Sindh Univ. Res. Jour. (Sci. Ser.) Vol.44 (3) 441-444 (2012)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Analysis With Special Reference To Genomic Sequence On Amazon Cloud

F. N. MEMON, Z.U.A. KHUHRO, M.U.R MAREE, I.A. KOREJO, A. P. HARRISON*

Institute of Mathematics and Computer Science, University of Sindh, Jamshoro

Received 09th March 2012 and Revised 07th July 2012

Abstract: Bioinformatics is an important field that mainly helps to interpret biological data using computing tools with the help of other fields such as statistics, mathematics, and chemistry. Due to the generation of enormous amount of biological data, bioinformaticians require high power computing resources. Cloud computing provides on-demand computing facilities that can expand and/or shrink as required. Cloud computing is attracting bioinformaticians to analyze vast amount of biological data currently available and it is expected that it will be the popular choice of bioinformaticians in near future. This paper presents the application of cloud computing in bioinformatics with special reference to genomic sequence.

Keywords: Bioinformatics, Cloud computing, Amazon's Cloud, Genomic sequence

1. INTRODUCTION

Large scale scientific research projects, such as Human Genome Project (Sawicki, 1993) or others, in various research laboratories are generating an ever increasing amount of biological data. Hence the biologists need to organize and interpret such enormous amount of data other than the manual way.

Bioinformatics provides the application of computer science in the field of biology with the support of statistics, mathematics and chemistry. Databases, artificial intelligence, data mining, simulation, algorithms and web technologies are considered as some of the important areas of computer science that play a critical role in the field of Bioinformatics.

Bioinformatics helps to organize huge amounts of biological data into databases, such as Ensembl (Hubbard , 2002) and GenBank (Benson, 1997), in order to access and to analyze them efficiently. (Parker, 2010) demonstrated the growth of genomic data provided by Ensembl over the last decade. As the biological data is growing towards Petabyte scale (Bateman, 2009), high-power computing and storage is required accordingly. Cloud computing fulfills the needs of such computing facilities and large scale storage.

Cloud computing can simply be think as on-demand computing where computing facilities are sold by suppliers as any other commodities such as electricity (Memon, 2010).

1.1 Cloud Computing for Bioinformaticians

Bioinformaticians usually deal with large data sets for data mining and analysis. Thus, they require high-powered computing and storage resources. These high-powered computing resources are not necessarily required for all the time but for a short period of time to perform some tasks. Cloud computing helps bioinformaticians to avoid purchase of high power computers that are decreasing in value by the time (Memon, 2010) and are not being used all the time. Cloud computing enable its users to use computing resources without maintaining their own clusters or networks. Thus, they don't need to buy the high-powered resources, but a simple and inexpensive terminal including input and output devices along with the computing power that is necessary to connect to the cloud. It is expected that cloud computing will be a popular choice among the bioinformaticians in the near future.

Ensembl and NCBI (Sayers, 2011) now provide some of the large publicly available data sets on Amazon cloud (http://aws.amazon.com/publicdatasets/). It will help many users to access these public data sets freely without worrying to locate, download and manage such data.

Huge amounts of genomic and related data are being produced every day that will lead to a new era in petabyte scale data (Bateman, 2009). Ensembl is one of the examples of genomic data provider for more than a decade and **(Fig. 1)** is showing that their data is growing rapidly during this period (Parker, 2010). The bioinformaticians require an ever-increasing amount of

++Corresponding author, Farhat Naureen Memon, farhatnaureen@gmail.com
*Departments of Mathematical Sciences and Biological Sciences, University of Essex, UK

computational resources in order to perform mining of large biological data sets. It is therefore expected that they will increasingly adopt the cloud computing solutions. It is also possible that in the future the cloud will become a free worldwide resource for academic purposes.

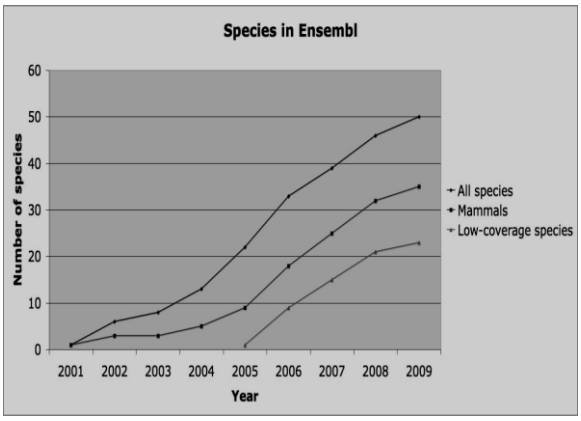

Fig.1: Ensembl is providing the genomic data of different species. The figure is showing the increase in number of species whose genomic data is available on Ensembl website since 2001. The figure is adopted from (Parker, 2010).

2. MATERIAL AND METHOD

A sequence of nucleotides having frequent occurrences of runs of guanines is capable of forming unusual four-stranded structures called G-quadruplex structures (Gellert, 1962). Further details of G-quadruplex structures are available else where (Memon, 2011). The G-quadruplex structures may be involved in various biological processes (Darnell, 2001; Zahler, 1991 and Siddiqui-Jain, 2002). It has demonstrated that quadruplex structures can play a role in gene expression and provide opportunities for a new class of anticancer therapeutics and drug targets (Neidle, 2006; Qin, 2008; Yadav, 2008). Qin, 2008) mentioned that the identification of such sequences helps identifying the biological role of G-quadruplex structures.

Since Ensembl provides genomic sequences and its data is publicly available on cloud, a cloud is explored to analyze Ensembl genomic sequences of Homo_Sapiens (Human).

2.1 Platform selection of Cloud

Among the different providers of cloud computing, a cloud has been selected to use through Amazon's platform. The users of Amazon cloud are not required for any long term commitment and they will pay only for the resources they use. The main reason for selecting Amazon cloud among the other platforms is that Ensembl genomic data set is available as a public data set at Amazon's cloud which is freely available to the general public.

Although Amazon provides the flexibility to work with Windows and Linux environment, the later environment is used during this analysis. Among the various services provided by the Amazon Web Services (AWS), Amazon EC2 and Amazon S3 have been explored along with one of the Amazon public data sets.

- *Amazon Simple Storage Service (S3):* Users use this service to store their own data. This data can be kept private for their own use or public to share with other users. S3 provides a highly scalable solution to store and retrieve data at high speed and low cost. The data is stored in the form of objects that are stored in buckets. Each bucket is identified by a unique key.

- *The Amazon Elastic Compute Cloud (EC2):* EC2 provides an environment to run virtual servers on demand. EC2 enables the users to get high computing power in the cloud that can be resized any time. The users of EC2 have the complete control over the computing resources.

An Amazon Machine Image (AMI) is required to use Amazon EC2 service. An AMI is an encrypted machine image (a file) that consists of all the information which required for launching an instance of user's software and it is stored in Amazon Simple Storage Service (S3). Users can either create a private or a public AMI. A public AMI can then be used by anyone in its original form or with some modification.

Users can launch one or more than one instances for an AMI and can administer these instances as they do on their own server. Further documentation is available at http://aws.amazon.com/documentation/.

Amazon provides different instance types according to the need of computing resources required. Throughout this study, a public AMI (ami-b55dbbdc) which is based on some bioinformatics tools along with Ubuntu Linux 8.04 and a small instance (a default instance type) have been used. A small instance consists of 1.7 GB memory and one EC2 compute unit (1 virtual core with 1 EC2 compute unit). Further details on instance types and their specifications are available at http://aws.amazon.com/ec2/instance-types/.

Once an instance of the AMI has been launched, data and computing tool to analyze that data are uploaded on the cloud. Since the Ensembl's genomic sequences (DNA sequences) of various species are already available at Amazon's cloud, only a computer

program that finds the possible G-quadruplex forming sequences is uploaded to the cloud.

To analyze these sequences on local machines, it is required to first download them from Ensembl website to a local machine. Being a large data set, the downloading process takes some time. Since the Ensembl data set is already available at Amazon cloud, the use of this public data set reduces time for downloading and/or uploading. So the advantage of using this public data set is taken.

To use the public data set, a user has to create its volume that is simply a copy of that data at the user's virtual machine. By creating a volume of a public data set, the original data is safe and could not be destroyed by mistake.

3. RESULTS AND DISCUSSION

The Homo_Sapiens (Human) genomic data provided by Ensembl is analyzed to find possible G-quadruplex forming sequences (PGQFS). **(Fig. 2)** shows the processing time to find the possible G-quadruplex forming sequences in each chromosome of Homo_Sapiens (Human).

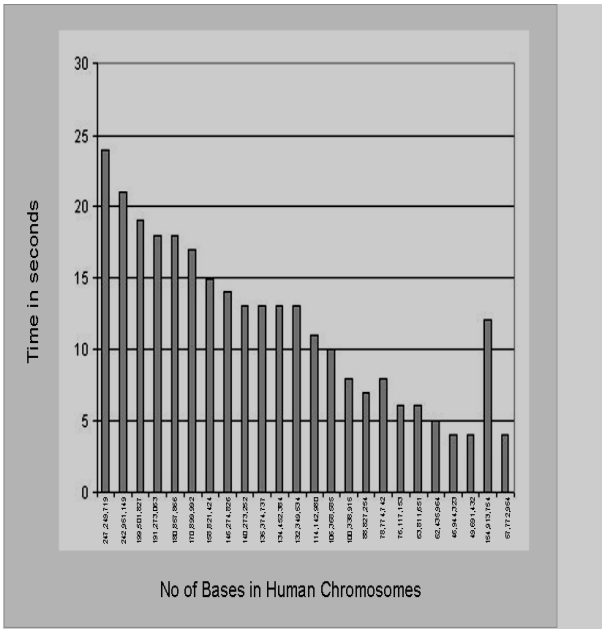

Fig 2: A graphical view to illustrate the processing time to find possible G-quadruplex forming sequences in each chromosome of Homo_Sapiens (Human) on Amazon cloud using small machine.

(Table 1) presents more details that include chromosome name/number, the size of each chromosome in bases, the total number of G-quadruplex forming sequences in each chromosome and the processing time.

Table 1: The chromosome name/number, number of possible G-quadruplex forming sequences (PGQFS), size of chromosomes (in bases), and processing time to find PGQFS in each chromosome of human are listed.

Chromosome	No. of Bases in chromosome	No. of possible G-quadruplex forming sequences	Time (in sec.)
1	247,249,719	175,131	24
2	242,951,149	149,051	21
3	199,501,827	105,700	19
4	191,273,063	86,836	18
5	180,857,866	99,071	18
6	170,899,992	91,671	17
7	158,821,424	108,987	15
8	146,274,826	83,421	14
9	140,273,252	91,570	13
10	135,374,737	95,383	13
11	134,452,384	99,129	13
12	132,349,534	98,393	13
13	114,142,980	46,857	11
14	106,368,585	62,863	10
15	100,338,915	59,306	08
16	88,827,254	92,806	07
17	78,774,742	93,204	08
18	76,117,153	38,873	06
19	63,811,651	108,865	06
20	62,435,964	62,214	05
21	46,944,323	28,044	04
22	49,691,432	49,044	04
X	154,913,754	115,767	12
Y	57,772,954	9,150	04

The entire human genome is analyzed in less than 5 minutes. However, for even larger data sets, the performance can be improved by using another instance type. Using public data set is found quite comfortable and it saves a user to locate and download the data. The original data is also safe from damages due to the fact that a user works on a copy of the original data.

It is expected that Ensembl and others will continue their practice to provide their data as public data sets on Amazon or other platforms. It will particularly help bioinformaticians with less computing resources to analyze large biological data without purchasing high power computing facilities.

Security is one of the serious issues of cloud computing. The users of cloud are cautious about protection of their data especially on a public cloud. A slowed down or lost of connection may also cause the users to lose or damage their data. Thus, it is important to regularly maintain backups of public data.

Although there are certain risks of using cloud computing, the cloud computing has many benefits for their users that make it successful. For instance, a user can expand or stretch a service at any time as much as required. Users can avail the cloud computing at anywhere as long as a computer and an internet connection are available. It also reduces the need of IT personnel who can keep the software up to date. Users are also free from the hardware maintenance such as constant server updates. The providers are responsible for the maintenance.

4. CONCLUSIONS

Bioinformatics deals with biological data using computing tools along with the help of statistics, mathematics and chemistry. The amount of biological data is increasing day by day towards PetaByte scale. Thus, Bioinformaticians require high power computing resources to interpret such data. Instead of maintaining individual clusters/network, bioinformaticians can use cloud computing which provides computing resources as a commodity and these resources can be expanded and shrunk according to the requirement of the job.

Although there are certain risks of cloud computing, it is becoming popular due to its benefits. Individuals or user groups of any size can use the cloud computing. It is expected that cloud computing will greatly be adopted by bioinformaticians as well as academicians.

Amazon provides some of the biological data sets of Ensembl, NCBI and others that are known as Amazon Web Services (AWS) public data sets and are freely available to the public. Ensembl genomic data of Homo_Sapiens (human) is analyzed at Amazon's platform to find possible G-quadruplex forming sequences. The entire human DNA sequence was analyzed in less than 5 minutes on a small instance type. This performance can be improved when a better instance type is used.

REFERENCES:

Bateman, A. and M. Wood (2009) Cloud computing, Bioinformatics, **(25):** 1475-1475.

Benson, D.A., M.S. Boguski, D.J. Lipman and J. Ostell (1997) GenBank, Nucleic Acids Research, (**25):** 1, 1-6.

Darnell, J., K. Jensen, P. Jin, V. Brown, S. Warren and R. Darnell (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function, Cell, **(107):** 4, 489-499.

Gellert, M., M. Lipsett and D. Davies (1962) Helix formation by guanylic acid, Proceedings of the National Academy of Sciences of the United States of America, **(48):** 12, 2013.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down and others (2002) The Ensembl genome database project, Nucleic acids research, **(30)** 1, 38-41.

Memon, F.N., A.M. Owen, O. Sanchez-Graillet, G.J.G. Upton and A.P. Harrison (2010) Identifying the impact of G-Quadruplexes on Affymetrix 3' Arrays using Cloud Computing, Journal of Integrative Bioinformatics, **(7):** 2, 111-112.

Memon F.N. (2012) Study of the effects caused by the G-Quadruplex Structures on high-throughput nucleic acid measurements. A Ph.D. thesis submitted at the University of Essex, UK.

Neidle, S. and S. Balasubramanian, (2006). Quadruplex nucleic acids. Royal Society of Chemistry.

Parker, A., E. Bragin, S. Brent, B. Pritchard, J. Smith and S. Trevanion (2010) Using caching and optimization techniques to improve performance of the Ensembl website, BMC bioinformatics, **(11):** 1, 239.

Qin, Y. and L. H. Hurley, (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions, Biochimie, **(90):** 8, 1149-1171.

Sawicki, M.P., G. Samara, M. Hurwitz and E. Passaro Jr (1993) Human genome project, The American journal of surgery, **(165)** 2, 258-264.

Sayers, E.W., T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen and others (2011) Database resources of the national center for biotechnology information, Nucleic Acids Research, **(39):** (suppl 1), D38-D51.

Siddiqui-Jain, A., C.L. Grand, D.J. Bearss, and L.H. Hurley, (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription, Proceedings of the National Academy of Sciences of the United States of America, **(99):** 18, 11593Pp.

Yadav, V., J. Abraham, P. Mani, R. Kulshrestha, and S. Chowdhury, (2008) QuadBase: genome-wide database of G4 DNA occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes, Nucleic Acids Research, **(36):** 1, D381-D385.

Zahler, A., J. Williamson, T. Cech and D. Prescott, (1991) Inhibition of telomerase by g-quartet dma structures, Nature, **(350):**718-720.