# Efficient Pattern Detection in DNA Sequences using Frequent Itemset Mining and Random Forest

**Vishakha Hole[1] Deepika Nimbalkar[2] Sajida Shaikh[3] Komal Londhe[4] Prof. Ranjana Kedar[5]**
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]KJ College of Engineering and Management Research, Pune, India

*Abstract*— The information contained in the human genome is akin to a blueprint of the human body. This is due to the fact that the genes contain valuable information about the various processes required in the human body. This information is necessary to allow for the replication and replacement of the damaged cells in the body. The genes can also allow for the effective treatment of various diseases that can be vital for the human survival. Therefore, for this purpose the vast amount of information must be extracted from the genes through the DNA. It is a very extensive procedure to be performed manually, thus the Frequent Itemset Mining paradigm comes to the rescue. The Frequent Itemsets can allow for the identification of the defective or the frequently occurring genes effectively. Therefore, in this publication, the DNA patterns are extracted using the frequent Itemset Mining approach through the Linear Clustering and Entropy estimation to achieve the candidate sets. The resultant candidate sets obtained are then effectively classified through the Random Forest Classification. The methodology has been experimented extensively to reveal that it achieves significant improvements over the conventional approaches.

***Keywords:*** DNA Sequence, Pattern Mining, Frequent Itemset Mining, Random Forest Classification

## I. INTRODUCTION

Life on this planet earth has been in the making for a long time. Earth, our planet is in what is described as the Goldilocks zone where the heat of the Sun is adequate enough to keep the water in liquid form. If our planet was too far away all the water will not get enough energy and freeze up and closer than this would cause the oceans of the world to dry up because of the heat. Therefore, our planet has the proper conditions for a living organism to thrive. due to having the proper ingredients of life especially water single-celled organism's starter appearing and evolved through the ages into the biodiversity that we see now. The Genetic diversity of each and every organism is due to frequent mutations and overlapping as the population progresses. Therefore, genes are an important part of any organism's body.

Genes contain valuable information about the various processes that go on inside an organism's body. The genes contain codes for various proteins that are required to be manufactured for the proper functioning of a cell. The cells slowly build up into specialized organs inside an organism and form the body of the organism. Therefore, it can be extrapolated that the genes are basically the blueprint of an organism as the organism is usually duplicated by a single cell. This information is highly valuable as it holds the key to the various processes and the structure of the organisms. Human beings also have similar chemistry with which the genes dictate various features that are seen in humans.

This gene is located in the cells specifically in the nucleus and is the valuable information holders in the form of gametes that are used for meiosis division. These genes allow for the expression of various characteristics in humans such as eye color, ear lobes, etc. Genes contain hereditary information as they are a functional and physical unit of heredity. The genes themselves are made up of DNA or deoxyribonucleic acid. There are four different bases of DNA such as a b c and g. The various combinations and sequences of these proteins in the DNA form a particular gene. This gene has important information stored inside them that facilitates the creation of the various proteins that I needed for the survival of a human being.

Such genes also contain a lot of information in the form of genetic diseases. This information is highly valuable as it can be studied to determine the causes of various infections and diseases. These diseases can be predicted utilizing the various combinations that occur in the genes using the information that is stored. But the problem is that every single division will allow for a new combination of genes and there are currently millions of genes that are responsible for various characteristics displayed in humans. The Human Genome Project has been designed for this very categorization of all the genes that occur in humans which is a massive number. Therefore, such a large data set would only be processed by the use of big data and frequent itemset mining.

Frequent itemset mining is highly useful and an application for this purpose. The frequent itemset mining paradigm utilizes large amounts of data for the purpose of forming frequently encountered items. This frequent item can then be used to form a general idea of how the various items are interconnected to each other and the strength of those connections. Therefore, the application of this parrot for the purpose of mining frequent itemsets in genes can help identify the various inconsistencies in the genes. These inconsistencies are basically the various defects in the gene that can be analyzed effectively to help find the root cause of a particular disease or ailment easily.

The Random Forest Classifier is one of the most useful classification techniques that utilize the paradigm of a decision tree on a small scale to provide effective classification. The classification labels are provided as the byproduct of processing it through a specialized hierarchy of decision trees. Therefore, the random forest classifier can be imagined as a collection of a hierarchy of decision trees that collectively looks and behaves like an intricate forest. Thus, the random forest classifier is named as such due to the extensive collection of decision trees that provide an accurate and complete classification of the data. This property of decision trees is highly useful for the purpose of

 

implementing this classifier in the frequent itemset mining of the genes.

Along with the use of frequent itemset mining a linear clustering algorithm is also the suitable choice for performing clustering mechanisms on the frequently generated itemsets. The linear clustering along with the entropy estimation would help weed out the inconsistencies in the data this candidate set which is generated can be used for others to achieve classification labels. What the generation of classified labels the random forest classifier is one of the best for this purpose as it utilizes the extensive form of a decision tree to achieve efficient and effective labeling of the candidate sets. This allows for the effective mining of the entire DNA sequence through the use of frequent pattern mining.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

## II. LITERATURE SURVEY

B. Yimwadsana [1] explains that DNA is a popular research topic in biological computing. Using the biological DNA method of computing is done in DNA computing it utilizes synthesized DNA sequences to form DNA structures. DNA molecules or biological substances (sensor) are the well-known applications of DNA computing. In the DNA computing model DNA logic circuit is one of the popular models. Deoxyribonucleic acid (DNA) forms the structure of humans and animals. There are four types of DNA Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The proposed paper fulfilled the efficient design of DNA sequences for a DNA-based logic AND gate.

J. Zrimec [2] narrates the Deoxyribonucleic acid (DNA) forms the structure of humans and animals. Regulatory DNA regions driving gene expression, DNA replication and bacterial horizontal gene transfer are some of the basis of the DNA melting bubbles. Bubble genesis is labored and strained by DNA duplex stability and thermally induced duplex destabilization (TIDD). In DNA neighboring the melting site, the local DNA melting is assisted by intrinsically low duplex stability of the melting site. Thus, the proposed paper predicts 6 orders of magnitude shorter than that of PBD.

E. Tevanyan [3] elaborates on the sequencing technologies revealing different types of DNA secondary structures at the genome-wide scale. Nucleosome positioning is one of the mechanisms of transcription regulation. 1kB centered region on DNA structures is enough to analyze an association of nucleosomes and DNA structures. Three types of patterns are there in nucleosome profiles around DNA structures revealed a) the region around a structure is nucleosome-free b) structure is surrounded by nucleosomes from both sides c) a structure is surrounded by from one side. The paper aims to study the distribution of the discovered patterns in different tissues.

Z. Elyazghi [4] specifies that the DNA molecule is the process of determining the exact order of nucleotides DNA sequencing. The clip ends based on Phred quality scores, then recall N peaks by finding the local maxima at each ambiguous site. Vector sequence, polyA tails, or other unrelated sequences are the sequences of clones from DNA libraries. The main panel is composed of the following tabs chromatogram, sequence info, alignment, report, multiple corrections, instructions. It also allows the process of multiple data and saves results as a multi-fasta file. In the future implementation of further applications such as phylogenetic analysis is in the process to increase the utility of our program.

M. Sarkar [5] defines DNA computing is one of the powerful emerging technology to solve the hard-computational problem by computing using DNA molecules. DNA computing is inherent massive parallelism and extremely high data density and an extremely powerful, energy-efficient emerging technology. In the proposed paper they have used well known algorithmic solutions that exist for conventional computing architectures using a suitable ALU. Logic operations including NOT, OR, AND, XOR, NOR, NAND, and XNOR; compare, shift, etc. are used to perform a complete ALU (Arithmetic and Logic Unit).

Kiran [6] presents partial encryption of medical images based on dual DNA add up with a merged chaotic map is proposed. It is necessary to secure health care information transmitting between the different medical centers and computational time also important. To replace the traditional computer algebraic operation a new algebraic operation of DNA sequence has been introduced by DNA Computing. By matric computing for R, G, B component DNA addition operation to realize a DNA sequence. In proposed paper map selection is an important step in any digital chaotic encryption.

A. George [7] proposes complex biological systems that can be controlled using fuzzy logic operations with the help of linguistic rules. Minimum, maximum, and fan-out gates are the building block of the proposed paper a detailed design, analysis, and kinetic simulation of each gate was carried out. The pre-defined rules to design the fuzzy inference engine carried by using minimum and maximum gates are cascaded. For designing nanostructures and circuits for future medical applications they have used Deoxyribonucleic as it is the best option. The proposed paper can be used for the design of more complex non-linear mathematical functions.

Raposo [8] explains the working of DNA and DNA encodes the genetic information of most living beings. DNA is described by the textual sequence of its subsidiary molecules known as nucleotides not be described by its atomic structure. Nucleotides can be characterized in four parts adenine (A), cytosine (C), guanine (G) and thymine (T). DNA is one of the important topics for the researcher in the coming years. Some of the available software for DNA and RNA thermodynamics calculations are Visual OMP, UNAFold, RNAstucture, Vienna RNA Package. Thus, to handle the DNA molecules that mitigate somehow the complexity and resolution issues.

M. Sarkar [9] specifies that DNA Computing is an attractive topic for the researcher from inception in 1996. DNA performs on different important mathematical calculations. Examples of such mathematical functions are the greatest common divisor (GCD) and the least common

 

multiple (LCM). Using Euclid's Algorithms that take several stages roughly proportional to the natural logarithm of the larger number of GCD can be found on a conventional machine. Two DNA sequences merge via hydrogen bonding between each Watson-Crick complementary base pairs (A with T, and C with G), forming DNA double helix. Thus, in the proposed paper algorithm to apply DNA computing to perform LCM and GCD.

T. Huang [10] expresses an electroactive DNA intercalator is used for the real-time DNA amplification. There are two important most significant tools for many important applications are Nucleic-acid amplification and analysis techniques. The most major process amplifies gripping nucleic-acid fragments for increasing the amount of nucleic-acid analyte are polymerase chain reaction (PCR). The relationship of DNA intercalated redox reporters and the PCR cycles of DNA segments is demonstrated by electrochemical. Thus, the proposed paper to develop a rapid, low-cost and real-time machine for point-of-care diagnostics, environmental, food monitoring and popularize it in several new fields.

H. Kasahara [11] presents a powerful tool for diagnostic procedures in bacterial and viral infections called a Polymerase chain reaction (PCR). The proposed paper implemented a new electrical technique for the rapid detection of DNA amplified by PCR using dielectrophoresis (DEP). Dielectrophoresis (DEP) is an electrokinetic movement of a dielectric particle settled in the non-uniform electric field is used to manipulate the small materials such as DNA in biological applications. Thus, the frequency dependence of the microbeads labeled with different lengths of DNA is described in the proposed paper.

X. Zheng [12] expresses in DNA computing, the problem information is stored by DNA molecules and a computing system can be realized more reliable in two possible ways. Two ways are one way is to make the computing system free of errors and the second is a way to make sure the computing system is more reliable is to include redundancy components in the computing system and which can be achieved by the hardware and software. Thus, the RRNS are introduced into the DNA arithmetic operation and the DNA procedures for parallel arithmetic operation and it is realized error detection, even error correction, in DNA computing.

## III. PROPOSED METHODOLOGY

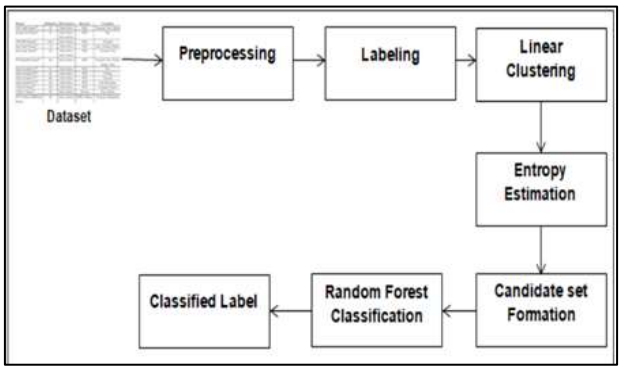


Fig. 1: Proposed model System Overview

The Proposed model for efficient pattern mining in DNA Sequence is depicted in the figure 1. And the steps that are involved in the building of the model are described in detailed with the below steps.

*1) Step 1: Data set Collection and Streaming-*

This is the preliminary step of the proposed model, where a DNA sequence dataset is being collected from the reputed Dataset repository like Kaggle from the URL: https://www.kaggle.com/trungv2/dna-tf-bound-classification.

This Dataset contains spreadsheets of the DNA sequence of ATGC, Where A stands for Adenine, T Stands for Thymine, G stands for Guanine and C stands for cytosine. These ATGC are actually represented the ribonucleic acids whose different patterns give different characters for the body cells.

So in the proposed model a spreadsheet that contains these DNA sequences are fed as the input to the proposed model. Once this spreadsheet is accessed by the proposed model, then the whole spreadsheet is read in the form of the objects using the help of the JXL external API of the Java programming language. As the spreadsheet is read, then it is streamed into a list data Structure for the further process of Pre-processing as mentioned in the next step.

*2) Step 2: preprocessing and Labeling –*

The obtained data list is then traversed for each of the rows to collect the DNA Sequence patterns. Once this pattern is received in the form of string, then this is tokenized for each of the four characters to decide them as the DNA blocks. The obtained DNA blocks are evaluated for the unique data and, then they are numbered based on the list indices in which they are stored.

*3) Step 3: Linear Clustering-*

Here in this process each of the unique labeled data is searched in the complete data string. The rows that actually contain the labeled DNA sequence is added into the cluster of that DNA label, this process eventually generates a linear cluster for the abstract DNA patterns of the labels which is depicted in the algorithm 1.

ALGORITHM 1: DNA Sequence Linear Clustering

//Input : Lablled List $L_L$ ,Data $D_L$
//Output: Linea Cluster $L_{CL}$
1: Start
2: $L_{CL=}\emptyset$
3: for i=0 to Size of $L_L$
4: $SN_L = \emptyset$ [$SN_L$ = Single Cluster]
5: SEQ1=$L_{L[i]}$
6: for j=0 to Size of $D_L$
7: $TMP_{LST}=D_{L[j]}$ [ Temporary List]
8: SEQ2= $TMP_{LST[0]}$
9: if (SEQ1=SEQ2), then
10: $SN_{L=}$ $SN_L$+ SEQ2
11: end if
12: end for
13: $L_{CL}$ = $L_{CL}$ + $SN_L$
14: end for
15: return $L_{CL}$
16: stop

 

*4) Step 4: Entropy Estimation-*

The obtained linear cluster from the past step is subjected to estimate the entropy for each of the labeled sequences using Shannon Information gain equation as shown in equation 1. Shannon information gain equation provides a distribution factor for the each of the labeled DNA Sequence. This distribution factor comes in a range of 0 to 1. If the value of any DNA sequence which is nearer to 1 indicates the importance of the DNA sequence, whereas if the values is nearer to 0 then it indicates the non-importance of the DNA sequence.

$$IG = -\frac{P}{T}\log\frac{P}{T} - \frac{N}{T}\log\frac{N}{T} \quad \text{_____(1)}$$

Where

P= Number of the cluster which contain Labeled DNA Sequence

T= Total Number of clusters

N= T-P

IG = Information Gain of the labeled DNA sequence.

*5) Step 5: Candidate set Formation –*

This is the most important part of the proposed model, here each of the labeled DNA sequence which is associated with the respective Shannon information gain values is sorted in descending order using the bubble sort algorithm. Once the labeled DNA sequences are sorted, then some top N sequences are selected for the further mutation process using frequent item set mining.

In frequent item set mining each of the labeled DNA sequence is associated with all other sequences to a form maximum frequency of 2. So this gives rise to form a frequent item sets which are containing maximum of 2 frequencies of DNA sequences. And then these DNA sequences are again counted in the respective linear cluster for their association to mark them with the respective weights in a double dimension list of two columns.

These counted frequent itemset list is subject to extract the top frequent itemsets for the further process of classification using random forest classification model.

*6) Step 6: Random Forest classification-*

The extracted frequent itemsets that are being generated are utilized for the formation of the tree structure using the Random Forest classification model. The root frequent itemset is set using the first candidate set that is encountered. The relative support values are then used to enable the assignment of the positions of the subsequent candidate sets.

The aforementioned steps are run continuously in a recursive approach to generate the complete and comprehensively sorted sub-tree through the algorithm 2given below.

ALGORITHM 2: Sub tree Formation

//input: Frequent itemset List $FI_L$
0: Start
1: $T_R$=Ø [$T_R$ = Tree]
2: $R_N$=Ø [$R_N$ = Root Node]
3: $R_N$= $FI_{L[0]}$ [ Root Node]
4: for i=0 to size of $FI_L$
5: $R_V$= $R_{N[1]}$
6: $TMP_{NODE}$= $FI_{L[i]}$ [ Temporary List]
7: VAL = $TMP_{NODE[1]}$
8: if (VAL <$R_V$ ), then
9: $T_R \Rightarrow TMP_{NODE\ (LEFT\ CHILD)}$
10: else
11: $T_R \Rightarrow TMP_{NODE\ (RIGHT\ CHILD)}$
12: end else
13: end for
14: return $T_R$
15: Stop

After formation of the sub tree the pre order approach is utilized for traversing the candidate sets. The nodes are traversed by initially visiting the ROOT, then the LEFT CHILD and then Finally the RIGHT CHILD. This procedure is essential for the collection of the candidate sets with similar support to generate the clusters of the frequent itemsets which are reliant on the aforementioned support. The obtained results are providing the best possible classification DNA patterns for the given input of the dataset.

## IV. RESULTS AND DISCUSSIONS

The presented system for the extraction of the frequent patterns in the DNA sequence have been implemented using the NetBeans IDE and coded in the Java programming language. For the execution of the proposed technique, a development machine is utilized which is equipped with Intel i5 processor realizing the processing tasks along with 500 GB of storage and 4GB of physical memory.

Extensive Experimentation was executed to evaluate the performance metrics of the proposed technique. For the purpose of measurement of the accuracy of the proposed system, the Precision and Recall performance metrics was analyzed which accurately underlines the performance characteristics of the proposed technique. The performance characteristics were analyzed to determine that the mining of frequent patterns in DNA sequences executed through the Random forest Classification proposed in this paper is implemented satisfactorily.

### *A. Performance Evaluation based on Precision and Recall*

Precision and Recall can allow for the extraction of useful information pertaining to the performance of the proposed system. The above metrics are one of the most in-depth and insightful procedures that are utilized for the extraction of the absolute performance of the system. Precision in this evaluation extracts the relative accuracy of the presented methodology by calculation of the accurate values regarding the level of precision accomplished by the technique.

Precision in this system is being measured as the ratio of the combined sum of all the proper DNA patterns identified along with the improper DNA patterns identified. Therefore, the assessment of the values of precision achieves an insightful evaluation of the effectiveness of the presented methodology.

The Recall parameters that are implemented for evaluating the absolute accuracy of the approach is significantly contrasting from the precision parameters given above. The Recall parameters are measured through the ratio of the number of DNA patterns identified versus

 

the total number of proper DNA patterns not identified. This approach towards the evaluation provides useful understanding as it calculates the absolute accuracy of the approach. Precision and recall are mathematically elaborated in the equations given below.

Precision can be briefly elaborated as below

- A = The number of proper DNA Patterns Identified
- B= The number of improper DNA Patterns Identified
- C = The number of proper DNA Patterns not Identified

So, precision can be defined as

Precision = (A / (A+ B)) *100

Recall = (A / (A+ C)) *100

The equations given above are implemented for performing intensive evaluation on the proposed technique through the analysis of the results of the Random Forest Classification. The evaluation of the experimental results is tabulated in table 1, given below.

| No of Trials | Proper DNA Patterns Identified (A) | Improper DNA Patterns Identified (B) | Proper DNA Patterns not Identified (C) | Precision | Recall |
|---|---|---|---|---|---|
| 38 | 33 | 3 | 2 | 91.66666667 | 94.28571429 |
| 54 | 50 | 3 | 1 | 94.33962264 | 98.03921569 |
| 79 | 75 | 3 | 1 | 96.15384615 | 98.68421053 |
| 119 | 114 | 2 | 3 | 98.27586207 | 97.43589744 |
| 131 | 122 | 5 | 4 | 96.06299213 | 96.82539683 |

Table 1: Precision and Recall Measurement Table for the performance of Random Forest Classification

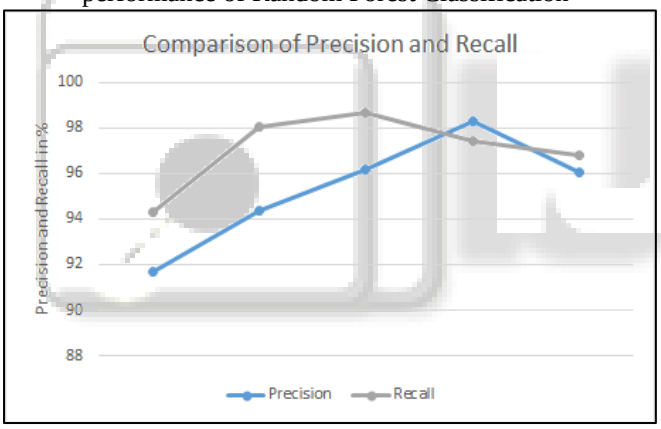


Fig. 2: Comparison of Precision and Recall for the performance of Random Forest Classification

The graph above represents that the Random Forest Classification in the proposed system attains appropriate values of precision and recall for the achieving the objective of DNA pattern identification. The methodology achieved a precision of 95.29% and Recall of 97.05% and this precision and recall performance metrics indicates that the Random Forest Classification model is realized with increased reliability and efficiency in the proposed technique.

These experimental outcomes illustrate that the Random Forest Classification module is performing precisely in the proposed technique and is extremely reliable in its execution. The Random Forest Classification is one of the essential modules in the proposed technique and the successful and accurate implementation of this module provides a significant improvement to the performance of the proposed frequent pattern mining mechanism for the DNA sequences which enhances the operation even further.

## V. Conclusion and Future Scope

The proposed methodology for the effective Frequent Itemset mining of the patterns in the DNA sequence has been elaborated in this publication. Genes are a highly important aspect of the human body as they contain valuable information related to human development and survival. It also contains the recipe for the ingredients or the proteins that make up the cells in the body. The DNA contains a vast amount of information that cannot be extracted through the conventional approaches. This is the reason why the genes must be studied in detail to unravel the information about the various processes in the human body. Therefore, the Frequent Itemset mining has been implemented to achieve the extraction of the frequent patterns in the DNA effectively. The methodology implements Linear Clustering and Entropy Estimation along with the Random Forest Classification of the candidate sets that are formed to reduce the time taken for execution. The experimental evaluation of the technique unravels the large potential of the approach which significantly improves on the previous approaches.

For future research prospects, the proposed methodology can be executed on the real-time DNA data. The presented methodology can also be implemented into an API for easier integration.

## References

[1] Boonsit Yimwadsana, Paramita Artiwet," On Optimizing DNA Sequence Design for DNA Logic AND Circuit" Proceedings of TENCON 2018 - 2018 IEEE Region 10 Conference (Jeju, Korea, 28-31 October 2018.

[2] Jan Zrimec and Aleš Lapanje," Fast Prediction of DNA Melting Bubbles using DNA Thermodynamic Stability" IEEE TRANSACTIONS ON JOURNAL NAME, MANUSCRIPT,2009.

[3] Elen Tevanyan, Maria Poptsova," Recognizing Patterns of Nucleosome and DNA Structures Positioning" ,IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018

[4] Zakaria Elyazghi, Loubna El Yazouli, Khalid Sadki, and Fouzia Radouani," ABI Base Recall: Automatic Correction and Ends Trimming of DNA Sequences" IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 16, NO. 8, DECEMBER 2017

[5] Mayukh Sarkar y, Prasun Ghosal y, Saraju P. Mohanty," Exploring the Feasibility of a DNA Computer: Design of an ALU using Sticker Based DNA Model" DOI /TNB.2017.2726682, IEEE Transactions on NanoBioscience,2009.

[6] Kiran, Parameshachari, Panduranga H T," Partial Encryption of Medical Images by Dual DNA Addition using DNA Encoding" Proceeding International Conference on Recent Innovations in Signal Processing and Embedded Systems (RISE-2017) 27-29 October 2017

[7] Aby K. George, and Harpreet Singh," DNA Implementation of Fuzzy Inference Engine: Towards DNA Decision-Making Systems", IEEE TRANSACTIONS ON NANOBIOSCIENCE, VOL. 16, NO. 8, DECEMBER 2017

 

[8] Adriano N. Raposo and Abel J.P. Gomes," Computational 3D Assembling Methods for DNA: A Survey" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS,2015.
[9] Mayukh Sarkar, Prasun Ghosal," Mathematics using DNA: Performing GCD and LCM on a DNA Computer" IEEE International Symposium on Nanoelectronics and Information Systems 2016
[10] Tsung-Tao Huang, Jun-Sheng Wang, Yu-Hsiang Tang, and Chun-Ming Chang," Electrochemical Real-Time DNA Amplification and Detection on a Microchip" Proceedings of the 10th IEEE International Conference on Nano/Micro Engineered and Molecular Systems (IEEE-NEMS 2015).
[11] Hiromichi Kasahara, Zhenhao Ding, Michihiko Nakano, and Junya Suehiro," Effect of DNA length on dielectrophoretic characteristics of DNA-labeled microbeads" 978-1-4799-7800-7/15, IEEE 2015.
[12] Xuedong Zheng, Bin Wang, Changjun Zhou, Xiaopeng Wei, Qiang Zhang," Parallel DNA Arithmetic Operation with One Error Detection Based on 3-moduli Set" IEEE Trans. Nanobiosci., vol. 11, no. 1, pp. 62-69, Mar. 2012.



 