BioHPC Suite for Next Generation Sequencing Data and Analysis Pipeline Management

http://BioHPC.net/

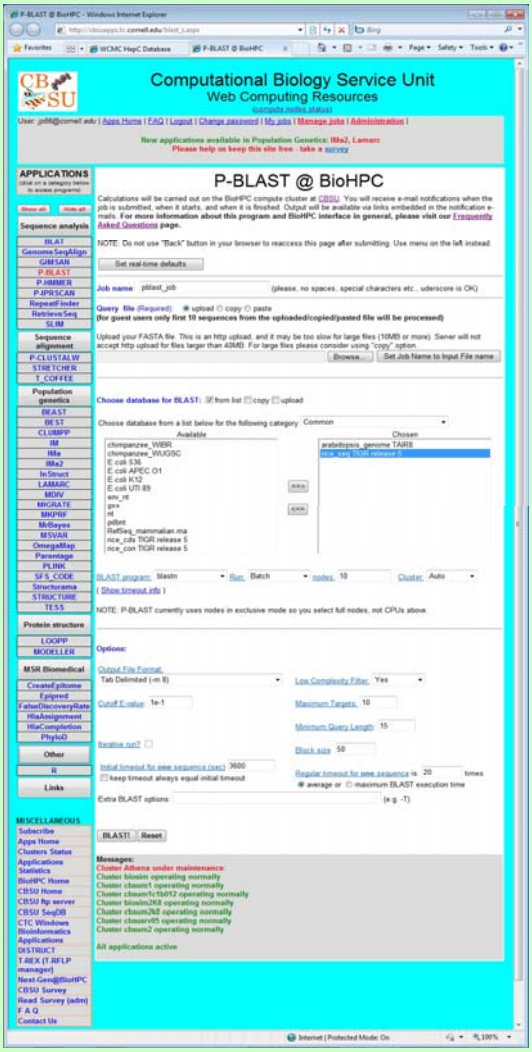
Robert Bukowski, Jaroslaw Pillardy, Qi Sun, Mary Howard, George Grills[a]

Computational Biology Service Unit, Microsoft HPC Institute

[a]Life Sciences Core Laboratories Center, Cornell University, Ithaca, N.Y., 14853

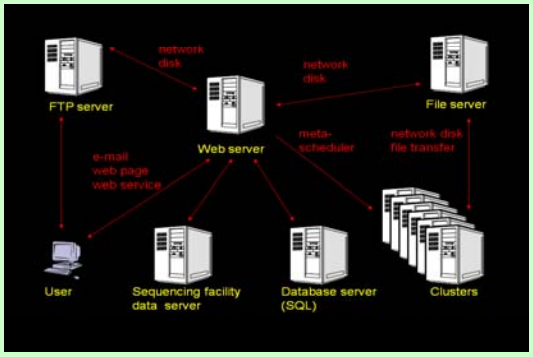
INTRODUCTION

One of the challenges of **High Performance Computing (HPC)** is the user accessibility. At the **Cornell University Computational Biology Service Unit**, which is also a **Microsoft HPC institute**, we have developed a suite of computational biology applications for HPC (**BioHPC**) that allows researchers from biological laboratories to submit their jobs to the parallel cluster through an easy-to-use web interface. They don't need to deal with parallel job submission, queues, clusters – knowing the application, parameters and input is all that is required. Recently, a **web service** layer has been added which allows job control through other clients, such as MS Excel or perl scripts. Through web services, BioHPC is being integrated with the **Microsoft Biology Foundation** platform.



ARCHITECHTURE

The system consists of a **web server** running the interface (ASP.NET C#), **Microsoft SQL server** (ADO.NET), **compute clusters** running Microsoft Windows, **ftp server** and **file server**. Two local compute cluster schedulers are supported (CCS and HPC Server 2008), remote clusters can be used via JSDL/HPC Profile. **JSDL connection** is now implemented and linked to Athena cluster at Microsoft (Redmond, WA) and biosim (Cornell).



Users interact with their jobs and data primarily by a **web browser** and **e-mail**. Jobs are submitted through our active web pages, fully compatible with all the popular web browsers supporting DOM and Javascipt. Interfaces to various application are standardized, users can choose the cluster, number of nodes or allow the system to determine it based on the best load balance and node availability.

Distributed cluster resources in a user-transparent way. Notification e-mails sent to users upon job submission, start, and completion contain links for job progress monitoring, job cancellation and restart, and results retrieval (by http or ftp). Job data files are private – they can be accessed only by a user who submitted a given job. Job and data control functions can also be performed via **web service interface** which enables developers to build custom clients independent of the web browser. The number of applications offered through the web service (currently 15) is growing.

Besides the job control interface, **BioHPC** also features a built in **user and data management system** which can limit software and/or database access to specified users. There is also an **administrative interface** allowing for easy management of jobs, clusters and applications with automatic e-mail notification of possible problems.

APPLICATIONS AND USAGE

Applications offered by BioHPC cover various aspects of computational biology: **data mining/sequence, protein structure prediction and modeling, population genetics, phylogenetics, association analysis/statistic, MSR Biomedical applications.** The system is flexible and can be easily customized to include other software. It processes over 40,000 job submissions a year, many of them **parallel.**

BioHPC jobs have been submitted by 11 471 users from 83 countries, the majority (57% by CPU time used) coming from the USA, with 52% of the utilized CPU time from the USA coming from New York State. These users include 257 users from Cornell, 2,580 users from *.edu* domains representing 426 unique *.edu* institutions, and 4,813 users from *.com* domains (including 4,191 users with Yahoo, Gmail and Hotmail e-mail addresses).

TAIR, the major database of the plant model organism Arabidopsis, and **SGN**, the international tomato genome database, are both using our system for data analysis.

The **BioHPC** source code is **freely available**. The "packaged" version of the interface can be downloaded from **BioHPC.net** and installed locally with any Microsoft CCS or HPC 2008 cluster.

ABOUT CBSU

Computational Biology Service Unit (CBSU) of the Cornell University Life Sciences Core Laboratories Center was initiated by the Tri-Institutional collaboration among Cornell University, Weill Cornell Medical College, Rockefeller University, and Memorial Sloan-Kettering Cancer Center. In February 2006 CBSU became Microsoft HPC Institute charter member. CBSU is Cornell core facility for computational biology. BioHPC development is now partially funded by **Microsoft Research** and CBSU is a **Microsoft Biology Initiative** partner.

The BioHPC installation at CBSU is currently using 5 Microsoft Windows based local compute clusters totaling 976 cores. The local nodes use Microsoft Server 2003 with CCS and Microsoft Server 2008 with HPC Server 2008. 80 CPU cores of the remote cluster Athena (located in Redmond, WA) are also available via JSDL, courtesy of Microsoft.

NEXT GENERATION SEQUENCING SUPPORT

We are currently implementing a new module of BioHPC designed to support analysis of **next generation sequencing** results. The module consists of several components:

Run Manager: connects to the sequencing facility and automatically detects finished sequencing runs for which base calling has been completed. It then configures the run in BioHPC database and sends an invitation to the facility manager to approve the results for distribution to users. Once approved, the results (read files) are asynchronously transferred to BioHPC file server and catalogued there for further use. Once the transfer is complete, all users assigned to distributed lanes are automatically notified by an e-mail message containing download links.

Lane Browser: allows users to browse their sequencing read files (Illumina lanes) catalogued at BioHPC. The browser displays lane annotation information and allows the file owner to grant additional users access to a file. Read files obtained outside of the Cornell sequencing facility can also be uploaded and catalogued at BioHPC.

Reference Manager: allows users to upload and catalogue reference genome files and annotation files needed in downstream data analysis.

Pipeline Manager (under development): will allow users to construct and run various analysis pipelines using sequencing reads and reference files stored at BioHPC as input. While default parameters will be provided, steps of each pipeline will be individually configurable by a user. Users will interface with pipeline manager using our specially constructed web interface or using a web service layer. Computationally intensive steps will be run on CBSU clusters linked to BioHPC (the tools comprising a pipeline will also be available as stand-alone applications). The web service interface will allow pipelines to be controlled from any client application, such as the **MBF platform** or the **Illumina Genome Studio**, or **Trident** scientific workflow workbench.

The new module is currently geared to handle mainly **Illumina** sequencing results, but extensions are possible.