data and projecting it across expected global climate change models where the models run in real time and one can slide forward in time or change parameters and see how model predictions change in response to global climate change. In reference to Ed's project, it is nice but it would just be another avenue that iPlant would have to pursue. He would like to see educational requirements exist as first class citizens along with the scientific requirements. He suggested focusing on an application that runs a very specific model and then iPlant can make it slick.

Myers suggested that the group could exploit models with "what-if" scenarios. White added that class activities would have to be low time commitment and things that are science fair type of projects. Stapleton said that the buzz phrase is "authentic experience", someone other than you, the student, has to care about the work/experiment you are doing. Welch commented that the group could have workshops for teachers where some information on plants is presented with this information is assisted by models. Myers proposed using the programming skills of younger kids to help with open-source of DSSAT and to look into Google Summer of Code (maybe next year?). He also said that the climate change aspect of things could be a major hook.

Welch was thinking that the group could develop things that run models that run in a distributed model (like OpenGrid). There would be educational materials delivered and could use existing models. What kids in the school see is simulations of plants growing and carrying out growth in conditions/parameters specified. It would be almost entirely web-centric other than training of the teachers. Researchers could get actual scientific calculations and would be much more public and service-oriented. Myers said another hook might be using simulations of plant growth to get people interested in growing their own food.

Welch summarized the discussion stating that the group is \looking for a variety of models to target high-schoolers, possibly earlier, that would involve models running in an OpenGrid type environment (users would see the outputs so it would be like running on their machine) that they can interact with and that has educational materials that goes with it. These models should address issues that are of interest: climate change, plant processes, and local food growth.