Partnership for International Research and Education
A Global Living Laboratory for Cyberinfrastructure Application Enablement

Finding Repeats and Signatures in DNA Sequences Using MPI Clusters

Students: Camilo A. Silva (FIU), Michael Robinson (Ph.D., FIU), and Guangyuan Liu (M.S. CAS, China)
FIU/FAU Advisor: Dr. S. Masoud Sadjadi and Dr. Giri Narasimhan (FIU, Miami, USA)
PIRE International Partner Advisor: Dr. Hector Duran, Universidad de Guadalajara, Mexico

National Science Foundation

I. Research Overview

MOTIVATION

Homeland Security applications with regard to microbial detection in environmental samples are among the many other possible applications of our project.

The main characteristic of genomic data is its large size. For example, the National Institutes of Health (NIH) sequence database has a total of over 65 billion nucleotides.

One of the most striking features of genomic DNA sequences is the extent to which repeated substrings occur in the genome. In *C. elegans* with a genome sequence of 100.2 million nucleotides over 7,000 families of repeat sequences have been identified. Families of repeat sequences account for about one third of the human genome. Repeat sequences come in many different flavors and are responsible for different functions and diseases. Finding repeats has applications in finding defective genes, and in forensic DNA fingerprinting. It also allows us to find differences between genomes.

BACKGROUND

An approach by Afgan and Bangalore utilizes grid computing to implement BLAST. Our results for scalability and load balancing performs better, probably because of the nature of the problems being tackled and because of our dynamic approach to load balancing.

There are a number of programs available for finding repeats, such as RepeatMasker a popular software for finding repeats. However, the problem is clearly compute-intensive and, creates costly bottlenecks in large-scale analysis.

APPROACH

We have implemented an algorithm based on suffix arrays for finding repeats and unique signatures. This was then ported to run on MPI clusters with good results. This algorithm also searches for direct repeats and other variants such as inverted repeats and complemented inverted repeats.

The implementation on a single computer appears to run much faster using the suffix array and the run on MPI provides a significant speedup. It searches for single patterns in a sequence of roughly a few million bases in less than a second.

We tested our data on five strains of the bacteria *Pseudomonas aeruginosa*, each containing over six million nucleotides. The serial implementation completed its task averaging 30 hours for every pair of genomes. The performance of the MPI implementation was compared to that of the serial one. All nucleotide sequences were acquired from NCBI.
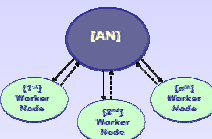
ALGORITHM

SERIAL ALGORITHM
Our serial algorithm consists of three modules. The first one creates an index from the genomic sequences provided. The second one (Search) helps in searching for the presence of a probe in the genomic sequence by searching in the created index. The third one (Signature Search) searches for "signatures" or probes present in one genome but absent from the others. This serial implementation was used as an "off-the-shelf" program in our cluster implementation.

CLUSTER ALGORITHM
The administrator node (AN) sends a task message indicating two data files, their indexes, and programs to the worker nodes (this communication is indicated by solid lines), which execute the programs on the data files and report the results to the administrator node (this communication is indicated by dotted lines).

Fig. 1 *Communication Paradigm:*
The administrator node [AN] will send messages containing information about the task assignment to each worker node. The AN keeps track of all completed and to-be completed tasks.

The worker node receives the message and disseminates it—by finding out what files to open in order to carry out the process and which output file to create for the results. After the task assignment is completed, the worker node sends a completion message to the AN, which will determine to send any available tasks to the worker node.

FAULT HANDLING ALGORITHM
The fault handling algorithm introduces both self-healing and self-optimizing attributes that specifically target the solution to the issue of improving time-processing efficiency whenever any worker node has scarce resources.
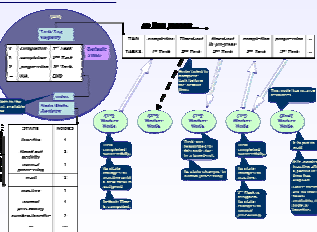
IMPLEMENTATION

Fig. 2 *Self-management Implementation:*
During runtime, the AN will be supervising the task completion efficiency in each worker node. At first, a *tag task registry* and *node state registry* are initialized with different integer values representing different states for both the tasks assignments and nodes. There is a *task list* containing the tasks being assigned—the next task to be assigned is pointed by an *index*. When the first task completion is received by AN, a *default-time* variable is computed, which will help determine the time threshold for a task to be completed by a node within the state of *normal processing*. Any node that takes longer than the *default-time* to complete a task process will be categorized in the *timed-out activity* state. AN will reassign such task to another node. If such node efficiently completes the task, AN will interrupt the process of the *timed-out* node and it will be put to *wait* in order to give time to recuperate the scarce resources. A random time (not more than the *default-time*) will elapse and the *waiting* node will be "awaken", and its state is changed to *awaken-inactive*—meaning it is ready to be assigned with more tasks. If a node finishes processing, its state is *inactive* until a new task is assigned to it (by then, its state will switched to *normal-processing*).

RESULTS

TASK CONTROL SYSTEM
The following figures present the results of the process run on the GCB cluster, showing a linear speedup and a run time decrease whenever a worker node is added.
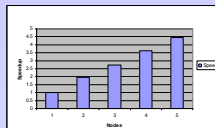

Fig. 3 *Speed-up:*
Speed-up as a function of the worker nodes. In essence, the more worker nodes added, the faster the completion of all tasks would be when all resources are used effectively.
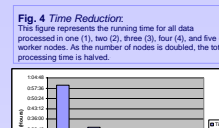

Fig. 4 *Time Reduction:*
This figure represents the running time for all data processed in one (1), two (2), three (3), four (4), and five (5) worker nodes. As the number of nodes is doubled, the total processing time is halved.
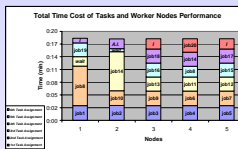
FAULT HANDLING SYSTEM

It is important to note how the whole application performs when there are some worker nodes with resources shared among more than one process. As shown in the graph, there are different time states for each node. The states of each node are *wait*, *inactive* (*I*), *awaken-inactive* (*A.I.*).

Fig. 5 *Fault Handling Test Assessment:*
The worker nodes were assigned with different tasks. Not all nodes had all resources available at all times. It can be seen that both nodes one (1) and two (2) had scarce resources at different periods of time. As expected, those *timed-out* tasks were reassigned appropriately to any worker node available (in this case to node four (4), since it was the first one available). Interestingly, node one was able to complete its timed-out task assignment of job8, a couple of minutes after it was reassigned to node four. Consequently, node four's assignment of job8 was interrupted by the AN, and then, the *timed-out* task from node two (job14) was reassigned to it.

ACCOMPLISHMENTS

This project contributes in both areas of Bioinformatics and High Performance Computing. We have been able to achieve our goals successfully in both domains.

We were able to find the differences between five strains of *Pseudomonas aeruginosa* for probes of length 10; this process will allow us to find other differences using any probe length.

The MPI cluster implementation allowed us to reduce the amount of time needed to process online search queries. We have also provided some levels of scalability and load balancing.

FUTURE WORK

An interesting topic of a broader implementation and usage of this application would be to create a dynamic web interface of the application. Such user friendly web interface would allow the user to create a job list for the cluster to process.

Adding more self-management behavior to the parallel search engine.

II. International Experience: Guadalajara, Mexico.

CITY LIFE

Guadalajara, Mexico: During the summer of 2008, Camilo Silva had the opportunity to visit Mexico, a neighboring country of United States. The first thing that he experienced from his trip was the city life. Guadalajara is one of the oldest cities in America—it is a Mexican epicenter of tradition and cultural flavors of this beautiful country. It is characterized by busy people that like to live their lives at a slow pace—not worrying about the future.

Its architecture is unique, namely, colonial and antique, although there were state of the art malls packed with the latest technology (e.g., hand touch directories). Guadalajara is a socio economic stratified society—the nice places are only located in "high-class" neighborhoods.

The public transportation is abundant and inexpensive. The food variety is gastronomical in nice restaurants, which are abundant. Tourism is not uncommon in this city.

City Life: These pictures are from Guadalajara's Downtown area. This area is characterized by the antique cathedrals and beautiful plazas. The colonial architecture of this area is unique and well preserved. A beautiful theatre, Teatro Degollado, is located in the main plaza of the city (lower right picture).

CULTURAL LIFE

Guadalajara is known for its tequila and Mariachis. It did not matter how little Camilo knew from Mexico, what made his experience remarkable was the friends that he was able to make during his stay. He made good friends, some of which he still talks to.

Camilo's cultural life experience at Guadalajara, was solely revealed or exposed to him by different people that he met. He was able to visit the town of Tequila, the Degollado Theatre (city's oldest theater), a Zoo, many restaurants, malls, karaoke's, soccer games, etc!

The best things that he brought back from this trip were the many pictures that he took!

CAMPUS LIFE

Camilo spent many days at the University of Guadalajara (UdeG)—in one of its many campuses, CUCEA, which is a beautiful campus irradiated by friendly people and quintessential landscaping.

Campus Life: UdeG's cyber-forest is the perfect place to study: free internet and a relaxing atmosphere all around.

OUR RESEARCH NETWORK

The most significant experience of this research project was the global and diverse collaboration that was carried at all times. Specifically, thanks to the partnership between Mexico and China this project was able to grow significantly by having a professional solidarity and synergy to solve a complex problem in a team-effort environment.

Camilo's visit to Mexico was an excellent opportunity to understand the importance of global collaborations by being actively involved in research alongside people from other nations. Nonetheless, he sought the need to become an independent researcher by finding solutions to specific problems and sharing results with his teammates.

During summer 2008, Camilo's research partners spanned over different parts of the world. Michael (the PhD student of this project) was in Miami, FL along with Dr. Sadjadi and Dr. Narasimhan. Guangyuan was collaborating from Beijing, China, while Dr. Hector Duran and Camilo were in Guadalajara, Mexico. Thus, our research network ranges from Mexico, United States, to China!

Michael Robinson   Dr. Sadjadi   Dr. Narasimhan   Guangyuan Liu   Dr. Duran   Camilo Silva

Research Life: The research experience in Mexico provided Camilo with the opportunity to work in a different cultural environment.

III. Acknowledgements

Cultural Life: Guadalajara is a city that has plenty of diversions and cultural events to offer. One of the most interesting places to visit is the little town of Tequila, where tequila is fermented and distilled.

There are plenty of nice restaurants that not only serve exquisite Mexican dishes, but also provide live entertainment!

The Theatre Degollado showcased a modernized ballet presentation of *Don Quixote de la Mancha.*

Karaoke's provide an excellent opportunity to showcase one's brave soul and singing performance—a great plan to do with other people that like to sing!

The City's Zoo has different kinds of animals: polar bears, elephants, monkeys, tigers, wolves, exotic birds, reptiles, and domestic animals.

Soccer games are always packed with tons of fun and excitement—a nice plan to do once in a while.

Sunday walks at the City's downtown Plaza are always remarkable since there are always live performances on the streets ranging from capoeira dancers to mariachis! The best thing is that on every Sunday all museums are free of entrance!