Using Workflow to Build an Information Management System for a Geographically Distributed Genome Sequencing Initiative

David Hall John Miller Jonathan Arnold Krys Kochut
Amit Sheth Michael Weise

September 3, 1999

Abstract

Genome projects are very demanding undertakings and are often carried out collaboratively by multiple research centers. There are many different types of tasks that must be performed both by researchers and automated tools. These include such activities as shotgun sequencing, sequence finishing, sequence similarity searches, data annotation, oligonucleotide synthesis, library construction, and data submission. These individual tasks are organized into workflows that carry out a particular function, such as providing an annotated sequence of a cosmid clone. The individual tasks may be carried out at one or more of the participating institutions. A single workflow may be spread across multiple research centers. Creating software systems to support distributed workflows presents developers with a number of challenges, such as coordinating the execution of applications running on different systems, transporting data between systems, integrating legacy applications, providing recovery mechanisms, and creating user interfaces. Additionally, there will likely be frequent change to the organizational procedures, especially in the early stages of a genome project. This paper discusses using a general purpose workflow management system (WfMS) to address these challenges in the implementation of an information system that manages a geographically distributed genome project. A prototype application built with the METEOR WfMS is described which is running on several systems at the University of Georgia.

1 INTRODUCTION

Genomics is a discipline that investigates biological problems by looking at entire genomes or large numbers of genes at one time. Immediate goals of a typical genome project include the creation of high-resolution physical and genetic maps, determining the complete DNA sequence of the genome, and identifying, mapping and determining the function of all genes (Collins, *et al.*, 1998). Tracking the large number of material samples and managing the analysis of data generated in a high throughput sequencing laboratory is a challenge. Laboratory information management systems (LIMS) have been used in different types of analytical laboratories to automate many aspects of data analysis, sample tracking, and scheduling of experiments (Steele, *et al.*, 1999).

Due to high cost and the need for a large number of personnel, genome projects are often carried out collaboratively. Within a genome project individual research centers may specialize in different analytical or production tasks. There may be specialized hardware at one of the centers or a person who is an expert at a particular analytical task. For example, one center may provide custom DNA sequencing primers for all research centers. The execution of individual tasks, such as shotgun sequencing or annotation of genes, can be logically organized into workflows. Each workflow implements a particular research activity. For example, Figure 1 shows a flow chart for a workflow that carries out sequencing and the annotation of genes for a cosmid clone. The tasks in a single workflow may be distributed across multiple research centers.

Workflow applications automate the execution of workflows. A workflow application consists of a number of different individual tasks that are performed in a prescribed order. The tasks are carried out on data objects that "flow" through the system. The tasks may be existing legacy applications which may or may not require human interaction. They may also be simple operations where data is acquired from or sent to a person or database. For example, such a task may consist of a user filling in an electronic form.

Workflow applications may be developed from scratch. However, this presents developers with a number of challenges. The execution of applications possibly running on multiple systems must be coordinated. These applications may be running on different computing platforms as well. Data delivery between applications must be implemented. Other issues that should be addressed include recovery from system failures, user interface design, and the implementation of run time monitoring tools. Additionally, it is likely that a workflow application in a domain such as high-throughput genomics will require frequent modification, especially early in its life cycle as organizational procedures evolve. One class of tools for helping software developers create workflow applications are programming libraries. For example, Goodman (1998) has developed an object library written in Perl for developing workflows for analytical laboratories.

Another type of tool for creating workflow applications are workflow management systems (WfMS's). These software packages typically provide tools for the design, execution, and monitoring of workflow applications. They may also provide debugging and workflow simulation tools (Cichoki, *et al.*, 1998). Systems such as the METEOR WfMS (Sheth, *et al.*, 1997) provide a "drag and drop" graphical design tool for creating the workflow application and designing user interfaces. A workflow enactment system then provides an environment for the execution of the application. The enactment system takes care of data delivery, task invocation, and maintains work lists for users. It may also provide error recovery and run time monitoring of the application. This paper discusses building a workflow application for a geographically distributed genome sequencing project using the METEOR WfMS and describes a prototype system.

2 DEVELOPING A DISTRIBUTED LIMS USING METEOR WfMS

2.1 Introduction to the METEOR WfMS

The METEOR Workflow Management System (WfMS) provides an automated framework for managing intra- and inter-enterprise organizational processes (Krishnakumar and Sheth, 1995; Sheth, *et al.*, 1996). Two enactment services for METEOR model have been developed: ORBWork, a fully distributed CORBA-based workflow enactment system (Kochut *et al.*, 1999) and WebWork, a distributed workflow enactment system relying solely on Web technology (Miller, *et al.*, 1998). ORBWork features adaptivity, scalability, and reliability, while WebWork features ease of installation and use (*e.g.,*, Web servers are the only required infrastructure).

An enactment service provides the command, communication, and control (C^3) for individual application *tasks* participating in a worklfow. Tasks are the run-time instances of intra- or inter-enterprise applications. Today they typically run independently or are tied together in ad-hoc ways. WfMS's tie these tasks together in a cohesive fashion. The main components of a METEOR enactment service are workflow schedulers, task managers, and (application) tasks. *Task managers*, as the name suggests, are used to control the execution of tasks (*e.g.*, when they execute, where they get their input, what they do when they fail, and where to send their output). To establish global control as well as facilitate recovery and monitoring, the task managers communicate with *workflow schedulers*. It is possible for scheduling to be either centralized or distributed, or even some hybrid between the two (Miller, *et al.*, 1996).

METEOR workflow applications have been tested and deployed by several organizations, *e.g.*, Medical College of Georgia (MCG), Connecticut Healthcare Research and Education Foundation (CHREF), Advanced Technology Institute (ATI), Microelectronics and Computer Technology Corporation (MCC), Bellcore, Boeing and Visigenic.

2.2 Implementation of Prototype

A prototype workflow application has been developed that encapsulates a set of sequence analysis tasks. This prototype will ultimately be the genesis for a complete workflow application that will manage the sequencing and mapping projects being carried out by the Fungal Genome Initiative (Bennet, 1997). This prototype application is based on a workflow to produce contiguous genomic sequence of a cosmid clone with annotation of predicted genes. A flow chart describing the laboratory and analytical steps for such a workflow is shown in Figure 1. Currently, this workflow omits some tasks that should be present in a production system. However, we believe it is sufficient to test the suitability of the overall software design and to gain early feedback from users.

The workflow begins by sequencing shotgun clones from a cosmid. Sequencing is assumed to be done using automated sequencers. Base calling on the output from the automated sequencers is then carried out. The resulting sequences are assembled automatically. Human editing and inspection of the assemblies is performed and if physical or quality gaps remain in the cosmid sequence, primers are designed for direct sequencing of the cosmid. After direct sequencing of the cosmid all sequences are again assembled. This cycle repeats until a single contiguous sequence remains for the cosmid. At this point annotation begins. Prospective genes are identified from the cosmid. A search for similar sequences is then carried out using the BLAST program (Altschul, it et al., 1997). Human annotation using the output from the gene finding and BLAST programs is then performed.

Workflow applications are designed around the structure of the organization that they support. The following organizational structure was the basis for the prototype design. Sequencing takes place at two geographically distributed centers. Sequence assembly is done at each center. Each center has a *finisher*, who is responsible for editing sequence assemblies and designing primers to close gaps in the cosmid sequences. Synthesis of new primers is done at a single location that serves the entire enterprise. There is a person called the *submitter* whose role is to inspect all newly assembled and/or annotated sequences and send them to a high-throughput sequencing public database. Likewise there is an *annotator*, whose role is to annotate sequence using the results of the gene finding and BLAST programs. These last two applications are run on special hardware at a central location. Again this organizational structure is a simplification of the actual structure of a collaborative sequencing initiative. We feel it is complex enough for testing of the system design strategy as it contains a number of tasks that must be coordinated across different physical locations.

Figure 2 is a screen shot of the graphical design tool, METEOR Builder, displaying the top level hierarchical design of the application that implements the workflow in Figure 1 within the organizational structure described in the previous paragraph. Each icon in the design represents either a single task, or a compound task which contains its own subworkflow. Figure 3 shows the design of the ASSEMBLE subworkflow. A legend and explanation of some of the design tool icons is given in Table 1. Table 2 lists the individual legacy applications encapsulated in each of the main subworkflows in Figure 2. As mentioned in section 2.1 there are currently two different enactment services for METEOR. We felt that WebWork would be adequate for prototyping and testing and eliminated the need for purchasing CORBA middleware. The prototype workflow is spread over several networks at the University of Georgia. Although this is not a large geographical distribution, the performance of the application over a larger area should be similar. Several applications in the workflow, such as *Consed*, require human interaction via the X Windows system. The application runs on a server but the the graphical output may be redirected to any machine running an X server. We found the performance of these applications to be good when the client and server are on the same network. Network latency may make graphical interaction with these applications difficult when the client and server are on different networks, however. This is a design constraint that must be taken into consideration when the production system is built.

Figure 4 shows the logical design of the data object that flows through the workflow. This object consists of all data from a single cosmid clone. Initially the object contains only the trace files from the sequanators. As the object flows among the different tasks the remaining fields are filled in. Note that it may take more than one pass through the application to completely fill in a data object. (As a simplification for the prototype, the back arrows in Figure 1 do not appear in the the METEOR design tool. Instead, a new data object is initialized for each cycle through the application.) Some of the data fields in Figure 4 may contain very large amounts of data. For instance, each individual trace file is on the order of 500 kilobytes. Because of this pointers to files are used wherever possible in the application and data fields which are not needed by the downstream tasks are deleted. For example, in the prototype the only applications that use the trace files are *Phred* and *Consed*. The trace data was not passed to any downstream tasks. However, in future versions of the application as more annotation components are developed it may be necessary to pass the trace data on to different servers, depending on the organizational structure of the initiative. This will be a design challenge which may require increasing network bandwidth as a 40 kilobase cosmid with 5-fold coverage will have around 200 megabytes of associated trace data.

The application is designed to be automatically launched at specified time intervals. The application first checks for the existence of new trace files. If new trace files are found then a new data object is initialized for each cosmid containing new sequences. This data object then enters the workflow. In the compound task ASSEMBLE_x (Figure 2) base calling is carried out on new trace files. Next all sequences from the cosmid are assembled. The data object is then placed on the worklist for the task consed. The task *consed_setup* in Figure 3 is a Web-based task where the researcher responsible for assembly editing can retrieve the worklist for consed. The network address of another machine can also be specified for redirecting the output of consed. Worklists are displayed in a browser as hyperlinked cosmid names. After clicking on a cosmid name consed will be launched with the corresponding data. After termination of consed, the data object will be sent to the next task, which writes all new trace reads and assembly data to the project database. If quality or physical gaps remain after editing the assembly the user may decide to design primers. If any primers were designed these are sent to the worklist for the oligonucleotide synthesis personnel. The data object is always sent to the task SUBMISSION, which is a compound task that is responsible for sequence submission to the high throughput sequencing database at NCBI. If no quality or physical gaps remain in the cosmid consensus sequence, then the data object is sent to the GENE and BLAST tasks for gene finding and similarity searches, respectively. The data object is then placed on the worklist for the annotator.

3 CONCLUSION

Many of the activities associated with a genome project contain workflows. For example, Figure 1 shows a workflow for generating annotated sequences of clones. Likewise other activities, such as constructing physical maps, generating EST's, and carrying out functional studies on genes, contain workflows. The high cost and high labor requirements for a genome project often necessitate the collaboration of different geographically distributed research centers. Many of the challenges of building an information management system to manage a geographically distributed research project can be addressed by a workflow management system.

WfMS's such a METEOR provide easy to use tools for workflow application design and implementation. They also provide run time enactment systems that handle the invocation of legacy applications, the delivery of data between tasks, and recovery mechanisms. Workflow application built with METEOR can be built on any number of systems as long as each system has a routable network address. As workflow is an active area of research and development new features, such as improved error recovery and transactional semantics, are continually being added to WfMS's. As these capabilities appear existing workflow applications can be upgraded to incorporate these new features.

A prototype workflow application to manage the generation of annotated clone sequence was developed using METEOR/WebWork. This prototype uses several systems at the University of Georgia to emulate the combined computing resources of a model geographically distributed genome project. This application is Web-based, meaning inter-task communication and distribution of data are done utilizing Web technology. Worklists and interaction with the workflow application are all done through a browser. Workflow tasks, such as consed, which have their own graphical user interface require client machines to have an X windows server installed. A significant challenge to future more complete versions of the application is the transport of very large data objects, such as sets of sequanator trace files, over a network. Addressing this problem may necessitate increasing network bandwidth.

REFERENCES

1. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

2. Bennet, J.W. (1997). White paper: genomics for filamentous fungi. *Fungal Genetics and Biology*, **21**, 3-7.

3. Cichocki, A., Helal, A., Rusinkiewicz, M., Woelk, D. (1998). *Workflow and Process Automation: Concepts and Technology.* Kluwer Academic Publishers.

4. Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998). New goals for the U.S. human genome project: 1998-2003. *Science*, **282**, 682-689.

5. Ewing, B., Green, P. (1998). Base calling of automated sequencer traces using Phred II: error probabilities. *Genome Research*, **8**, 186-194

6. Goodman, N (1998). The LabBase system for data management in large scale biology research laboratories. *Bioinformatics*, **14**, 562-574.

7. Gordon, D., Abajian, C., Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Research*, **8**, 195-202.

8. Kochut, K., Sheth, A., Miller, J. (1999). Optimizing workflow. *Component Strategies*, **1**, 45-57.

9. Krishnakumar, N., Sheth, A. (1995). Managing heterogeneous multi-system tasks to support enterprise-wide operations. *Distributed and Parallel Databases*, **3**, 155-186.

10. Miller, J., Sheth, A., Kochut, K., Wang, X. (1996). CORBA-based run-time architectures for workflow management systems. *Journal of Database Management*, **7**, 16-27.

11. Miller, J., Palaniswami, D., Sheth, A., Kochut, K., Singh, H. (1998). WebWork: METEOR2's web-based workflow management system. *Journal of Intelligent Information Systems*, **10**, 185-215.

12. Sheth, A., Kochut, K., Miller, J., Worah, D., Das, S., Lin, C., Palaniswami, D., Lynch, J., Shevchenko, I. (1996). Supporting state-wide immunization tracking using multi-paradigm workflow technology. *Proceedings of the 22nd International Conference on Very Large Data Bases*, 263-273.

13. Steele, T.W., Laugier, A., and Falco, F. (1999). The impact of LIMS design and functionality on laboratory quality achievements. *Accreditation and Quality Assurance*, **4**, 102-106.

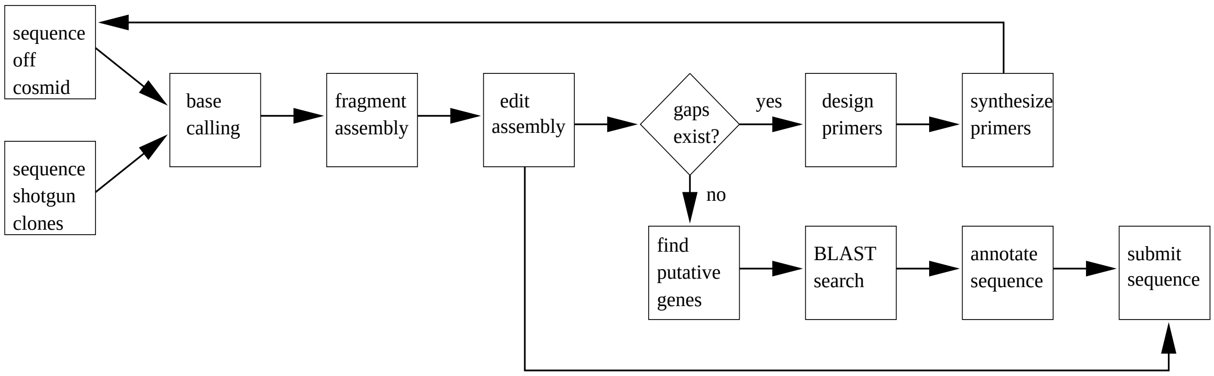


Figure 1: Flow chart description of a simplified workflow to produce contiguous genomic sequence of a cosmid clone with annotation of putative genes.

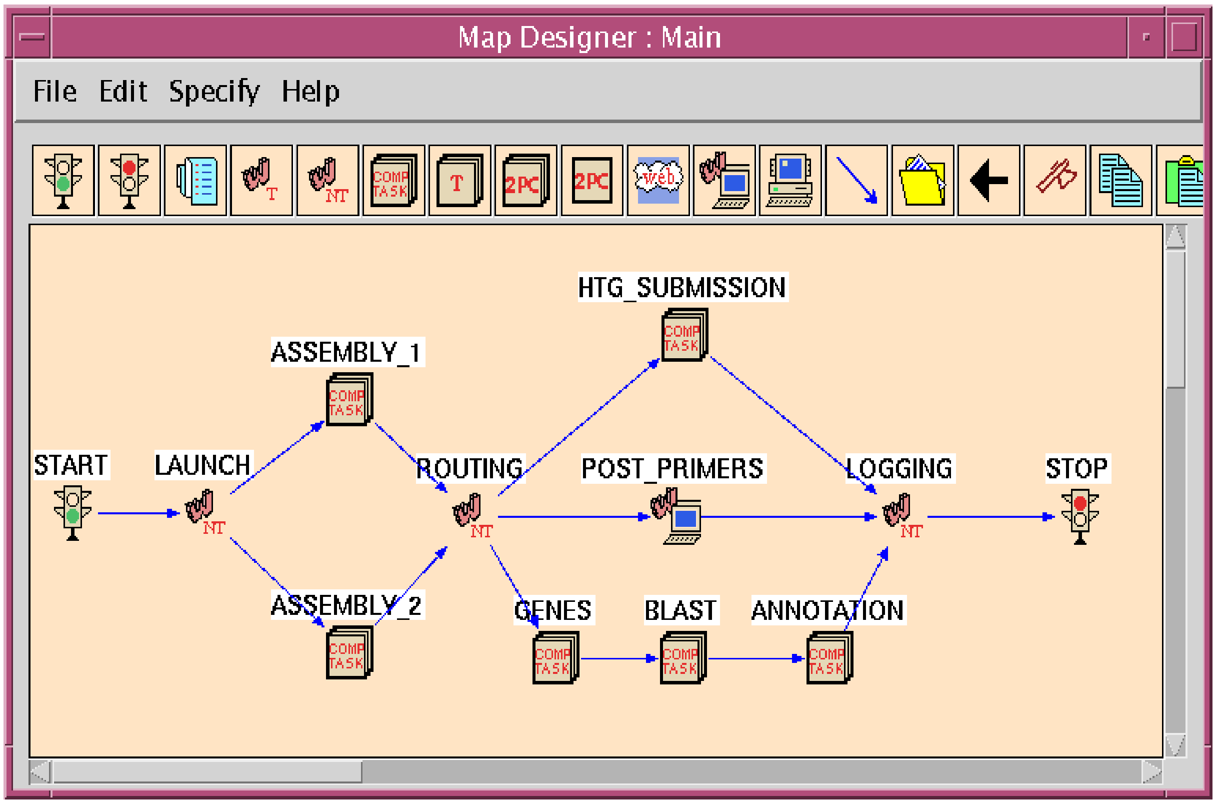


Figure 2: Screenshot of the graphical METEOR workflow design tool displaying the top level hierarchical design of the prototype application.

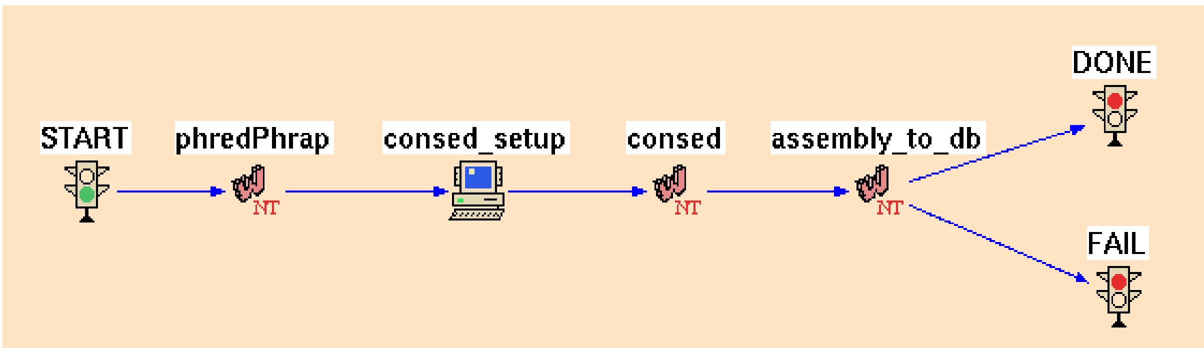


Figure 3: Screenshot of METEOR design tool showing ASSEMBLE sub-workflow.

GeneFlow Data Object
cosmid name
accession number
traces
sequence fragments
consensus sequence
predicted genes
annotation

Figure 4: GeneFlow Data Object

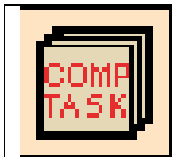
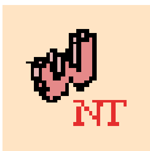
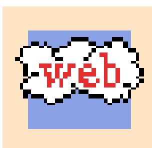
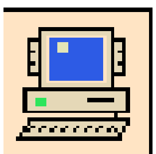
	Compound Task
	Nontransactional Task
	Web Form
	Human Computer Task

Table 1: Legend of some of the METEOR Designer icons

Subworkflow	**Applications**	**Purpose**
ASSEMBLY	Phred (Ewing and Green, 1998) Phrap Consed (Gordon, *et al.*, 1998)	Base calling Sequence Assembly Editing assembled sequence Primer design
HTG_SUBMISSION	Sequin	Submission of high throughput sequence to NCBI
GENES	ffg	Searching for predicted genes
BLAST	BLAST (Altschul, *et al.*, 1997)	Sequence similarity search

Table 2: Legacy applications embedded in subworkflows