PathologyXGrid Implementation for the InterProScan Bioinformatics Software

Cody K Bumgardner, University of Kentucky

Available at: https://works.bepress.com/cody/1/
XGrid Implementation for the InterProScan Bioinformatics Software

InterProScan is an essential step in the *Epichloë festucae* genome project’s pipeline. Complete determination of the genome sequence for *Epichloë festucae*, requires substantial programmatic and computing efforts. Due to the running time requirements the computational volume of some of its subtasks exceed the power of a traditional single processor computation and high performance approaches, such as grid computing, are needed.

### Biology

*Epichloë festucae* (anamorph *Neotyphodium*), a fungal endophyte of cool-season grasses, represents a model plant symbiont for studying mutualistic and parasitic effects.

1. **Asexual stage:** resembles life of mutualistic *Neotyphodium* endophytes
   - benign colonization of intercellular spaces of all above-ground plant organs
   - seed transmission through benign infection of embryo
   - production of bioprotective alkaloids, improved drought stress tolerance

2. **Sexual stage:** resembles life of antagonistic *Epichloë* species
   - inflorescence production induces external hyphal growth
   - perithecia bearing stromata envelop and choke the inflorescence
   - affected tiller produces few or no seeds

3. **Epichloë festucae** dual life cycles
   - dual growth mode: production of benignly infected seeds and stromata on different tillers of the same host plant!

4. **Alkaloids**

### Background

InterProScan*, developed by the European Bioinformatics Institute, is a bioinformatics program which allows the user to discover positive gene function. InterProScan employs a suite of 13 programs which identify motifs and domains in gene product sequences because each program has its own strengths and weaknesses. The use of the whole suite will help ensure that we discover the functions of the vast majority of the genus.

The analysis of our sequence with InterProScan can take up to three minutes for a protein sequence and up to 17 minutes for a DNA sequence. For the *Epichloë festucae* genome project tens of thousands of sequences will need to be analyzed, which makes analysis on a single computer prohibitive. Therefore it was decided to implement InterProScan on a computational grid.

For our computational grid architecture, Apple’s XGrid was chosen. Apple’s XGrid architecture was selected because of the availability of the Apple computers on campus; and the system administrators were eager to flex the computational muscle of the Apple computer labs.

### Implementation

Apple’s XGrid consists of three components. The main component is the Controller. The Controller accepts jobs and is in charge of distributing jobs to the worknode of the grid, the Agents. The Agents are responsible for running jobs assigned to them by the Controller and then letting the Controller know when they are finished and ready for more work. Finally we need a way for a user to interact with the grid, which in the XGrid, is the responsibility of the Client. Through the Client the user lets the grid know what processes need to be run.

In our implementation the user passes in a file of protein sequences or DNA sequences, the file is then parsed and entered into our sequence database, which is visible to all of the agents. When an agent becomes available the controller instructs the agent to run the perl script ‘runiprscan.pl’ whenever an agent is available for work. When an agent runs InterProScan the first time it needs to download a copy of the protein databases. This process takes about 12 minutes and only needs to be done the first time the agent runs InterProScan, or when new versions of the protein databases are available.

The protein databases used by InterProScan reside on each agent. So when an agent runs InterProScan for the first time it needs to download a copy of the protein databases. This process takes about 12 minutes and only needs to be done the first time the agent runs InterProScan, or when new versions of the protein databases are available.

### Results

Preliminary results indicate a phenomenal decrease in execution time. With a computational grid we are able to do in one night what would take years with a single computer.

The following graphs compare the number of sequences that can be run per hour to the number of processors working on the sequences. As you can see the more processors that are added to the grid the more sequences we can process per hour.

### References


---

V. Cody Bingham
Department of Computer Science
University of Kentucky
cody@uky.edu

Wayne E. Beech
Academic Technology
University of Kentucky
wayne@uky.edu

Uljana Hesse
Department of Plant Pathology
University of Kentucky
jurek@cs.uky.edu

Christopher D. Van Horn
Department of Computer Science
University of Kentucky
christopher.vanhorn@uky.edu

Wayne E. Beech
Academic Technology
University of Kentucky
wayne@uky.edu

Jerzy J. Jaromczyk
Department of Computer Science
University of Kentucky
jaromczyk@cs.uky.edu

Christopher Schardl
Department of Plant Pathology
University of Kentucky
schardl@uky.edu

This project has been partially supported by the Kentucky Biomedical Research Infrastructure Network, funded by grant P20GM103401-04 from the National Center for Research Resources. Further support provided by grants NSF grant EF-0521661 and USDA-NRI grant 2005-35319-16141.