EST-PAC^{HPC} – a web portal for high-throughput EST annotation and protein sequence prediction

Adam K.L. Wong[1], Andrzej M. Goscinski[1], Christophe Lefèvre[2.3]

[1]*School of Information Technology, Deakin University, Geelong, Australia*
aklwong, ang{@deakin.edu.au}
[2]*Institute for Technology Research and Innovation (ITRI), BioDeakin, Deakin University*
[3]*Victorian Bioinformatics Consortium, Monash University*
clefevre@deakin.edu.au

Abstract – **Expressed Sequence Tags (ESTs) are short DNA sequences generated by sequencing the transcribed cDNAs coming from a gene expression. They can provide significant functional, structural and evolutionary information and thus are a primary resource for gene discovery. EST annotation basically refers to the analysis of unknown ESTs that can be performed by database similarity search for possible identities and database search for functional prediction of translation products. Such kind of annotation typically consists of a series of repetitive tasks which should be automated, and be customizable and amenable to using distributed computing resources. Furthermore, processing of EST data should be done efficiently using a high performance computing platform. In this paper, we describe an EST annotator, EST-PAC^{HPC}, which has been developed for harnessing HPC resources potentially from Grid and Cloud systems for high throughput EST annotations. The performance analysis of EST-PAC^{HPC} has shown that it provides substantial performance gain in EST annotation.**

Keywords: Expressed Sequence Tag, High Throughput EST Annotations, EST Data Mining, Grid and Cloud Computing, Performance Evaluation.

BIOCOMP 2011

I. Introduction

High-end computing facilities such as grids and clouds [8] are the key to enabling bioinformatics projects in the next generation sequencing era. Different research groups both nationally and internationally could benefit by sharing research data, computing platforms and experiment results cost-effectively. Many research laboratories will have many terabytes if not petabytes of data to transfer, store and analyse. Handling and analysing such huge amount of genomic data require fast and reliable computer networks as well as a huge amount of computation power and storage. Although high-end supercomputers are now easily available to a broad scientific community, users without in depth I.T. knowledge are often forced to cope with many low-level details when using those machines for scientific investigations. Cloud technologies in particular promise to provide seamless access to high performance computer clusters through the abstractions of services and brokers that hide the details of the underlying software and hardware infrastructure. In this paper, we describe an expressed sequence tag (EST) annotator, EST-PAC^{HPC}, which has been developed for EST annotation on a high performance computing platform.

BLAST [13] is probably the most worldwide used bioinformatics tool for sequence alignment and BLAST searching of ESTs is a key component task of EST annotation. A typical EST annotation procedure often needs to perform BLAST searching for a large volume of ESTs repeatedly on different genomic databases. Thus, such procedure should be executed on a HPC platform to leveraging the power of parallel processing. There are a number of programs and hardware solutions for efficient high-throughput BLAST searching in Grids [4] and Clouds [5]. However, there is a lack of generic software solutions for personalized management, presentation and mining of the search results. For this reason, downstream analysis remains a task to be solved in ad hoc ways by different users. On the other hand, other EST annotators [17] have concentrated on providing an intergraded annotation and data mining environment but have failed to handle the high throughput computational requirement of EST annotation. EST-PAC^{HPC} is a fully functional EST annotator, which performs using HPC resources potentially from various grid and cloud systems.

The rest of this paper is organized as follows. Section 2 provides the background knowledge of EST annotation. It also describes the EST-PAC and EST-PAC^{HPC} software packages. Section 3 explains the approach taken by EST-PAC^{HPC} for high throughout BLAST searching. Section 4 covers the performance evaluation of EST-PAC^{HPC} for EST annotation using BLAST searching on a HPC platform. The experimental test-bed, workload construction and results of the performance evaluation are discussed. Finally, Section 5 presents the conclusions and our future work.

II. Background

An expressed sequence tag (EST) is a short DNA sequence, usually 200 to 500 nucleotides long, that is generated by sequencing the transcribed cDNA sequence of an expressed gene. ESTs were used for the first time as a primary resource for human gene discovery [1]. Since then, there has been an exponential growth in the generation and accumulation of EST data, with approximately 69 million ESTs now available in public databases (GenBank 01 March 2011, all species). Since ESTs can provide significant functional, structural and evolutionary information, there are a lot of worldwide biological projects and laboratories that continually produce ESTs for different researching tasks. Many EST sequencing projects are underway for numerous organisms and extensive computational strategies have been developed to organize and analyse both small- and large-scale EST data for gene discovery, transcript and single nucleotide polymorphism analysis as well as functional annotation of putative gene products [17].

With the decreasing cost of DNA sequencing technology and the vast diversity of biological resources, researchers increasingly face the basic challenge of annotating a huge amount of EST data from a variety of species. EST annotation basically refers to the analysis of unknown ESTs that can be performed by database similarity search for possible identities and database search for functional prediction of translation products. Such kind of annotation typically consists of a series of repetitive tasks, which should be automated, and all these operations should be self-installing, platform independent, easy to customize and amenable to using distributed computing resources. Furthermore, processing of EST data should be done efficiently on high performance computing platforms.

A. EST-PAC

EST-PAC was developed as a web oriented multi-platform software package for EST annotation which can run on a single compute-server [19]. It has integrated the BLASTALL suite [13], EST-Scan2 [10] and HMMER [7] in a relational database system accessible through a web portal. The system allows users to customize annotation strategies and provides an open-source data-management environment for research and education in bioinformatics.

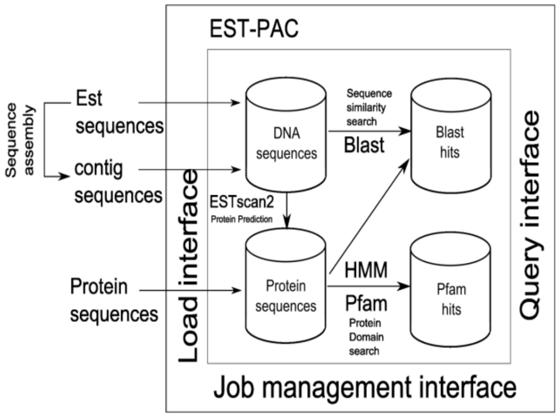


Fig. 1 Workflow and interfaces available in EST-PAC

The core of EST-PAC consists of an open source relational database management system that uses Structured Query Language, MySQL 5 [11], and a number of PHP 5 [15] programs, which allow the storage and management of ESTs using a web interface. The workflow of ESTs annotation is shown in Fig.1. User login is available for visualization and query, with additional privileges to run annotation tools. Sequences in FASTA format [9] are loaded into the database through a web interface and annotation tasks can be requested. A set of continuously running programs checks the database and extracts sequences to be processed using the BLASTALL suite, ESTScan2 or, HMMER.

The coding content of the EST can be evaluated with the Hidden Markov Model approach of ESTScan2 and the predicted translation products can then be compared against protein sequence databases. A report can be obtained from a web query page. As all results are stored in a relational database, users are able to query on every value returned by the annotation process. An interface is also available to assist the construction and storage of database queries. In addition to the public databases which can be downloaded and installed locally or accessed through web based blast services such as NCBI [12], users have the possibility to create their own databases from EST-PAC in order to make more precise and relevant comparisons.

B. EST-PAC^{HPC}

We have extended EST-PAC into EST-PAC^{HPC} which can utilize HPC resources such as computer clusters in Grids; and with a potential of using clouds resources for bioinformatics computing. The web portal approach of EST-PAC^{HPC} has enabled biologists who are not IT specialists to benefit directly from the use of high-performance computing technology. EST-PAC^{HPC} supports both high throughput and high performance computation of the selected bioinformatics applications.

To achieve high throughput computation, bioinformatics jobs from many users can run on different processors of a cluster concurrently. This solution has shortened the service waiting time. To achieve high performance computation, many of the submitted bioinformatics jobs can run on multiple processors of a cluster as parallel applications.

As shown in Fig. 2, an Apache [3] web-server with MySQL database system and PHP language script interpreter form the backbone of the EST-PAC^{HPC} Bio-Server, which is currently providing computation services to the Bioinformatics Research Group [6] at Deakin University. The heart of EST-PAC^{HPC} lays on its novel job-scheduling mechanism that integrates transparently with most of the existing cluster and grid resource managers such as PBS [16] and Sun Grid Engine [18]. Currently, the openMPI [14] parallel programming environment is adopted for parallel computation.

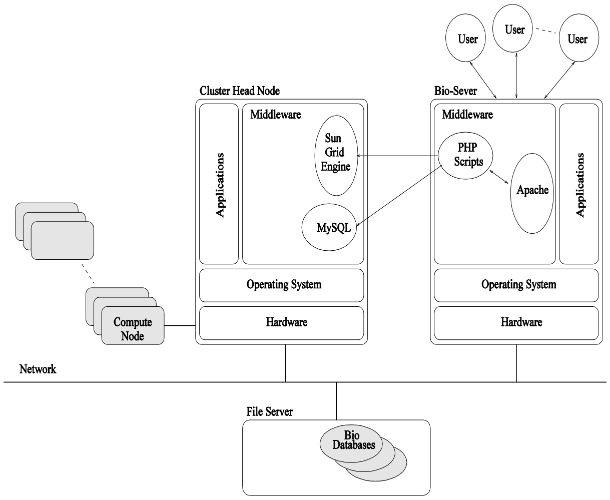
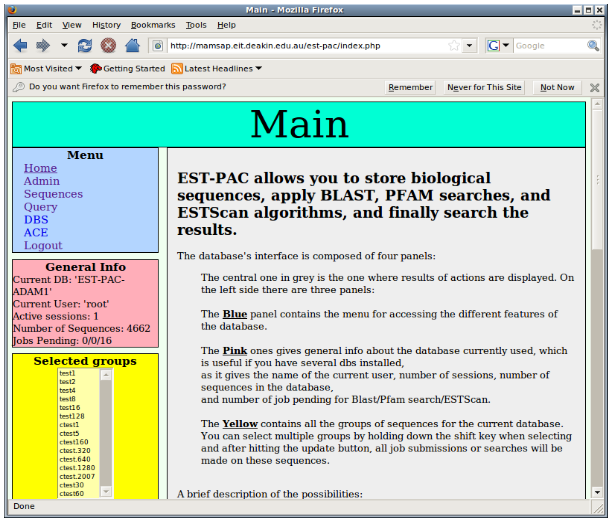


Fig. 2 Architecture of EST-PAC^{HPC} running as a bioinformatics computation server at Deakin University

A web-portal interface is provided in EST-PAC^{HPC} to release biologists from performing tedious I.T. tasks such as hardware setup, software installation and configuration as well as data management. Most importantly, it hides completely the details of bioinformatics application deployment in the underling HPC platform. Fig. 3, Fig. 4 and Fig. 5 are the snapshots extracted from EST-PAC^{HPC} web portal; each of them shows the main functions, EST sequence handling and EST annotation correspondingly.



Do you want Firefox to remember this password?

General Info
Current DB: 'EST-PAC-ADAM1'
Current User: 'root'
Active sessions: 1
Number of Sequences: 4662
Jobs Pending: 0/0/16

EST-PAC allows you to store biological sequences, apply BLAST, PFAM searches, and ESTScan algorithms, and finally search the results.

The database's interface is composed of four panels:

The central one in grey is the one where results of actions are displayed. On the left side there are three panels:

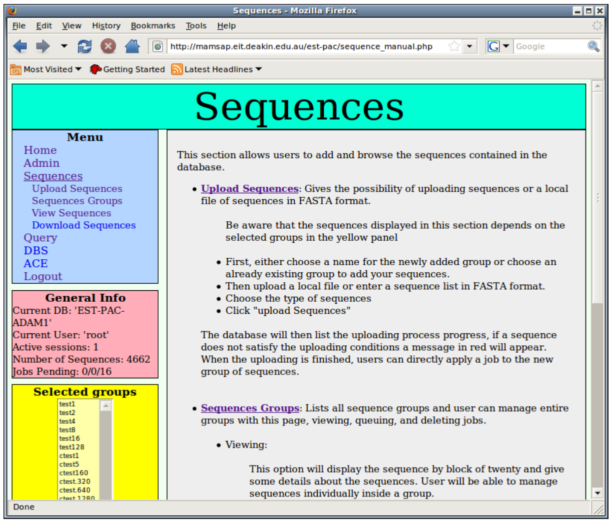
The **Blue** panel contains the menu for accessing the different features of the database.

The **Pink** ones gives general info about the database currently used, which is useful if you have several dbs installed,
as it gives the name of the current user, number of sessions, number of sequences in the database,
and number of job pending for Blast/Pfam search/ESTScan.

The **Yellow** contains all the groups of sequences for the current database. You can select multiple groups by holding down the shift key when selecting and after hitting the update button, all job submissions or searches will be made on these sequences.

A brief description of the possibilities:

Fig. 3 Main page of the EST-PACHPC web portal



General Info
Current DB: 'EST-PAC-ADAM1'
Current User: 'root'
Active sessions: 1
Number of Sequences: 4662
Jobs Pending: 0/0/16

This section allows users to add and browse the sequences contained in the database.

- **Upload Sequences**: Gives the possibility of uploading sequences or a local file of sequences in FASTA format.

  Be aware that the sequences displayed in this section depends on the selected groups in the yellow panel

  - First, either choose a name for the newly added group or choose an already existing group to add your sequences.
  - Then upload a local file or enter a sequence list in FASTA format.
  - Choose the type of sequences
  - Click "upload Sequences"

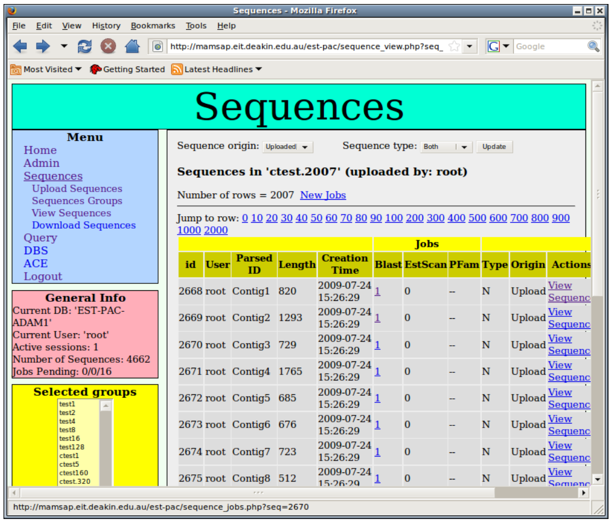
  The database will then list the uploading process progress, if a sequence does not satisfy the uploading conditions a message in red will appear. When the uploading is finished, users can directly apply a job to the new group of sequences.

- **Sequences Groups**: Lists all sequence groups and user can manage entire groups with this page, viewing, queuing, and deleting jobs.

  - Viewing:

    This option will display the sequence by block of twenty and give some details about the sequences. User will be able to manage sequences individually inside a group.

Fig. 4 Web page showing major functions for EST sequence handling



General Info
Current DB: 'EST-PAC-ADAM1'
Current User: 'root'
Active sessions: 1
Number of Sequences: 4662
Jobs Pending: 0/0/16

Sequence origin: Uploaded   Sequence type: Both

Sequences in 'ctest.2007' (uploaded by: root)

Number of rows = 2007 New Jobs

Jump to row: 0 10 20 30 40 50 60 70 80 90 100 200 300 400 500 600 700 800 900 1000 2000

					Jobs					
id	User	Parsed ID	Length	Creation Time	Blast	EstScan	PFam	Type	Origin	Actions
2668	root	Contig1	820	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2669	root	Contig2	1293	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2670	root	Contig3	729	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2671	root	Contig4	1765	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2672	root	Contig5	685	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2673	root	Contig6	676	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2674	root	Contig7	723	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc
2675	root	Contig8	512	2009-07-24 15:26:29	1	0	--	N	Upload	View Sequenc

http://mamsap.eit.deakin.edu.au/est-pac/sequence_jobs.php?seq=2670

Fig. 5 Web page showing EST annotation job deployment

III. HIGH THROUGHPUT EST ANNOTATION IN EST-PAC ^{HPC}

As mentioned in Section II, the core operation in EST analysis is a database similarity search, which assigns possible identities to the unknown ESTs. This can be done by the BLAST program. There are many publicly available resources for users to carry out BLAST operations that can be found such as a free resource from the National Centre for Biotechnology Information, NCBI-Blast [13] on the one side, to a pay-per-use resource from Windows Azure Blast [5] on the other.

The NCBI-Blast server, which is backed by a high-end supercomputer, provides a real time and high performance BLAST service to users. However, this free service has restricted users from carrying out high throughput BLAST searches. A submitted BLAST job of more than 50 sequences will be penalized in term of its dispatch time as the server is shared world wide. Besides, users' search results will not be kept in the NCBI databases indefinitely. The pay BLAST service from Windows Azure seems to be a flexible and cost-effective solution for carrying out high-throughput BLAST despite that it is still a trial service from Azure. Nevertheless, those service providers do not mean to provide EST annotation service to users. The downstream result analysis remains a task to be solved in ad hoc ways by users.

A. *EST annotation: An ad hoc approach*

Assuming BLAST is used to carry out the sequence similarity search, the basic steps of performing a high-throughput EST annotation are as follows:

1. Obtain and prepare copies of known genomic databases.
2. Obtain and prepare ESTs (short sequences).
3. Carry out BLAST search of ESTs on the known genomic databases.
4. Post-process BLAST search results: this refers to i) store results into DBMS system for further data mining process; and ii) visualize BLAST search results for ESTs analysis.
5. If necessary, repeat step 3 and 4 by replacing BLAST with different search tools such as EST-Scan2 and HMMER for coding region detection of DNA and protein sequence alignment.

Fig. 6 shows a workflow of high-throughput EST annotation via an ad hoc approach. As can be seen, users (mostly biologists) have to cope with many low-level details of BLAST parallelization as well as handling of annotation post-processing, which is tedious and time consuming.

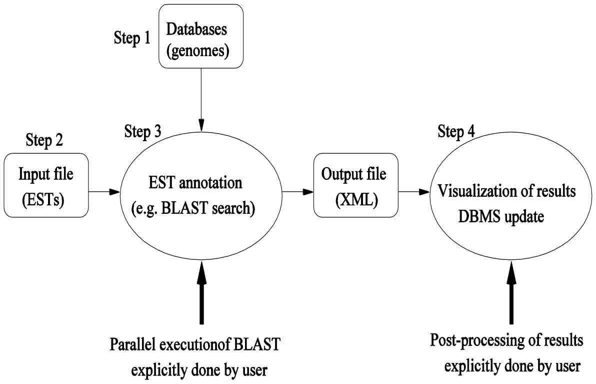
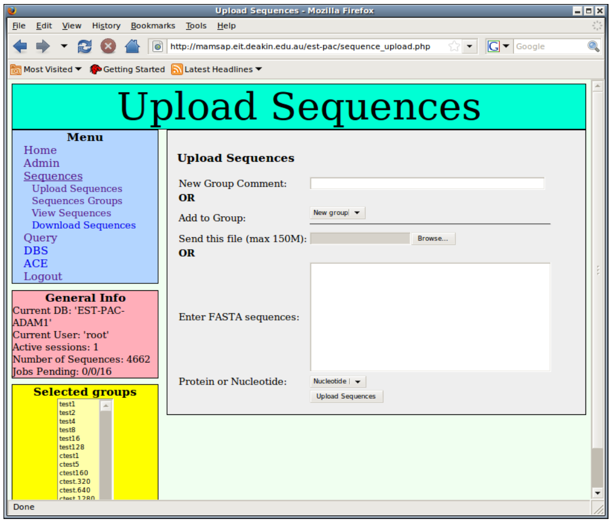

Step 1
Databases (genomes)
Step 2
Input file (ESTs)
Step 3
EST annotation (e.g. BLAST search)
Output file (XML)
Step 4
Visualization of results DBMS update
Parallel executionof BLAST explicitly done by user
Post-processing of results explicitly done by user

Fig. 6 Workflow of high-throughput EST annotation

B. *EST annotation: The HPC-PAC^{HPC} approach*

The web-portal interface provided in EST-PAC^{HPC} has simplified the tasks of Step 1 and Step 2 as described in the previous subsection (See Fig. 7). Once EST data are uploaded to the system, users can easily deploy an EST annotation, as corresponding to Step 3, via the web-portal (See Fig. 8). The running of BLAST searches on a computer system, e.g. HPC clusters, is completely transparent to users. The current implementation of EST-PAC^{HPC} has provided a job-scheduler, which can be integrated to most of the existing cluster and grid resource managers such as PBS and Sun Grid Engine, thus harnessing HPC resources potentially from various grids and clouds. Results of the BLAST searches are permanently stored in the MySQL DBMS and can be visualized in real time via the web-portal (See Fig. 9).



Upload Sequences

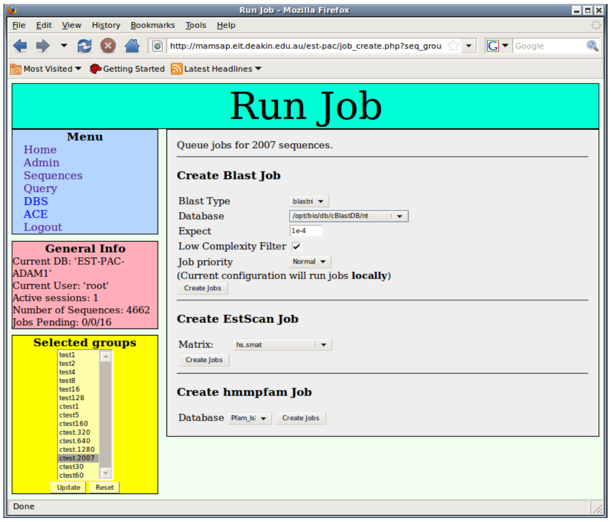
General Info
Current DB: 'EST-PAC-ADAM1'
Current User: 'root'
Active sessions: 1
Number of Sequences: 4662
Jobs Pending: 0/0/16

Selected groups
test1
test2
test4
test8
test16
test128
ctest1
ctest5
ctest160
ctest.320
ctest.640
ctest.1280

Upload Sequences

New Group Comment:
OR
Add to Group: New group
Send this file (max 150M):
OR
Enter FASTA sequences:
Protein or Nucleotide: Nucleotide

Fig. 7 Web page showing uploading of EST data



Run Job

General Info
Current DB: 'EST-PAC-ADAM1'
Current User: 'root'
Active sessions: 1
Number of Sequences: 4662
Jobs Pending: 0/0/16

Selected groups
test1
test2
test4
test8
test16
test128
ctest1
ctest5
ctest160
ctest.320
ctest.640
ctest.1280
ctest.2007
ctest30
ctest60

Queue jobs for 2007 sequences.

Create Blast Job

Blast Type: blastn
Database: /opt/bio/db/cBlastDB/nt
Expect: 1e-4
Low Complexity Filter ☑
Job priority: Normal
(Current configuration will run jobs **locally**)

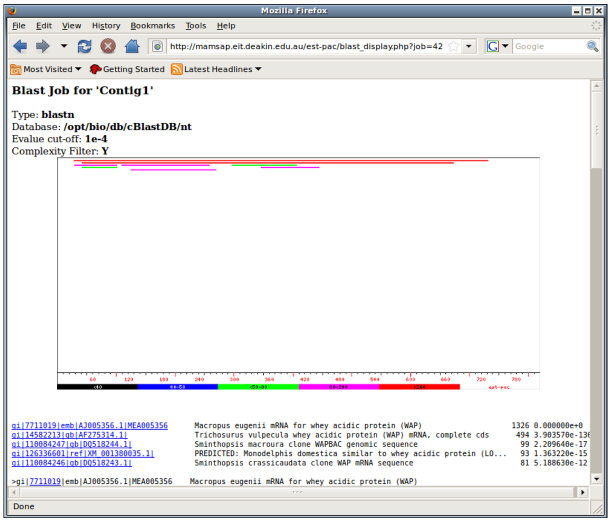
Create EstScan Job

Matrix: hs.smat

Create hmmpfam Job

Database: Pfam_ls

Fig. 8 Web page showing EST annotation job deployment



Blast Job for 'Contig1'

Type: **blastn**
Database: **/opt/bio/db/cBlastDB/nt**
Evalue cut-off: **1e-4**
Complexity Filter: **Y**

```
gi|7711019|emb|AJ005356.1|MEA005356     Macropus eugenii mRNA for whey acidic protein (WAP)                  1326 0.000000e+0
gi|14582213|gb|AF275314.1|              Trichosurus vulpecula whey acidic protein (WAP) mRNA, complete cds    494 3.903570e-136
gi|110084247|gb|DQ518244.1|             Sminthopsis macroura clone WAPBAC genomic sequence                     99 2.209640e-17
gi|126336601|ref|XM_001380035.1|        PREDICTED: Monodelphis domestica similar to whey acidic protein (LO...  93 1.363220e-15
gi|110084246|gb|DQ518243.1|             Sminthopsis crassicaudata clone WAP mRNA sequence                      81 5.188630e-12

>gi|7711019|emb|AJ005356.1|MEA005356    Macropus eugenii mRNA for whey acidic protein (WAP)
```

Fig. 9 Web page showing a BLAST search result of an EST

IV. Performance Evaluation

The previous section has demonstrated how accessibility and user-friendliness could be improved by using EST-PAC^{HPC} for EST annotation. In this section, we provide a performance evaluation of the system that aims to show how HPC resource can reduce the time of high-throughput EST annotation and thus to improve the productivity of biological researchers.

A. Experimental Testbed

The HPC hardware used in this study is a cluster of 10 compute-nodes (each with 8 CPU cores @ 1.6GHz and 8GB of RAM) where the EST-PAC^{HPC} web-portal and the MySQL DBMS are running on two separated workstations. The compute-nodes are networked by Cisco (Topspin) 10Gbs Infiniband PCI network interface. All of the compute-nodes are running the Linux CentOS operating system. Also, the Sun Grid Engine resource manager is used.

B. Experiments

Two major sets of experiment were performed. In the first experiment set, we calibrated our HPC cluster for its speedup performance by using some constructed workloads. The workloads were made up of mouse ESTs ranging from 2000 to 10000 sequences. BLAST search of these workloads were than performed against the mouse genome database. In the second experiment set, we measured the job escape time, which is the total execution time of BLAST search plus the total database storage time of BLAST search result. The experiment was performed for each of the entries shown in TABLE 1 for two cases. The first is to run on a computer server without HPC cluster. The second is to run on a computer server which is backed by a HPC cluster.

C. Results

Fig. 10, Fig. 11 and Fig. 12 have captured the outcome of the first experiment set, that is, a calibration of our 10 nodes (80 CPU cores) HPC cluster for high-throughput EST annotation. Fig. 10 shows a linear incremental relationship of EST annotation time of jobs against the EST sizes when the EST annotations were run on a single computer server. Fig. 11 shows that the job escape time of EST annotations decrease as the number of CPU used increase; and the job escape time of EST annotations increase as the EST size increase. Finally, Fig. 12 is the speedup representation of Fig. 11. The speedup is defined here as follow:

$$\mathit{Speedup} = \frac{\mathit{Job\ Escape\ Time}\ \ (\mathit{Single\ Computer\ Server})}{\mathit{Job\ Escape\ Time}\ (\mathit{Computer\ Server\ wiht\ HPC\ Cluster})}$$

TABLE 1 Specifications of Input EST Sequences and Search Databases

Unknown ESTS			Genomic Databases	
Source	**Nucleotide Length**	**Total Sequences**	**Source**	**Storage Size**
Mouse	65	500000	Mouse	3.5 GBytes
Mouse	65	1000000	Mouse	3.5 GBytes
Seal	Upto 2500	11232	All-organism	13 GBytes
Wallaby	Upto 2000	14837	All-organism	13 GBytes

This calibration has also shown that reasonable speedup of EST annotation is achievable even for data sets of small number of EST sequences. However, we believe that there is still room for improvement in the speedup, especially in handling the concurrence of DBMS update operations.

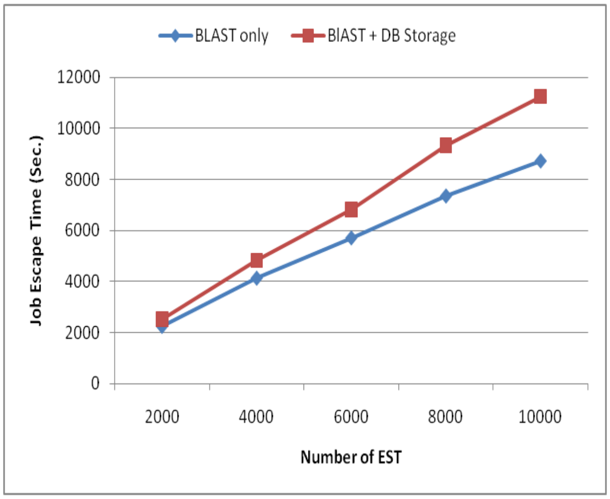


Fig. 10 Job Escape Time of EST annotation against EST size

(Single computer server)

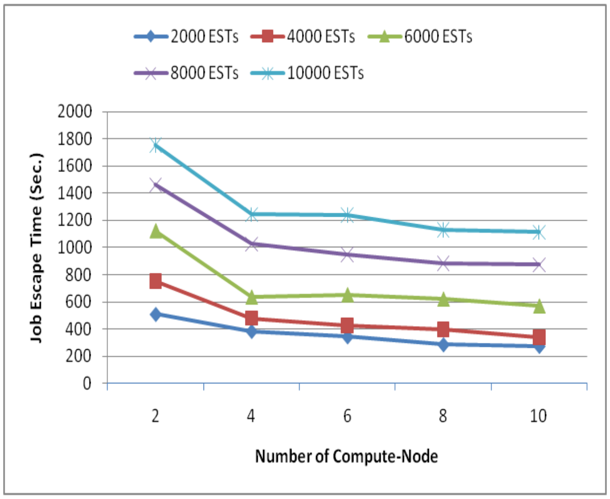


Fig. 11 Job Escape Time of EST annotation against No. of Compute-Node with different EST sizes (Computer server with HPC cluster)

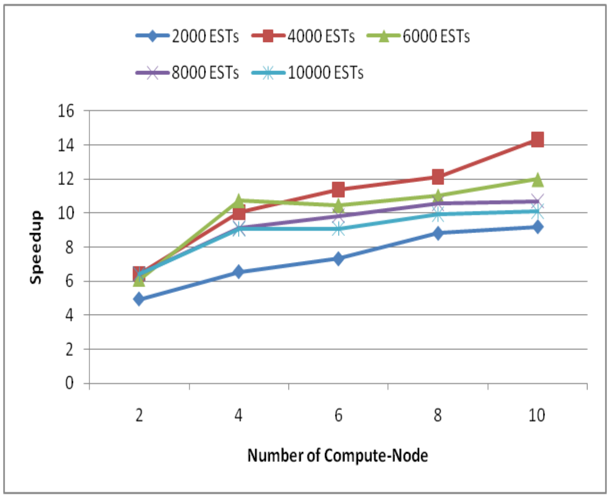


Fig. 12 Speedup of EST annotation against No. of Compute-Node with different EST sizes (Computer server with HPC cluster)

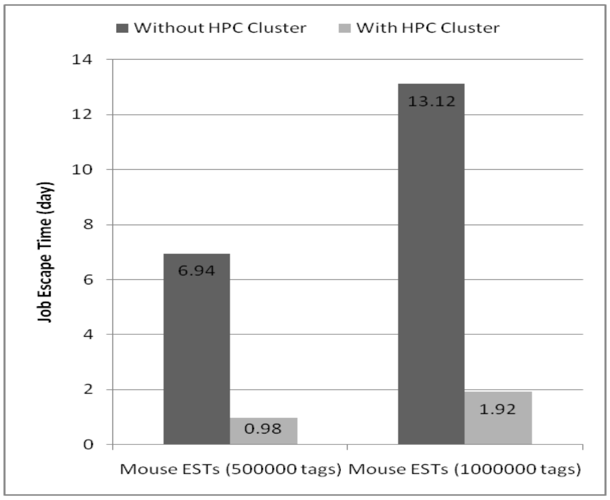


Fig. 13 Job Escape Time of Mouse EST annotations against Mouse Genomic Database

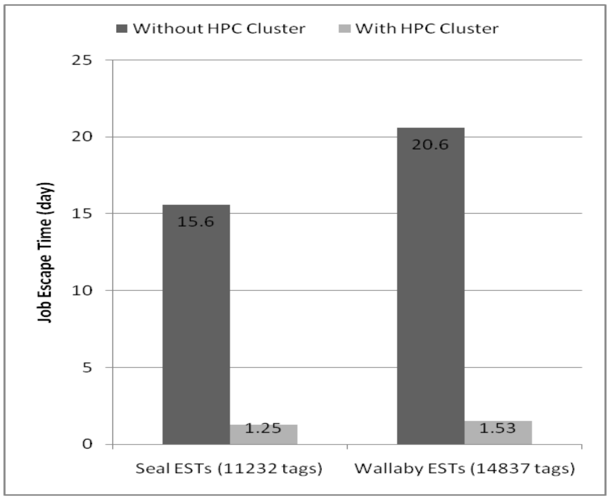


Fig. 14 Job Escape Time of Seal and Wallaby EST annotations against All-organism Genomic Database

Fig. 13 and Fig. 14 present the results obtained from the second experiment set. A promising improvement has been achieved in each of the EST annotations. Job escape time for the annotation of the Wallaby ESTs can be reduced from 20 days to less than 2 days.

V. Conclusions and Future Work

We have extended the EST annotation software package EST-PAC to EST-PAC^{HPC} which can harness HPC resources potentially from various grid and cloud systems for high throughput EST annotations. The performance gain is substantial. The web-portal based approach of EST-PAC^{HPC} can remove the burden of biologists from performing tedious I.T. tasks such as hardware setup, software installation and configuration as well as data management. Even more, it also hides all the details of high performance computing from the users. In conclusion, EST-PAC^{HPC} provides an open framework for rapid prototyping of data mining and on-line visualization of sequence data, presenting an expandable data-management environment for research and education in bioinformatics.

Currently, we are extending the job-scheduling mechanism and the HPC job scheduler of EST-PAC^{HPC} to make it become cloud-enabled. Preliminary work has begun to study Amazons Elastic Compute Cloud (EC2) for HPC [2].

Reference

[1] Adams MD, Kelley JM, Gocayne JD, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science Vol. 252, Issue 5013, pp.1651–1656. Jun 1991.

[2] Amazon. Amazon ec2 high performance computing. http://aws.amazon.com/ec2/hpc-applications/. Last access: April 2011.

[3] Apache HTTP Server Project. http://httpd.apache.org/. Last access: April 2011.

[4] Arun Krishnan. GridBLAST: a Globus-based high-throughput implementation of BLAST in a Grid computing framework: Research Articles. Concurrency and Computation: Practice & Experience archive, Vol. 17, Issue 13, pp. 1607-1623. John Wiley and Sons Ltd. UK. November 2005.

[5] Azure NCBI Blast. http://research.microsoft.com/en-us/projects/azure/azureblast.aspx. Last access: April 2011.

[6] Bioinformatics Research Group, Deakin University. http://mamsap.eit.deakin.edu.au/wiki/index.php/Main_Page. Last access: April 2011.

[7] Eddy SR: Profile hidden Markov models. Bioinformatics. Vol. 14, pp. 755-763. 1998.

[8] Gartner. Gartner highlights five attributes of cloud computing: http://www.gartner.com/it/page.jsp?id=1035013, last access: April 2011.

[9] HUPO-PSI Standard FASTA Format. http://en.wikipedia.org/wiki/FASTA_format. Last access: April 2011.

[10] Iseli C, Jongeneel CV, Bucher P: ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol. pp. 138-148. 1999.

[11] MySQL. http://dev.mysql.com. Last access: April 2011.

[12] National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/. Last access: April 2011.

[13] NCBI Blast. Basic Local Alignment Tool from NCBI: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome, last access: April 2011.

[14] OpenMPI. http://www.open-mpi.org/. Last access: April 2011.

[15] PHP. http://www.php.net. Last access: April 2011.

[16] Portable Batch System. http://www.openpbs.org/. Last access: April 2011.

[17] Shivashankar H. Nagaraj, Robin B.Gasser and Shoba Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics. Vol. 8, No. 1, pp. 6-21. Advance Access Publication. May 23, 2006.

[18] Sun Microsystems Inc. Sun Grid Engine. URL: http://gridengine.sunsource.net/. Last access: April 2011.

[19] Yvan Strahm, David Powell and Christophe Lefèvre. EST-PAC a web package for EST annotation and protein sequence prediction. Source Code for Biology and Medicine 2006.