IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 3, March 2015.
www.ijiset.com
ISSN 2348 – 7968

Detection of Copy Number Variations in Genomic regions

Rituparna Sinha[1], Prianka Kundu[1]

[1] Information Technology Department, Heritage Institute of Technology,
Kolkata, West Bengal, India

Abstract

Copy number variation (CNV) is a form of structural variation caused by duplications or deletions of a large DNA segment, which may have a vital impact on human health, causing many neurological diseases as well as cancer. Detection of genomic regions with copy number variation is a challenging task. Our method is based on next generation sequencing (NGS) technology's read depth approach. We have used smith waterman algorithm for alignment of the NGS based short reads and calculated the read count per window. The read count data may suffer from mappability bias which may lead to false detection of variants. To deal with this issue, we have introduced a k-mer based normalization technique. The smoothened read count data is then transformed to its corresponding statistical measure. Finally, we have applied clustering technique to identify the regions with duplications and deletions. Our approach achieves higher accuracy than existing packages with a low FPR value as well our method is applicable to both small and large genomic datasets.

Keywords: *CNV, NGS, Mappability bias, Structural Variations, Z-score, K-means.*

1. Introduction

Copy number variation (cnv) [2] is a type of structural variation (sv) [3] in the mammalian genome. It is also called genomic variation which refers to the duplication or deletion of the DNA segment. The number of copies of genomic DNA is referred by copy number. Structural variations i.e duplications or deletions can affect the DNA which is larger than 1kbp. Although there are other forms of structural variations but copy number variation is the major contributor of various diseases such as autism, schizophrenia, alzheimer disease, cancer, etc [8,9]. In earlier days fluorescence in situ hybridization (FISH) [10] and array comparative genomic hybridization (aCGH) [1] were used to detect copy number variations (cnv).

However these methods suffer from certain limitations i.e., they could not identify regions with translocations and inversions (another form of structural variations). The reason is that this is only chromosomal rearrangements and do not result any loss or gain of the genetic material. In addition to this array CGH [1] could not detect single base pair changes in DNA. Furthermore, these techniques can be affected by noise due to cross -hybridization between probe and target sequences. Now a days many methods of high throughput sequencing have been discovered named as next generation sequencing(NGS) [11,12]. NGS technology brought a revolutionary change in the field of life science. Next generation sequencing(NGS) [11,12] are also known as the high throughput sequencing methods that include the concept of massively parallel sequencing of millions of DNA fragments in a single sequencing run.

Next, we discuss some of the existing methods to detect structural variations. SegSeq [4] is a method which compares the tumor samples to the reference to detect copy number aberrations. SegSeq includes segmenting the chromosome. i.e it partitioned the genome into fixed size windows. The ratio between tumor sample read counts and reference read counts are observed for each and every window. If the ratio is greater than 1.5 then segments are called gains and if the ratio is below 0.5 then segments are called loss. However, in window based approaches window size is fixed and should be defined first. This can be a major disadvantage because the performance of the algorithm can be affected by the size of the window. rSW-seq [5] is another method which is based on the modification of smith waterman algorithm that is why it is called recursive smith waterman-seq. The read ratio i.e (total number of tumor reads)/(total number of normal reads) is observed. Copy number alterations is there where the read ratio varies from the expected read ratio. It uses the concept of moving window to generate the read ratios along the genome. Here the window size does not need to be specified previously. This is the main advantage of rSW-seq over seg-seq.

Another copy number alterations (CNA) detection method is CMDS [6] i.e correlation matrix diagonal segmentation.,

ISSN 2348 – 7968

which is a computationally efficient method. The intensity ratio is calculated between a target sample and a reference sample. Based on this the Pearson correlation coefficient is calculated between the chromosomal sites and this value is normalized by fishers transformations. If the value of the correlation coefficient is greater than zero then this result a square block along the diagonal. If the value of the correlation coefficient is zero then no square block is identified. In this way the recurrent copy number alteration is detected by searching the square block.

In our work, we propose a computationally efficient and statistically powerful approach to identify copy number variations (CNV) in a genome. Our method is based on NGS technology's read count based approach. We have used smith waterman algorithm for alignment of the short reads and calculated the read count per window. The read count data may suffer from mappability bias which may lead to false detection of variant. Hence the false positive rate (FPR) would increase. To reduce this we have introduced a normalization technique. This normalization is done on the read count data using some statistical measure. In our approach, to eliminate this mappability bias we have introduced a statistical technique. We have divided the reference genome into overlapping *k*-windows of 36bp (*k*-mers). Then, we have used smith waterman alignment algorithm again to align these reads back to the reference genome. Next, by using the concept of posterior probability we have calculated the probability of a read to get mapped to the reference genome. The reads which are uniquely aligned the probability of that window will be 1. Based on this we have smoothed the read count of those windows where multiple alignment occurs.

Next, we have used a statistical measure to convert the smoothened read count data into its corresponding Z-Score value. Finally, we have applied clustering technique to identify the regions with duplications and deletions. Our method is easily applicable for large data sets. In order to evaluate the performance of our method we compared it to the correlation matrix diagonal segmentation (CMDS), where our method performed better with respect to both large and small input size and with low false positive rate.

2. Methods

Our approach goes through a pipeline of processes, described below.

2.1 Input Data

Our work is based on next generation sequencing (NGS) where we have taken a standard reference sequence of size n. We have taken s number of samples. Each sample is associated with multiple reads generated from massively parallel sequencing technology (NGS). Reference sequence is then divided into w number of windows.

2.2 Alignment of short reads

Next, we have used Smith waterman algorithm [7] for alignment of the short reads back to the reference genome. We have divided the reference genome into w number of windows. By using alignment we can find similarity between query sequence and different database sequence. Smith waterman algorithm is a local sequence alignment algorithm, where similar sequence or even dissimilar sequence can be compared using local alignment method. Furthermore, smith waterman algorithm can find the local region with high level of similarity and can also align two partially overlapping sequences. In our approach, we have assumed the match score as +2 and the mismatch score as -1. However, the formula for matrix filling is.

$$M_{i,j} = \text{Maximum} [M_{i-1,j-1} + S_{i,j}, M_{i,j-1}+W , M_{i-1,j}+W,0] \quad (1)$$

After filling the entire matrix we traced back for an optimal alignment. Next, we have calculated the read count (number of reads mapped to a particular window) per window, where read count means how many reads are getting aligned to a particular window. In case of duplications, more number of reads will get mapped to a window. So, the read count of that window where duplication occurs will be automatically high. Similarly, the windows affected by deletion event, will have a low read count value.

2.3 Mappability bias

The read count data may suffer from mappability bias [14,15] which may lead to false detection of variant. Hence the false positive rate (FPR) would increase. To reduce this we have introduced a normalization technique. This normalization is done on the read count data using some statistical measure.

Mappability bias is the major bias that can affect the structural variation. Furthermore, due to the short read length a small number of reads are mapped to multiple positions. This can also happen due to the repetitive regions in the reference genome. However, some methods completely ignored this ambiguously mapped reads. Some methods use the technique of randomly assigning an ambiguous read to any one of the possible alignment positions. The problem is that this strategy suffers from false positive [17]. Failure to eliminate the mappability bias will lead to increase read densities within regions with higher mappability [16]. This can lead to spurious results. The amount of non-unique sequence in a genome directly affects read count.

ISSN 2348 – 7968

In our approach, to eliminate this mappability bias we have introduced a statistical technique. We have cut the human reference genome into overlapping *k*-windows of 36bp (k-mers) [19]. The term k-mers refers to all the possible substring of length k. However, we have used smith waterman alignment algorithm again to align these reads back to the reference genome. Next, by using the concept of posterior probability we have calculated the probability of a read to get mapped to the reference genome. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred. We have chosen a predefined threshold value for this purpose and the probability of a read to get mapped to the reference genome is calculated. If a particular read is repeated n times then the probability of those window where the read is repeated will be 1/n means at those windows multiple alignment occurs i.e reads are ambiguously mapped. Furthermore, the reads which are uniquely aligned the probability of those window will be 1. Based on this we have smooth the read count of those windows where multiple alignment occurs.

2.4 Normalizing the read count data

We have used a statistical technique called moving average to normalize the read count data.

We have calculated the Z- score of the normalized read count data using the following formula..

$$Z=(x-\mu)/\eth \qquad (2)$$

Where x is the value of the read count data of w window. μ is the mean of the read count data and ð is the standard deviation. For each sample we transformed the smoothened read count data into its corresponding Z-Score value. The genomic regions having duplications will have a high value of Z-Score and the regions having deletions will have a low value of Z-Score.

2.5 Clustering

Clustering is a process where the data items are partitioned to meaningful subclasses i.e the similar data items are grouped in one cluster and the dissimilar data items are grouped in another cluster. Homogeneity and separation are two main principle of clustering. Homogeneity means that the elements of the same cluster are maximally close to each other. Separation means that the elements in separate cluster are maximally apart from each other. We have used clustering to cluster the genomic regions into homogeneous group. We have considered three cluster here i.e the genomic region where the duplication has occurred belongs to one cluster, the genomic region where the deletion occurs belongs to another cluster, and the genomic region where no structural variations are identified belongs to a separate cluster. In our approach, we have used K-means clustering algorithm [20] for this purpose. We have used it for the analysis of single samples. The main idea of K-means clustering is to define k centroids. We have created 3 cluster so k=3. We have considered each point to our data set and associate it to the nearest centroid. This is an iterative algorithm. When all the data have been assigned, we have to recalculate the position of k centroid and this process continues until convergence [21]. This algorithm aims to minimize an objective function.

$$\sum_i \sum \| X_i - C_j \|^2 \qquad (3)$$

where i={1 …k} and $\| X_i - C_j \|$ is the distance between a data point X_i and the cluster center C_j. We have considered the Z-scores as the data points and Z-scores is partitioned to k clusters. However, where structural variations i.e duplication occurs the Z-scores of those window will be automatically high and they formed a cluster. Where deletion occurs the Z-Scores of those window will be low and they formed another cluster. Finally, the rest of the windows where there is no structural variation belongs to a separate cluster.

3. Results

In our work, we have simulated human reference DNA sequence by generating a reference sequence of size 100000 base pair of ATCG. We had generated 5 samples, among which in 3 samples, we had introduced duplications & deletions. We introduced 2 duplications, one from position 300-2000 bp and another from 8000-9500 bp in one sample. In the second sample we introduced only one duplication from position 500-2500bp. In the 3rd sample we had introduced a deletion event from position 7500-9000bp. To make the experiment more realistic we did not introduce any structural variations in the other 2 samples. Each sample is divided into fixed 36 bp short reads (Figure 1). Next, we have aligned the short reads with smith waterman algorithm, considering several parameters, i.e., match score as +2 and mismatch score as -1. All the reads (considering S=36 bp) were aligned back to the reference genome (considering w=100 bp fixed window size) and got the match score. Match score which is maximum considered as the best alignment. The read count per window were then calculated (Figure 2). We observed that the false positive rate would increase. Due to this we have developed a simple statistical method to correct the mappability bias. The reference were sliced into overlapping k windows (k=36) and the reads were again aligned back to the reference genome using smith-

ISSN 2348 – 7968

waterman algorithm. Next, we have calculated the probability of a read to get mapped to the reference genome. For example, if a particular read is repeated 30 times then the probability of those window where the read is repeated will be 1/30 means at those windows multiple alignment occurs i.e reads were ambiguously mapped. Probability 1 means that the reads were uniquely aligned. Our algorithm selects those windows for bias correction where multiple alignment occurs. Based on this observation, our algorithm uses a simple statistical correction to adjust the read count for the bias. After correction of the mappability bias we have calculated the Z-score (figure 3) from the read count data.

Once the read count were adjusted for each window in the genome, our algorithm applies K-means clustering to divide the genome into same contiguous region with the same copy number. We have chosen the parameter K as K=3. That means our algorithm have created 3 clusters. One is for duplication i.e the windows where the copy number gain occurs, forms one cluster. The second cluster is where the copy number loss i.e deletion occurs. Finally, those windows where there were no variation forms the third cluster (Figure 4).

CCAGACGCCTTGAATTAGTCAGCCGAGTACTCTAAA
CGCTTAGAGCGCATGGACATTGAAATCGTCCACCAA
ACATAGTCCGACCATCCACCCACGCGGTTGGCTAGC
TCCAGGGAGGCCGGACAAACCGAGTAGACGAATCAA
TAGGAGCCGTCACCCGGAGGTGACCTTTGGATCGAA
TTGCCAAAGTTAAGCAATAGACGAGGATCCGAAAGT
GTGAAAGGGTAAGGCACACTCATCTAGTAGTGCAAT
CGCGGTGAAGGTTATTTGTCAAATGCCGATAGCTGC
CTCGCTTACCGAACTACTGATGATAGATCGCACCTC
GTGCGAATAATGGACGTATCGTTTACGATGTTATAC
GACGTTGTGTTGCCTAGGTGACTTGATTTGTGATGC
ACAGGGAACTGCTACTCTACACGCGGTACACGTCAG
GATGGATAATTTTACAGCGAGAGTGAAATGGACGCT
TGCGGCATAGTCTATTTACACGATAGGAAATAGAAC
AATACTTTGGGGGTCACCTACAATACTCAGTGAGAG
GGGGGGTCGCCAGTTTTGTAATGTGAGTGGGACCGG
ATAGACCGGTCCCAGTAAGAAGAAGACAATTATAGT
CGCAATGTTGTCTAACCAGGGATGTCTATAAACGTC
GCAGGAGATATGCACGGCTTCGCTGAGGATATTCGG
TACTAGGCCATTGGTAGGCTAAGGTCGTTCATGCAA
AACCTTGAGGGATAGAAATAGGCCACTGAGATACTT
TCGTAGCATTGACGCGGACCGTCGTATTTGGACATG
ATAGCGAACCAAGGGCCAAACTACCTCTTCTCGCTG
CTACCCGGGAGTGCGGCCGGCGGGATTCTTGATTAG

Fig 1: Samples are sliced into 36bp reads. These are called the short reads.

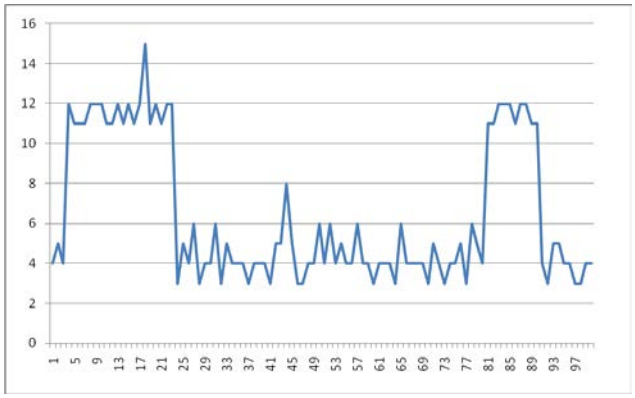

Fig 2: The read count per window is shown by this diagram. The X axis denotes window and Y axis denotes the read count value. In the first sample we have added two duplications. One from position 300-2000 and another one is from 8000-9500. Figure shows that the read count value where the duplication occurs was high.

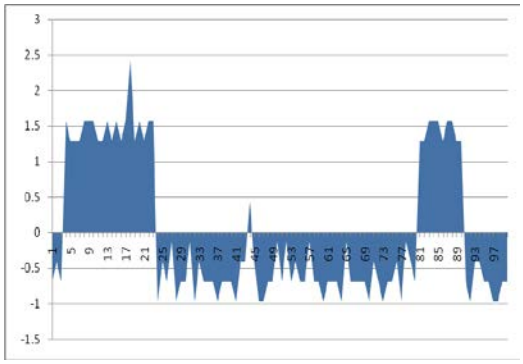

Fig 3: Figure shows the Z-score corresponding to the read count value where X axis denotes the window and Y axis denotes the corresponding Z-score.

ISSN 2348 – 7968

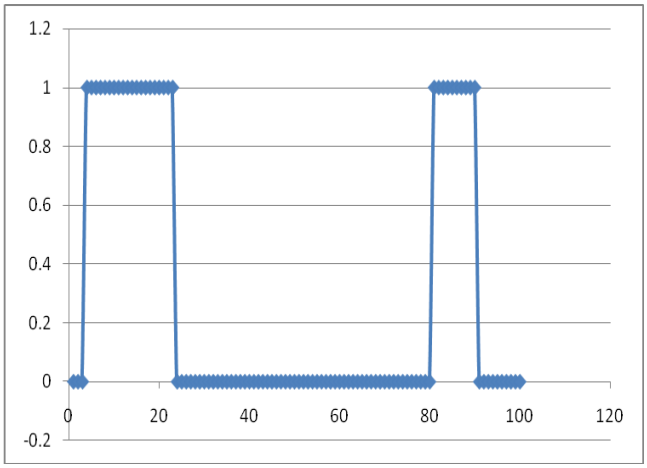

Fig 4: Clustering divided the genome into same contiguous region with the same copy number. We have added two duplications. One from position 300-2000 and another one is from 8000-9500. Figure shows that where duplication or copy number gain occurs forms one cluster and the rest belongs to another cluster.

4. Comparison

In order to evaluate the performance of our algorithm, we compared it to the CMDS (correlation matrix diagonal segmentation). CMDS takes the copy number data of m physically ordered chromosomal sites of n individual samples as an input. Then pearson's correlation coefficients is calculated between the chromosomal sites. If the value of the correlation between the sites is greater than zero then it will cause a positive correlation and it will result a square block along the diagonal of the matrix. The copy number regions can be detected by searching the square block. CMDS performs a diagonal transformation and calculate a RCNA score based on the correlation values. Though CMDS is statistically powerful approach but the statistical power of it depends on the block size. The block size needs to be specified for CMDS analysis. If the block size is not optimal then the false positive rate (FPR) would increase as well as the method will loss its sensitivity.

CMDS and our algorithm both showed >90% of sensitivity at detecting large sized alterations. With low sequencing depth our algorithm showed low false positive rates compared to CMDS.

We have calculated F-Score which considered both the precision and sensitivity to compute the score. The higher the F-Score, the better predictive power of the classification procedure. We have added 5 duplications in both the algorithm and observed that the false positive of CMDS is comparatively higher than our method. We also calculated the true positive rate (TP/TP+FP) and precision (TP/TP+FN). We observed that our method shows 83% sensitivity in detecting copy number variation (Table 1). Furthermore, the F-Score of our method is comparatively higher than CMDS.

Table 1: The overall Precision, Sensitivity and F-Score obtained by the two compared algorithms.

	Precision or positive predictive value	Sensitivity Or true positive rate (TPR)	F-Score [2*(precision*sensitivity/ precision +sensitivity)]
CMDS	0.62	0.714	0.66
Our Method	0.833	0.833	0.83

5. Conclusion:

In this work, we propose a computationally efficient and statistically powerful approach to identify copy number variations (CNV) in a genome. Our method is based on NGS technology's read count based approach. We have used smith waterman algorithm for alignment of the short reads and calculated the read count per window. The read count data may suffer from mappability bias which may lead to false detection of variant. Hence the false positive rate (FPR) would increase. To reduce this we have introduced a normalization technique based on k-mer approach as described in Method Section. Next, we have used a statistical measure to convert the smoothened read count data into its corresponding Z-Score value. Finally, we have applied clustering technique to identify the regions with duplications and deletions.

We compared our method with CMDS (correlation matrix diagonal segmentation) which is a popular copy number variation detection method. Correlation matrix diagonal segmentation (CMDS) is a computationally efficient and statistically powerful approach for detection copy number alterations. In CMDS, We observed that If the block size is

ISSN 2348 – 7968

not optimal then the false positive rate (FPR) will increase as well as the method will lose sensitivity in the estimation of RCNA score. With low sequencing depth our algorithm showed low false positive rates compared to CMDS. We have calculated the sensitivity and F-Score in both the methods. Our method shows >84% sensitivity in detecting copy number variations. Furthermore, our algorithm shows higher F-Score that means better predictive power of the classification procedure.

The limitations of our algorithm is that clustering of the genomic regions done using k-means algorithm is sensitive to the initial selection of centroids. It is also sensitive to outlier data points. As a solution, genetic algorithm can be used to overcome the issues.

6. References:

[1] Lai WR, Johnson MD, Kucherlapati R, Park PJ, "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data", Bioinformatics 2005, 21:3763-3770.

[2] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, Zhongming Zhao, " Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives", IEEE International Conference on Bioinformatics and Biomedicine Philadelphia, PA, USA. 4-7 October 2012.

[3] Junbo Duan, Ji-Gang Zhang, Hong-Wen Deng, Yu-Ping Wang, "Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies", PLoS ONE 8(3): e59128.

[4] Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES, "High-resolution mapping of copy-number alterations with massively parallel sequencing" Nat Methods 2009, **6:**99-103

[5] Tae-Min Kim1, Lovelace J Luquette1, Ruibin Xi1, Peter J Park1,2,3*, "rSW-seq: Algorithm for detection of copy number alterations in deep sequencing data", Kim et al. BMC Bioinformatics 2010, 11:432

[6] Qunyuan Zhang, Li Ding, "CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data", Advance Access publication December 23, 2009, Vol. 26 no. 4 2010, pages 464–469.

[7] Smith waterman, Waterman MS, " Identification of common molecular subsequences" J Mol Biol 1981, 147:195-197.

[8] Frohling S, Dohner H," Chromosomal abnormalities in cancer", N Engl J Med 2008, 359:722-734.

[9] Albertson DG, Collins C, McCormick F, Gray JW, "Chromosome aberrations in solid tumors", Nat Genet 2003, 34:369-376

[10] Pinkel D, Albertson DG, " Array comparative genomic hybridization and its applications in cancer", Nat Genet 2005, 37(Suppl):S11-S17

[11] Mardis ER," The impact of next-generation sequencing technology on genetics", Trends Genet 2008, 24:133-141

[12] Jorge S Reis-Filho, "Short communication Next-generation sequencing", Breast Cancer Research 2009, 11(Suppl 3):S12

[13] Gu¨ nter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arne´ Clevert, Andreas Mitterecker, Ulrich Bodenhofer and Sepp Hochreiter," cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate", Nucleic Acids Research, 2012, 1–14.

[14] Christopher, Miller1, Oliver Hampton, "ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads", PLoS ONE 6(1): e16327.

[15] Schraga Schwartz., Ram Oren, Gil Ast, "Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads", January 31, 2011, PLoS ONE 6(1): e16685.

[16] Derrien T, Estelle J, Marco Sola M, Knowles DG, Raineri E, Guigo R, Ribeca, "Fast computation and applications of genome Mappability",PLoS ONE 2012, 7:e30377.

[17] Lee H, Schatz MC: "Genomic dark matter: the reliability of short read mapping illustrated by the genomemappability score", Bioinformatics 2012, 28:2097–2105.

[18] Koehler R, Issac H, Cloonan N, Grimmond SM "The uniqueome: a mappability resource for short-tag sequencing", Bioinformatics 2011, 27:272–274.

ISSN 2348 – 7968

[19] Wentian Li1, Jan Freudenberg1 and Pedro Miramontes2, "Diminishing return for increased Mappability with longer sequencing reads: implications of the k-mer distributions in the human genome", 3 January 2014, Li et al. BMC Bioinformatics 2014, 15:2.

[20] K. Alsabti, S. Ranka, and V. Singh " An Efficient K-Means Clustering Algorithm.".

[21] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE, IEEE TRANSACTIONS ON "Pattern analysis and machine intelligence" ,VOL. 24, NO. 7, July 2002

Rituparna Sinha, has completed M.Tech in the year 2008, from National Institute of Technical Teachers Training and Research Institute (NITTTR) , in the Department of Computer Science and Engineering. Currently, she is employed as an Assistant Professor , in the Department of Information Technology at Heritage Institute of Technology (Kolkata).
She has 8 years of Teaching experience. Her research interest include Bio-Informatics and Data Mining. She has 1 International Journal with title "A Tutorial on Spatial Data Handling" and one International Conference with title "Hierarchical Data Structures for Accessing Spatial Data".

Prianka Kundu, has completed B.Tech in the year 2011, from Camellia Institute of Technology in Computer Science and Engineering Department. Currently she is pursuing M.Tech (Final Year) from Heritage Institute of Technology in the Department of Information Technology. She worked as a software trainee in Novel Research and Development India private ltd.