ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Efficient DNA Sequence Analysis for Reduced Gene Selection Using Frequency Analysis

A.Surendar[1], M.Arun[2]
[1]School of Electronics, Vignan's University, India.
[2]School of Electronics, VIT University, India
***Corresponding author: E-Mail: surendararavindhan@gmail.com**

ABSTRACT

The modern trends in bioinformatics are focused to improve the sophisticated solution to the medical industries. The human society has more impact from the diseases found and they highly reflect from the side of DNA. Any human disease and the person who affect from any disease is based on their DNA sequence. Not only that there are number of biological information can be used to perform any analysis on the disease. The biological information has wide range of applications and can be used for any purpose. In general the biometrics is larger in size and to perform any validation of such wide dimensional features, there require efficient fast applications. We propose an FPGA based multilevel sequence similarity identification to reduce the computational overhead, time and data complexity, GENIE, UNI, dbGap are the benchmarked database considered for the validation of the proposed method. Performance of the proposed method is compared with existing Bloom filter, PSR, SRA and DFT.

KEY WORDS: Gene selection, frequency analysis, DNA sequencing, multilevel sequencing, Bioinformatics

1. INTRODUCTION

The genetic theory has higher impact in the common society and it has the rule for anything happening in the world. For many reasons, the activities of the people have great relationship with their genetic. The behavior and property of the human depending on the presence of the gene. So the problem of gene detection has been applied for many issues. For example, in a university there are number of students present but among them only selective students has great records. When you analyze their gene pattern, it is noticeable that they have higher memory and ideology. Similarly, there are number of cases notified from the DNA sequence.

The medical domain has great focus on DNA analysis and they apply the DNA property for many purposes. Any disease on the human can be treated with the DNA sequence and by identifying the gene which has the property of tamper resistance for particular disease it can be treated. Also, in case of breast cancer, there are certain genes which has higher impact on occurrence in cancer. So in order to identify and predict the list of genes, they require some efficient approaches.

Let's discuss with an example, as the given sequence is "MEKLLDEVLAPGGPYNLTVGSWVRDHVRS IVEGAWEVR", according to the territory structure it can be framed as: CBCCCBCCCCAACCCCCCACCCCCA CCACCBCC, which is the complete pattern. From this pattern sequence identified, you can generate number of sequences of small size. Anyway in order to identify the subset of gene sequence with higher importance, it is necessary to compute the frequency. As like, text clustering, any sequence of DNA has been selected according to their frequency and weight.

The Frequency of any sequence Si, can be computed as follows:

$$\text{Sequence Frequency sf} = \frac{Number\ of\ occurrence\ of\ sequence\ in\ class}{Total\ number\ of\ sequencce\ of\ class} \quad \text{--- (1)}$$

The sequence frequency represents the influence of the sequence in particular class of gene set. Similarly the influence of the sequence in other class can be computed using the below formula.

$$\text{Inverse Sequence Frequency Isf} = \frac{Number\ of\ occurrence\ of\ sequence\ in\ other\ class}{Total\ number\ of\ sequencce\ of\ other\ class} \text{ --- (2).}$$

The equations 1 and 2, represents how the influence of the sequence in same class and other class can be computed. Further the selection of the sequence as a representative for the class can be performed by computing the weight for the sequence. The sequence weight can be computed as follows:

Sequence weight sw = Sf×ISF --- (3).

Using the sequence weight computed the method can select the sequence with higher influence and used to perform prediction and gene selection minimization.

2. METHODS EXPLORED

There are number of methods has been discussed earlier for the problem of gene selection and minimization. This section discuss about some of the methods.

To improve the stability of feature selection under varying samples, the author proposed a sample weighting technique in (Christopher, 2012), which uses microarray data. The method improves the performance gene selection and increases the stability of gene selection also. The performance of the method has been evaluated and compared with the performance of support vector machine and relief classifiers.

To perform microarray cancer classification, a Relevant and Significant Supervised Gene Clustering algorithm is discussed in (Grigorios, 2012). The method uses mutual information based supervised gene clustering

(MSG) algorithm to form the reduced gene clusters for cancer classification. The efficiency of the method has been evaluated with different micro array cancer data sets and compared with the classifiers like Naïve bayes, K-nearest rule, and SVM.

Cancer subtypes prediction using Gene-Expression using Feature Selection and Transductive SVM has been discussed in (Alachiotis, 2011). The method adapts both gene selection and transductive support vector machine (TSVM) to predict the gene sets. The method identifies the genes using TSVM to improve the accuracy of prediction. The performance of the method has been compared with standard SVM technique.

To identify uncovered gene pathways which characterize the cancer heterogeneity an efficient method has been proposed in (Gabriel, 2011), which uses the sparse statistical method. The method specifies set of pathway activities which are identified from the micro array data using Sparse Probabilistic Principal Component Analysis (SPPCA). The method also generates an association between gene-gene related to the cancer phenotypes.

Predicting metastasis of breast cancer has been discussed in (Surendar, 2016), and performs a comparison of classification performed by different methods and analyzes the results. For the prediction of metastasis, the method uses voting approach. The method has produced efficient results than other methods.

The survivability of breast cancer diagnosis has been approached using embedded genetic algorithm in (Pall Melsted, 2011). The shapely value based feature selection technique use include and remove memetic operators. The entire algorithms feature selection has been optimized using the genetic algorithm. The gene selection algorithm selects a subset of genes from the high dimensional data set using the genetic algorithm. The method performs the differentiation based ranking of genes to select them. The method use four different classifiers to improve the quality of gene selection.

In (Arun and Krishnan, 2011), identifies the pattern of genes present in the breast cancer patients. Using the pattern identified from the gene set available, the method selects the subset of genes in form of pattern. The selected pattern represents the gene selection. A comparative analysis has been performed with various gene classification approach in (Che, 2009). The author presents a comparative study on various classification algorithms and support vector machine.

Identification of a Comprehensive Spectrum of Genetic Factors for Hereditary Breast Cancer in a Chinese Population by Next-Generation Sequencing (Yoginder, 2008), discussed to classify a complete spectrum of genetic factors for genetic breast cancer in a Chinese population, we did an analysis of germline alterations in 2,165 coding exons of 152 genes related with genetic growth using next-generation sequencing (NGS) in 99 breast cancer patients from relations of cancer patients irrespective of growth types.

Genomic prediction of disease occurrence using producer-recorded health data: a comparison of methods (Lysaght, 2006), discusses of single-trait then two-trait sire models was examined using BayesA and single-step approaches for mastitis and somatic cell notch. Variance mechanisms were projected. The comprehensive dataset was alienated into exercise and authentication sets to perform perfect comparison. Projected sire upbringing values were used to approximation the amount of daughters probable to develop mastitis. Predictive ability of each model was assessed by the sum of $\chi 2$ values that associated foretold and observed facts of daughters with mastitis and a number of wrong forecasts.

Gene Change Profiling of Breast Growths for Scientific Decision Creation: Motorists and Travelers in the Cart Beforehand the Mount (Surendar and Arun, 2016), discusses topical advances cutting-edge molecular summarizing allow for a rapid and relatively cheap assessment of manifold changed genes or gene crops from small quantities of tumor tissues or gore. The test ahead is how to incorporate these consequences into clinical practice correctly, and in what way to provide patients with the finest possible yet evidence-based care. Herein, the author provides a brief overview of genetic mutation profiling with a focus on next-generation sequencing (NGS) and possible clinical utility.

Iterative Sequence Frequency Analysis Based Gene Selection: In this method, the gene sequence has been split into number of tiny sequences and for each size a set of patterns are generated as discussed in previous chapters. Then the method computes the sequence frequency and inverse sequence frequency for each class. Using computed frequency values, the method compute the sequence weight. Based on computed weight, the method selects the most weighted sequence. This will be iterated for number of times to reduce the size of gene selection. The entire approach has been split into number of stages namely preprocessing, sequence generation, Frequency Analysis, Gene-selection. Each will be discussed in detail in this section.

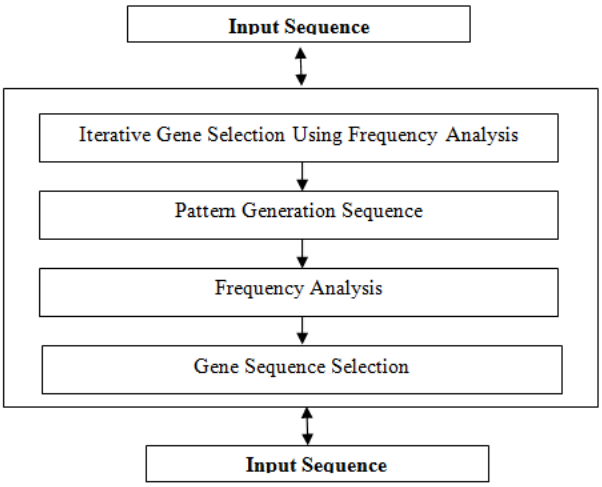

Figure.1.Architecture of proposed gene selection approach

The Figure 1 shows the architecture of gene selection approach and shows the functional components in detail.

Preprocessing: In this stage, the method reads the sequences present in the sequence set and for each sequence, the presence of all genes is verified. If any of the sequence has been identified as incomplete then it will be considered as noisy and removed from the sequence set. The noise removed sequence set will be used to perform gene selection.

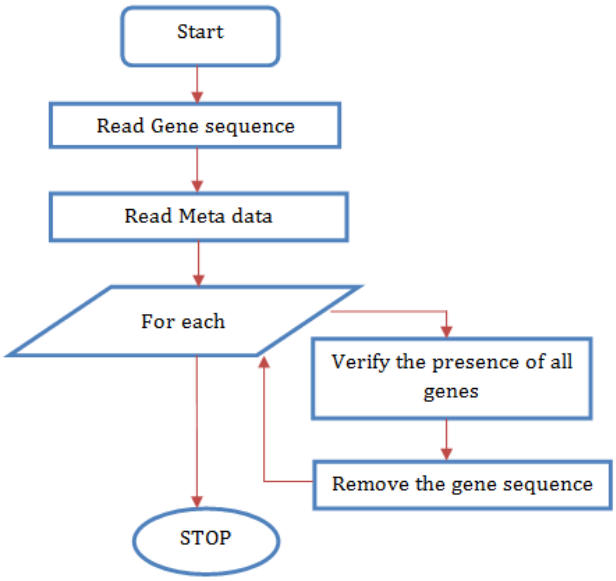

Figure.2.Flow chart of preprocessing

The figure 2 shows the flow chart of preprocessing algorithm and shows the details steps.

Pseudo Code of preprocessing:

Input: Gene Sequence Set Gs

Output: Gene Set Ges

Start

Read Meta data Md.

Identify unique genes Ug = $\int_{i=1}^{size(Md)} \sum Gi(Md(i)) \nexists Ug$

For reach gene sequence Gsi

If Gsi contains all genes

Else

Remove the sequence.

End

End

Stop.

The above discussed algorithm identifies the presence of all the genes from the meta data and removes the incomplete genes.

Sequence Pattern Generation: The method generates patterns from 1 to N dimension. For each pattern generated the method computes the number of occurrence in all the sequence set available in the data set. Using the number of occurrence value computed and the details of the sequence set, the method computes the protein impact value. The computed protein impact value will be stored in the matrix.

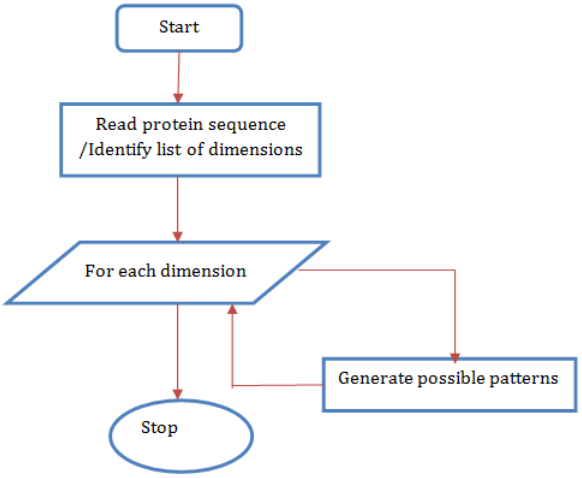

Figure.3.Flow chart of sequence pattern generation

The Figure 3 shows the flow chart of sequence pattern generation and shows the detailed stages. The flow chart of protein impact matrix generation and shows the detailed steps.

Pseudo Code of Protein Sequence Pattern Generation:

Input: sequence S

Output: Sequence Set Ss.

Start

Read sequence S.

Identify the dimension of the sequence Sdim = $\sum Genes \in S$

For each dimension size Dsize

Generate possible patterns Ps.

End

Stop.

The above discussed algorithm generates the possible sequences for the protein sequence given.

Frequency Analysis: In this stage, the method computes the sequence frequency for each pattern from the pattern set. Similarly the method compute the inverse sequence frequency for each of them. Using both the frequency values, the method compute the sequence weight. Based on computed weight the method selects the most weighted sequence.

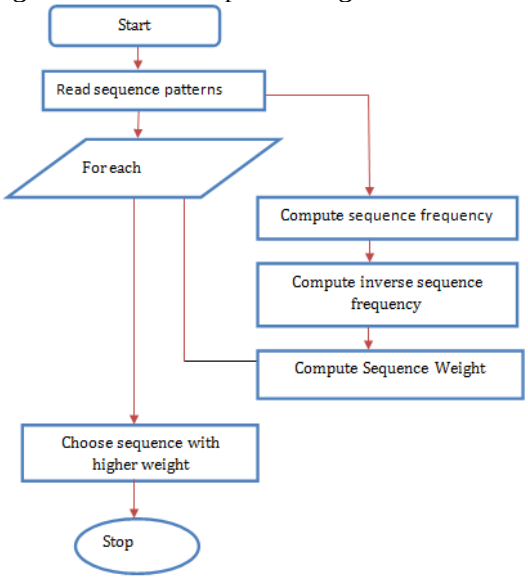

Figure.4. Flow chart of frequency analysis

The Figure.4 shows the flow chart of frequency analysis and shows the detailed stages in frequency analysis.

Pseudo Code of Frequency Analysis:

Input: Gene sequence set Gs, data set Ds

Output: Selected sequence Sels.
Start
Read gene sequence Gs.
Read gene data set Ds.
Ps = generate sequence pattern.
For each sequence Pi from Ps
Generate sequence frequency Sf.
Generate inverse sequence frequency Isf.
Compute sequence weight sw.
End
Choose sequence with higher sequence weight.
Stop.

The above discussed algorithm computes the sequence weight and selects the most weighted sequence to perform sequence selection.

Gene selection: In this stage, the method performs frequency analysis and obtains the sequence. The method iterates the frequency analysis to reduce the size of sequence. This will be iterated for number of times till the size gets reduces. The selected sequence will be used to produce the sequence at each time.

3. RESULTS AND DISCUSSION

In this paper we discusses about the iterative frequency analysis to perform gene selection. For the given sequence the method computes sequence weight to select the sequence and iterates the process for number of times to reduce the gene selection size.

Iterative Frequency Analysis Based Gene Selection

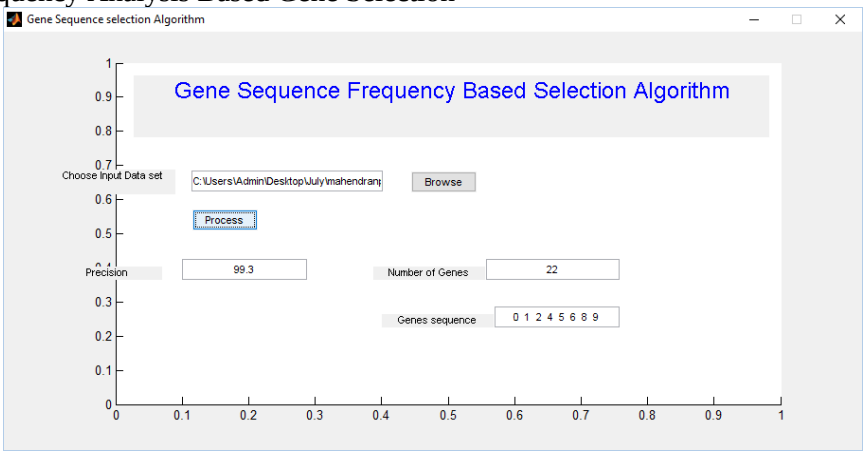

Figure.5.Result of gene selection

The Figure 5 shows the result of gene selection produced by the proposed frequency analysis approach.

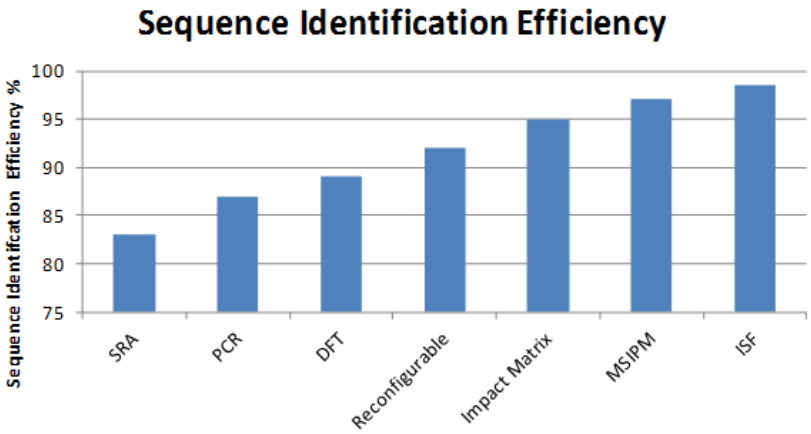

Figure.6. Comparison of sequence identification efficiency

The Figure 6, shows the comparison of sequence identification efficiency produced by different methods and it shows that the proposed method has produces higher efficiency than other methods.

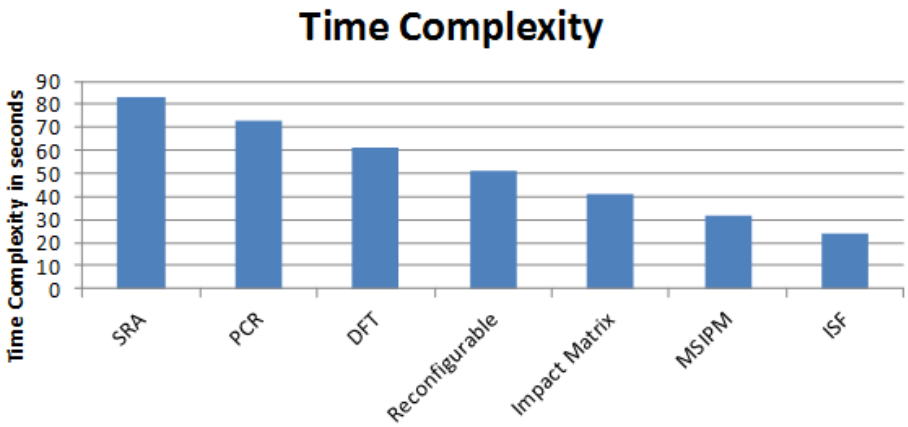

Figure.7. Comparison of time complexity

The Figure 7, shows the comparison of time complexity produced by different methods in identifying the sequence and the result shows that the proposed method has produced less time complexity than other methods.

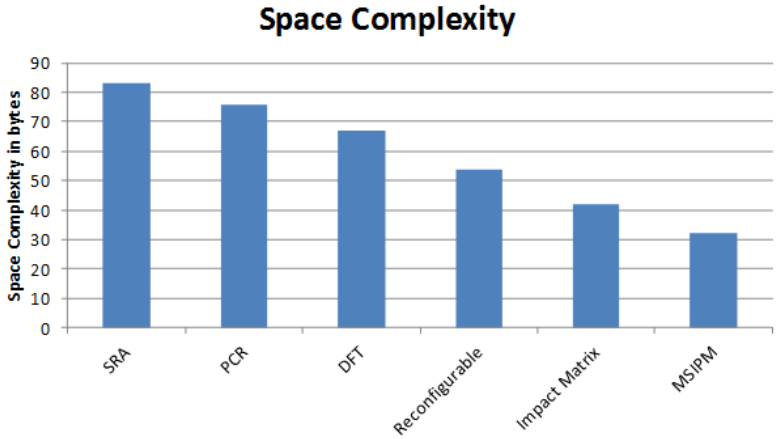

Figure.8. Comparison of space complexity

The Figure 8, shows the comparison of space occupied by the different methods and it shows that the proposed method has produced less space complexity than other methods.

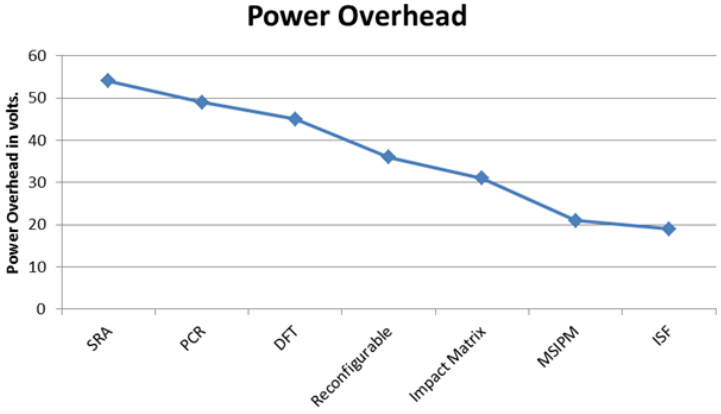

Figure.9. Comparison of power overhead

The Figure 9, shows the comparative results on power overhead produced by different methods and the result shows that the proposed method has produced less power overhead than other methods.

4. CONCLUSION

This paper presents the detailed information about the results produced by various methods. Each method has been implemented and evaluated for their efficiency. The method has produced efficient results on gene sequence selection and has produced efficient results. In this paper, we have improved the efficiency, time and space of Efficient DNA Sequence Analysis for Reduced Gene Selection Using Frequency Analysis by using a hardware software co-design approach. In addition, we compare proposed method with Bloom filter, PCR, SRA, DFT and Phylogeny Aware. As FPGA-based designs exhibit high performance for parallel computing and fine-grained pipelining,

REFERENCE

Surendar A, Arun M and Basha AM, Micro Sequence Identification of Bioinformatics Data Using Pattern Mining Techniques in FPGA Hardware Implementation. Asian Journal of Information Technology, 15, 2016, 76-81

Alachiotis N, Berger S A and Stamatakis A, Accelerating Phylogeny-Aware Short DNA Read Alignment with FPGAs, 19th IEEE Annual International Symposium Field-Programmable Custom Computing Machines (FCCM), 2011.

Arun M and Krishnan A, Functional Verification of Signature Detection Architectures for High Speed Network Applications, International Journal of Automation and Computing, Springer, 9(4), 2011, 395-402.

Che S, Boyer M, Meng J, Tarjan D, Sheaffer JW, Lee SH and Skadron K, Rodinia, A Benchmark Suite for Heterogeneous Computing, In Proceedings of the IEEE International Symposium on Workload Characterization (IISWC), 2009, 44-54.

Christopher MA, Thomas K F Wong, Lam TW, Hon WK, Sadakane K and Yiu SM, An Efficient Alignment Algorithm of Searching Simple Pseudoknots over Long Genomic Sequence, IEEE/ACM Transactions on Computational Biology and Bioinformatics, DNA sequencing, Biomedical Research, Special Issue, ISSN 0970-938X, 9(6), 2012. S75-S79.

Gabriel F Villoente, Mark Oliver L Ouano, Mary Grace C Dy Jongco, and Emilyn B Escabarte (2011), FPGA Based Agrep for DNA Microarray Sequence Searching, International Conference on Computer Engineering and Applications, IACSIT Press, 2, 2011.

Grigorios Chrysos, Agathoklis Papadopoulos and Geore Petihakis, Opportunities from the Use of FPGAs as Platforms for Bioinformatics Algorithms, Twelfth IEEE International Conference on Bioinformatics and Bioengineering (BIBE), 2012.

Lysaght P, FPGAs in the decade after Von Neuman Century, DATE06 Conference proceedings, Munich, Germany, 2006.

Pall Melsted and Jonathan K Pritchard, Efficient counting of k-mers in DNA sequences using a bloom filter, BMC Bioinformatics, 12, 2011, 333.

Surendar A, Arun M, FPGA based multi-level architecture for next generation, 2016.

Yoginder S Dandass, Shane C Burgess, Mark Lawrence and Susan M Bridges, Accelerating String Set Matching in FPGA Hardware for Bioinformatics Research, BMC Bioinformatics 2008, 9:197 doi:10.1186/1471-2105-9-197, April 2008.