Developing A Scalable Platform For Next-Generation Sequencing Data Analytics Over Heterogeneous Clouds and HPCs : A Case for Transcriptomes and Metagenomes

Shayan Shams[1,2,3], Nayong Kim[1,2], Ming-Tai Ha[4], Shantenu Jha[4], Jian Tao[1], Ramesh Subramanian[2], Vladmir Chouljenko[2], K. Gus Kousoulas[2], Ram J. Ramanujam[1], Seung-Jong Park[1,3,*], Joohyun Kim[1,2,*]

1. Center for Computation and Technology, LSU, 2. Center for Experimental Infectious Disease Research, 3. Division of Computer Science & Eng., LSU, 4. Electrical and Computer Engineering Busch Campus, Rutgers University

Abstract

We present **a distributed computing platform** with which significant roadblocks in Next-Generation Sequencing (NGS) data analytics, associated with ever-growing and noisy data sets, can be effectively resolved. Merits and advantages of the platform are exemplified with the specific application of **de novo genome sequence reconstruction**, a fundamentally critical task for studies of transcriptomes and metagenomes.

The core underlying framework is built upon the pilot system, Radical Pilot, allowing a developer to rapidly implement pipelines in which accessing and utilizing **heterogeneous distributed computing environments** are natively supported. This aspect is further strengthened with the **distributed application runtime environment** that aims effective management of massive distributed workloads and data processing tasks over heterogeneous resources. Consequently, the platform becomes an efficient distributed parallel task management system leveraging **high-end HPC technologies as well as emerging Hadoop-based software models**. The DOCKER container technology is also available as options, implying all together increasing the repertoire for flexible and scalable runtime environment scenarios.

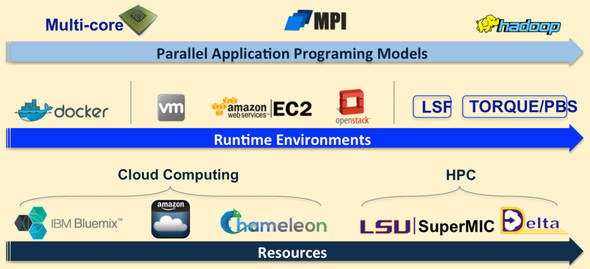
Based on the platform, a novel pipeline has been developed for metagenome and transcriptome research. The pipeline is capable of the novel de novo genome sequence reconstruction with **Multiple Assembler Multiple Parameter (MAMP)**.

Summary

- A pipeline tool for transcriptome profiling tasks and metagenome assembly can
  - Process massive data volumes thanks to distributed computing models.
  - Increase the accuracy of analytics against the nature of noisy data by developing a novel method with MAMP, a kind of ensemble learning methods.
- Distributed Application Runtime Environment (DARE) is a set of programs built upon Radical Pilot and open source bioinformatics tools for NGS data analytics, primarily for effectively executing large scale workloads and data processing.
- Heterogeneous computing resources such as our local HPCs (SuperMic (Intel cluster) and DELTA (IBM Power8 cluster)) and cloud systems (Amazon EC2, OpenStack-based Chameleon, and IBM Bluemix) are successfully integrated.
- Multiple common local runtime environments such as those provided in HPCs with local schedulers (PBS/TORQUE and LSF), virtualization environment (EC2 and OpenStack), and the DOCKER container technology are seamlessly utilized for execution scenarios for the pipeline

I. Distributed Platform Architecture

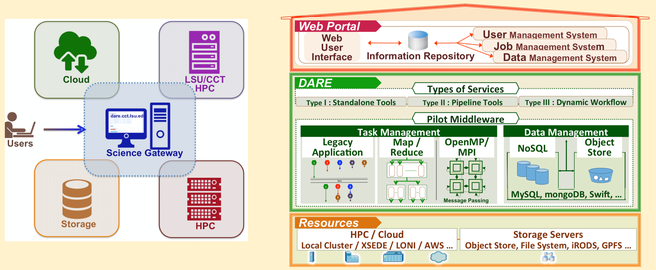
A. Hierarchical structure of heterogeneous distributed resources, their local runtime environments, and common programming models of NGS applications



LSU Center for Computation & Technology

SC16

B. Distributed Application Runtime Environment

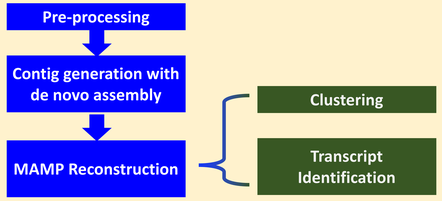


II. Genome Sequence Reconstruction

Question : Challenges in de novo transcript and metagenome assembly with NGS Data

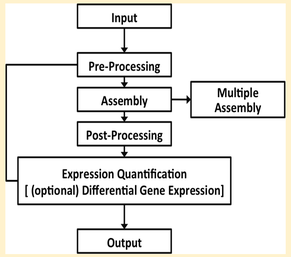
- Genome reconstruction without a reference sequence is fundamental but challenging.
- NGS raw data are intrinsically noisy, so analysis are error-prone and difficult to be evaluated.
- Data sets for NGS are often larger than the capacity of a single node solution and will grow in a rapid pace.

Solution : Multiple Assembler Multiple Parameter (MAMP) facilitated by distributed computing techniques



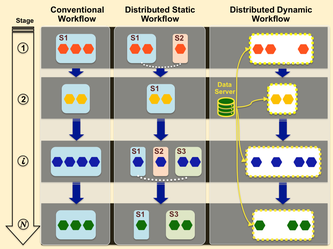
III. Scalable Transcriptome and Metagenome Pipeline

Overall Pipeline Workflow



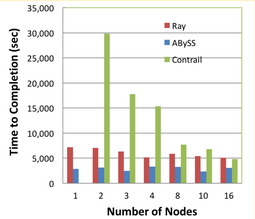
Schematic of the pipeline for transcriptome/metagenome analysis

Distributed Application Workflows



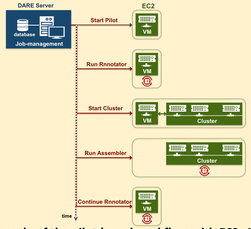
The pilot system (Radical Pilot) for efficient task-level and data-level parallel executions of sub tasks with distributed resources. Three different patterns on how these sub tasks belonging to each stage of a pipeline are managed

Scale-out Performance of Distributed Application



Large-scale data set can be processed with distributed parallel application models such as MPI (Ray and ABySS) and Hadoop (Contrail). The full support of these programming models, regardless of different local runtime environment (HPC/PBS, HPC/LSF, EC2, OpenStack, Bluemix, and DOCKER), is the key for our platform.

Heterogeneous Distributed Resources



An example of the pilot-based workflow with EC2. A wide range of options for scale-out and scale-across scenarios with multiple distributed resources and corresponding local runtime environments are supported in a systematic manner

MAMP-based De Novo Genome Sequence Reconstruction Method

- Contig generation using multiple de novo assemblers (Ray, ABySS, Contrail, etc) along with multiple k-mers
- Main strategy for merging those contigs with two steps, graph-based clustering and consensus-based merging. This scheme allows also various flexible optimization strategies for a broad range of related problems.
- Clustering is implemented with the weighted edge graph (DIME)
- Merging can be done with VMATCH/minimus2, CAP3, and DIME
- Promising method for controlling false negatives and false positives, and notably computationally demanding due to massive heterogeneous tasks required

MAMP Performance

Method Category	Assembler used (Clustering and Merging method)	Precision, Recall, F_1 (Nucleotide-level)	Weighted k-mer recall, kc score	RSEM-eval (higher is better)
Standalone	Trinity	0.51 (7), 0.35 (6), 0.42 (6)	0.84, 0.83 (2)	-236,024,028.9(6)
Single	Ray (V-m2 : VMATCH/minimus2)	0.84 (1), 0.26 (7), 0.40 (7)	0.86, 0.86 (1)	-239,866,658.0 (7)
Assembler	ABySS (V-m2)	0.82 (2), 0.42 (5), 0.55 (5)	0.79, 0.78 (4)	-225,154,884.4 (4)
Multiple K-mer	Contrail (V-m2)	0.78 (5), 0.43 (3), 0.56 (2)	0.84, 0.83 (2)	-222,219,796.1 (1)
Multiple	Ray + Contrail (V-m2)	0.78 (5), 0.43 (3), 0.56 (2)	0.78, 0.77 (5)	-225,887,718.1 (5)
Assembler	Ray+Contrail+ABySS (V-m2)	0.79 (3), 0.44 (1), 0.57 (1)	0.77, 0.76 (6)	-224,433,916.1 (3)
Multiple K-mer	Ray+Contrail+ABySS (Cap3)	0.79 (3), 0.44 (1), 0.56 (2)	0.77, 0.76 (6)	-224,236,495.7 (2)
	Ray+Contrail+ABySS (MAMP) 1. Before Clustering 2. Clustering/CAP3 3. Clustering/V-m2	 N/A, 0.47, N/A 0.79, 0.44, 0.56 0.80, 0.40, 0.52	 0.93, 0.92 0.78,0.78 0.79,0.78	 -223,943,006.9 -229,153,749.0

Conclusions

- Our distributed platform over heterogeneous HPCs and Clouds is an efficient solution for developing tools that are capable of overcoming challenges in NGS data analytics.
- New pipeline for transcriptome and metagenome, along with a novel approach for de novo sequence reconstruction with MAMP, underscores benefits of the platform

References

1. Maddineni, Sharath, et al. "Distributed application runtime environment (DARE): a standards-based middleware framework for science-gateways."*Journal of Grid Computing* 10.4 (2012): 647-664.
2. Shams, S., Kim, N., Meng, X., Ha, M. T., Jha, S., Wang, Z., & Kim, J. A Scalable Pipeline For Transcriptome Profiling Tasks With On-demand Computing Clouds. *HiCOMB2016, IEEE IPDPS* (2016)
3. Ragothaman A, Boddu SC, Kim N, Feinstein W, Brylinski M, Jha S, Kim J. Developing ethread pipeline using saga-pilot abstraction for large-scale structural bioinformatics. BioMed research international. 2014 Jun 9;2014.

Acknowledgement

We are grateful for collaborations and constructive discussions from Zhong Wang (JGI), and Xuan Guo (UTK). This work is supported in part by the funding from NIH P20 GM103458-10. We also thank the partial support from LBRN.