11/10/2009
I think that we need to consider more fully what our advantages really are in altering the future of PGP and the way that our data is processed.

The Hybrid Model

1. **The state of trait-o-matic**
    a. **Trait-o-matic as a module on top of an abstraction layer**
        i. T-o-m is just an example of a mr-function (Map Reduce Function) that can be accessed as a job by any Freegol
2. **The need for expandability with open source software**
    a. **Master – Node relationship such as in Free Factories**
        i. Software such as Slurm can track which nodes are available for processing.
    b. **Minimizing the amount of data-transfer**
        i. Transferring only results when necessary (people are working on this currently wrt trait-o-matic)
3. **Taking advantage of parallelism**
        i. We keep talking about 100,000 genomes – Whatever is done on one will need to be done on all of the rest, so we can capture the speed and efficiency gains of parallelism
            1. There is a need for scheduling all of these different machines together to synthesize their outputs.
4. **Creating a system for modular applications on the genomic network layer**
    a. **Creating a standardized query language for running experiments on top of this layer**
        i. By working on expanding Trait-o-matic in a particular direction, we are shaping the way in which research on genomics is going – but we can't predict where scientific inquiry will lead us with such a massive data set.
        ii. Query languages such as SQL allowed for the abstraction to allow normal users to utilize the power of relational databases without having to understand the workings of the hardware that the system operates on
            1. Maintenance tasks are handled by the system itself
            2. A language has been developed for biological network data (PQL) based off of SQL[i]. It serves to
5. **Eventual developments**
    a. **New components**
        i. **Extension of trait-o-matic to process polygenic traits**
        ii. **Automated Genomic Annotation**
            1. At this point, it's quite easy to determine what are coding regions, introns, exons, etc.
            2. In the future, people may come up with novel ways to categorize DNA, particularly non-coding DNA

iii. **Network Reconstruction**

1. Considering how difficult it is to parameterize a signaling, metabolic, or genetic network, tools may be developed that take advantage of the deployment of such tools on large cluster

iv. **Generation of hypotheses based on previously observed biological information**

1. For example, the paper referenced below makes hypotheses about genes encoding orphan enzymes in Saccharomyces cerevisiae: "enzymes catalyzing biochemical reactions thought to occur in yeast, but for which the encoding gene(s) are not known"

v. **Coordination with robotic experimenters (such as ADAM)[ii] who can validate claims made by this system**

vi.

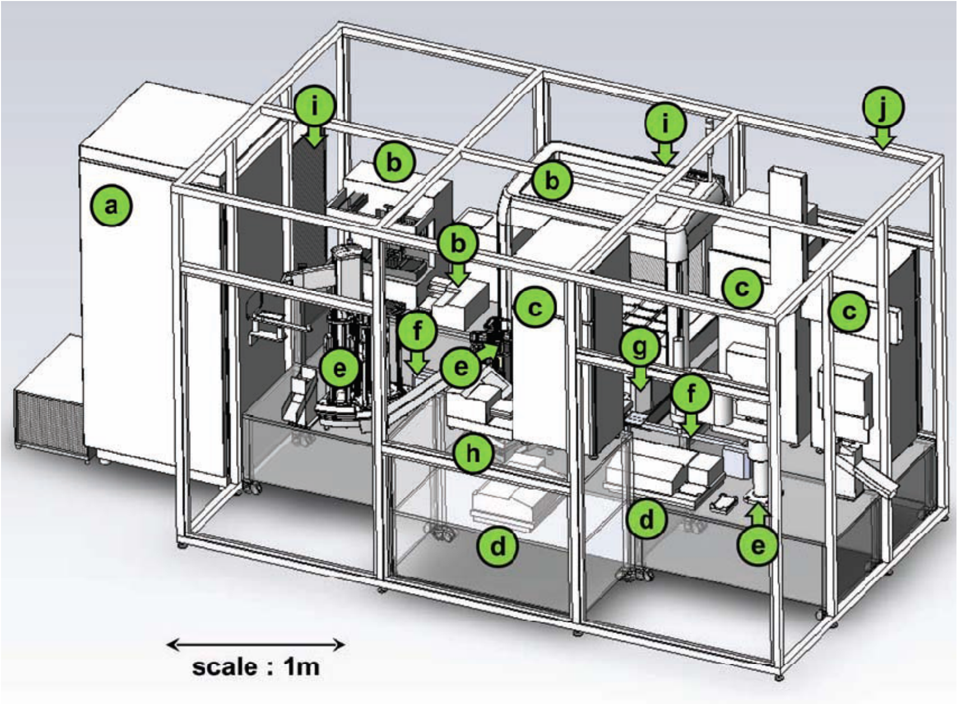

Fig 1: The robot scientist "Adam", which generates and validates hypotheses about *Saccharomyces cerevisiae*, specifically about orphan genes.

1. This experimentation is only possible because of an ontology and logical language developed to describe Adam's experiments

Fig 2: A model for a modular node

That said, I think that for the purposes of this class, improving the Trait-o-matic component to begin to analyze polygenic diseases and traits would be a great start and much more realistically achieved. Also, creating a tool that accesses the particular reference for an SNP or something and gives a synopsis to the user of the trait-o-matic system.

1. **Clustering and categorization of trait-o-matic results**
    a. **MeSH**
        i. We now have a way of ascribing particular tags to the traits, from where we can calculate "distance" along the MeSH tree. Also, traits share some
    b. **Machine learning algorithms**
    c. **Pre-categorization vs. Post-categorization**
        i. We only have a limited sets of disorders and diseases that we check for – we should probably preprocess them into categories rather than performing this sorting once our results already come in.
            1. If we do it beforehand it will probably be $O(N^2)$
2. **Previewing of reference information**
    a. **Resources such as pubget that automatically retrieve the original PDF of a reference – together with an OMIM or SNPedia that is parsed to include references, it is simple to acquire the original PDF of a reference**
    b. **PDF-reading libraries for python. These will allow us to perform text-search and to isolate the paragraph containing mentions of a particular SNP/etc**

[i] Lesser, Ulf. A query language for Biological Network Data. *Informatics* **21** (Supplement 2): ii33-ii39. (2005)

[ii] King, RD *et al.* The Automation of Science. *Science* **324**: 85-89 (2009)