A Hybrid Approach for Sequence Alignment over Genome Data using Compressive Sensing and HBLAST

Prabhat Gupta
(UIT) RGPV Bhopal, India

Rajeev Pandey
(UIT) RGPV Bhopal, India

Anjna Deen
(UIT) RGPV Bhopal, India

S. P. Pandey
RBS Engineering Technical Campus, Bichpuri, Agra, India

ABSTRACT
Medical data is an exponential growth in all the hospitality service area. Genome is an special type of data which deals with the small unit of medical cells. Various matching operation over the genome data is required because of some medical issues arise in various cases. DNA matching, sequence matching, pattern analysis and matching is so called requirement in this area. There are some techniques such as BLAST, HBLAST, RMAP is involved and performed by past researcher. The technique use pre-processing and other filteration , sequence finding is performed. Past approach finds limitation where the large data processing, sequence detection and combine score generation for overall data processing is not performed. In this paper proposed approach is given which work towards the enhancement of previous approach extended with compressive sensing usage for pre-fetching of data and its filteration. It make use of compressive sensing with which a noise removal, filtering process is executed and thus a refined data is observed for Hadoop processing Mapping approach. Our proposed technique executed with different data set of sequence, count of data present in millions and it gives an effective results while comparing with existing scenario. A further implementation on security usage can performed by us.

Keywords
BLAST, Compressive sensing, big data, protein DB, sequence alignment

1. INTRODUCTION
Sequence alignment [1] is an important aspect while dealing with genome dataset which arise in medical field. Data field and protein oriented data which deals in the large data sequence alignment and processing need to be investigated. A proper filteration process and proper processing steps can be more useful to process over large data such as NCBI [16].

The rest of the paper is organized as follow: section 2 depicts related work problems and discussion about them, section 3 depicts proposed work and its work flow, section 4 explain experiment and result analysis while section 5 conclude the complete work.

2. RELATED WORK AND PROBLEM DEFINITION
As per the literature survey is performed with different techniques and different result from the algorithms were monitored such as BLAST ,HBLAST and Data matching , feature extraction and other different technique for sequence alignment technique on large amount of structured data packets available dataset our monitoring is performed[2,3] .

The following are the monitored points which identified as problem and further analyzed and performed further with enhancements.

1. Previous technique such as BLAST & other Sequence alignment algorithm for the processing model generation but still the obvious problem occur with the technique is in generating better result and sequence derivation and finding better result in processing time is lagging in the traditional BLAST algorithm. This technique persist better result than existing but still enhancement is required which is provided by the proposed procedure[4].
2. Previous technique basic classification doesn't perform a better data classification due to lacking of number of rules thus a better probability model can't get generated using the technique [10].
3. In previous technique distribution is used because of that the data of the topics varies which determine the drawback of different entities than proposed work which include Compressive Sensing search and distribution algorithm [5].
4. In the existing distribution independent proportion among component is found thus there is no relation with the other topics is found, where as in new technique normal distribution is used, which provide relation among the topics and provides a flexible framework for the process.
5. Previous Technique make use of original data usage processing with some filtering process to remove the unwanted data. But still the processing of complete data is required at running end.
6. Algorithm process data is need to be investigate and format processing pre-filtering is needed which can save the time and utilization of memory can be saved over the large data process [9].

Thus in order to proposed a better prediction model using classification and further combine approaches requirement is to further acquire an scheme which contribute on getting better outcome and system, here our proposed methodology Compressive Sensing is utilize scheme in place of traditional Sequence alignment approach[6].

3. PROPOSED METHODOLOGY
As per our observation about the previous technique and their disadvantage in different terms and scenario's. Our work present a new approach which is productive and consumes highvalue and thus computational better result over the large number of available dataset.

Big data and Hadoop [8] plays an important role in dealing with the large number of process and big data processing in different scenario. There are different features in Blast algorithm for separation and sequence finding is performed. The existing algorithm which is taken for the further improvement work is HBLAST , which is phase based Blast algorithm to process the large number of data and data packets from genome dataset taken from standard repository. HBLAST outperform as compare with the existing YARN and other scheduler so far. Thus the requirement lead us to perform further improvement to the HBLAST technique. In this dissertation our work perform to create a new compressive sensing HBLAST [7] algorithm which make use of existing scenario and improve it with balanced distribution technique. The proposed technique is based on CS based and data process Sequence alignment based scenario which make use of all the resources properly and outperform the complete distribution using the provided Hadoop based algorithm [8].

The proposed work implemented and tested in Java technology and upon processing data range from 500 MB to 1 GB there are performance monitored in terms of computation processing time, throughput with Hardware configuration as 8 GB RAM, 1 TB Harddisk and the competitive results were monitored. Thus the result as compare to existing technique were monitored which are 1.1x improved while comparing with existing algorithm [2].

Our work propose a new algorithm Compressive Sensing Based prediction model which utilize a new logistic normal distribution technique, which give a relation between the topics and also provide a flexible environment for the complete process and thus it generate a better prediction model for data sequence generation.

Compressive Sensing :

- Take m random measurements: One can think of random sensing as a sampling using a different vector space basis. The new basis has minimal properties (isotropy, incoherence), and some common examples include Gaussian vectors, random binary vectors, randoms rows of a fixed unitary matrix, etc [13,14].
- Reconstruct original signal, f, using linear programming: Minimize the L_1 norm (sum of the inner products between f & the new basis vectors) instead of the usual energy metric (the L_2 norm). Call the L_1 minimizer $f^{\#}$.
- L_1*magic* and exact reconstruction: Unlike the L_2 norm, L_1 promotes sparsity. Moreover, it turns out that $f^{\#}$ also minimizes MSE^4 [15].

The proposed algorithm is described below:

1. Loading of all the available data & packets from the created given message which are participating for the communication from the genome dataset AA file which is downloaded from the standard website.
2. Loading the complete node dictionary pair from the dataset.
3. Perform the particular algorithm as per selected by the user for further execution such as existing or proposed
4. Clustering Approach from existing algorithm.

   Partitioning based clustering methods are an alternative to hierarchical clustering. Many of these standard methods (K-means, Self Organizing Maps, etc) use the Euclidean properties of the object space, for example for defining the cluster centers. This is not appropriate for sequence patterns. Alternatively, graph theory based methods could be used, for example. The K-medoids clustering could also be readily used for pattern clustering. K-medoids is similar to K-means, only it uses as cluster centers the original data objects. Instead of moving cluster centers to the center of gravity of a particular cluster, the object which appears to be most central to the cluster is chosen.

   A consensus sequence representing all patterns in a cluster, a list of sequences matched by patterns, an alignment of the sequences matched by patterns, a position weight matrix generated from that alignment, or its graphical representation by sequence logos. These features are implemented in Expression Profiler with a combination of EPCLUST, URLMAP, and SEQLOGO packages. The position weight matrices are described in the next subsection and a strategy for identifying them is outlined.
5. Perform data move operation and matching operation if any single match is obtained and conclude that further using model for the data shifting either it is working or not.
6. Perform model and match operation if atleast 2 or more dictionary match is performed by the system.
7. Obtaining parameter wise data for the history model.
8. Observing the values and thus it effect computation time and efficiency for the complete scenario.
9. Exit.

Algorithm PseudoCode:
Proposed Sequence alignment Compressive Sensing Based Technique:

Input: Node data Qi,

Output: algorithm process, Metadata, score values.

Steps:

```
Active either BLAST or Compressive Sensing

While(true) do{

    data distribution{p1,p2…..pN};

    dictionaryRequest();

    If(scorematching()==1)

    {

    Recognition();

    Perform Compressive Sensing model()

    {
```

Data collection;

Feature noise detection();

Noise alteration scheme through wavelet;

Data filtering by feature();

}

Compute the prediction values;

{

Result computation;

}

Set status=finish and exit;

}If(scorematching()>=2)

{

Re-distribution;

{

Computing parameter upon distribution;

}

Set status=finish and exit;

}

The above pseudo code shows the complete explanation about how the process is executing. The detail about the compressive sensing which is used in conversion of large amount of data in some short indexing format.

The algorithm used to process data in fast and secure format which is the requirement of any data processing in between the server and user interface. The algorithm which is BLAST (Basic local alignment search tool), which is based on sequence search which help in finding and searching the sequence matching content over it.

Sequence pattern processing include following step procedure in algorithm and computation matching:

Sentence Length Cutoff Feature: If a sentence is longer than the pre specified threshold value is more important than shorter sentence.

FixedPhrase Feature: If Sentence containing any fixed phrases like "this letter","In conclusion" or following immediately after heading containing a keywords like "conclusion","results","summary" are more important.

Paragraph Feature: If a paragraph containing more than one sentence than importance of sentence is based on position, whether it is paragraph initial or paragraph final or paragraph medial. Paragraph initial sentence is, more important than Paragraph final sentence.

Thematic Word Feature: The most frequent content words is known as thematic words. A sentence is scored based on function of frequency.

Uppercase Word Feature: Proper names are important. e.g. " ASTM (American Society for Testing Materials)".This feature is computed with the constraints that an uppercase thematic word is not sentenceinitial and begin with a capital letter. Actions are: TFIDF, entropy, mutual information and statistics. Another Statistical approaches used for keyword extraction are: TFIDF, entropy, mutual information and statistics [11].

Direct lexical chain span score of a word: It is computed same as the lexical chain span score except that it is considered the words which are directly related with the word in the lexical chain. Author applied these four features with a corpus consists of 155 abstracts and got 45% precision in the extraction of keywords.

Relative Utility: Relative Utility measure to overcome the problem of the Precision and recall based evaluation. Suppose a manual summary contain sentences 1, 2, 3, and 4 from a document. There are two systems S1 and S2, creates summaries consisting of sentences 1, 2, 4 and 1, 2, 3.It can be possible that two sentences in one document are equally important. Using Precision and Recall, S 1 can rank higher than S 2. Judges to judges, ranking of sentences are varies.

Content based Measure: Content based evaluation mainly focuses on extracted summaries where comparison is done among words.

The flow where the processing system unit is given step by step from the starting of query as input and then intermediate process, further the execution and processing of data, further data is generated.

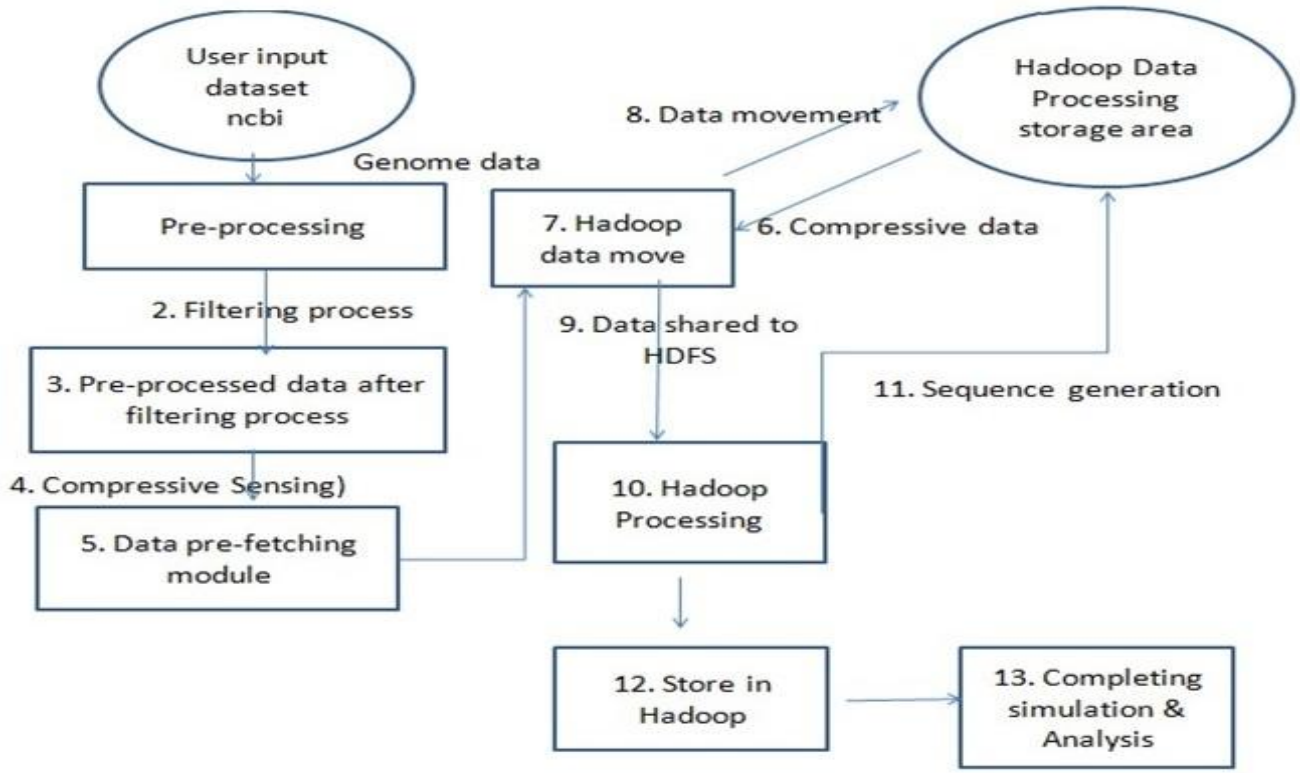

Figure 1: Flow diagram of complete processing proposed architecture

The diagram above shows the complete flow which shows the process executed from first data load to thirteen which is comes to simulation complete and result analysis phase.

With our pattern discovery approach it is possible that two different pattern prefixes (branches in the pattern trie giving alternative pattern representations) have exactly the same set of occurrences. As each branch would be expanded independently, the subtrees below these branches would be identical. In general, it is not possible to discover such patterns without comparing the sets of occurrences. In our algorithm we maintain the lists of locations of each pattern, and the problem can be solved by keeping track of which sets of positions have been studied earlier. This can be done by maintaining the set of position lists in the data structure, from where the lookups (whether exactly the same set of positions has occurred elsewhere in the pattern trie) can be made. We have experimented with this feature and found that for some data sets with conserved complex motifs we can get a considerable speedup to the pattern discovery process. This area of research will need to be investigated further [12].

In this our work is to define the problem definition of the existing presented algorithm in the scenario and the proposed work derived by our work definition .thus the work present by us is highly efficient for prediction model, Further performed work is investigating dataset, query input for the processing with system available and outperform result parameter monitoring.

4. EXPERIMENT SETUP AND RESULT ANALYSIS

An ubuntu Machine with 16.04 version in 1 TB HDD and 8 GB of RAM is configured to perform the HDFS simulation over the large data which is downloaded from ncbi and thus an sequence alignment is performed.

A major role in gene regulation in eukaryotic organisms is played by specific proteins, called transcription factors. By binding to sequence-specific sites in the DNA, called transcription factor binding sites, they influence the transcription of a particular gene. The transcription factor binding sites are located in promoter regions. In yeast these regions are predominantly (but not exclusively) in the immediate vicinity of the gene (typically less than a thousand basepairs upstream of the translation start site). These sites are specific DNA sequences of length from about 5 to 25 nucleic acids, and in yeast they are usually located within a few hundreds of nucleotides upstream from the gene.

The inspection of the patterns that occur substantially more frequently in upstream regions than in random (i.e., having a high rating R(; S+; S)) showed, that many of these can be regarded as “simple” (i.e., easily compressible) sequences (e.g., AAAAAAATA). This is not unexpected, as upstream regions have higher A-T content, while genes rarely contain long simple sequences. Among other patterns, one of the top scoring was a pattern AAAGCGAAA. Matching this pattern against the transcription factor database TRANSFAC (Wingender et al. 1996) we found that it was similar to the binding site for the yeast transcription factor URS1 (Turi & Loper 1992). The pattern given for URS1 in TRANSFAC is AAACGAAACGAAACGAAACTAA. This pattern has only one single match in the entire yeast genome, which means that it cannot be a generic actual binding site. The pattern AAAGCGAAA, on the other hand, has 119 matches in the upstream regions of length 600, and a total of 222 matches in the full yeast genome of 12Mbp. For more detailed analysis and patterns see (Brazma et al. 1998b).

Here we have demonstrated our work in various respects and observed the result and measure the results based on the experiment performance we have observed and notified that

CSBased Hblast technique can be more better when we are using HBLAST technique before applying the BLAST technique, while dealing with the fast system before the user can draw the things [1].

Dataset

Data file which can be simulate can be given as input and to process in the flow where the further distribution among the node can be perform.

A dataset is derived from the ncbi website. This is the website containing genome dataset in all the requirement format. From this web a dataset taken by us is in .AA format and further index format is converted.

https://www.ncbi.nlm.nih.gov is the web resource on which complete dataset can be found of different size and used for sequence alignment [16].

Performance Measures

Computation Time(parameter name): A training time of a dataset in Java is computed with the help of start and end time class variables defined in the tool and here as we load the dataset and verifies the eligibility and taking their features for consideration or not is the time taking process to identify and to load the images and selection of password comes under training time of a dataset, extracting the properties and making them in process format is training time [12].

CT= Finishing time – initializing time ----------(1)

Throughput

This term can be defined as the amount of work which can be done over the period of time, which shows the efficiency of data processing.

T = Amount of work/ Time period----------(2)

Clustering Performance

The result of the expression profile clustering is sensitive to the choice of the distance measure in the expression profile space, as well as on the clustering algorithm itself. Apparently there is no single “right way” of clustering the expression profiles, since various elements in each profile may be influenced by some particular regulation aspects and regulation is usually not based on a simple on/off switching.

Given a set S of N sequences, a subset C S of size n, and a pattern that occurs in k sequences from C, we can calculate the probability of such an event from the binomial distribution. Note that in this application we count as an occurrence only the fact that pattern matches a sequence, regardless in how many places it matches. Thus the number of occurrences in a set of sequences is the number of sequences from that set that contain at least one match by pattern.

>gi|6325162|ref|NP_015230.1| Ypl095cp

Length = 456

Score = 13.5 bits (23), Expect = 0.57

Identities = 4/6 (67%), Positives = 5/6 (83%)

Query 9 IFTSGY 14

+F SGY

Sbjct 1 MFRSGY 6

As per the observed result and experiment setup , technique is implemented. The proposed and existing technique is performed with the above post which aredata packet & distribution among the multiple node matching which in result given by the Sequence alignment algorithm performed with the system and following output results were monitored:

In the table present below is a statistical comparison of the values which are retrived as time taken by the different process algorithm, throughput and other parameter can be observe.

Table 1: Data distribution for different data set and sequence finding time

Dataset (sequence in millions)	Proposed Technique	Existing Technique
Dataset 1	1225ms	2111ms
Dataset 2	2781ms	3230ms
Dataset 3	13215ms	15117ms

The above table represent the number of data values from the data and algorithm is performed.

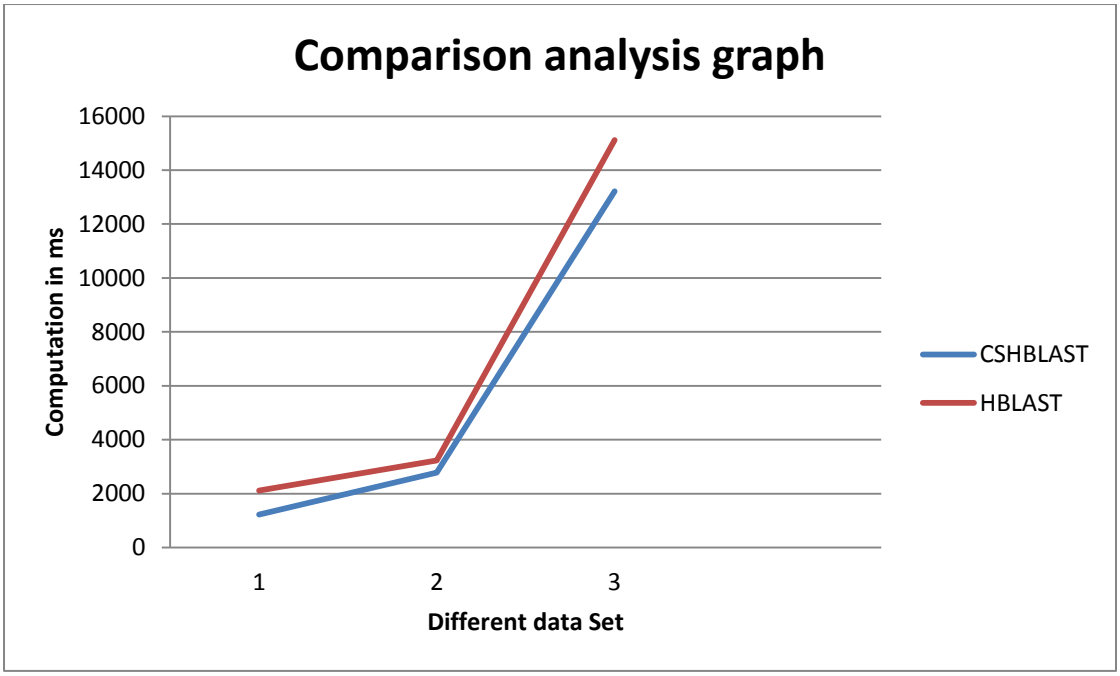

Figure 2: Comparison Linegraph for Computation Time

In the table present below is a statistical comparison of the values which are retrived as throughput by the different process algorithm , throughput and other parameter can be observe.

Table 2: Data distribution for different datasets

Dataset (sequence in millions)	Proposed Technique	Existing Technique
Dataset 1	3.55	3.32
Dataset 2	5.1	4.51
Dataset 3	6	5.89

The above table represent the number of data values from the data and algorithm is performed.

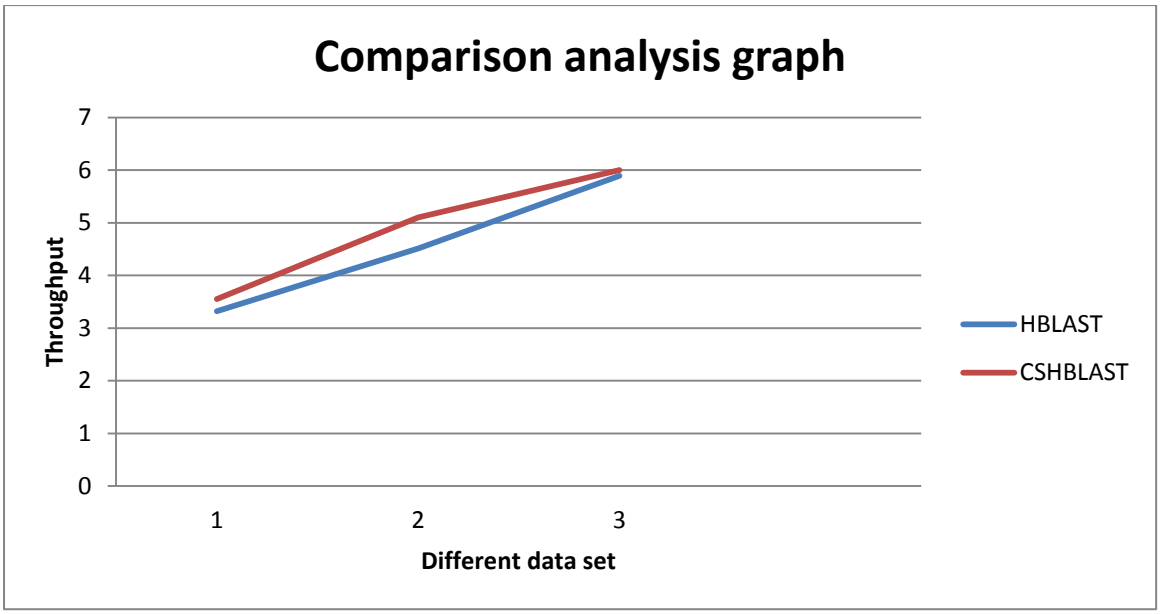

Figure 3: Comparison Linegraph for Throughput

In the above graph drawn x axis as data from which post were extracted for the query processing for specified dataset and line graph is printed using the chart library provided by the Microsoft and further analysis can easily performed thus the Compressive Sensing based approach outperform the best .

The graph representation shows the efficiency of our proposed algorithm work and it outperform effective parameter value.

5. CONCLUSION AND FUTURE WORK

Sequence alignment over genome data is gaining a huge popularity over the large dataset and finding matching pattern sequence in between popular dataset. Pattern detection, pattern matching often provide the proper information , score information and different value. Previous approaches such as BLAST, HBLAST, RMAP and other approach make it effective processing but still while dealing with large data its tedious process. In this paper a proposed technique which use HDFS system of Hadoop file sharing and processing mechanism . An approach for the pre-filtering and processing approach is used by us which is compressive sensing. The algorithm make use of detection technique and find optimized data from large genome dataset. Further an Mapping scheme which is available in Hadoop component is used for processing data in millions set of sequences. An experiment is performed with HDFS installation over Ubuntu 16.x machine having 1 TB of HDD and 8 GB of RAM. Processing computation time and throughput is computed over different range of data. Driven output shows the proposed technique is effective while comparing with existing RMAP and BLAST scenario. A security concern to the data traffic and transformation is not evaluation , hence security applicability and threats can be study in future work analysis with continuation .

6. REFERENCES

[1] Miss. Anju Ramesh Ekre1 , Prof. Ravi. V. Mante, "Hadoop Based Clustering System For Genome Sequencing", 2016 Second International Conference on Science Technology Engineering and Management (ICONSTEM).

[2] Aisling O'Driscoll, Vladislav Belogrudoy, John Carroll, Kai Kropp, Paul Walsh, Peter Ghazal, Roy D. Sleator, "HBLAST: Parallelised sequence similarity – A Hadoop MapReducable basic local alignment search tool" ScienceDirect, Elsevier, Journal of Biomedical Informatics 54 (2015) 58–64, https://doi.org/10.1016/j.jbi.2015.01.008.

[3] Fritz J Sedlazeck,Philipp Reschender, Arndt von Haeseler, "NextGenMap: Fast and Accurate read mapping in highly polymorphic genomes," Bioinformatics advanced access, August 2013.

[4] Kun Sun, Yuet-Ping Yuen, Hauting Wang, Hao Sun, "The Online diagnosis system for Sanger Sequencing based genetic testing," BigComp, 2014.

[5] Quan Zou, Xu-Bin Li, Wen-Rui Jiang, Zi-Yu Lin, Gui-Lin Li, Ke Chen, "Survey of MapReduce frame operation in Bioinfromatics," Briefings of Bioinformatics, Feb 2013.

[6] Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z., ... Chen, H., 2012. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS One 7 (7), e40968. http://dx.doi.org/ 10.1371/journal.pone.0040968.

[7] Dimitrios Milioris, "Classification in Twitter via Compressive Sensing", IEEE International Conference on Computer Communications (INFOCOM), Apr 2015, Hong Kong, Hong Kong SAR China.

[8] Prabhat Gupta, Rajeev Pandey, Anjna Deen, "A Survey on Sequence Alignment Based on Hadoop/MapReduce Big Data" International Journal of Computer Technology & Application, Vol 8(4), 441-447 July-August 2017 Available online@www.ijcta.com.

[9] Vink, M., Raemaekers, M., van der Schaaf, A., Mandl, R., Ramsey, N., 2007. Pre-processing and Analysis.

[10] Hsu, C., Chang, C., Lin, C., 2010. A practical guide to support vector classification. Tech. rep. Department of Computer Science, National Taiwan University.

[11] Pelle Jakovits, Satish Narayana Srirama, "Evaluating MapReduce Frameworks for Iterative Scientific Computing Applications," 2014, pp. 226-233.

[12] Zefeng Zhang, Hao Lin, Bin Ma, "ZOOM Lite: nextgeneration sequencing data mapping and visualization software," Nucleic Acid Research, vol. 38, pp. w743-w748, June 2010.

[13] D. Donoho, "Compressive sensing", in IEEE Trans. on Information Theory, Vol. 52, No. 4, pp. 1289–1306, April 2006.

[14] L. M. Aiello et al., "Sensing Trending Topics in Twitter", in IEEE Transactions on Multimedia, Vol. 15, Iss. 6, pp. 1268–1282, Oct 2013.

[15] G. Tzagkarakis, D. Milioris and P. Tsakalides, "Multiple-Measurement Bayesian Compressive Sensing using GSM Priors for DOA Estimation", in 35th Int. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP), Dallas, TX, Mar. 2010.

[16] https://blast.ncbi.nlm.nih.gov/Blast.cgi

[17]