ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 2, August 2012

# All Pairs Snag for Genome Sequence Unearthing Using Cloud

L.Muthuvel, S.M.Krishna Ganesh
Department of Computer Science and Engineering
St.Joseph College of Engineering and Technlogy, Chennai

***Abstract*—In Cloud computing era, which emphasis us to go for utility computing to use the resources available in a public network on demand and pay for only use of the resources and own nothing. In this paper we have planned to build an application over Cloud to solve all pairs problem by applying bio-informatics, genome sequence detection. With this experiment we decided to make a study on Genome sequence Detection using ordered sequence detections.***

***Index Terms*—Cloud Computing, WGS (Whole Genome Sequence), GSD (Genome Sequence Detection), Windows Azure.**

## I. INTRODUCTION

In recent 10 years, Internet has been developing very quickly. The cost of storage, the power consumed by computer and hardware is increasing. Cloud computing, a new kind of computing model, is coming. This word is a new word that appears at the fourth season, 2007[1]. It is an extend of changing with the need, that is to say the manufacturer provide relevant hardware, software and service according to the need that users put forward. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud. Cloud Computing distribute computation task on the resource pool which consists of massive computers, accordingly, the application systems can gain the computation strength, the storage space and software service according to its demand. This specified cloud computing environment can be created in many ways. In that many ways' one is windows azure. Windows azure cloud development platform. By using the windows azure we can develop web fabrics and storage. The sequencing and assembly process is done as shown in the figure below.

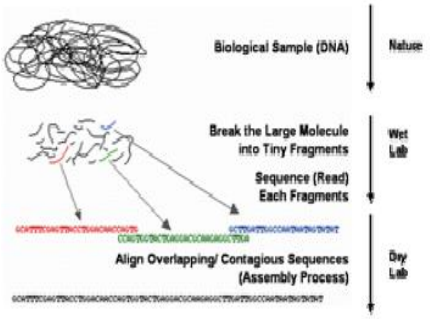


**Fig 1: The Sequencing and Assembling Process**

## II. RELATED WORK

All pairs problem were addressed and solved by map reduce method and hadoop. Hadoop is an open source Apache project written in Java and designed to provide users with two things: a distributed file system (HDFS) and a method for distributed computation. It's based on Google's published Google File System and Map Reduce concept which discuss how to build a framework capable of executing intensive computations across tons of computers. [6]The above mentioned techniques have been implemented in "Cloud-MAQ: The Cloud-enabled Scalable Whole Genome Reference Assembly Application" this give motivation for me to proceed further with genome sequence ordered detection method.

The All-Pairs problem is easily stated as[9]" All-Pairs (set A, set B, function F ) returns matrix M:Compare all elements of set A to all elements of set BVia function F, yielding matrix M", such that M [i,j] = F(A[i],B[j]).Variations of the All-Pairs problem occur in many branches of science and engineering, where the goal is either to understand the behavior of a newly created function F on sets A and B, or to learn the covariance of sets A and B on a standard inner product F. We are working with two user communities that make use of All-Pairs computations in Bio Informatics.

## III. WINDOWS AZURE

Windows Azure is Microsoft's operating system for the cloud. It lets your applications scale up or down depending on the demands of your business. Free your developers to flex their creative muscles on a platform that already speaks their languages as like (DOT).Net, PHP, Java or Ruby.[2] Wherever their creativity takes them, in the language they choose they have the power. Plus with a pay-for-use business model, you won't be wasting your money on services you thought you might need but never got around to using. Now that's a win-win situation. That's cloud power.

The Windows Azure platform is a flexible cloud–computing platform that lets you focus on solving business problems and addressing customer needs. No need to invest upfront on expensive infrastructure. Pay only for what you use, scale up when you need capacity and pull it back when you don't. It handles all the patches and maintenance — all in a secure environment with over 99.9% uptime [4]. An application that is designed to be a hosted service in Windows Azure consists of discrete scalable components built with managed code and XML configuration files that define how the service should run.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 2, August 2012

Windows Azure currently supports the following types of roles:

### A. *Web Role*

A web role is a role that is customized for web application programming as supported by IIS 7 and ASP.NET. The benefit of using this type of role is that the IIS setup is done for you. This role is best used for providing a web based front-end for your hosted service [3]. It is not suited for long running processes.

### B. *Worker Role*

A worker role is a role that is useful for generalized development, and may perform background processing for a web role. When you have a need for a background process that performs long running or intermittent tasks, you should use this role

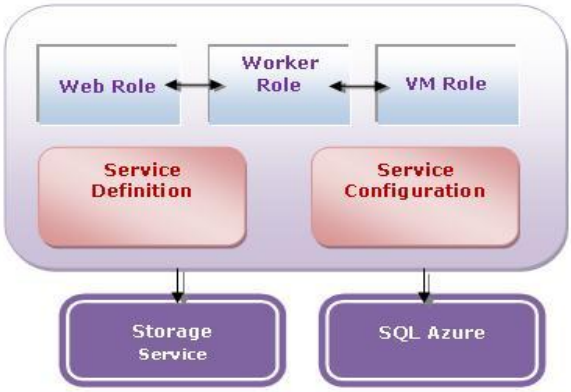


**Fig. 2. Components of a Windows Azure environment**

### C. *VM role*

A VM role is a special type of role that enables you to define the configuration and updates of the operating system for the virtual machine [2].While a web role and a worker role run in a virtual machine, the VM role is the virtual machine, which gives you full control of operations. When you have long and complicated installations in the operating system or special setup issues, you should use this role. This role is especially suited for migrating existing applications to run as hosted services in Windows Azure.

## IV. DATA SET USED

The data set used sample as Escherichia coli, Oryza sativa japonica and the following figure shows Model of successive binary fission in E. coli.

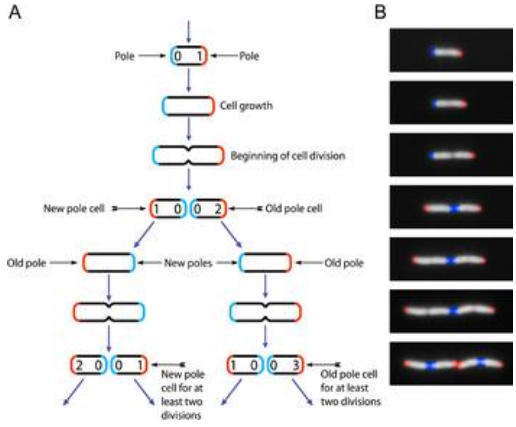


**Fig 3. Block Diagram**

## V. IMPLEMENTATION

Whole genome sequencing, complete genome sequencing, or entire genome sequencing, is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and for plants the chloroplast as well. Almost any biological samples even a very small amount of DNA or ancient DNA can provide the genetic material necessary for full genome sequencing. Such samples may include saliva, epithelial cells, bone marrow, hair (as long as the hair contains a hair follicle), seeds, plant leaves, or anything else that has DNA-containing cells. Because the sequence data that is produced can be quite large (for example, there are approximately six billion base pairs in each human diploid genome), genomic data is stored electronically and requires a large amount of computing power and storage capacity. Full genome sequencing would have been nearly impossible before the advent of the microprocessor, computers, and the Information Age.

From the WGS we are extracting the sequence with reference to order such as two order, three order and four order respectively. [7]This Detection is done with reference of Kaplan's Algorithm which specifies the conditions of Detection. On executing the application we have a login screen, which authenticate the cloud user to make use of the apps. On the screen using the UI we need to give the WGS as input from .txt file. and select the order of Detection as two order or three order or four order. As of user selection UI will show the nucleotide options as AG,AC,AT… for two order, AAG,AAC,AAT,… for three order and AAAG,AAAC,AAAT, for four order. Then we get the output screen as Two column of UI with WGS and GSE with its count of nucleotide. Then we do compare of WGS and GSE with its count

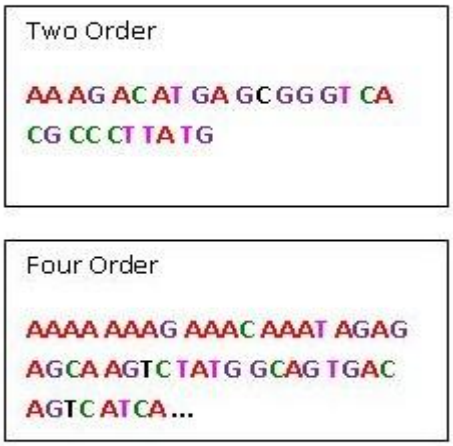


**Fig. 4. Sequential Order of Genome sequence Detection [7].**

The actual process is the nucleotide of order AG will look of its appearance on WGS and as soon as found the sequence it will start to extract the sequence till it's reverse order is found as GA. The scheduled process is done till the end of file.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 2, August 2012

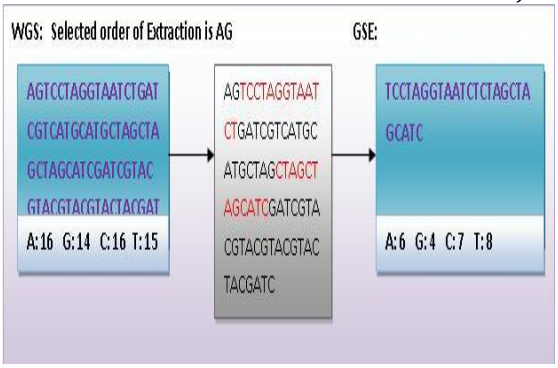


**Fig. 5. Genome Sequence Detection from WGS**

## VI. RESULTS

Genome Sequence Detection is implemented and tested with some sample genome sequences from NCBI and the results are as follows

### *A. Escherichia Coli*

The Pie graph represents Escherichia coli nucleotides combination and composition of AGCT in Whole genome sequence.

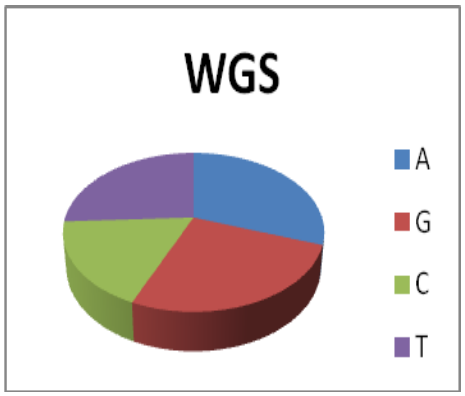


**Fig. 6.Pie Chart Result of Escherichia Coli**

The comparison graph of WGS and GSD is below which indicates the difference between the combination of Nucleotides present in Escherichia coli.[8] the results are likely to be in histogram order.

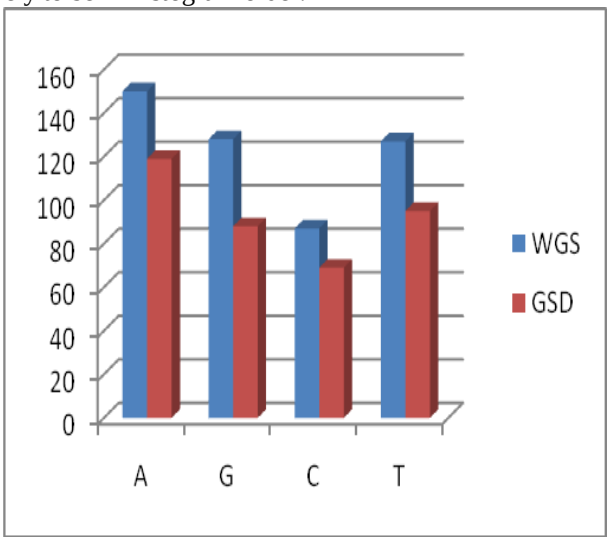


**Fig. 7.Comparison Result Of Escherichia Coli WGS and GSD**

### *B. Microbacteriaceae Bacterium-*

The Pie graph represents Microbacteriaceae bacterium nucleotides combination and composition of AGCT in Whole genome sequence

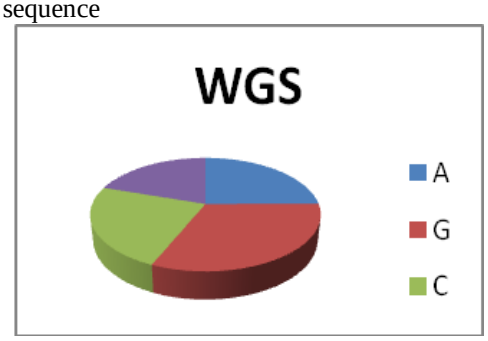


**Fig. 8.Pie Chart Result of Microbacteriaceae Bacterium**

The comparison graph of WGS and GSD is below which indicates the difference between the combinations of Nucleotides present in Microbacteriaceae bacterium.

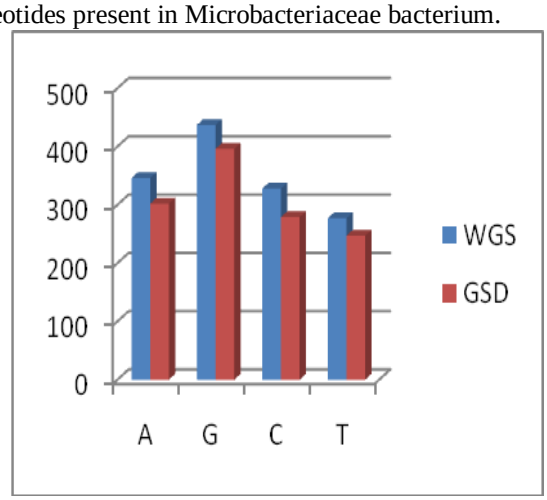


**Fig. 9.Comparison Result of Microbacteriaceae Bacterium WGS and GSD**

### *C. Oryza Sativa Japonica*

The Pie graph represents Oryza sativa japonica nucleotides combination and composition of AGCT in Whole genome sequence

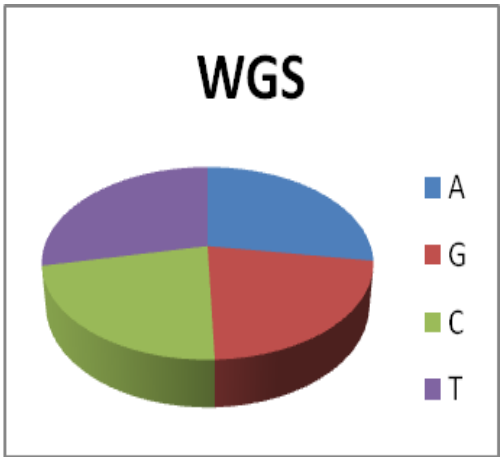


**Fig. 10.Pie Chart Result of Oryza Sativa Japonica**

The comparison graph of WGS and GSD is below which indicates the difference between the combinations of Nucleotides present in Oryza sativa japonica.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 2, August 2012

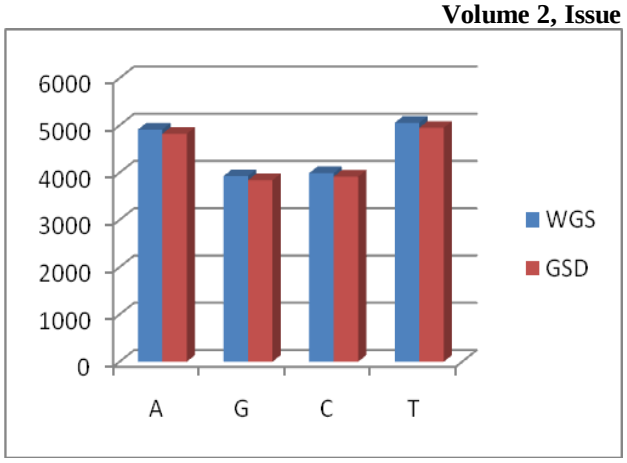


**Fig. 11 .Comparison Result of Oryza Sativa Japonica WGS and GSD**

### *D. Thermoanaerobacterium*

The Pie graph represents Thermoanaerobacterium nucleotides combination and composition of AGCT in Whole genome sequence

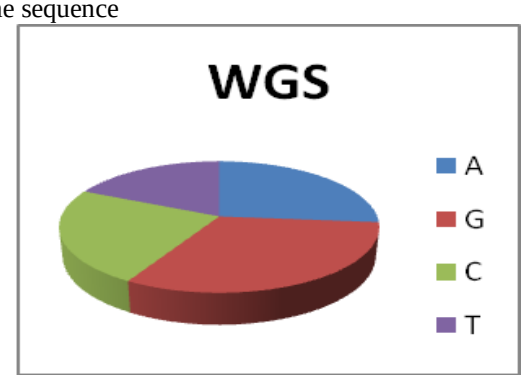


**Fig. 12.Pie Chart Result of Thermoanaerobacterium**

The comparison graph of WGS and GSD is below which indicates the difference between the combinations of Nucleotides present in Thermoanaerobacterium

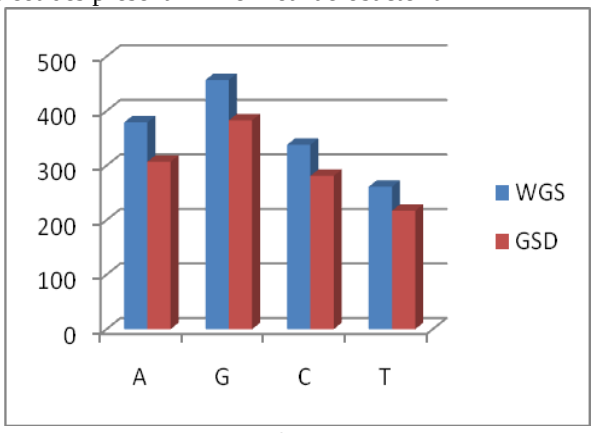


**Fig. 13.Comparison Result of Thermoanaerobacterium WGS and GSD**

## VII. CONCLUSION

This application is developed under windows azure platform with Microsoft visual studio 2010 free version. This application is used only for genome sequence detection and in future we can develop application to find the repetition of genome sequence in ordered sequences and hence the bioinformatics applications can be developed in large number to study and learn the properties of genome sequences which can be made available updated over internet.

## REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica,M. Zaharia. Above The Clouds: A Berkeley View of Cloud Computing. Technical Report No. Ucb/Eecs-2009-28, University Of California at Berkley, USA, Feb. 10, 2009.

[2] D. Chappell. Introducing the Azure Services Platform. White Paper, Oct. 2010.

[3] Microsoft, 'Transforming It with the Windows Azure Platform', White Paper, Nov. 2010.

[4] Http://Msdn.Microsoft.Com/En-Us/Library/Gg432976.Aspx.

[5] Http://Www.Microsoft.Com/En-Us/Cloud/Cloudpowersolutions/Development-And-Hosting.Aspx#Tab3-Tabs.

[6] Asoke K Talukder Et Al"Cloud-Maq: The Cloud-Enabled Scalable Whole Genome Reference Assembly Application"978-1-4244-7202-4/10/$26.00 ©2010 IEEE.

[7] Zimao Li, Lusheng Wang, Member, IEEE, and Kaizhong Zhang, Member, IEEE"Algorithmic Approaches For Genome Rearrangement: A Review"IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 36, No. 5, September 2006.

[8] Anca Bucur, Jasper Van Leeuwen, Nevenka Dimitrova, and Chetan Mittal"Alignment Method for Spectrograms of DNA Sequences"Ieee Transactions On Information Technology In Biomedicine, Vol. 14, No. 1, January 2010.

[9] Christopher Moretti, Student Member, IEEE, Hoang Bui, Student Member, IEEE, Karen Hollingsworth, Brandon Rich, Patrick Flynn, Senior Member, IEEE, And Douglas Thain, Member, IEEE" All-Pairs: An Abstraction For Data-Intensive Computing On Campus Grids" IEEE Transactions On Parallel And Distributed Systems, Vol. 21, No. 1, January 2010.

## AUTHOR BIOGRAPHY

**Muthuvel Laxmikanthan** has finished his Bachelor of Computer applications at N.M.S.S Vellaichamy nadar college, Madurai Kamaraj university,Madurai,India, Master of Computer Applications at Arulmigu Kalasalingam college of engineering, Affiliated to Anna University, Master of Technology in computer science and engineering from Kalasalingam UniversityVirdhunagar, Tamilnadu, India, And also Master of Business Administration in Human Resources at Madurai Kamaraj university He has worked as a Software Engineer at INDUS TEQSITE,Chennai. Currently he is pursuing his Master of Science in Psychology at Barathyiar University, Coimbatore, India. The Interest in Research area is Cloud Computing, Data mining, Life span Psychology, and Forensic Psychology. And so as various fields. Working as assistant Professor in an Engineering College.