Poster H-59

XGrid Implementation of the InterProScan for the Epichloe festucae Genome Project

Authors:
Wayne E. Beech (*Computing Services, University of Kentucky*)
V. Cody Bumgardner (*Department of Computer Science, University of Kentucky*)
Uliana Hesse (*Plant Pathology, University of Kentucky*)
Ch. David Van Horn (*Department of Computer Science, University of Kentucky*)
Jerzy W. Jaromczyk (*Department of Computer Science, University of Kentucky*)
Christopher Schardl (*Plant Pathology, University of Kentucky*)

Short Abstract: We present our XGrid implementation for InterProScan needed in determining the complete sequence of a model endophyte, Epichloe festucae, commonly found in grasses of genera Lolium (ryegrasses) and Festucae (fescues). Preliminary results indicate a phenomenal decrease in genome annotation time as we observe an almost linear speedup.

Long Abstract:
Background: Mutualistic symbioses such as the fungal endophytes of grasses, are extremely important in all of the worlds ecosystems. We are engaged in determining the complete sequence of a model endophyte, Epichloe festucae, commonly found in grasses of genera Lolium (ryegrasses) and Festucae (fescues). Understanding approximately 10,000 genes, and thereby understanding the biological significance of the sequences of nucleotides that make up the endophyte genome, requires intensive computational analysis. 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.

Methods: InterProScan, see [Zdobnov and Apweiler], an essential step in the Epichloe festucae project's pipeline, is a bioinformatics program which allows the user to discover putative gene functions. 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 helps ensure that we discover the functions of the vast majority of the genes. The analysis of one sequence with InterProScan can take up to three minutes for a protein sequence and five to six times more for a DNA sequence. For the Epichloe festucae genome project tens of thousands of sequences will need to be analyzed, which makes analysis on a single computer prohibitive. To address this issue it was decided to implement InterProScan on a computational grid; Apple's XGrid was chosen as our architecture. The grid was implemented in our local Apple computer lab with about 250 processors.

Implementation: Apple's XGrid consists of three components: The Controller accepts jobs and is in charge of distributing jobs to the workhorse of the grid, the Agents. The Agents are responsible for running jobs assigned to them by the Controller. Through the Client the user lets the grid know what processes need to be run; in our case the processes are instances of

InterProScan.

In our implementation the user passes in a file of protein sequences or DNA sequences, the file is then parsed and entered in to our sequence database, which is visible to all of the agents. The client tells the controller to run the program (the perl script 'runiprscan.pl') whenever an agent is available for work. When an agent becomes available the controller to run the program (the perl script 'runiprscan.pl') whenever an agent is available for work. When an agent becomes available the controller instructs the agent to run the script which grabs the next available sequence to be annotated from the sequence database, runs InterProScan on the sequence, and then places the output in the sequence database.

The protein databases used by InterProScan reside on each agent, which we found to be more efficient. So when an agent runs InterProScan for the first time it needs to download a copies of the protein databases. This process takes about 12 minutes on our system 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 genome annotation time. There is very little parallel overhead when adding processors to the grid. The size of the grid has no effect on the number of sequences a processor in the grid can resolve and we observe an almost linear speedup.

References: Zdobnov E.M. and Apweiler R. InterProScan an integration platform for the signature recognition methods in InterPro Bioinformatics, 2001, 17(9): pp. 847-848.

Acknowledgments: The project has been partially supported by the Kentucky Biomedical Research Infrastructure Network, funded by grant 2P20RR01648104 from the National Center for Research Resources. Further support provided by grants NSF grant EF-0523661 and USDA-NRI grant 2005-35319-16141.