



Featured Application:

UABgrid Dynamic BLAST: Searching Nucleotide and Protein Databases Using SURAGRID

ABSTRACT: BLAST is a database search application for matching protein and nucleotide sequences. Maximizing the throughput of searches is key to improving research results. This distributed implementation of BLAST uses the Dynamic BLAST Meta-scheduler to select appropriate grid resources for select query strings. Globus is used for job staging, submission and retrieval. ncbiBLAST performs the computations. Jobs are submitted using a web-based interface that leverages campus identity credentials via Pubcookie and manages grid authentication on behalf of the user via MyProxy, providing a simplified user authentication experience.

Understanding the Science

Molecular biologists at the University of Alabama at Birmingham (UAB) use gene databases published by the National Center for Biotechnology Information (NCBI) and locally developed databases to guide their search for genetic similarities in research subjects. The NCBI Basic Local Alignment Search Tool (BLAST) is a broadly available tool that finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. NCBI BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. BLAST finds similarities between a short query sequence and a large database of infrequently changing information such as DNA sequences. With the ever-growing size of the search databases, the need to increase the speed of query searches becomes critical. Grid computing environments provide an ideal development platform to take advantage of distributed computational resources. The UAB Department of Computer and Information Sciences in conjunction with the UAB Office of Information Technology have developed a grid meta-scheduler and web-based interface called UABgrid Dynamic BLAST to improve the performance and simplify the execution of BLAST searches for the UAB research community across grid resources on their own campus and on external grids such as SURAGRID .

Application Characteristics

Maximizing the throughput of BLAST searches is key to improving research results. Each search consists of a string that represents a gene sequence of interest. The normal operation of NCBI BLAST treats each search string as an independent process, depending only on the target search database. The aggregate speed of many searches can be improved by maximizing the number of processors available to perform individual searches. As several BLAST parallelization techniques

exist, the variability of BLAST algorithm implementations fits in many categories of grid application implementation and makes a challenging and yet flexible grid problem. By adding the compute power of SURAGRID resources to the institutional resources of UABgrid, UAB Dynamic BLAST users can benefit from increased resource heterogeneity and maximize the number of their simultaneous search processes, thus reducing the time required to match all sequences of interest.

SURAGRID Deployment

In order to use UABgrid Dynamic BLAST or any other web-based user interface application, in UAB's grid computing environment, UABgrid, UAB users must register via a web-based registration interface. This registration interface is controlled by a Pubcookie web single sign-on service that is integrated with UAB's BlazerID-based identity management system. By leveraging this institutional identifier the web interfaces to the UABgrid can rely on it to validate users, receive a very reliable assertion of membership in the UAB community, and minimize end-user authentication events in UAB's distributed application environment.

Grid Workflow

The UABgrid registration interface verifies the Dynamic BLAST user's identity, initializes their grid credentials in a MyProxy certificate store (where UABgrid Dynamic BLAST retrieves the proxy in order to submit jobs to grid resources), and allocates them basic resources in the UABgrid portal. UABgrid credentials are standard PKI certificates signed by the UABgrid Certificate Authority (CA). If the Dynamic BLAST user at UAB wishes to use grid resources on SURAGRID rather than or instead of those on UABgrid, their authentication and authorization to use SURAGRID resources is facilitated by the trust relationship between the UABgrid CA and the SURAGRID Bridge CA (BCA). UAB established trust relationship by cross-certifying their UABgrid CA with the SURAGRID BCA. This arrangement provides the basis for identifying

UABgrid Dynamic BLAST users to all SURAggrid resource sites, lets them transparently leverage their UABgrid user certificate across SURAggrid resources and leaves authorization decisions in the hands of resource owners (who control UABgrid user access via their resource's grid mapfile.)

Even though the Globus technologies on which UABgrid and SURAggrid are based on well established protocols and technologies, new users may still experience difficulties, for example, submitting jobs or understanding how to access grid resources. The very heterogeneity in type, location and ownership of grid resources that combine as the overall strength of grids is also the very thing that needs to be hidden from UABgrid Dynamic BLAST and other grid users. UAB's Dynamic BLAST developers have minimized some of these difficulties by providing a web-based user interface that automates the intricacies of resource authentication, algorithm selection, data distribution, resource acquisition, and job submission for the user - which allows them to remain focused on application specific parameters and data interpretation.

UABgrid Dynamic BLAST was designed to allow the user to retain the control they traditionally have over execution options when executing a BLAST search while avoiding any additional requirements imposed by execution in a grid environment. Dynamic BLAST determines the grid job submission requirements by analyzing job parameters, comparing these to historical run data and modifying the job submission process accordingly. The application then creates a grid job submission script and the job is submitted using the Globus-based GridWay¹ meta-scheduler which, when given the appropriate parameters, searches for an optimal grid resource to run the job(s) on. Once the jobs have completed, Dynamic BLAST retrieves all partial results from the grid resources and gathers them into a single data file for the user to access via GridFTP or which will be presented to them the next time they log into the job submission portal. The current set of available resources includes two compute clusters at UAB that are part of the UABgrid and one compute cluster at Old Dominion University that is part of SURAggrid. This approach has maximized the use of existing application code: stock versions of BLAST algorithms are run on each resource, as are standard Globus technologies for job control, and the SURAggrid BCA cross-certification that facilitates the leveraging of UAB users' institutional identifiers.

Lessons Learned

There are two main areas where UAB developers ran into issues when grid-enabling Dynamic BLAST: identity management and job submission. The primary challenge in the area of identity management was system integration, that is, getting consistent identities across a distributed set of resources. UAB developers overcame these problems by leveraging the institutional BlazerID

¹ <http://www.gridway.org/>

identifiers that have a high level of assurance which enabled them to create PKI identities that could be trusted by all other cross-certified SURAggrid members. While these technologies make using distributed SURAggrid resources possible, much coordination is still needed to initialize support for Dynamic BLAST on new resources. This translates into manpower and funding issues that are not yet resolved, UAB is, however, working with initial SURAggrid sites to streamline and document the Dynamic Blast resource configuration process, which may reduce some overhead.

Unlike efforts focused on improving the performance of traditional BLAST software on specific compute resources, the primary challenge of modifying BLAST for job submission in a grid environment is to increase throughput and efficiency in a pool of very dynamic and heterogeneous grid resources. Using the GridWay meta-scheduler simplified the development of this application since it automatically manages resource selection and data staging and allowed UAB's developers to focus on BLAST-related performance optimization and usability issues. Their initial grid BLAST implementation didn't use GridWay, thus they spent the majority of their grid-enabling work on resource management and coordination issues.

Conclusion

The UABgrid Dynamic BLAST application makes it easier for molecular biologists to expand the number of nucleotide and protein searches they can perform in parallel by leveraging available resources on SURAggrid. Dynamic BLAST turned into a powerful tool used to perform these BLAST searches by transparently exploiting the power of grid resources, regardless of their size, and provide the user with a simple interface, minimal adjustment requirements and a convincing performance increase. By leveraging opportunities for BLAST optimization using the grid, the current implementation has provided a solid proof of concept for leveraging inter-institutional resources to expand the research capacity of local researches. Future development plans include migrating UABgrid to myVocs+GridShib technologies that will make it easier for select researches from any institution to access Dynamic BLAST and the SURAggrid resources that help power it.


For more information on UABgrid Dynamic BLAST contact:

John-Paul Robinson – jpr@uab.edu
Purushotham Bangalore - puri@cis.uab.edu
Enis Afgan - afgane@uab.edu



or see:

<http://uabgrid.uab.edu/dynamicblast>

 SURAggrid is developing an infrastructure for access to heterogeneous resources and support for a dynamic and diverse application set.

For more information, or to join SURAggrid:

- <http://www.sura.org/SURAggrid>
- maryfran@sura.org