International Journal of Education and Science Research Review
Volume-11, Issue-2 March-April-2024 E-ISSN 2348-6457
(Peer-Reviewed, Open Access, Fully Refereed International Journal) P-ISSN 2349-1817
www.ijesrr.org Impact Factor - 7.02 Email- editor@ijesrr.org

APPLICATIONS OF PLSA IN NEXT-GENERATION SEQUENCING DATA ANALYSIS

Anita
Research Scholar, Dr. K. N. Modi University, Newai, Rajasthan

Dr. Amit Kumar Sharma
Assistant Professor, Dr. K. N. Modi University, Newai, Rajasthan

ABSTRACT

The Next-Generation Sequencing (NGS) technology has brought about a revolution by making it possible to sequence DNA and RNA at a high throughput. A tremendous amount of data has been generated as a consequence of the utilization of this method. One of the most significant procedures is the analysis of NGS data, which requires the utilization of sophisticated computing tools in order to evaluate the complex information that is produced. This is one of the most crucial operations. One such efficient statistical method is called Probabilistic Latent Semantic Analysis (PLSA), and it is utilized for the aim of comprehending and modeling the underlying patterns that are present in NGS data. PLSA is a methodology that has proven to be useful. PLSA is often used in the field of topic modeling, and it is amenable to modification in order to discover and evaluate latent patterns that are present within genomic datasets. This is done in order to help with the process of topic modeling. Researchers are able to uncover patterns of gene expression, variation distribution, or microbial diversity that were previously unknown by applying PLSA to data obtained from next-generation sequencing (NGS). Consequently, this becomes PLSA an exceptionally valuable instrument for applications such as the identification of disease genes, the investigation of evolutionary processes, and the practice of customized medicine. The incorporation of PLSA into the analysis of next-generation sequencing (NGS) data not only enhances the comprehension of biological processes, but it also makes it possible to generate interpretations of complex genomic data that are more accurate and insightful, thereby expanding the scope of what can be accomplished in contemporary genomics research. Both of these benefits are a result of the incorporation of PLSA.

Keywords: - annotation, agriculture, bioinformatics, disease diagnosis, health and web tools

INTRODUCTION

In the field of metagenomics, for example, PLSA can be applied to discern between different microbial communities based on the individual genetic signatures of each of these communities. This makes it possible to have a more in-depth understanding of the manner in which environmental conditions influence the variety of microorganisms inside the environment. In a similar fashion, PLSA can be applied in the field of transcriptomics to assist in the discovery of distinct gene expression profiles across a number of contexts or time points. This can result in the provision of substantial insights into the regulatory processes associated with the regulation of biological responses.

PLSA's capacity to cope with noisy and incomplete data, which are prevalent challenges in genomic investigations, is yet another significant advantage of using PLSA in the analysis of NGS data. This is because

PLSA is able to handle these types of data. The probabilistic architecture of PLSA permits robust modeling, even when uncertainties are present. As a result, it is a reliable instrument for the analysis of complex biological datasets because of its ability to accomplish this.

The field of genomics is expected to continue to undergo development, and it is projected that the incorporation of advanced analytical approaches such as PLSA with data obtained from next-generation sequencing (NGS) will become increasingly significant. In addition to the fact that these methodologies have the potential to enhance the interpretative capabilities of genomic investigations, they also make new avenues of research possible. This is especially true in sectors such as precision medicine, where it is vital to have a comprehensive understanding of the tiny nuances of genetic variation in order to construct therapies that are personalized to the individual patient. The conclusion is that PLSA provides a powerful framework that may be utilized to unlock the full potential of data obtained from next-generation sequencing (NGS). Researchers are now able to explore and comprehend the complexity of the genome at a degree of detail that has never been seen before because to this approach.

In addition to its applications in gene expression analysis and metagenomics, the applicability of PLSA in near-genomic sequencing (NGS) data processing extends to a number of other significant domains. For instance, in the field of cancer genomics, where the identification of driver mutations and the classification of tumor subtypes are of the utmost importance, PLSA can be applied to uncover latent genetic and epigenetic traits that differentiate between normal and malignant cells. This is demonstrated by the fact that PLSA can be used to identify these characteristics. As a result of the study of somatic mutations, copy number variations, and other genomic abnormalities, PLSA is able to provide assistance in the identification of the molecular signatures that are associated with specific forms of cancer. It is because of this that it is feasible to improve the development of treatment regimens that are more successful and suited to the individual.

Basic preparations

This section provides a concise explanation of the subject models that were utilized in the proposed study, including NNMF and PLSA. In addition to this, it provides an explanation of the data mining techniques that are applied, such as distance matrix analysis and zwei-way hierarchical clustering.

Models of the Subject

The generation of a text corpus using topic models is characterized by the fact that each document is associated with one of the 119 strains and that each document has an equal amount of words. NNMF, PLSA, and LDA are some of the topic models that are incorporated in Mallet. These models are utilized for the purpose of corpus modeling1 and for obtaining the topic mixture distributions and themes of each strain.

the non-negative matrix factorization of the matrix An example of a mathematical model that folds high-dimensional vectors into low-dimensional vectors is the NNMF, also known as the NMF. Due to the fact that the vectors are not negative, NNMF is able to factor them into a form that is both lower-dimensional and non-negative.

Take into consideration a matrix X that, when the two matrices T and M are generated, X is equal to TM. Due to the fact that the NNMF possesses a clustering property, the matrices T and M are used to express the following information concerning matrix X:

- X, also known as the document-word matrix, is a representation of the input words that are found in particular documents.
- Based on the documents, the themes, which are also referred to as clusters, are referred to as T (Basis vectors).
- The participation weights for the themes in each report are represented by the letter M in the coefficient matrix.

By employing "Scikit-learn," a free piece of software included in the machine learning library, a Python code is constructed for the purpose of executing NNMF. This code is used to actualize NNMF and obtain the strain-topic proportion matrix from the input matrix of position and base at each and every nucleotide polymorphism (SNP) of each strain.

Latent Semantic Analysis using Probabilistic Probabilities Through the use of the probabilistic technique, the PLSA is able to successfully solve the dimensionality reduction problem. In addition to the LSA, the PLSA incorporates a probabilistic approach to the handling of subjects and words.

In the PLSA model, the entry of the document-term matrix is denoted by the notation P (d, w) for each document d and word w associated with it. In addition, each and every document is made up of a mixture of subjects, and each topic is made up of a collection of words. Figure 4.1 provides a visual depiction of the PLSA topic model in its many forms. An additional probabilistic dimension is added to the following assumptions by the PLSA model:

- The probability that subject c is included in a given document d is denoted by the symbol P (c|d).
- The likelihood that a word w will come from a particular subject c is denoted by the symbol P (w|c).

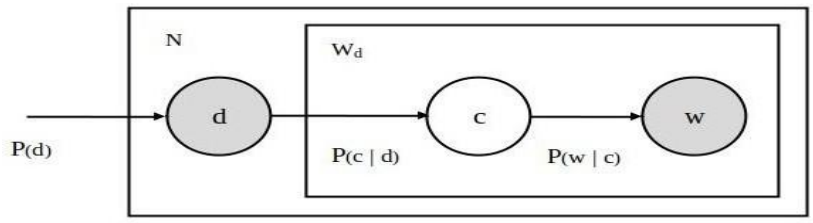

Figure 1: Graphical representation of PLSA topic model.

In mathematical terms, the equation (4.1) represents the combined probability of a certain text and word when they are considered together.

$$P(d,w) = P(d)\sum_{c} P(c|d)P(w|c) \quad \text{.....(1)}$$

The structure of the document is explained by the equation (4.1), which is based on the distribution of the topics contained inside that text. The parameters of the model are denoted by the letters P (d), P (c|d), and P (w|c) in this equation. The corpus may be used to find the value of P (d). There are two types of

multinomial distributions: P (c|d) and P (w|c). Both of these distributions may be trained with the Expectation-Maximization (EM) technique. It is possible to discover the most likely parameter estimates of a model by employing a technique known as the EM. Following this line of reasoning, the amount of parameters is identical to the sum of cd and wc. There is a clear correlation between the amount of documents and the quantity of parameters that occur. A further feature of PLSA is that it is a model that generates new studys.

In addition to this, the PLSA method is implemented in Python code by making use of the EM algorithm. A total of one hundred emphasizes are carried out in order to accomplish the EM computation, and the log-likelihood convergence is set to a value of one hundred percent. In addition, the strain-topic proportion matrix is produced by utilizing the input, which consists of a grid of location and base at each SNP of each strain.

The Mining of Data

Data mining techniques such as two-way hierarchical clustering and distance matrix analysis are utilized in order to get a better understanding of the connections that exist between the various subjects.

A clustering method that takes into account hierarchical structure in both directions

Through the utilization of this technique, it is feasible to simultaneously group the rows and columns of a matrix. The Euclidean distance approach is a technique that can be applied for the goal of computing the diversity that exists between the subjects and the strains. A visual representation of the subgroups of samples and the connections between the various themes is provided by the result. It was determined that the NCSS program would be the most suitable tool for carrying out the two-way clustering heat map.

An examination With the help of the Distance Matrix In order to assess the degree of uniqueness that exists between the strains of strain-topic combinations that are produced from the fliC SNPs data, the Euclidean distance can be applied for the purpose of this research. This research aims to determine the degree of uniqueness that exists between the strains. Through the utilization of the "dist" function, which is a component of the "stats" package of the R programming language, it is possible to ascertain the Euclidean distances that exist between each of the strains. There is a program called "heatmap" that is a part of the "gplots" package of R. This software is used when it comes to the implementation of heat maps.

In the process of analyzing the NGS data, both PLSA and the findings of the experiments were applied.

For the purpose of preparing the dataset, the sequences of the NGS data are originally retrieved from the NCBI database. This is done in the experimental study with the intention of preparing the dataset. After that, the sequences that have been collected are locally aligned with reference fliC by utilizing the Basic Local Alignment Search Tool (BLAST) tool. This comes after the sequences have been obtained. In conclusion, the CLUSTAL tool is applied for the purpose of using multiple sequence alignment. The composition of the SNP is extracted from the sequences that have been matched, and it is considered to be a file of "a bag of words." This is done after all has been said and done. The NGS dataset that has been prepared has been subjected to PLSA topic modeling, and the results of this application have been examined in relation to and contrasted with other topic models, such as LDA and NNMF. The metrics that are utilized for the purpose of performing model evaluation are the NMI, ARI, NID, and NVI. Data mining techniques, such as topic analysis, two-way hierarchical clustering, and distance matrix analysis, are also utilized for the purpose of information retrieval. These approaches are utilized throughout the process. Furthermore, a graphic representation of the progression of the

technique that was recommended can be found in Figure.

The first step is to prepare the dataset.

The National Center for Biotechnology Information (NCBI) database was accessed in order to acquire 119 Salmonella O antigen group B strains and their reference genes for the purpose of constructing the NGS dataset. A total of 75 strains of the 'S. Agona' strain, two strains of the 'S. Saintpaul' strain, fourteen strains of the 'S. Heidelberg' strain, one strain of the 'S. Paratyphi B' strain, two strains of the 'S. Schwarzengrund' strains, one strain of the 'S. Stanley' strain, one strain of the 'S. Typhimurium var.5-' strain, one strain of the 'S. 4, 12:i:-' strain, and 22 strains of the 'S. Typhimurium' strains are included in this collection of strains. Additionally, the BLAST algorithm was applied in order to determine which sequence particles exhibited the highest degree of similarity to the fliC reference gene. This was done in order to identify the sequence particles. It is generally agreed that the genes "S. Agona SL483," "S. Heidelberg SL476," "S. Newport SL254", "S. Schwarzengrund CVM19633," "S. Paratyphi B SPB7," "S. Typhi CT18," and "S. Typhimurium LT2" are the ones that are considered to be the reference fliC genes.

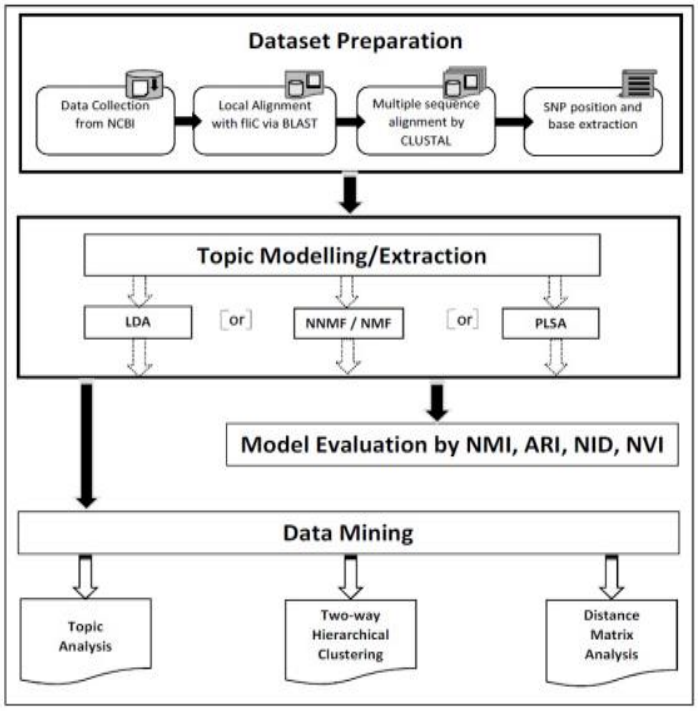

Figure 2: The workflow of the NGS data analysis has been customized to include the utilization of PLSA topic modeling.

The sequences that were included in the newly produced dataset were aligned by utilizing a number of different sequence arrangements, such as MUSCLE or CLUSTAL. This led to the development of a dataset of adjusted sequences, which included inclusions and cancellations, which are also referred to as indels. At each and every location, the single nucleotide polymorphisms (SNPs) were collected, and each strain has its very own unique folder of words that contains not only the SNPs but also information about its location.

Analysis of the Subject Area

The subjects that were produced from the PLSA were stratified in order to ensure that the collection of words that share similar characteristics was categorized. The PLSA topic model was utilized in order to ascertain the 'ten' probable phrases that were associated with the most frequent occurrences across five distinct subjects. In the case when the number of topics was fixed at five (T0, T1, T2, T3, and T4), these phrases were displayed in table. Words are arranged in a hierarchy, from most likely to least likely, according to the likelihood value of each phrase, and this hierarchy is present in every subject. A few of the ten words that featured the most frequently across all five disciplines were quite extraordinary. Each subject has its own distinctive word arrangement, and some of these terms were particularly noteworthy. In addition, every strain has its own record that details a variety of topics and compares the odds of each one. Additionally, the topic mixture coefficients of the strains that were of the same serotype were seen to be comparable to one another. By utilizing the strains that have been gathered, it is possible to compute the realistic expectations of the themes for each serotype. This is something that may be done effectively. It is shown in Figure 3 that the topic number is set to 4, and the topic distribution is demonstrated to be dispersed across the different serotypes. It is clear from looking at the graphic that the probability of topic distributions varied depending on the serotype. It is possible to make a note of this particular fact.

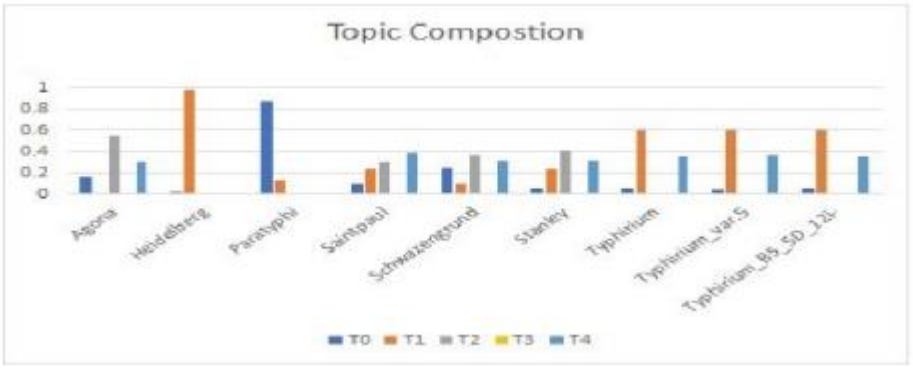

Figure 3: For the purpose of determining the topic distribution of the various serotypes for subject number 5, the PLSA was utilized. Five themes are reflected from the beginning of time until the fourth time.

A look at the distance matrix and its analysis

The strain-topic matrix, which is formed by the PLSA topic modeling and consists of 119 strains, is subjected to an analysis. This analysis is carried out. Figure 4.5 depicts the similarities in subject combinations that are discovered between each pair of strains. These commonalities can be detected in some strains. There are a variety of shades of blue to red that are included in the spectrum, and these shades are used to indicate the estimates of the Euclidean distance, which vary from 0 to 1. This data mining technique is also used to examine the findings of strain-topic combinations, which are representative of the SNPs of all strains. In addition, this technique is used to study any strain-topic combinations. Additionally, the PLSA topic modeling is applied to the SNPs corpus, and the topic numbers (K) are assigned as 2, 5, 15, 30, 50, and 75, respectively. This is done in order to improve the accuracy of the results. To shed light on the impact that the variation in the topic number has on the biological significance that is finally produced, this is done in order to shed light on the impact that this variation has. The efficiency of the PLSA distance matrices heat map may be shown in Figure, which displays the outcomes for a wide range of various subjects.

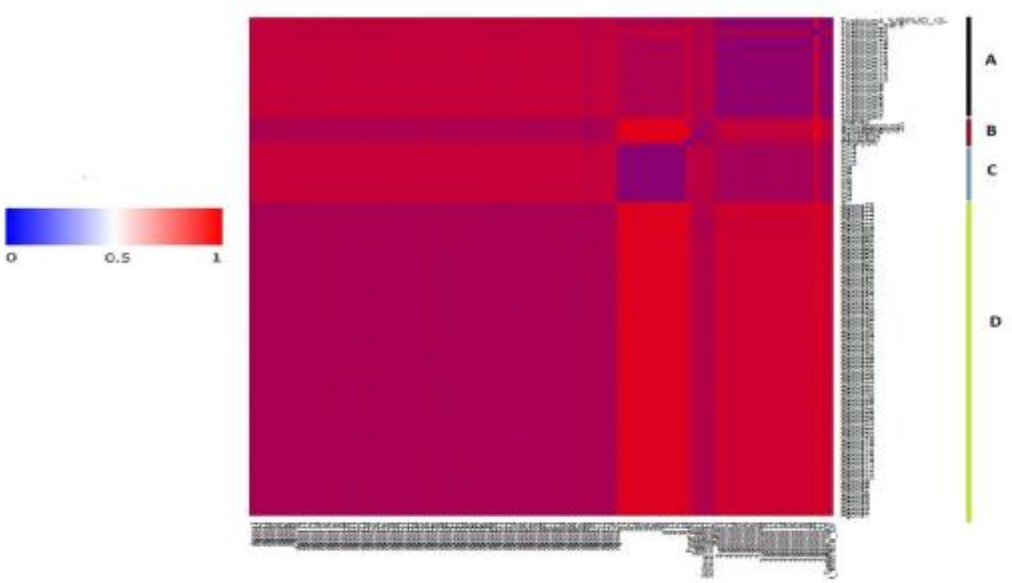

Figure 4: Within the framework of the PLSA model, the distance matrix of strains that have themes that include five is utilized. The histogram, which represents the varied degrees of similarity among point mixtures that are created by each match of strains, varies in color from blue to red and depicts the spectrum of characteristics that are present. There is a group that contains strains that are associated with Typhimurium, another group that contains strains that are associated with Saintpaul, another group that contains strains that are associated with Heidelberg, and a group that contains strains that are associated with Agona.

CONCLUSION

The application of computational approaches in the analysis of biological data has brought about a significant transformation in the field of biology. Researchers are now able to process and interpret large volumes of complicated data with a precision and efficiency that has never been seen before because to the utilization of new algorithms, machine learning, and statistical models. As a result of these methodologies, patterns, correlations, and trends that may be disguised in traditional analyses can be identified more easily, which ultimately leads to deeper insights into the mechanisms and processes that are involved in biological processes? Furthermore, computational techniques make it possible to integrate data from a wide variety of sources, including genomic, proteomic, and metabolomic data, which contributes to a more comprehensive knowledge of biological systems. As the complexity of biological data continues to increase, the role of computational techniques will become increasingly pivotal in driving innovations and discoveries, which will ultimately lead to an advancement in our knowledge of life sciences and an improvement in applications in fields such as personalized medicine, drug development, and environmental biology.

REFERNCSE

1. Irene M. Ong (2007) COMPUTATIONAL TECHNIQUES FOR INFERRING REGULATORY NETWORKS https://pages.cs.wisc.edu/~ong/io_thesis.pdf

2. Matthew R. Pocock (2003) Computational Analysis of Genomes https://www.sanger.ac.uk/theses/pocock-thesis.pdf

3. Xintao Wei (2010) Computational approaches for biological data analysis SBN:978-1-124-21198-5 Order Number:AAI3422315

4. Jingyi Li (2013) Statistical and Computational Methods for Analyzing High-Throughput Genomic Data https://escholarship.org/content/qt9c54z306/qt9c54z306_noSplash_1e06ce06e067d54b6a981cc5b5e9ce5b.pdf?t=mtff28

5. Byron, Kevin, "Big data analytics in computational biology and bioinformatics" (2017). Dissertations. 17. https://digitalcommons.njit.edu/dissertations/17

6. Tyagi, Anuj. (2020). Biological Databases.

7. Kaur, Parampreet & Singh, Ashima & Chana, Inderveer. (2021). Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions. Archives of Computational Methods in Engineering. 28. 1-37. 10.1007/s11831-021-09547-0.

8. Solanki, Vivek & Arora, Neha & Hashmi, Faiz & Raghuvanshi, Durgesh. (2020). Computational of Bioinformatics. 128-131.

9. Francisco A. Gómez Vela (2021) Computational Methods for the Analysis of Genomic Data and Biological Processes ISBN978-3-03943-771-9 (Hardback) ISBN978-3-03943-772-6 (PDF) https://doi.org/10.3390/books978-3-03943-772-6

10. Patil, Sonali & Durve-Gupta, Annika. (2022). Bioinformatics and Its Application in Computing Biological Data. 10.1007/978-981-19-6506-7_8.

11. Khalid Raza (2019) Application Of Data Mining In Bioinformatics Indian Journal of Computer Science and Engineering Vol 1 No 2, 114-118

12. Nehul Singh (2021) Overview of Some Computational Techniques for Bioinformatics file:///C:/Users/skdrg/Downloads/Overview_of_Some_Computational_Techniques_for_Bioi.pdf

13. Abdi, H., and Valentin, D. (2007). Multiple correspondence analysis. Encyclopedia of Measurement and Statistics 2, 651–657.

14. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods.

15. Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.

16. Botstein, D., and Fink, G.R. (2011). Yeast: an experimental organism for 21st Century biology. Genetics 189, 695– 704.

17. Breiman, L. (1997). Arcing the edge (Technical Report 486, Statistics Department, University of California at …).

18. Breiman, L. (2001). Random Forests. Mach. Learn. 45, 5–32.

19. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees. Monterey, Calif., USA: Wadsworth.

20. Breitling, R. (2010). What is systems biology? Front. Physiol. 1, 9.