On Development of Portal for Deploying Bioinformatics Applications in Cluster and Grid Environments

Chiou-Nan Chen[1], Kuan-Ching Li[2*], Yaw-Ling Lin [2], and Hsiao-Hsi Wang[3]

[1] Department of Computer Science

National Tsing Hua University

Hsinchu 300, Taiwan

cnchen@cs.nthu.edu.tw

[2] Department of Computer Science and Information Engineering

Providence University

Shalu, Taichung 43301, Taiwan

{kuancli,yllin}@pu.edu.tw

[3] Department of Computer Science and Information Management

Providence University

Shalu, Taichung 43301, Taiwan

hhwang@pu.edu.tw

Received 12 June 2007; Accepted 20 August 2007

Abstract. Over last few years, interest on biotechnology has increased dramatically. With the completion of sequencing of the human genome, such interest is likely to expand even more rapidly. The size of genetic information database doubles every 14 months, overwhelming explosion of information in related bioscience disciplines and consequently, overtaxing any existing computational tool for data analysis. There is a persistent and continuous search for new alternatives or new technologies, all with the common goal of improving overall computational performance. Grid infrastructures are characterized by interconnecting a number of heterogeneous hosts through the internet, by enabling large-scale aggregation and sharing of computational, data and other resources across institutional boundaries. In this research paper, we present BioPortal, a user friendly and web-based GUI that eases the deployment of well-known bioinformatics applications on large-scale cluster and grid computing environments. The major motivation of this research is to enable biologists and geneticists, as also biology students and investigators, to access to high performance computing without specific technical knowledge of the means in which are handled by these computing environments and no less important, without introducing any additional drawback, in order to accelerate their experimental and sequence data analysis. As result, we could demonstrate the viability of such design and implementation, involving solely freely available softwares.

Keywords: Grid Portal, bioinformatics applications, Cluster and Grid environments

1 Introduction

The merging of two rapidly advancing technologies, molecular biology and computer science, has resulted in a new informatics science, namely bioinformatics. Bioinformatics includes methodologies on processing molecular biological information, in order to speedup researches in molecular biology. Modern molecular biology is characterized by huge volume of biological data. Take the classic molecular biology data type, the DNA sequence, for instance, major bioinformatics database centers including GeneBank, the NIH (National Institute of Health) genetic sequence database and its collaborating databases, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, these data have reached a milestone of 100 billion bases from over 165,000 organisms [3]. Common operations on biological data include sequences analysis, protein structures predication, genome sequences comparison, sequence alignment, phylogeny tree construction, pathway research, visualization of sequence alignment results and placement of sequence databases. The most basic and important bioinformatics task is to find the set of homologies for a given sequence, since sequences are often related in functions if they are similar.

* Correspondence author

Genome research centers, such as the National Center for Biotechnology Information (NCBI) and the European Molecular Biology Laboratory (EMBL), they host enormous volume of biological information in their bioinformatics database. They also provide a number of bioinformatics tools for database search and data acquisition. With the explosion of sequence information available to researchers, computational biologists face the challenge to aid biomedical researches, that is, to invent efficient toolkits to enlarge the use of available computational cycles. Sequence comparison, multiple sequence alignment and phylogeny tree construction are the most fundamental works in biomedical research. There have been many abundant examples of bioinformatics applications that are able to provide solutions for these problems in biomedical research. Some of most extensively utilized applications for these research activities include BLAST [4][5], ClustalW [6][7] and Phylip [8].

However, bioinformatics applications typically are distributed in different individual projects and they require high performance computational environments. Biomedical researchers need to combine many works to conclude their investigation. For instance, in the south of an Asian area, once farms with many dead chickens are reported, biologist may need to identify whether it was infected by H5N1 influenza virus urgently. After obtained the chicken's testimony and RNA sequence, biologist may use BLAST tool to search and acquire other influenza virus sequences from the public database. ClustalW tool is required to compare and investigate their similarity, so then construct the phylogenic tree using Phylip tool. In the above situation, biomedical researchers need these bioinformatics applications. They may download a local version to their own computer or use them in individual server, but either one is complicated and inefficient way, due to a number of drawbacks that any similar solution may bring. Therefore, an efficient and integrated bioinformatics portal is necessary, in order to facilitate biomedical researches.

Grid computing has irresistible potential to apply supercomputing power to address a vast range of bioinformatics problems. A computational grid is a collection of distributed and heterogeneous computing nodes that has emerged as an important platform for computation intensive applications [9][10][11]. They enable large-scale aggregation and sharing of computational, data and other resources across institutional boundaries. It offers an economic and flexible model for solving massive computational problems using large numbers of computers, arranged as clusters embedded in a distributed infrastructure [12][13][14][15].

In this research paper, we integrate several important bioinformatics applications into a novel user-friendly and biologist-oriented web-based GUI portal on top of PCGrid grid computing environment [16]. The major goal in developing such GUI is to assist biologists and geneticists to access to high performance computing, without introducing additional computing drawbacks to this attempt, as to accelerate their experimental and sequence data analysis.

The remainder of this paper is organized as follows. In Section 2, a number of bioinformatics application tools available are introduced, while in Section 3 is introduced the experimental grid computing platform PCGrid, a computing environment built by interconnecting a number of computational resources located inside Providence University Campus. In Section 4, it is discussed the BioPortal bioinformatics portal workflow and implementation. Finally, in Section 5, conclusions and future works are presented.

2 Bioinformatics Applications

Molecular biologists measure and utilize huge amounts of data, of various types. The intention is to use these data to:

1. reconstruct the past (e.g., infer the evolution of species),
2. predict the future (e.g., predict how some genes affect a certain disease),
3. guide bio-technology engineering (such as improving the efficiency of drug design).

Some of the concrete tasks are so complex that intermediate steps are already regarded as problem in their own and constructed an application for it. For instance, while the consensus motif of a sequence in principle determines its evolution function, one of the grand challenges in bioinformatics is to align multiple sequences among to conclude their consensus pattern and predict its function. Sequence comparison, multiple sequence alignment and phylogeny tree construction are fundamental works in biomedical research and bioinformatics. The most extensively applications for these works include BLAST, ClustalW and Phylip. BLAST is a sequence comparison and search tool, ClustalW is a progressive multiple sequence alignment tool, and Phylip is a program for inferring phylogenic tree.

The BLAST (Basic Local Alignment Search Tool) application is a widely used tool for searching DNA and protein databases for sequence similarity to identify homologs to a query sequence [20]. While often referred to as just "BLAST", this can really be thought of as a set of five sub-applications: blastp, blastn, blastx, tblastn, and tblastx.

Five sub-applications of BLAST perform the following tasks:

1. *blastp*: compare an amino acid query sequence against a protein sequence database,

2. *blastn*: compare a nucleotide query sequence against a nucleotide sequence database,
3. *blastx*: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database,
4. *tblastn*: compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands),
5. *tblastx*: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

BLAST tool plays an extremely important role in NCBI GenBank database. It not only provides sequence database search, but also include many toolkits for sequence comparison. BLAST is based on Smith-Waterman local alignment algorithm [17][18], which basically identifies the best local alignment between two sequences by using dynamic programming and tracing back metrology through the sequence matrix. The mpiBLAST is a parallelized version of BLAST, developed by Los Alamos National Laboratory (LANL) [19]. The mpiBLAST segments the BLAST database and distributes it across cluster computing nodes, permitting BLAST queries to be processed on a number of computing nodes simultaneously. The mpiBLAST-g2 is an enhanced version of LANL's mpiBLAST application [21]. This enhanced application allows the parallel execution of BLAST on a grid computing environment.

ClustalW is a general purpose multiple sequence alignment program for DNA or proteins, and it produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can clearly be seen. ClustalW is one of the most popular sequences alignment packages, and it is not only a multiple sequence alignment package, but also a phylogenetic tree construction tool. The progressive alignment algorithm of ClustalW is based on three steps:

1. Calculating sequence pairwise similarity,
2. Construction of guide tree,
3. Progressive alignment of sequence.

In the first step, all pairs of sequences are aligned separately, in order to calculate a distance matrix giving the divergence of each pair of sequences. As next step, the trees are used to guide the final multiple alignment processes that are calculated from the distance matrix of step 1 using the Neighbor-Joining method [22]. In the final step, the sequences are progressively aligned according to the branching order in the guided tree. ClustalW-MPI [24] is a parallel implementation of ClustalW. All three steps have been parallelized in order to reduce the global execution time, and it runs on distributed workstation clusters as well as on traditional parallel computers [23]. The only requirement is that all computing nodes involved in ClustalW-MPI computations should have installed MPI.

Phylip is an application for inferring phylogenies tree. The tree construction algorithm is quite straightforward, and it adds species one by one to the best place in the tree and makes some rearrangement to improve the result.

3 The PCGrid Computing Infrastructure

The PCGrid grid-computing platform, standing for The Providence University Campus Grid platform, consists basically of five cluster platforms located in different floors and laboratories inside the College of Computing and Informatics (CCI) of this university. The project of constructing such grid infrastructure is aimed to increase Providence University's computational power and share the resources among investigators and researchers in fields such as bioinformatics, biochemistry, medical informatics, economy, parallel compilers, parallel software, data distribution, multicast, network security, performance analysis and visualization toolkit, computing node selection, thread migration, scheduling in cluster and grid environments, among others.

The PCGrid computing infrastructure is formed by interconnecting the cluster computing platforms via Gigabit Ethernet (1Gb/s), as illustrated in Fig. 1.

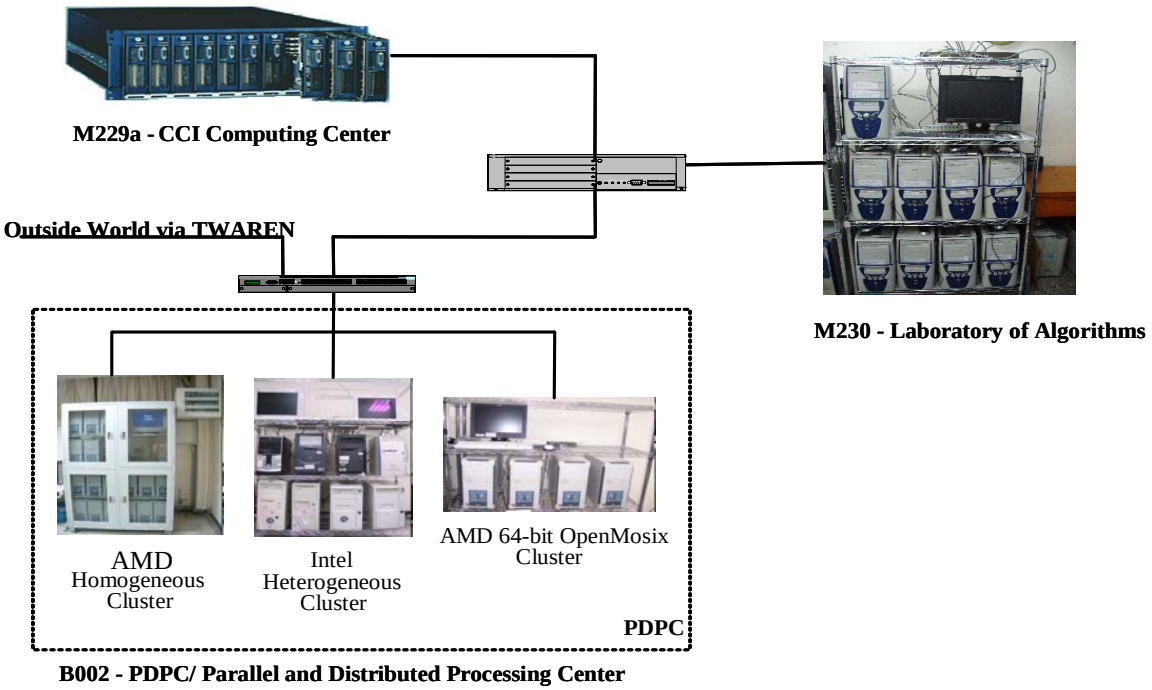

Fig. 1. The PCGrid grid computing infrastructure

The first platform is AMD Homogeneous Cluster, consisting of 17 computing nodes, where each node contains one AMD Athlon 2400+ CPU, 1GB DDR memory, 80GB HD, FedoreCore4 OS, interconnected via Gigabit Ethernet. The second cluster is Intel Heterogeneous Cluster, built up using 9 computing nodes with different CPU speed and memory size, FedoraCore2 OS, interconnected via Fast Ethernet. The third cluster platform consists of 4 computing nodes, where each computing node has one AMD 64-bit Sempron 2800+ CPU, 1GB DDR memory, 120GB HD, FedoreCore4 OS, interconnected via Gigabit Ethernet. The fourth cluster platform is IBMCluster, consisting of 9 computing nodes, where each node contains one Intel P4 3.2GHz CPU, 1 GB DDR memory, FedoraCore3 OS, 120GB HD, interconnected via Gigabit Ethernet. The fifth computing system is IBMBlade, consisting of 6 computing blades, where each blade has two PowerPC 970 1.6 GHz CPUs, 2GB DDR memory and 120GB HD, SUSE Linux OS, interconnected via Gigabit Ethernet. At present moment, the total storage contains more than 6TB of storage space.

4 BioPortal: a Portal for Bioinformatics Applications

We have integrated most fundamental computing applications in biomedical research and bioinformatics inside BioPortal: sequence comparison, pairwise or multiple sequence alignment and phylogeny tree construction, all in a complete workflow. We also provide an additional feature to biologists to choose automatically computing nodes to execute their parallel applications, by setting the number of computing nodes. The BioPortal will take care of selecting best computing nodes that fits users' requested computation, as described in subsection 4.1.

BioPortal is built using Grid Portal and integrated to a multi-function Grid Portal ScalaPortal, providing a single web interface to all computational resources available. The Grid Portal ScalaPortal includes also functions RSS, documentation repository, instant messaging, and information about users' recent accesses, and additional functions are still under development, to fulfill users' need, and therefore, to make this a complete and integrated environment for Grids.

Fig. 2 shows the homepage for BioPortal. The biologist can use bl2seq (a BLAST toolkit for two sequence comparison) to compare their own sequence with other sequences that was acquired from a bioinformatics database by blastcl3 (a NCBI BLAST client), while figures 3 and 4 show the web interface screenshot of Bl2seq and Blastcl3 respectively.

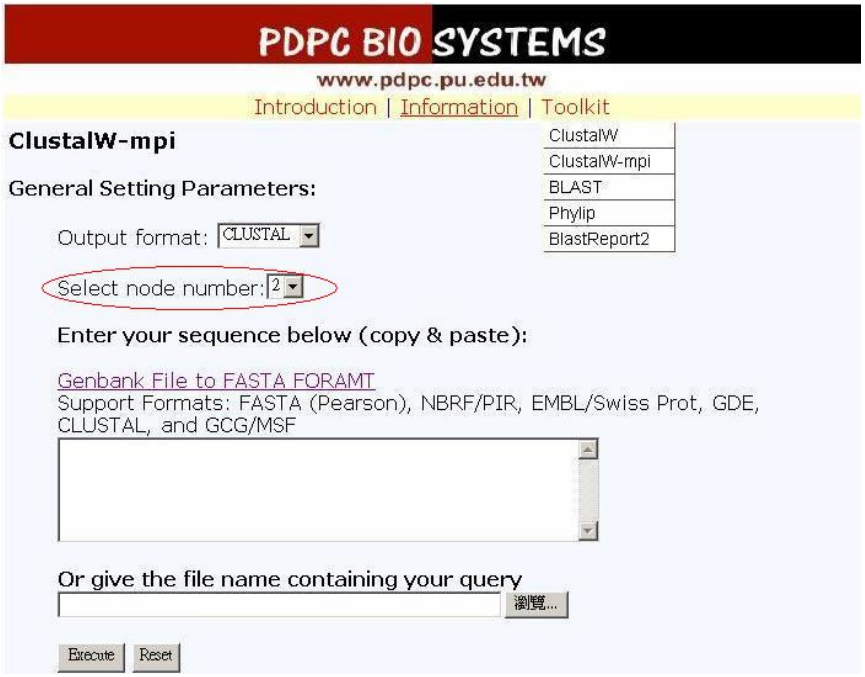

Fig. 2. BioPortal web-based GUI screenshot

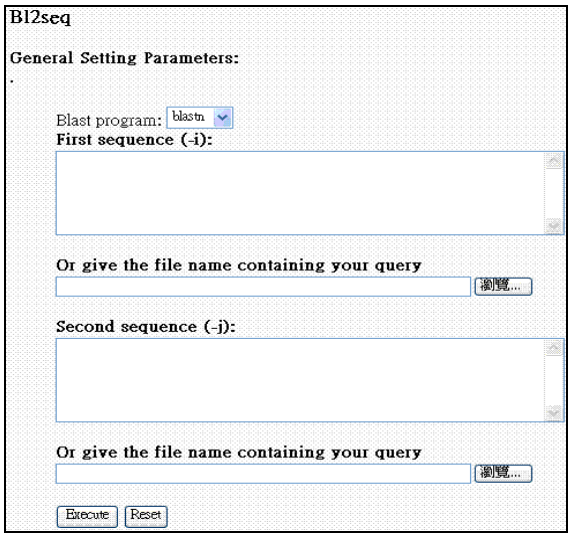

Fig. 3. bl2seq interface

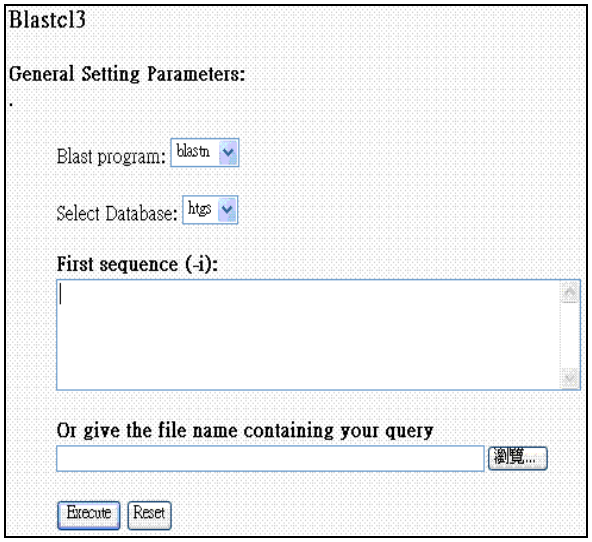

Fig. 4. Blastc13 interface

4.1 Selecting Computing Nodes

There are two ways to select computing nodes in PCGrid grid computing platform, either manual or automatic. In the manual process, the developer chooses the computing nodes based on CPU activities, depending on the status (busy or idle), as shown in Fig. 5 and Fig. 6. If the developer persists in selecting a computing node showing RUNNING (that is, CPU in use), this job will be queued, and its execution will only be started when all selected computing nodes are idle. The alternative way to select computing nodes is automatic. All computing nodes in PCGrid platform are sorted and ranked, so that the developer selects a given condition, if he would like to select a number of computing nodes according to their speed (and idle) or he would like to select a number of computing nodes with higher network bandwidth.

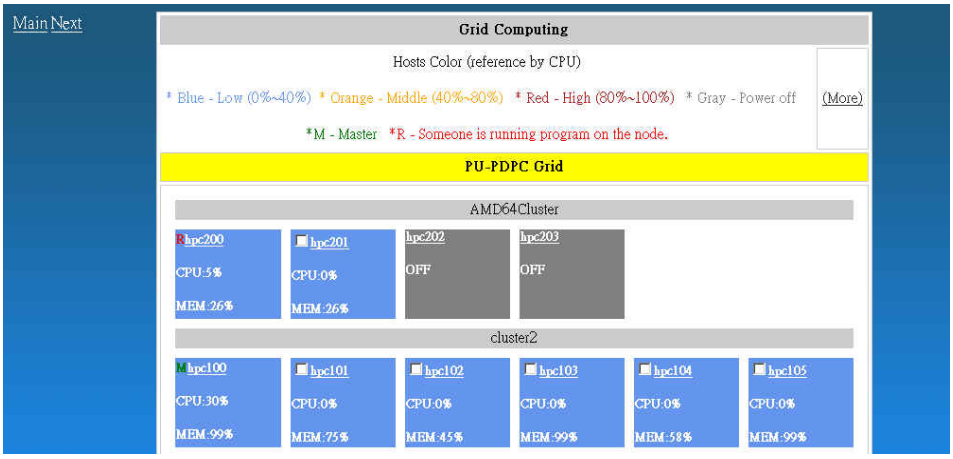

Fig. 5. Computing Node manual selection simple mode

All jobs submitted by any user are ranked according to user credentials, his level of priority inside the queue. The higher a user's credentials; highest is the priority to execute this user's applications in our computing platform. The queue is re-ranked every time a job is submitted to our grid platform.

Fig. 6. Real-time display of all computing nodes status in complete mode

Biologists make use of ClustalW-MPI to perform multiple sequence alignment with a number of sequences, and then construct corresponding phylogenic tree using Phylip directly. Biologists do need not to copy the alignment result from the ClustalW-MPI and paste to Phylip to get the phylogeny tree, since our system provide a "shortcut" button in order to facilitate similar procedures. Fig. 7 shows the web interface of ClustalW-MPI integrated with Phylip. We also develop a data format translation tool to ease biologist's usage. Biologist can input GeneBank data format, and our translation toolkit can transform it to legal FASTA format for ClustalW-MPI, as in Fig. 9. Detailed description of all bioinformatics services available in our BioPortal is listed in Table 1, while Fig. 8 shows the complete workflow of the BioPortal.

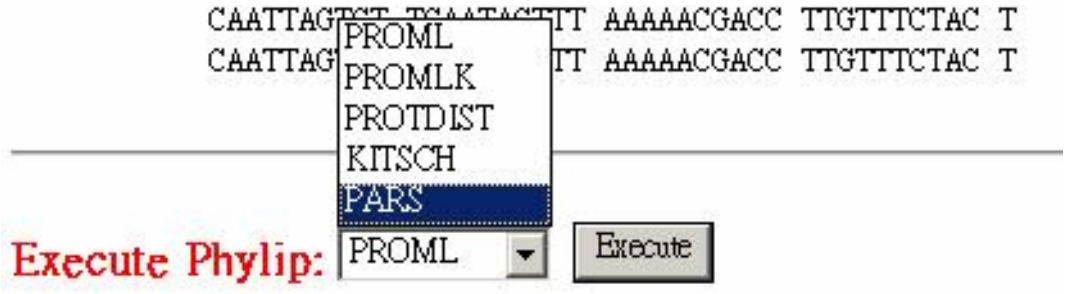

Fig. 7. Using Phylip application to construct phylogenic tree, directly from the output generated by ClustalW-MPI

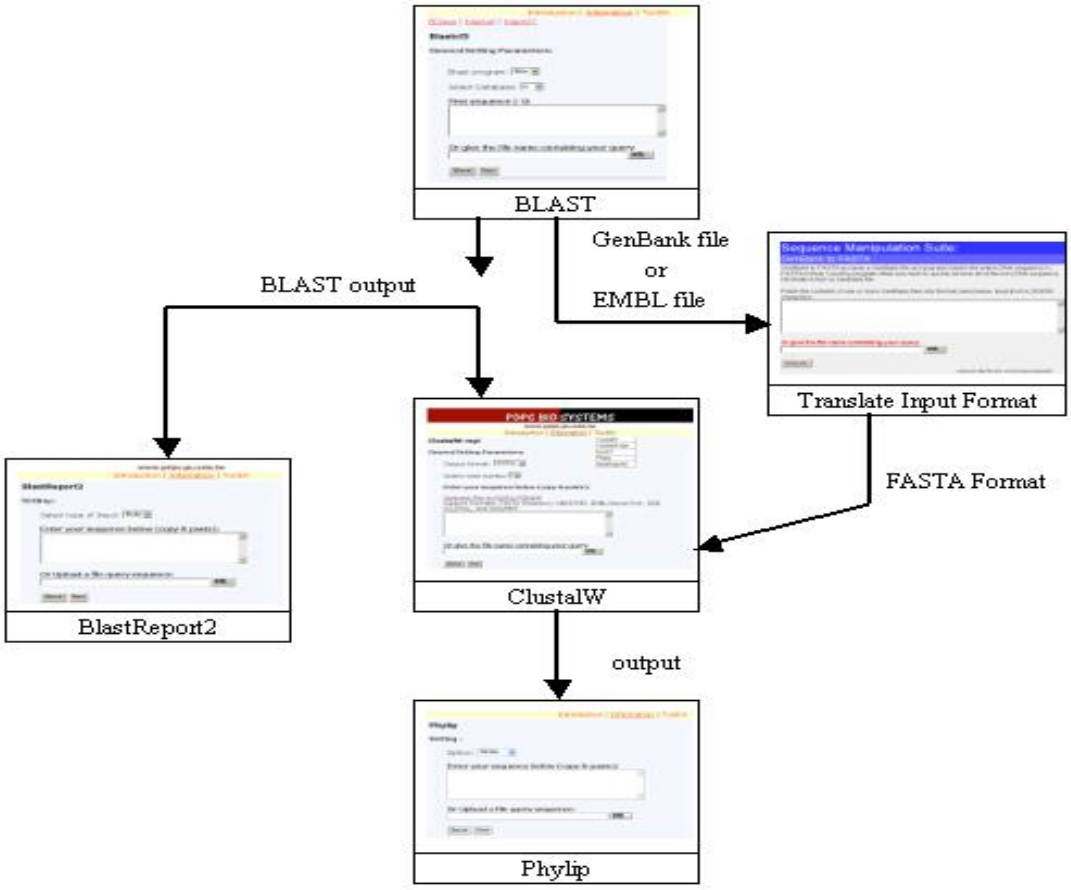

Fig. 8. BioPortal web-based GUI complete workflow

Table 1. List of bioinformatics applications provided by BioPortal

Application Tools	Description
mpiBLAST-g2	An enhanced parallel application that permits parallel execution of BLAST on Grid environments, based on GLOBUS and MPICH
Bl2seq	This application performs comparison between two sequences, using either blastn or blastp algorithms
Blastall	This application may be used to perform BLAST comparisons
BLASTcl3	A BLAST software client running on local computers that connects to BLAST servers located at NCBI, in order to perform searches and queries of NCBI sequence databases
Formatdb	It is used to format protein or nucleotide source database before these can be utilized by Blastall, Blastpgp or MEGABlast
BlastReport2	A Perl script that reads the output of Blastcl3, reformats it to ease its use and eliminates useless information
ClustalW-MPI	Parallel version of a general purpose multiple sequence alignment application for DNA or proteins, by producing meaningful multiple sequence alignment of divergent sequences
Phylip	Set of applications that performs phylogenic analyses

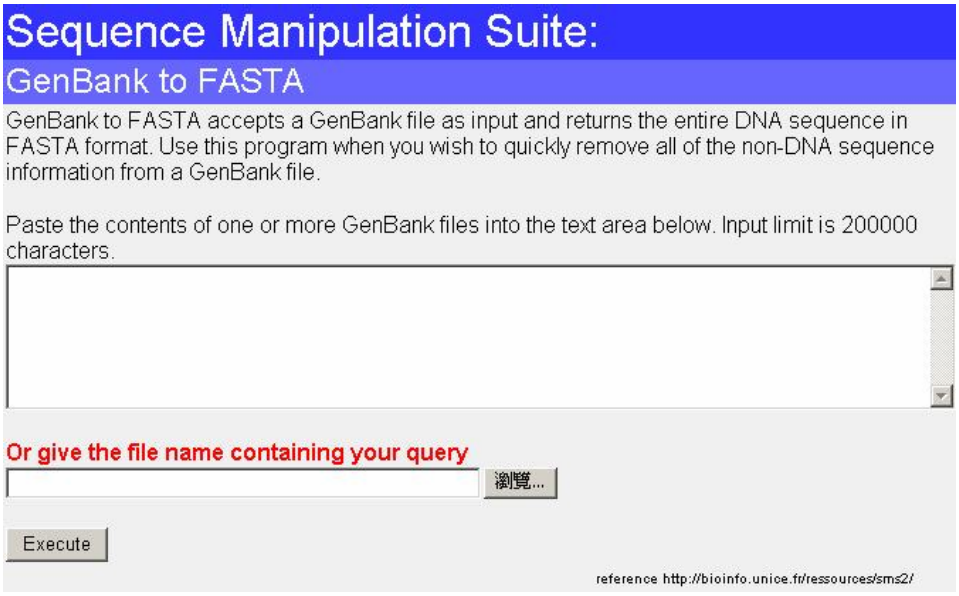

Fig. 9. Sequence data transformation toolkit

5 Conclusions and Future Work

We have constructed a campus scale computing grid platform, as also implemented a portal providing a number of well-known bioinformatics application toolkits. Not only to provide easy access of bioinformatics application toolkits to biologists and geneticists, but also large amount of computational cycles in an easy way. This portal contributes three fundamental molecular biology activities: sequence comparison, multiple sequence alignment and phylogenic tree construction, all integrated in a friendly and easy-to-use web-based GUI portal. We have solved many data inconsistency problems and finally integrated a number of different tools that are able to co-operate all together. This BioPortal not only facilitate biomedical researcher investigations and computational biology courses in graduate-level, as also it demonstrates a well-succeeded combination of high performance computing with the use of grid technology and bioinformatics.

As future work, several directions of this research are ongoing. One of goals is to develop a one-stop-shop bioinformatics portal, to provide efficient and economic computational power and cycles to biomedical researchers. At the present moment, we are in the process of integrating other well-known bioinformatics applications into this BioPortal, for instance, applications for protein structure predication and protein visualization. We expect to continuously develop on top of grid technology, so that in near future, researchers will not only be able to seamlessly utilize PCGrid computational resources, but also expand on demand to larger scale grid computing platforms, such as regional or national grid platforms.

Acknowledgement

This paper is based upon work supported in part by National Science Council (NSC), Taiwan, under grants no. NSC93-2213-E-126-010 and NSC94-2213-E-126-005, and National Center for High-Performance Computing (NCHC), Taiwan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSC or NCHC.

References

[1] K.C. Li, H.H. Wang, C.N. Chen, C.C. Liu, C.F. Chang, C.W. Hsu, S.S. Hung, “Design Issues of a Novel Toolkit for Parallel Application Performance Monitoring and Analysis in Cluster and Grid Environments”, *Proceedings of The 8th IEEE International Symposium on Parallel Architectures, Algorithms, and Networks*, Las Vegas, USA, 2005.

[2] H.C. Chang, K.C. Li, Y.L. Lin, C.T. Yang, H.H. Wang, and L.T. Lee, "Performance Issues of Grid Computing Based on Different Architecture Cluster Computing Platforms", *Proceedings of The 19th IEEE International Conference on Advanced Information Networking and Applications*, vol. II, Taipei, Taiwan, 2005.

[3] Public Collections of DNA and RNA Sequence Reach 100 Gigabases, National Institutes of Health, August 22, 2005. (http:// www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html).

[4] S.F. Altschul, W. Gish, W. Miller, E.G. Myers, and D.J. Lipman, "Basic Local Alignment Search Tool", *Journal of Molecular Biology*, Vol.215, pp.403-410, 1990.

[5] S. F. Altschul, T. L. Madden, A. A. Schaeffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", *Nucleic Acids Research*, Vol.25, pp. 3389-3402, 1997.

[6] D.G. Higgins, P.M. Sharp, "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer", *Gene*, Vol.73, No.1, pp.237-44, Dec. 1988.

[7] J.D. Thompson, D.G. Higgins, T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", Nucleic Acids Research, Vol.22, No.22, pp.4673-4680, Nov. 1994.

[8] Joe Felsenstein, "PHYLIP (Phylogeny Inference Package)", version 3.5c. (http://evolution.genetics.washington.edu/phylip.html), 1993.

[9] B. Allcock, J. Bester, J. Bresnahan, A. L. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnal, and S. Tuecke. "Data Management and Transfer in High Performance Computational Grid Environments", *Parallel Computing*, Vol. 28, No.5, pp. 749-771, 2002.

[10] B. Allcock, J. Bester, J. Bresnahan, I. Foster, J. Gawor, J. A. Insley, J. M. Link, and M. E. Papka. "GridMapper: A Tool for Visualizing the Behavior of Large-Scale Distributed Systems", *Proceedings of 11th IEEE International Symposium on High Performance Distributed Computing* (HPDC-11), 2002.

[11] K. Czajkowski, S. Fitzgerald, I. Foster, and C. Kesselman. "Grid Information Services for Distributed Resource Sharing", *Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing* (HPDC-10), 2001.

[12] M. Baker, R. Buyaa, D. Laforenza, "Grid and Grid Technologies for Wide-Area Distributed Computing", available at http://www.csse.monash.edu.au/~rajkumar/papers/gridtech.pdf .

[13] F. Berman, A. Chien, K. Cooper, J. Dongarra, I. Foster, D. Gannon, L. Johnson, K. Kennedy, C. Kesselman, J. Mellor-Crummey, D. Reed, L. Torczon, and R. Wolski, "The GrADS Project: Software Support for High-Level Grid Application Development", *International Journal of High-Performance Computing Applications*, Vol.15, No.4, 2002.

[14] M. Chetty, R. Buyya, "Weaving computational Grids: How analogous are they with electrical Grids?" *Journal of Computing in Science and Engineering* (CiSE), Vol.4, No.4, pp. 61-71, July/August 2002.

[15] K. Czajkowski, I. Foster, and C. Kesselman. "Resource Co-Allocation in Computational Grids", *Proceedings of the Eighth IEEE International Symposium on High Performance Distributed Computing* (HPDC-8), pp. 219-228, 1999.

[16] K.C. Li, C.N. Chen, C.W. Hsu, S.S. Hung, C.F. Chang, C.C. Liu, C.Y. Lai, "PCGrid: Integration of College's Research Computing Infrastructures Using Grid Technology", *Proceedings of the 2005 National Computer Symposium*, Taiwan, 2005.

[17] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool", *Journal of Molecular Biology*, Vol. 215, pp. 403–410, 1990.

[18] T.F. Smith, M.S. Waterman, "Identification Of Common Molecular Subsequences", *Journal of Molecular Biology*, Vol. 147, pp. 195-197, 1981.

[19] Los Alamos National Laboratory (http://mpiblast.lanl.gov).

[20] Heshan Lin, Xiaosong Ma, Praveen Chandramohan, Al Geist and Nagiza Samatova, "Efficient Data Access for Parallel BLAST", *Proceedings of IEEE International Parallel & Distributed Symposium*, 2005.

[21] mpiBLAST-g2, Bioinformatics Technology and Service (BITS) team, Academia Sinica Computing Centre (ASCC), Taiwan. (http://bits.sinica.edu.tw/mpiBlast/mpiBlast-g2/README.mpiBLAST-g2.html)

[22] N. Saitou, M. Nei, "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees", *Molecular Biology and Evolution*, Vol.4, No.4, pp. 406-25, 1987.

[23] J.D. Thompson, D.G. Higgins, and T.J. Gibson, "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Positions-Specific Gap Penalties and Weight Matrix Choice", *Nucleic Acids Research*, Vol.22, pp. 4673-4680, 1994.

[24] K.B. Li, "ClustalW-MPI: ClustalW Analysis Using Distributed and Parallel Computing", *Bioinformatics*, Vol.19, No.12, pp.1585-6, 2003.