MT_20100216

iPG2P Modeling Tools Minutes
February 16, 2010; 8am to 5:30pm CST
Kansas City, KS

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

The meeting was convened at 8am CST.

Agenda/Presentations:

Time	Session/Topics for Discussion	Discussants	Action Items/Decisions
8 AM - 8:15 AM	Welcome & Goals of the Meeting	Steve	
8:15 - 9:35	Core Modeling Tools, Work Flows, Audiences & Discussion	4 Discussants (20 min each)	
9:35 - 10:00	Discussion		Group focus on: 1. parameter estimation 2. sensitivity analysis (change input parameter and see how it changes the output; what happens as you change multiple parameters) 3. interface with QTL mapping activity
10:00 - 10:20	**Coffee-Snack Break**		
10:20 - 12 Noon	Parameter Estimation (inc. estimation post-processing such as ANOVA runs)	All	Wish list: 1. A utility to handle the distribution of data amongst processors that is generic enough to be responsive to the different types of data 2. Different optimizations and algorithms a. Nelder-Mead: used as stand alone or local search algorithm b. Gradient Descent: used as stand alone or local search c. Population based algorithm
Noon - 1 PM	**Lunch**		
1:00 - 2:40	Sensitivity / Response Surface Analysis	All	Wish list: 1. Ability to supply level information with data to use during parameter estimation 2. Ability to tell computer to do separate estimations done for one of those level factors 3. Ability to output all of the data (including factor levels) for offline analysis 4. Bootstrap analysis in one level 5. Encourage people to do offline analysis 6. Support MC simulation in neighborhood of optimum either using a. DAKOTA - Design Analysis Kit for Optimization and Terascale Applications (http://www.cs.sandia.gov/DAKOTA/software.html) b. Keeping archives of data/trial solutions from which subsets can be selection by various criteria 7. Within CI, want ability to output predicted vs observed observations for analysis by ViVA 8. For models that are differentiable by computer, support eigenvector analysis of sensitivity matrices

2:40 - 3:00	**Afternoon Break**		
3:00 - 5:00	Conveying Outputs of Parameter Estimation and Sensitivity Analysis to Other Programs (e.g., Visualization and QTL Analysis)	All	Send visualization experts (Bernice, Greg) datasets with guidance as to what types of relationships you would like to see. Some examples of questions to ask of the data: • Does multivariate data fit into a smaller dimensional space? For example, is it a line that curves a certain direction? • Is it one data set or can it be subdivided? (is it one cloud or several separated ones) • What does it look like when you separate it out? (when you have subgroups)? Does subgroup correspond to labels • Want some characterization of the outer surface; reconstruct the geometry of these regions • Represent ensemble of time series generated by an ensemble of parameters or output variables
5:00 - 5:30	Recap and Action Items	Matt/Steve	
5:30 PM	**Adjourn**		

Notes/Summary

Welcome & Goals of the Meeting
The meeting opened with introductions of participants. Steve Welch stated that the goals of the meeting were to have focused discussions on modeling requirements and to begin to develop a workflow that address the different styles of modeling that were represented in the room.

Core Modeling Tools, Work Flows, Audiences & Discussion 4 Discussants (20 min each)
In order to have focused discussions on modeling, Welch thought it would be good if members in the room talked about their recent activities.

Audiences

Ann Stapleton presented the research that she has done regarding the audiences for modeling. She gave everyone a handout and asked for input. The updated handout can be found here .

Stapleton commented that the grand challenges are in plant biology; thus the audiences would include faculty and their associated postdoc students. However, a lot of faculty come from the single mutant paradigm, are ecologically minded, and are studying particular genes in particular pathways. Others are people who do modeling. There is a large group who also does QTL analysis but there is not much overlap between these people and those that perform single gene type analysis.

There exist strong connections between people who do QTL analysis and the modeling community but there are few connections between the modeling community and those that work in a single pathway/single gene analysis group. All groups focus on gene regulation networks (either top-down or bottom-up). A modeler will often ask what are the implications of the circuitry, what is the circuitry, what to predict for phenotype. Data will be combined, the model will be parameterized, and verification will be done.

White asked how does breeding or extension for growers fit into the model presented. He commented that the world of plant sciences has to get a lot closer to agriculture and that this group needs to decide what the real scope of the iPlant project is going to be. White provided the following example of how a grower might interact with websites and data that is already out there: A grower might go to a website for details of what to plant and when to plant for the location that they are at. They would ask what variety of plants to plant and would want to pull in short-term climate forecast. Jim Jones is working with OpenAg Climate to develop an open source platform for climate.

Workflows

During the Steering Committee meeting in January, the group started to talk about about getting people to work with real data and suggested the NAM dataset would be a good set to start with. White's presentation relating to his work with this dataset can be found here .

For phenology data, he was looking at time of anthesis and time of silking. With the NAM set, there are 26 true populations, 27 with association mapping and over 200 lines per each of 25 crosses. White commented that working with the NAM data was a testimony to the power of good data organization as he could get data in 3-4 days where before it would take weeks. He was able to get phenotypic information and defined provenance as where did the phenotypic information come from (when it was planted, where it was planted, how big of a plot, how many plants per meter, row spacing, weather conditions, water and nitrogen management, application of growth regulators, initial soil conditions). For the simplest phenology studies, one needs latitude and weather information. He could not find information on fertilizer and irrigation. He also needed to know about temperature regime. Some problems with the data included that it was only from the start of planting and ended at maturity end (insufficient data), which is a problem for optimization and parameterization. TexasA&M has daily weather for modeling and other decision support (might have experiment station data) and using their site can take care of formatting issues. NASA/POWER has a data set for solar radiation. For soil data, he used Google maps that highlight where NRCS has taken soil samples.

White stated that the above process is pretty typical of what he goes through when ramping up to work on modeling a new species. Data has to be imported, checked and formatted. As an example of checking, with weather, he uses simple checks to make sure the minimum is lower than

max. The DSSAT package tool has the ability to check weather data. White said it is easier for him to write a SAS program and start merging spreadsheets together. Once checked and formatted, he has a model-ready data set. However, DSSAT has the limitation of only allowing 999 treatments and with this dataset, there are 55k treatments. White wrote a Python script to simulate a user working at the terminal but would want to run all 11 possible experiments as a single model which would be important for optimizing the program and would also help with processing.

To White's knowledge, this type of workflow has only been done 3-4 times in the crop modeling community and it has never been done with QTL data and DSSAT. DSSAT is thinking about going open-source and Vaughn suggested that iPlant could support forward-looking parts of DSSAT.

White concluded that this was a great exercise is writing the workflow but unfortunately the science did not turn out as expected. Even if he had access to super-computing, it would not have made up for the model not being good. He felt that sensitivity analysis should have been done first.

Welch suggested that perhaps iPlant can take the lead in combining parameter estimation, sensitivity analysis, QTL modeling to see new things about a model.

Sensitivity Analysis

Chris Myers presented a shortened version of his presentation from the Steering Committee meeting and this can be found here . He has code called "SloppyCell". He suggested that it would be nice to have something like MapMan for dynamic visualization. With PathVizio and GPML, the topology is embedded in the model and they are not really modeling languages but more layout languages. Myers asked if there are ways to automate layouts of pathways; Is there a generic approach to developing visualization tools on top of modeling? He also suggested that perhaps something that was missed in the workflow that Welch created was model construction.

Steve Welch

Welch presented on his work/research that he has been doing on modeling languages and tools that exist for the modeling community. He suggested that perhaps SBML isn't expressive enough to represent ecophysiological models. Myers said that you could take the SBML model and not have the environment be part of it and still use the model. Other tools Welch added to the list include Systems Biology Workbench, JSim, and OpenMI.

Discussion
After the presentations, further discussion ensued. Stanzione asked what is the right interface for the most important set of users and who is that most important set of users. Myers responded that he doesn't use interfaces as they can be decoupled using SBML. With students who are just beginning to understand modeling, a GUI is nice. He suggested that perhaps there should be focus on standardizing tools. Welch added that to begin with, it is worth spending time to build a GUI. iPlant wants to make sure to hit practical scientists at the cutting edge first or those at the core of the GC who are attacking the G2P problem. However, there is a wide swath of people that will not have the same skill level. As Myers pointed out, in the room, there are different worlds of modeling represented and in a first pass, we can't accommodate all. With the StatInf group, they agreed on what the obvious problem was and just did that. In this working group, there is not consensus as to what the problem is. Stanzione asked if there was a common agenda; if not, can we provide a CI and tools that will connect tools and make them run faster?

Welch suggested that the group focus on 1) parameter estimation, 2) sensitivity analysis (change input parameter and see how it changes the output; what happens as you change multiple parameters) and 3) interface with QTL mapping activity. White wondered if perhaps the group is overweighting the QTL analysis and perhaps this is not something that people would be doing all the time. Welch added that QTLs are not mechanistic and there is an open scientific question as to whether QTL models can provide level of predictability of what is going on in non-constant environments. He suggested that a better approach might be to attack from both ends of the spectrum and the problem may not be as intractable as the group thinks. If the right set of tools is built, a solution to the G2P problem is possible in the time frame of iPlant. However, the key thing is to look at the synthesis of different kinds of things people are doing now. Myers added that the group hasn't really addressed model construction and hasn't really defined what modeling is or what people think it is.

Parameter Estimation (inc. estimation post-processing such as ANOVA runs)
Welch opened discussion up by asking if there is agreement that parameter estimation is part of the modeling process and that it is compute intensive. There were no arguments to that statement. Myers asked if most of the crop models are ODE or stochastic because if they are stochastic, then parameter estimation becomes horrific. Myers approach is all ODE but when everything is written out, there is no specification needed in SBML. His group hasn't figured out how to use tools in stochastic situations but there are people that have.

Welch stated that there is agreement that focus of parameter estimation should be on ODE type models. He suggested that an exemplar algorithm be included but asked with what characteristics. As an example, would the group want to do grid search stuff that is embedded in GenCalc (though he does not advocate this) or maybe take a population based approach. White stated that he has no preference as iPlant is not into algorithm development but in CI development. With GenCalc, there are things that are very useful that you don't often think of: hand edit parameter file, backup of parameter file, automatically update parameter file, etc.

Inputs should be the same and outputs should be the same. For his grid search algorithms, he specifies the max number of iterations, how much to reduce the size of the grid search in following iterations, selects initial parameters, and selects upper and lower bounds. For this, he suggested that the group should flesh out a standard set of inputs. Myers stated that when looking for optima, there's the objective function and problems of initial guess. For algorithms, he uses the Nelder-Mead Simplex method and has found that gradients are typically most useful. Optimization is fickle, as you don't know if you are not getting the solution because you are stuck in parameter space or because the model does not describe the data. Welch added that he and his collaborators have found to work well is a top-level algorithm to explore parameter space and something nested to explore local opportunities. He has been successful in using Nelder-Mead algorithm for local search and also in using it alone. His suggestion would be to do a combination like described above.

White suggested that the group have three different algorithms as test cases to show robustness of the system. He added the three might be 1) the Simplex algorithm, 2) Monte Carlo, and 3) grid search algorithm.

Myers stated that abstractions for optimizations are good. With having an abstraction, if your data is put in the iPlant framework, then the person can use the CI. Welch commented that more often than not, calculating the objective function is the expensive part. He has found that he runs one machine for optimization and uses his others for evaluation of the objective function. The length of his biggest run to date has been 3 days on

200 processors, done 4 times. Myers added that generating MC parameters is a long process; you have to first monitor what is the correlation time and this can go on for several days. Stanzione commented that the group wants to enable runs on processors but then there is the issue of managing the multiple runs. Myers said that with population-based optimization, he would like to have the management all hidden, as you don't need to be explicitly looking at it. Welch said that the data management should be generic. Myers added that it would be nice if the launcher were embedded in the code. Stanzione said that it would be possible to either take parameters to launch from the command line or can make it internal to the code.

Welch summarized saying that what is needed is 1) a utility to handle the distribution of data amongst processors that is generic enough to be responsive to the different types of data and 2) different optimization and algorithms: Nelder-Mead to be used as stand alone or local search algorithm, gradient descent the same way and population based algorithm. All of these would be off the shelf methods and would be a concrete suite of algorithms to use in the first release. This would operate where the model is a black box. Myers stated that for model developers, this would require them to bundle their models in a certain way but the problem is important and big enough that people would be willing to reorganize their work to use the iPlant resources.

Myers cautioned that part of the stumbling block is the finicky nature of the algorithms. Welch stated that population based methods do a much better job of exploring parameter space. Welch also added that there will be some people that like to supply their own algorithms and perhaps there will be some that will want to supply their own objective function. This could be a sub-branch where a user could provide a model that can be evaluated using least-squares and represents potential specialization. The user could supply the objective function with the model embedded in it. People could end up submitting models that do not relate to plants; they could submit just an objective function and could specialize to be least squares or even specializing the model to be something like SBML. A user will supply the code to be optimized. This would include the objective functions along with such things that the objective function has to call to calculate values. In an initial release, there will be the optimizations algorithms discussed above that can be implemented with the API, which would allow for users to supply their own algorithms later.

Stanzione stated that the solution is three-tiered. With the API, a user can use their own code and it will only crash their stuff. If a component is to be registered, it will go through automatic checks to make sure that it doesn't crash the framework. Tier 2 will be community rated. For those that are highly rated, they will move to tier 3 or the gold standard complete with validation testing.

Welch stated that the suite of test problems be a set of models listed on his ladder diagram plus some others. Biomodels.net has a framework to automatically validated models and perhaps iPlant should use that. He asked if the standards that they have would be sufficient to be scoped into this system.

Sensitivity / Response Surface Analysis

Welch opened the discussion about sensitivity analysis stating that it might be done either before or after parameter estimation. With White's recent work, he did estimation runs and then did a series of ANOVAs. White added that his analysis was done several times but this is not really seen in the modeling community. Someone cynical would ask if model is better than generic variation. He asked questions relating to where should he look to improve overall simulations; with populations, where is further research needed. White's real point is that once finished with basic fitting, there are basic analyses that need to be done. In any of the workflows, there is a need for post-calibration analysis.

Welch summarized saying that are various categories such as observed vs predicted, decomposing sources of variation in the results, and cross-validation processes that will need a fairly large data set. White added that he did bootstrap cross-validation to check the model. In general, people will either have independent data or the opportunity for cross-validation or they don't do either as they don't have enough data. White suggested that perhaps one should first do sensitivity analysis to check the model and then do parameter estimation (calibrate and validate until confident) and then test fit will all the observations. However, the robustness of parameter estimates is different than the robustness of predictions.

Welch stated that the group would want to have bootstrap estimation procedures to give info about parameter distribution and robustness and have it be offered as a before or an after step to test the model. White added that it would be good if the system can be set up so that the data can be exported to do further analysis. Statistics should be much more interactive. White suggested that you could generate a final set of simulated values from the optimal set of parameters and could export this data. He added that in his work, the optimizations would have gone better/faster if it were random to begin with. Each method for optimization talked about this morning has *a priori* information requirements (ranges, initial estimates, etc). The user should provide as much data about the dataset as possible such as factor level information and what factors govern resampling. Also allowed should be something to extract subsets of data, either ad-hoc or post-hoc. For resampling, it is necessary to pick data at random. White added that you have to optimize for each line and that the computer has to know that each factor is line level. Welch suggested that the machine resample on the category that is controlling. All leveling information should be supplied as this will feed back to the users to do ad-hoc analysis. The user should be able to say that this factor represents things that are to be optimized separately and that for resampling, this is the category that should be resampled. White added that one would want to import factor levels and look at sampling strategy later on.

Welch concluded that the decision is to allow the import of level information at whatever scheme the user wants so that they can export subsequent files to be used for analysis and possibly guide bootstrapping analysis later. Also the user should be able to identify one of the factors in any subgroup over which optimization is to be done (line #) and if they want, optimization can be done over the whole subset. Multi-category bootsrapping is deffered until later. With the above options, there will be support for post-hoc analysis (observed vs predicted for all factors). Stapleton suggested that export files should be organized such that factors are columns.

When models are in a sufficient format, derivatives can be taken symbolically and if the models aren't in such a format, a user should be able to make scatter plots of the response surface. Welch asked if MCMC is a possibility for estimating the distribution of parameters. Myers said that MC is good at picking numbers in an area. He also uses a Hessian method to get local optima. Welch added that bootstrapping is different from MCMC and is more reflective of the data itself. Stanzione stated that PECOS has spent a lot of money on DAKOTAH, a suite of analysis tools, and maybe it might be worth looking into incorporating one of the tools into the environment.

Welch summarized the discussion, listing the following as what is on the wish list:

1. Ability to supply level information with data to use during parameter estimation
2. Ability to tell computer to do separate estimations done for one of those level factors

3. Ability to output all of the data (including factor levels) for offline analysis
4. Bootstrap analysis in one level
5. Encourage people to do offline analysis
6. Support MC simulation in neighborhood of optimum either using
    a. DAKOTA
    b. Keeping archives of data/trial/solution that can be rejected by some criteria
7. Within CI, want ability to output predicted vs observed observations for analysis by ViVA
8. For models that are expressible by computer, support eigenvector analysis

Conveying Outputs of Parameter Estimation and Sensitivity Analysis to Other Programs (e.g., Visualization and QTL Analysis)
Welch stated that the group has already talked about reproducing inputs with all the level information along with model predictions but this is related more to parameter estimation. The question is what do to do with the parameters themselves and the sensitivity analysis. How can you visualize the relatively high-dimensional spaces?

Myers presented some of the visualizations that he has used in the past. If the clustering is curved, you'll see variance that looks much larger. He suggested that you could do projections but maybe related to linear combinations. One could explore parameter spaces but overlay local structure on them. Welch suggested multidimensional scaling.

The decision was to send the visualization experts (Bernice, Greg) a dataset with guidance as to what types of relationships you would like to see and they would come back with ways to visualize the data. Both Myers and White have datasets that can be used. Some examples of questions to ask out of the data

- Is there a smaller space aspect to it? Is it a line that curves a certain direction
- Is it one data set or can it be subdivided? (is it one cloud or multi)
    - What does it look like when you separate it out? (when you have subgroups)? Does subgroup correspond to labels
- Want to some characterization of the outer surface; reconstruct the geometry of these regions
- Ensemble of time series generated by an ensemble of parameters

Stanzione suggested that Welch write an English description of what the tools do that are represented on Welch's slide for further investigation by iPlant.