NEXT GENERATION SEQUENCING ANALYSIS FOR SALMONELLA FLIC GENE

R.Balaji[1], R.Balaji Rathinam[2], P.Hariharan[3]
Dept of Computer Science & Engg, Thiagarajar College of Engineering. Madurai

Abstract

Next Generation sequencing is a high throughput sequencing used to sequence both the DNA and RNA sequences. It is a research tool to address complex genomics beyond the classic DNA sequencing methods. The Next Generation Sequenced data obtained from the Pre-processed stage has to be processed with the best data mining strategies. Topic Modeling is used to discover the abstract topics that occur in a collection of documents. The workflow of the sequence analysis consist of Data Retrieval of the 119 strains of Salmonella from the NCBI database and Preprocessing of NGS data set with 119 strains of Salmonella FliC gene with their corresponding reference genes. Then the Data are processed using Pairwise sequence alignment of the AGTC sequence using Bioedit. The Corpus of text data obtained are analyzed using Topic Modeling using Latent Dirichlet Allocation Algorithm implemented in MALLET. The output topics obtained by the LDA Algorithm used to find out the gene relationship of the Salmonella gene.

Keywords: Next Generation Sequencing (NGS), Topic Modeling, Latent Dirichlet Allocation (LDA), Pairwise Sequence Alignment

I INTRODUCTION

Next Generation Sequencing is a high throughput sequencing used to sequence both the DNA and RNA sequences. It is mainly used for the research field . The greatest advancement in next generation sequencing have been increasing in speed and accurate results , resulting in deduction of manpower and cost. The advancement is mainly because of pairwise analysis and very high throughput measurements.

Sequence Alignment is the current development in the field of bioinformatics, which describes the detailed arrangement of DNA and Protein sequences and identify the regions of similarity of the AGTC sequences among them. It is used to infer structural, functional, genetic and evolutionary type gene relationship between the sequences.

Sanger sequencing was widely used traditional sequencing methods and Nowadays its mostly replaced by Next-gen sequencing. Sanger sequencing is very low paced sequencing methodology when compared to next-gen method and can sequence only a few thousand nucleotide genes for a short period. The next-gen sequencing is very accurate , easy to accessible and cost effective. It can sequence about billions of nucleotides words in a week. Sanger is high accuracy over long lengths but expensive so due to this proximity we process the genomics sequence using Next Generation sequencing. The Genomics consist of Salmonella FliC gene which consist of 119 strains of several genes with each of its reference strains.

The Analysis of the sequence is done by following methods which consist of Data Analysis i.e, retrieving the data from the NCBI Database and Pre-Processing of the data is done by using Pairwise Sequential Alignment. Pairwise Sequential Alignment is the process of extracting two sequences to perform a maximum levels of identity for the purpose of determining the degree of similarity and the possibility of the determination of gene-phenotype relationship . The pairwise Alignment is mainly used to obtain the global alignment of the two query sequences. To

identify the hidden information about our sequence using the known characteristics of the other sequence.

After obtaining the corpus of the AGTC sequence these words are to be processed by Topic Modeling using Latent Dirichlet Allocation algorithm to obtain the relationship between the topic and the words. Topic Modeling is a Statistical model used in text mining mainly used for discovery of hidden sematic structure in a text body of the topics .

In Latent Dirichlet Allocation algorithm each document is generated as mixture of topics and where the value mixtures proportions are distributed as Latend Dirichlet allocation. It is used to find out the per-document topic distributions and . per-topic word distributions.

Therefore , In this paper we propose LDA algorithm to process the large datasets of the NGS data and To Obtain the AGTC sequence as the Strains is used to elucidate the information to the context of the Salmonella fliC gene. To implement Topic Modeling in NGS data analysis and to understand and decode the hidden genetic information in the biological system.

II RELATED WORKS

In this section we present the important aspects of our research such as the AGTC sequence analysis and the alignment of the obtained gene sequences using the topic modeling for the gene relationship.

NGS data's[1] are vastly used in many research fields, the large amounts of the sequence produced by NGS technologies play a detailed challenge for data analysis and its interpretation. Next generation sequence data analysis [1] for the single nucleotide polymorphism. It involves very efficient throughput sequencing technologies that has minimal sequencing time and cost for processing the strains of the genome. The availability [2] of very large volumes of data inexpensively has greatly impacted genetic and medical research for the treatment of various syndromes such as in the treatment of cancer.

Most of these tools are used in the sequencing the strain categories, such as sequence alignment, genome characteristics and its genetic relationship detection [3].Very few research has been reported on mining strategies for greater next generation sequencing data to address bio-driven questions.

Sequencing follows the process of preprocessing, sequencing[4] and analysis of the Salmonella. The Pre-processing of the Salmonella gene is done by using the 119 strains of the Salmonella FliC gene. Primarily [5] used for NGS data preprocessing from the original gene sequence.

Pairwise sequence alignment [10] is a core intensive problem in bioinformatics that has helped researchers analyze biological sequences between the two strains. The research analysis has been used to implement the biologists detect pathogens and identify common genes. The genetic sequence database has been growing rapidly due to new sequences being discovered for the strains by pairwise method .The algorithms uses various methodology to efficiently find optimal or nearly-optimal alignments for the reference strains. In this paper, we present the local and global pairwise sequence alignment algorithms.

Topic modeling is an active research tool that has vast analytical applicability for presenting very large data sets in text mining [6-7] and data retrieval procedures [8]. The core concept is that a document is a mixture of latent topics, each is given by a distribution on words. Latent Dirichlet Allocation (LDA) [7] is the most popular topic modeling algorithm.

Topic Modeling [9] is a Statistical model used in text mining mainly used for discovery of hidden sematic structure in a text body of the topics .LDA processes a text corpus containing documents of AGTC sequence Latend Dirichlet allocation. It is used to find out the per-document topic distributions and . per-topic word distributions.

In this study, we propose a LDA topic modeling to analyze NGS large data sets. The NGS data set containing the Salmonella fliC gene was used to analyze the workflow and the function of this process.

The fliC gene encodes a Salmonella reference gene for each of the serotypes of Salmonella [11]. The procedure blast process was applied to the fliC gene-containing NGS sequences of 119 Salmonella strains of nine Salmonella serotypes. These sequences were retrieved from the database of the National Center for Biotechnology Information (NCBI) and then transformed into the files of documents on which the LDA algorithm was run and then the resultant topic analysis matrices were generated other data mining methods to elucidate the

hidden information within the content of DNA sequences.

III APPROACH

When we come to methodology, there are three steps : 1.Data Preprocessing 2.Data Processing using Pairwise sequential alignment 3.Topic Modeling using Latent Dirichlet Allocation Algorithm .

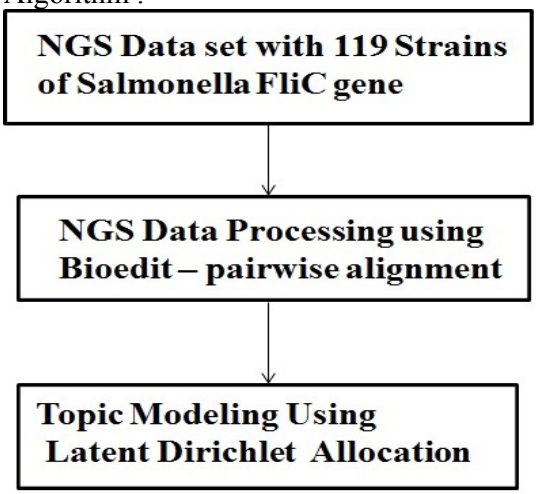

Fig 1.1.Methodology – workflow of NGS Analysis

Data Preprocessing:

In this process, we take the whole genome sequence of the Salmonella O antigen group B . These whole genome was obtained from the National Center for Biotechnology Information (NCBI) Database. The Salmonella 119 strains whole genome includes 75 strains of S. Agona , 14 strains of S. Heidelberg, one strain of S. Paratyphi B, two strains of S. Saintpaul, two strains of S. Schwarzengrund, one strain of S. Stanley, 22 strains of S. Typhimurium, one strain of S. Typhimurium var.5-, and one strain of S. 4, 12:i:- respectively.

The AGTC sequence for the 119 strains of Salmonella gene was retrieved and collected to form a data pool. For each of the 119 strains that consist of various strains the best matching reference gene was selected to perform the data processing using BLAST.

The fliC gene contains a Salmonella phase 1 antigen, and is considered to be one of the Salmonella serotype determinant genes to find the hidden gene relationship . The developed procedure was applied to the fliC gene containing NGS sequences of 119 Salmonella strains of nine Salmonella serotypes. These sequences were extracted from the database of the National Center for Biotechnology Information (NCBI) and were done BLAST operation and loaded into the files of documents .

No	Strain Group	Reference FliC gene
1	75 strains of S. Agona	S. Agona SL483
2	14 strains of S. Heidelberg	S. Heidelberg SL476
3	1 strain of S. Paratyphi B	S. Paratyphi B SPB7
4	2 strains of S. Saintpaul	S. Newport SL254
5	2 strains of S.Schwarzengrund	S. Schwarzengrund CVM19633
6	1 strain of S. Stanley	S. Typhi CT18
7	22 strains of S. Typhimurium	S. Typhimurium LT2
8	1 strain of S. Typhimurium var.5-	S. Typhimurium LT2
9	1 strain of S. 4, 12:i:-	S. Typhimurium LT2

Fig 1.2 . Reference FliC genes for the Strain group

Data Processing:

The retrieved sequences that constituted to form the data pool was used to obtain the AGTC sequence for the Salmonella gene. The Data processing is done by blasting the Salmonella serotype with their necessary FliC genes using BLAST operation in Bioedit tool.

The Basic Local Alignment Search Tool (BLAST) operation is done by Pairwise Sequential Alignment. The method uses local blast methodology to efficiently find optimal or nearly-optimal alignments for the reference strains.

Pairwise Sequential Alignment used in the analysis of genomes. used to decide if two genes are related structurally or functionally.. Alignments made can be local or global alignments. The sequence similarity is measured using Hamming Distance Score value that allows gaps i.e., insertions and deletions. The Hamming Distance score value of two strains is calculated by number of positions with mismatching characters. It is given by,

$$Score = \sum(identities, mismatches) - \sum(gap\ penalty)$$

So by this way will get the required AGTC sequence with necessary Hamming Distance score value.

Using Local BLAST: By creating the local database nucleotide file for the strain to be analyzed and the reference FliC gene will be loaded in the form of the file which will be in FASTA format or it can be loaded directly in

the form of the query. So, the Local BLAST arguments include Maximum number of hits report as 500 and Maximum number of alignments to show as 250. By using this Local BLAST operation the AGTC sequence will be obtained for the necessary reference genes.

Topic Modeling:

After the Data Processing the text corpus is generated with the AGTC sequence in each documents for the 119 strains of the nine Salmonella serotype. The Topic Modeling is used to obtain the relationship between topics and words.

Topic Modeling using Latent Dirichlet Allocation Algorithm is implemented in MALLET The basic assumption behind LDA is that each documents is a collection of a mixture of words and topics. But in the major research shows that we observe only documents and words, not topics – they are part of the latent structure of documents. LDA performs this concept by recreating the documents in the corpus by adjusting the important topics in documents and words in topics iteratively.

The Latent Dirichlet Allocation Algorithm is given by,

For each document:

(a) Draw a topic distribution, $\theta_d \sim Dir(\alpha)$,

Where Dir(·) is a draw from a uniform Dirichlet distribution with scaling parameter α

(b) for each word in the document:

(i) Draw a specific topic

$z_{d,n}$ ~ multi(θd) where multi(·) is a multinomial

(ii) Draw a word $w_{d,n} \sim \beta_{zd,n}$

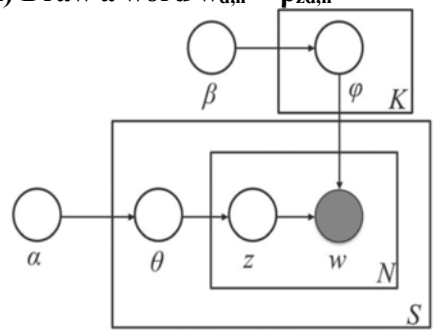

Fig 1.3 Graphical Representation of LDA model

Where,

S – number of strains in corpus, S={D1,D2,…,Ds}

K – number of topics , K={1…K}

N – number of words in one strain

α – prior on the topics

β – prior on distribution over words

θ – topic mixture proportion , S X K

φ – distribution over words, K X V

V – size of vocabulary

Z – particular topic generating a word

Gibbs sampling was used to posterior distribution of $\theta_s Z_{sn}$ and φZ_{sn}.

Dir represents Dirichlet distribution and multi represents Multinomial distribution.where initial values of $\alpha = 0.1$ and $\beta = 0.01$.

The Explanation of the LDA is given by,

- Go through each S, and randomly assign each word in the S to one of the K topics.
- for each document d ,Go through each word w in d.
- And for each topic t, compute :
  1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t
  2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w.
- Reassign w a new topic, where you choose topic t with probability
  p(topic t | document d) * p(word w | topic t)

Topic modeling provides accurate and convenient way to perform unsupervised classification of a corpus of documents.

IV RESULTS AND DISCUSSION

Here we present our results after the analysis of the Pairwise Sequential Alignment with the use of the Hamming Distance Score value to obtain the AGTC sequence for the 119 strains of Salmonella genes.

This is obtained by the use of BIOEDIT tool that was used to perform Local BLAST with arguments include Maximum number of hits report as 500 and Maximum number of alignments to show as 250.

```
Sequences producing significant alignments:

>
        Length = 4772697

 Score = 2809 bits (1417), Expect = 0.0
 Identities = 1495/1521 (98%)
 Strand = Plus / Plus
```

Fig 2.1 Hamming Distance Score value for Stanley str. 06-0538

The process is continued for all the 119 strains of the Salmonella FliC genes.

Then the AGTC sequence obtained is given by,

```
Query: 901      cttgctgcagcaggggttactggc
                ||||||||  ||||||||||||||
Sbjct: 2365954 cttgctgctgcaggggttactggt
```

Fig 2.2 Portion of AGTC sequence obtained for Stanley str. 06-0538

The AGTC sequence is obtained by the similar method for all the 119 strains of Salmonella FliC genes.

The LDA implementation in MALLET produces a result of Topic Distribution over words and the topic count is taken as 20 and optimized interval as 20. The Topic Composition over words by implementing LDA is given by,

	composition	value
Topic Analysis 1	15	0.431543698
Topic Analysis 2	12	0.302644803
Topic Analysis 3	6	0.063261142
Topic Analysis 4	5	0.055539088
Topic Analysis 5	16	0.017522825
Topic Analysis 6	2	0.011285781
Topic Analysis 7	13	0.010988779
Topic Analysis 8	19	0.010394775
Topic Analysis 9	11	0.009503769
Topic Analysis 10	4	0.009503769
Topic Analysis 11	9	0.009206767
Topic Analysis 12	18	0.008612763
Topic Analysis 13	8	0.008612763
Topic Analysis 14	0	0.008018759
Topic Analysis 15	14	0.007721757
Topic Analysis 16	7	0.007721757
Topic Analysis 17	3	0.007721757
Topic Analysis 18	17	0.007424755
Topic Analysis 19	10	0.00683075
Topic Analysis 20	1	0.005939744

Fig 2.3 LDA output for 20 topics with topic composition and the topic distribution values.

The results obtained are plotted in the form of a heatmap.2 function using gplots package in the R studio.

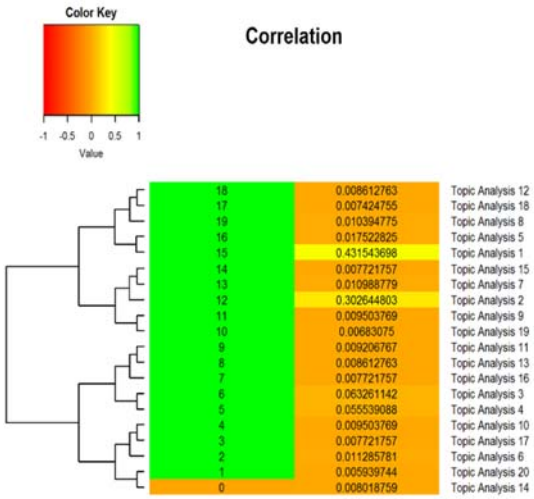

Fig 2.3 LDA output for 20 topics is plotted using heatmap.2 function in gplots package in R studio

Summary of the result is given by,

- Document has multiple topics with various probabilities, it means that the document indicates that all those topics in proportions are indicated by the probabilities.

V CONCLUSION

We have analyzed a procedure for Next Generation sequencing data by implementation of topic modeling, which is an active research in machine learning and has been mainly used as an tool to process large text corpora for data mining. The implementation consist of the following procedures: Data Retrieval, Preprocessing, Topic Modeling, and Data Mining using Latent Dirichlet Allocation (LDA) topic implementation. It is implemented by use of the Salmonella enterica strains. The results provides the implementation of topic modeling in NGS data analysis for the identification of the genetic information and the topic analysis of the AGTC sequence from NGS data, and identify the gene-phenotype relationships.

VI ACKNOWLEDGEMENT

We owe our special thanks and gratitude to Dr. S. Sudha, Assistant Professor, Department of Computer Science and Engineering, Thiagarajar College of Engineering for his guidance and support throughout our project

VII REFERENCES

1. Wei zhong Zhao James J. Chen Roger Perkins "A novel procedure on next generation sequencing data analysis using text mining

Algorithm" Zhao et al. BMC Bioinformatics (2016).
2. Metzker ML. Sequencing technologies - the next generation. Nature reviews Genetics. 2010;11(1):31–46.
3. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generationsequencing on genomics. Journal of genetics and genomics Yi chuan xue bao. 2011;38(3):95–109.
4. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS,Musser SM, Brown EW, Allard MW, et al. Identification of a salmonellosis outbreak by means of molecular sequencing. The New England journal of medicine. 2011;364(10):981–2.
5. "The next generation sequencing revolution and its impact on genomics". Cell. 2013; 155(1):27–38.Koboldt DC, Steinberg KM ,Larson DE, Wilson RK , Mardis ER.
6. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning. 2001;42:177–96.
7. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993–1022.
8. Blei DM, Jordan MI. Modeling annotated data. In: The Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. 2003. p. 127–34.
9. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 2003;3:993–1022. Blei DM ,Ng AY,Jordan MI.
10 .Pairwise sequence alignment algorithms: a survey, ISTA '09 Proceedings of the 2009 conference on Information Science, Technology and Applications , 2009, Waqar Haque, Alex Aravind, Bharath Reddy
11. Macnab RM. The bacterial flagellum: reversible rotary propellor and type III export apparatus. Journal of bacteriology. 1999;181(23):7149–53.