Project Title : Constructing index structures of biological sequence using Hadoop MapReduce.

Team Member : Rishi Pathak
UFID: 1926-9281
(rishipathak@ufl.edu)

Abstract : DNA sequence alignment is one of the most important application in Bioinformatics for identifying sequence similarity,developing homology model of protein structures,analysing gene expression, that may be a consequence of functional,structural or evolutionary relationship between the sequences.In order to accelerate the alignment for large sequences, sequence alignment task rely on precomputed indexes of sequence.Two of the most important index structures of biological sequence are **suffix array(SA)** and **Burrows Wheeler Transform(BWT)**. Serial computation of these structures for large sequences i.e.,human genome,takes a lot of time.I'm planning to construct these structures using an efficient framework called **Hadoop** in cloud using **MapReduce** programming model.

Plan of action :

1. Setup Hadoop framework.
2. Implementing the MapReduce parallel algorithm for constructing suffix array and BWT.
3. Comparing the end to end runtime by running on different cores,using machine leased from Amazon Elastic Compute Cloud(EC2),as a student member.
4. For evaluation,I will construct the suffix array of important genomes ranging in size from few millions to billion nucleotides.I will download the data from NCBI (National Center for Biotechnology Information)

List of research papers :

1. A View of Cloud Computing, Communication of ACM, 2010.
   by Michael Armbrust, Armando Fox, Rean Griffith,Anthony D. Joseph, Randy Katz, Andy Konwinski,Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica,and Matei Zaharia
2. Rapid Parallel Genome Indexing with MapReduce,Rohith K. Menon,Goutham P. Bhat,Michael C. Schatz
3. Jinesh Varia, Amazon Web Services - Architecting for the Cloud: Best Practices, 2010.
4. MapReduce: Simplified Data Processing on Large Clusters.
   Jeffrey Dean and Sanjay Ghemawat(jeff@google.com, sanjay@google.com),Google, Inc.
5. M. Burrows and D. Wheeler. A Block-Sorting Lossless Data Compression Algorithm. Technical report, 1994.
6. F. Kulla and P. Sanders. Scalable Parallel Suffix Array Construction. pages 543–546. 2007.
7. B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg. Searching for SNPs with cloud computing. Genome Biology, 10(R134), 2009.