A Grid based Mining Approach to Genomic Data Set

S. Jessica Saritha
Dept of CSE,
JNTUA CEP Pulivendula-516390
Andhra Pradesh, India

P. Govindarajulu
Department of CSE
Sri Venkateswara University
Tirupathi, Andhra Pradesh, India

ABSTRACT

An improvement to the processing efficiency of genomic data sequence for automated detection and diagnosis is presented in this paper. For the automation of genomic signal processing, the problem of representation, extraction and retrieval is proposed. In the current form of automated genomic processing system, the retrieval of the gene information depends on the representation of the gene sequence. The retrieval accuracy also depends on the training data sets used. To achieve e the accuracy of retrieval it is required to represent the informatics regions more accurately and extract the relevant matching faster. To achieve this objective in this paper, a grid based computing approach to the distribution genomic dataset is proposed, with sequence shuffled, information region prediction filtration. A faster and accurate retrieval is obtained by the usage of sequencing, filtration and grdification modeling.

Keywords

Data mining, genomic signal processing, spectral sequencing, region prediction, grid computing.

1. INTRODUCTION

Data mining is an open area application in various usages. Wherein past, mining approaches were limited to few areas of applications such as scientific application, security concerns, astronomical applications etc., in recent past it has been applied to more common usages such as web mining, medical applications, e-learning etc. Among various usage of data mining, the usage of these algorithms under medical applications is observed in current developments [1-4]. For the application of medical data mining the process of feature description, grouping, extractions are more focused. Most of the medical mining approaches are developed over medical images, developed for the automation of captured medical scan images for diagnosis [3, 4]. Where in focus is also made towards the improvement of mining architecture [4] by predictive logic. With the development of technologies, medical sciences are now applied to various advanced applications, such as forensic science, bio informatics, gene processing etc. In recent years processing of gene sequence for biometric applications has emerged. Using the gene sequence, various information, such as health disorder, genetic issues, identifications etc. are revealed. Though gene processing is an advanced and accurate approach, retrieving gene information in reference to a query is highly time consuming, as the conventional database are very large in count. The basic representations of gene sequence are variant with few combinations, which results in very large set of information's in the database. A large information, where results in better retrieval, they take very large time for a decision, which extends from few hours to a day or so. To improvise the efficiency of processing in genomic signal processing a mining approach following pair wise coding was proposed in [5]. However the logic suggested, basically focus on the mining of gene data set and result in a retrieval. A dimensionality reduction is suggested by the approach of pair wise coding. However even after a pair wise coding it could be observed that the information counts are very large. It is observed in a gene sequence that there are two basic regions, the 'intron' and exon' region. Where in exon region (protein coding regions) reveal the region of information's, intron regions should be removed. In the process of genomic signal processing filtration approach [6] to segregate exon region is presented. In the process of exon region segregation various works were developed in past [7]. These approaches are DSP based coding techniques which filter the numerically represented Gene sequence into filter signal based on period-3 periodicity. Anastassiou [8, 9] presented an optimized spectral content measure based on windowing DFT for exon detection. Vaidyanathan and Yoon [10] proposed the use of IIR anti-notch filter. Rao and Shepherd [11] proposed the use of Auto-regressive (AR) model for detection of 3-periodicity for DNA sequences. These representations are then coded for its equivalent binary representation. These representations are achieved via a 4 base binary representation of '1' and '0' coding for presence and absence of character [12]. Once the gene sequences are coded with exon region in binary sequence, these are recorded into a database for learning and to utilize in future application. However these recorded information's are very large data sets and mining information out is a large time consuming task. Various past approaches are made [13] to reduce the computational overhead, however due to the uneven periodicity of the binary pattern and distributed data set these methods get limited. With this objective in this paper a pattern aligned coding, with spectral similarity and a Grid based coding technique is proposed to overcome the mining overhead in genomic signal processing. To present the proposed work, this paper is outlined into 8 sections. An outline to the genomic signal coding, and method to region prediction is proposed in section 2. Section 3 outlines the representation approach to the gene sequence and conventional modeling of gene mining approach. The proposed approach of pattern alignment is outlined in section 4. A grid based mining approach to this aligned sequence is then presented in section 5. The experimental results for the developed mining approach are presented in section 6. A conclusion to the developed work is outlined in section 7.

2. REGION PREDICTION CODING

Genomic signal processing (GSP) is the engineering branch that studies the content of the genomic signal and explains the production of mRNA and proteins, which are carried out by the genome. Based on today's technology GSP focuses on obtaining the gene information from analyzing the gene. The processing of genomic signal is nothing but analyzing the gene, and processing it to extract various information which will be useful for the observation. The purpose of the GSP is to combine the theory and various methods to process the genomic information, so that it will be useful. Therefore, GSP includes different techniques regarding expression profiles: detection, prediction, classification, control and statistical modeling. GSP is a essential engineering branch that studies the analysis and process of genes which is based on a synthesis model which involves a meticulous mathematical approaches. As the information's are very important in analysis and predication, the information retrieval accuracy is of major importance. To obtain optimal retrieval accuracy, the gene sequence needs to be represented and processed for retrieval. Genome sequences are large and can range from a number of million base pairs in prokaryotes to billions of base pairs in eukaryotes. DNA consists of long sequences of four kinds of nitrogenous bases {A, C, G, T }.These sequences are clustered in the coding regions-exons(eukaryotic genes) and non-coding regions-introns (such as promoters, enhancers, silencers, etc.).The process of exon region prediction is outlined in [16]. Agene sequence consists of four bases and such a chain can be mapped to four signals. The string that is composed of four bases is mapped into four binary signals. The value of '1' is taken by the signal $b_A(n)$ in the case if A is present in the DNA sequence at index n. But if it is not the case that is if A is absent at index n the value of '0' is taken. For example, for 'CGTCGTGGAA' subdivision of DNAisgivenas"0000000011". Likewise the signals, $b_T(n)$, $b_G(n)$ and $b_C(n)$ can be obtained. Later the DFT of $b_A(n)$, $B_A(f)$ over W samples is found. DFT of $b_T(n)$, $b_G(n)$ and $b_C(n)$, is labeled as $B_T(f)$, $B_G(f)$ and $B_C(f)$ respectively. Period-three behavior is observed in several genes and is also found which is very useful for recognizing the coding regions. It is also observed that theperiod-3 property indicates the location of genes more accurately. For this reason, the $(f = N/3)$-DFT coefficient magnitude is frequently considerably larger when compared to the surrounding DFT coefficient magnitudes and this corresponds to a coding region of the gene with in. With this approach for exon prediction the concentric higher magnitudes are extracted. This predicted region reduces the coefficients to process. This is one initial approach for faster mining via dimensional reduction.

For a genome sequence with K characters, the transformed binary pattern B(n) is 4*K. These patterns after extraction reduces to (4*K)/M, where M is the number of peak coefficients selected after DFT threshold process[16]. The process of exon region prediction is carried out on Human prothrombin (F2) [24] gene sequence. For a selected sequence coefficients the obtained exon region estimates is as shown below.

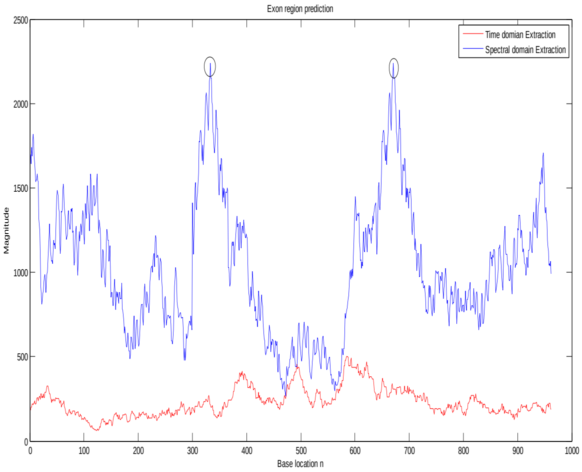

Figure 1: Derived Exon regions for window size of 270

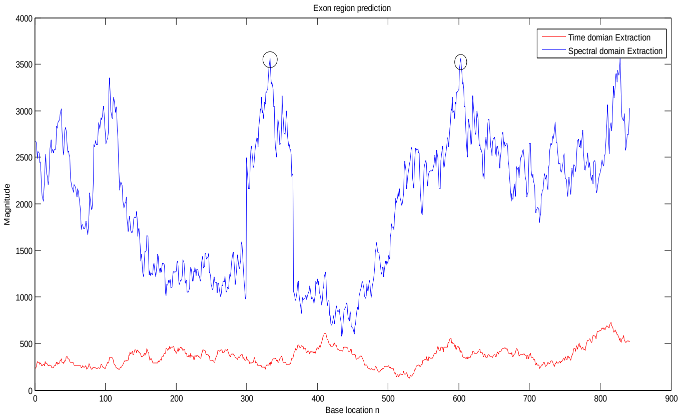

Figure 2: predicted Exon region for window size of 510

The spectral peaks could be observed more isolated in case of the DFT processed signal. The region of prediction is set to 60% of the peak value. A threshold is set and a value of 0 or 1 is coded based on the crossing limit of the threshold value. This binary sequence represents the first sequence representation of gene sequence after filtration. These sequences present the only proteins region of the gene sequence called exon. These binary patterns are recorded as gene sequence for A,C,T,G sequence in a data Base. For the future usage of these recorded data, it is required to have a mining system to extract the match details to represent genome information.

3. DATA MINING IN GENOMIC APPLICATION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining consists of collecting and managing data, analysis and prediction. The two major benefits of data mining are decision support and application development. In decision support, mining of a system can transform the data to reveal hidden information in the form of facts, rules and graphical representations. The extremely large amounts of data are, thus, compressed to highlight inner relationships between different elements. Towards the application of data mining for retrieval of genomic information from the data set, various data mining approaches were proposed in past. Among the well-known techniques of DNA-string matching are the Smith-Waterman algorithm [2], [3] for local alignment, the Needleman-Wunsch algorithm [4] for global alignment, Hidden Markov's model, matrix model, evolutionary algorithms for multiple sequence alignment [5] etc. were well known approaches. These works, though extremely valuable, have their limitations. The demerits include the use of complicated matrix algebra and dynamic programming, and the results of sequence matching are not free from pre-calculated threshold values. It is to be noted that none of the above-mentioned methods can be directly employed to identify the species from the structural signature of the genomes. To overcome this problem, in [17] a mapping approach based on neuro modeling called 'SOFM' is proposed. This approach presents a PCA based dimensional reduction technique with neural network for gene sequence prediction. The author presents the Eigen feature descriptor for gene sequence. A covariance for the normalized mean sequence is proposed [17].

$$cov(a_i, a_j) = C_{ij} = \frac{\sum_{k=1}^{64}(a_{ik} - \bar{a}_i)(a_{jk} - \bar{a}_j)}{(n-1)}$$

The covariance factor is defined as a distance measure for the gene sequence. This covariance factor is used for neural system learning and evaluation. Wherein this approach is an advanced approach towards applying advance learning approach for gene mining, the approach suffers with high computational complexity. The system applies the PCA approach for feature description and selection. As the gene sequence data bases are basically very large in dimensional, applying PCA over such data set is a computational overhead. In addition due to large data set the neuro learning and testing becomes very complex and slower. To optimize the mining performance in this paper we present a distributed approach for large data set mining by introducing a grid based mining approach with modified sequence representation approach.

4. PATTERN ALIGNED CODING

As the gene patterns are represented in binary sequences, which are buffered as a large data set. These patterns are varied by few bit transitions, and the correlative nature is very high among each sequence. Due to the self-correlative nature among multiple sequences the probability of correlative error and miss-classification is high. The correlative property in binary pattern is observed by two well-known approaches, Hamming distance (HD) or Levenshtein Edit Distance (LED)[18]. Wherein the hamming distance is a measure of co-relative pattern match count defined by,

$$d_{HD}(s, s_1) = \sum_{s_{1i} \neq s_{2i}} 1, \quad i = 1, \ldots \ldots m$$

Where, d_{HD} measures the number of varying positions intwo strings s_1 and s_2 of equal length m.

for example,

s_1= ”ATCCATGC” and s_2 = ”ATGGATAC”

s_1 : ”ATCCATGC”

s_2 : ”ATGGATAC” =>d_{hm} = 2.

However such sequences measurement is carried out for equal length sequencing. In gene sequence however, the pattern insertion, deletion, or substitution is observed. In such a case Hamming distance computation is not effective. In such case Levenshtein Edit Distance (LED) is computed. The LED is the minimum number of edit operations to transform a string s_1 of length n into a string $s_2$of length m. For the processed signal with insertion, deletion or updation the d_{LED} is computed as,

$$w(a_i, b_j) = \begin{cases} 0, & if\ a_i = b_j \\ 1, & if\ a_i \neq b_j, substitution \end{cases}$$

$$d_{LED}(i,j) = min \begin{cases} d_{LED}(i-1, j-1) + w(s_{1i}, s_{2i}) \\ d_{LED}(i, j-1) + 1 & insertion \\ d_{LED}(i-1,) + 1 & deletion \end{cases}$$

For example, for two given pattern sequences, S_1 and S_2 ,

S_1 = ATCTGA

S_2 = ATTA

Then d_{LED} = 2 , as C in S_1 is inserted at 3 and G is deleted at 5.

In the process of pattern matching the hamming distance provides the similarity index and LED provide the cross correlative counts. However, applying these approaches for measuring distances, the bit pattern plays an important role. It is observed that more the pattern transition exist the probability of pattern matching error get higher. The pattern matching errors could be minimized via updated sequencing of binary patterns. To develop such approach in recent past [19] a 1-D Local binary pattern (LBP) representation of signal is presented. A LBP is locally transformed binary pattern used for the optimization of pseudo random nature of a binary pattern. They are more dominantly used as a filtration approach in image processing, signal processing, speech processing etc.[19].A LBP sequence is generated as uniform patterns if they have at most two circularly bitwise transitions from 0 to 1 or vice versa, and non-uniform patterns if otherwise. In uniform LBP mapping, one separate spectral analysis using histogram bins is used for each uniform pattern and all non-uniform patterns are accumulated in a single bin. By grouping the non-uniform patterns into one label, the noise in non-uniform patterns is suppressed. The number of patterns is reduced significantly at the same time.

For example, “11110000”is a uniform pattern as *U* = 2, whereas “01010111” is a non-uniform pattern as *U* = 6. With the realignment of such pattern the error robustness is achieved, the LBP is hence robust to alignment error [20]. With this approach the LBP coding is applied over sequence of gene pattern, and the patterns are categorized into uniform and non-uniform patterns to reduce alignment and matching error. The spectral feature of the pattern distribution using Histogram bin is computed and the patterns are realigned to achieve lower transition matching error. For this local binary patterns the sequence are stored as a data set, and mining is then carried out over such binary pattern. However these patterns are very large data set, though the pattern are aligned to minimize the transition error, they need to processed with advanced technologies to mine the information faster.

5. GENOMIC GRID COMPUTING

Towards improving the mining performance for large data set, distributed computing have emerged as an optimal approach for fast processing system. In distributed computing, grid computing is an area of emergence, wherein distributed data are processed together to achieve the objective of faster processing. A basic model for the grid mining approach is as shown in Figure 3.

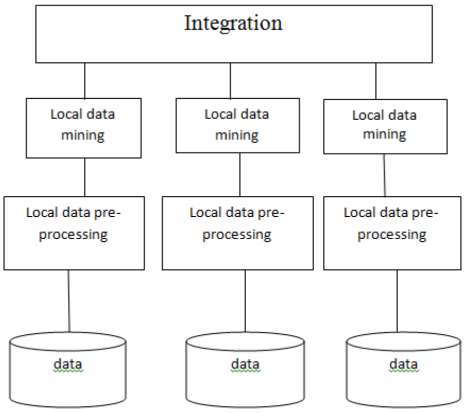
Integration
Local data mining
Local data mining
Local data mining
Local data pre-processing
Local data pre-processing
Local data pre-processing
data
data
data

Figure 3. A Data mining approach to grid based computing [15]

In the process of grid mining, various sets of data are buffered at different locations and all these data sets are controlled via a common node. An isolated processing is carried out and the observation are integrated to develop the final result. In the process of genomic signal mining, various mining approaches based on grid computing are under the research in present scenario. Various approaches such as BLAST [4], FASTA [4], Profile Search [4] etc. are few of them. In these approaches the mining approach finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. In such an approach the distributive data sets are processed in parallel to achieve the objective of fast mining. In [21] the approach of grid computing for gene expression detection is proposed. This approach develops a dimensional reduction technique using association studies. A new grid computing approach for gene mining was developed in recent past, termed Multifactor Dimensionality Reduction (MDR)[21]. This approach focus on developing new dataset based on genomic cross validation method. For a k-subset and the sets are labeled as high to low risk based on the case control ratio. The min max approach is defined for the computation of MDR algorithm. A multi locus analysis based on load and nodes is proposed. The Association approach is however defined based on the offered load for processing. The pattern of data passed is not observed in this processing. The inter-relation of data information is also an important factor of analysis in association based mining. Hence in this work, a modified association based system using information inter-relation for grid mining is developed.

In the process of grid data mining, the operation is majorly divided into 2 tasks. In the 1st task, it is required to bifurcate the volumetric data into lower sub data sets for distributed mining. In the 2nd task the mining process on distributed approach using parallel execution is made. For the segregation of the volumetric data set into sub-data set, the processing dataset is initially clustered into k-sub data sets. Wherein the clustering approach is performed by the spectral similarity method of the individual pattern defined by,

1. No of switching states out of all bits. i.e. how much number of times the bits changed their state out of all bits.

2. The symmetry property.

3. The transition time taken from one bit to next bit.

To merge or split subsets of patterns rather than individual patterns, the distance between individual patterns has to be generalized to the distance between subsets. Such derived proximity measure is called a linkage metric [14]. The type of the linkage metric used significantly affects associate algorithms, since it reflects the particular concept of closeness and connectivity. Major inter-cluster linkage metrics include single link, average link, and complete link. The underlying dissimilarity measure (usually, distance) is computed for every pair of patterns with one point in the first set and another point in the second set. A specific operation such as minimum (single link), average (average link), or maximum (complete link) is applied to pair-wise dissimilarity measures $d(c_1,c_2)$ defined by,

$$d(c_i c_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

$$where, (x_i\,, y_i\,) \in c_i \; and \; (x_j\,, y_j) \in c_j$$

Where x,y are the pattern sequences to be grouped subjected to,

$$\min\left(d(c_i c_j)\right)$$

Based on the pair-wise dissimilarity measure initial data clusters are created. An iterative operation is then carried out to optimize these sub-sets so as to achieve faster processing. The iterative optimization clustering starts with an initial partition. Estimates of this partition is then improved by applying a local search algorithm to the pattern. The iterative Refinement algorithm for sub set refinement is as defined below,

***Algorithm:* Refinement Process**

Input: The number of clusters K, and a database containing *n* objects from some R*p*.

Output: A set of K clusters, which minimizes a criterion function J.

Step 1. Begin with initial K centers/distributions as the initial solution.

Step 2. (Re) Compute memberships for the data points using the current cluster centers.

Step 3. Update some/all cluster centers/distributions according to new member-sips of the data points.

Step 4. Repeat from Step 2. Until no change to J or no data points change cluster.

Using this framework, iterative methods compute the estimates for cluster centers, which are rather referred to as prototypes or centroids. The prototypes are meant to be the most representative points for the clusters. The mean and median are typical choices for the estimates. The estimate computes a set of parameters that maximizes the likelihood of the chosen distribution model for a data. These data sets are then processed in parallel to retrieve information's from the data set. For the process of parallel execution of sub data sets different approaches were developed in past. Among various such approaches, "PAGE", a Parallel Computing Methods for Analyzing Gene Expression [24] was developed in recent past. This approach is developed as a ternary logic filter operation, where the processing data are divided into 3 base logics and a prediction value is computed using three phase of operations. Though the retrieval accuracy is comparatively higher, the system is computationally intensive. To minimize the computational iterations, a multilevel threading of cluster searching is proposed. In this approach an intermediate cluster match variable 'K-Temp'is introduced, which records the cluster transition and Mean deviations as a learning value for next process initialization. The system outline for the proposed system is as shown in figure 4.

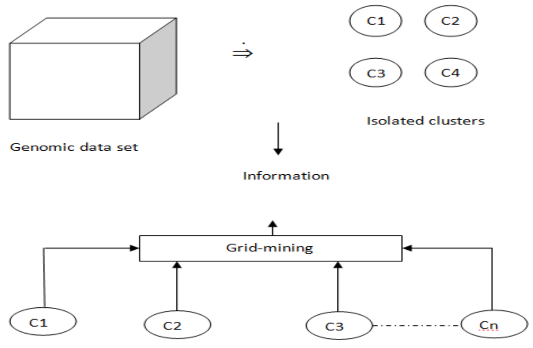

Figure 4: A Basic flow diagram of proposed sub-group grid mining approach

In the process of parallel processing, wherein multiple threads are created and executed simultaneously, the control over these threads could reduce the overhead. In this work, the thread operations are controlled via intermediate registers, which initialize or block the thread creation and execution. The proposed approach of grid mining is as outlined.

For, a given pattern, $P_1 \in M_{i,k}$ where $M_{i,k}$ is a set of genomic data with k-dimensions, apply a cluster segregation,

$M_{i,k} \Rightarrow C_{(i,k)_N}$. Where, N is the number of clusters which together consist of (i, k) observations. To mine a pattern, P_i, among $(C_{i,k})_N$ clusters,$T_{(i,k)_N}$ Threads and created. For each thread a parallel execution is computed, which computes the deviations for given observation over a cluster, defined by,

$$de_{i,N} = P_i - C_{i,N}$$

These deviations are then co-related over all observations and final information is retrieved. However the iteration taken for such operation is(i* k* N) which is comparatively high. This computational effort is minimized by introducing a 'K-Temp'1-value register, which hold the average deviation of all N-clusters,

$$A_{dev}(i,N) = P_i - \bar{C}_{i,N}$$

Where, $\bar{C}_{i,N}$ is the average value of each cluster, and among these, $A_{dev}(i,N)$the highest 3-dimentionals are considered for mining. As the comparative deviations are for the mean of the cluster, and highest three clusters are chosen as the mining groups, the retrieval accuracy is less affected. Only for these three selected clusters parallel threads are created and deviations are computed. This approach reduces the computation iteration by, (i*k*N)-(i*k+3) times, the retrieval accuracy however remain same as, highest three spatial similar clusters are considered.

To evaluate the suggested approach an experimental analysis is carried out and the obtained results are as outlined in following section.

6. EXPERIMENTAL RESULTS

To evaluate the proposed work, a simulation model is developed and comparative performance analysis is carried out. The proposed coding we applied to the gene sequences for human prothrombin. These sequences are extracted from Gene Bank database, form National Center for Biotechnology Information (NCBI) [24]. The evaluation is carried out over human prothrombin (base number 6654-7884, accession M17262 M33691), containing 2 exon regions. The observed observations are as illustrated,

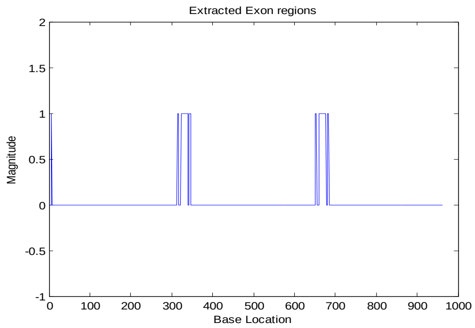

Figure 4: Represented Exon regions for the given Gene pattern

Figure illustrates the extracted exon regions segment. The considered gene sequence of above stated sequence, consist of two exon regions, which are extracted and represented in binary logic. The exon region prediction is developed based on the predication approach outlined in section II. The marked exon regions are represented in figure 4.

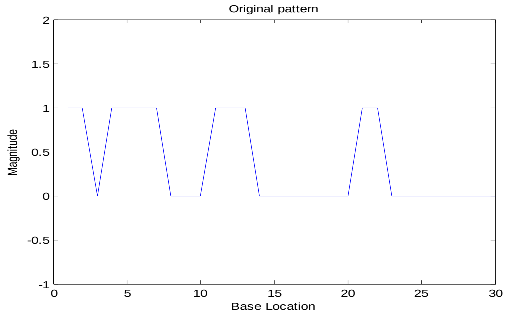

Figure 5: Enlarged portion of an exon region for figure 4

For the extracted regions, pattern alignment logic is carried out. Pattern alignment reduces the Levenshtein Edit Distance (LED) reducing the probability of bit mismatch error due to pattern insertion or deletion. Reducing the pattern variations reduces the hamming distance and hence the patterns are more appropriately represented with minimum correlation logic. The re-alignment approach reduces the correlation logics required, recuing the buffering memory overhead. This approach also overcomes the probability of LED errors, as the patterns are more uniformly arranged.

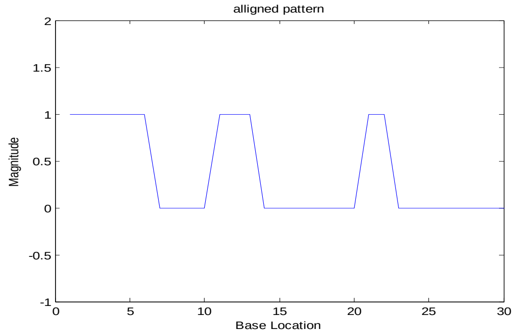

Figure 6: The aligned pattern for the pattern sequence

The realigned pattern for the read exon region is presented in figure 5. These realigned patterns have reduced hamming distance, and more uniformity than the actual pattern, reduces the random behavior of the pattern. This alignment reduces the probability of error introduction.

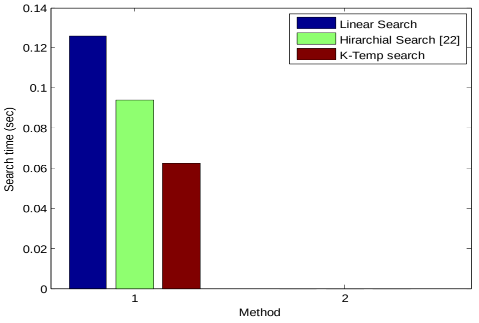

Figure 7: Comparative result of search time

With the pattern aligned, the sequences are represented with lower transition; these patterns are stored as a sequence of patterns in a database. For the evaluation of the developed work, patterns are extracted from Gene Bank database, form National Center for Biotechnology Information (NCBI) [24]. The database consists of 26928 base patterns, defined as a set of 10 character gene pattern, total consisting of 2692 patterns. These patterns are coded and realigned for exon region prediction and the extracted exon regions represented in a low transition binary patterns. The total database used is 13590 binary bit patterns. These patterns are then mined for a given test pattern. The search time are recorded based on the three benchmark searching algorithms, wherein a linear search is carried out for mining the query. A hierarchal mining approach is also developed, based on linkage clustering approach [22], and a k-Temp threading approach using grid mining is compared. The obtained search time for the three methods is presented in figure 7.

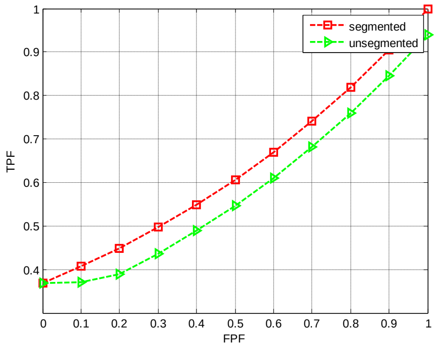

Figure 8: ROC curve for the proposed segmented Exon region

For the analysis of the developed exon region predication evaluation is carried out for segmented regions prediction and full length sequence. The exon region prediction is observed to be more accurate under lower segment length. AS the sequence pattern is lower segmented, the regions are more concentric and hence the energy concentrations, based on DFT coding results in more energy concentration at the exon regions. Due to the energy concentration, the patterns are more clearly concentrated at exon regions. The TPF (true positive fraction) for the segmented region is hence observed to be higher than non-segmented analysis.

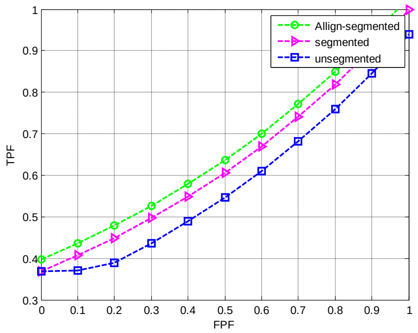

Figure 9: ROC for the sequence aligned pattern

Due to the segmented approach, the prediction accuracy is improved. However, when the extracted regions are binary represented, the random behavior leads to lower in accurate representation. To achieve a better pattern representation, pattern alignment logic is developed. When aligned with pattern realignment, the exon regions are well defined as in comparison to non-aligned logic.

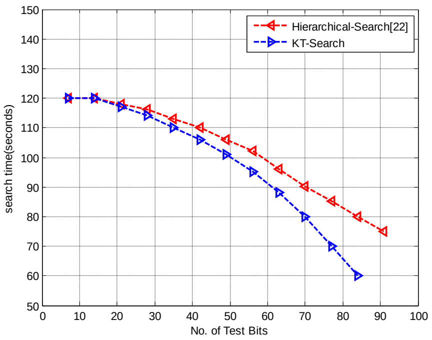

Figure 10: Observed search time performance for test bit variations

With all the above factors a simulation is carried out for query pattern mining over the stated data base and the effect of selective bit counts for matching is also evaluated. Figure 10 illustrates the obtained observation for the mining performance under linkage hierarchal coding and the proposed K-Temp grid mining. The search time, is observed to reduce with increase in bit pattern, due the probability of increase in more bit matching.

7. CONCLUSION

An optimization approach to gene data mining, using a modified representation and mining approach is presented. The proposed method develops a dimensional reduction approach by exon region prediction, optimizes the Levenshtein Edit Distance (LED) by a pattern alignment coding, and proposes a new k-Temp register based grid mining for gene data set. The region prediction based on DSP approach, under spectral domain representation reveals accurate exon region prediction. The mining approach is applied over the predicted regions only, as the regions are of reduced coefficient counts, they take less mining time, in comparison to full segment mining. These sequences are represented as binary patterns for dataset creation. These sequences consist of multiple transitions, which take larger correlation operation; as well the probability of misleading is more. The effect of insertion, deletion, or updation, could leads to misinterpretation. Hence the pattern alignment logic is proposed. The reduced number of transition, optimizes the mining effort, and hence results in faster processing. These patterns are mined for information retrieval. To minimize the speed of mining, a distributed mining, based on group modeling, and K-temp register is proposed. The retrieval performance with respect to accuracy and convergence is found more optimal than the conventional approaches.

8. REFERENCES

[1] Gir Won Lee ,Sangsoo Kim, “Genome data mining for everyone”, BMB Reports, Minireview, 2008.

[2] Doug Szajda, Michael Pohl, Jason Owen, Barry Lawson, “Toward A Practical Data Privacy Scheme for A Distributed Implementation of the Smith-Waterman Genome Sequence Comparison Algorithm”, NSF 2000.

[3] Georgios A Pavlopoulos, AnastasisOulas, Ernesto Iacucci, Alejandro Sifrim, Yves Moreau, Reinhard Schneider, Jan Aerts and IoannisIliopoulos, “Unraveling genomic variation from next generation sequencing data”, BioData Mining 2013.

[4] Anton James Enright, “Computational Analysis of Protein Function within Complete Genomes”, thesis, 2002.

[5] Peter Schattner, “Genomics made easier: An introductory tutorial to genome data mining”, Genomics 93, Elsvier, 2009.

[6] Inbamalar T M, Sivakumar R, “Filtering Approach to DNA Signal Processing”, IACSIT Conference, 2012.

[7] Yukinori Okada, Robert M. Plenge, “Entering the Age of Whole-Exome Sequencing in Rheumatic Diseases: Novel Insights into Disease Pathogenicity”, Arthritis & Rheumatism, 2013,

[8] Anastassiou D, “Frequency-domain analysis of bio-molecular sequences”, Bioinformatics, Vol.16, No.12, 2000.

[9] Anastassiou D, “Genomic Signal Processing”, IEEE Signal Processing Magazine, July, 2001.

[10] P.P.Vaidyanathan ,Byung-Jun Yoon, “The role of signal-processing concepts in genomics and proteomics”, Genomics, 2004.

[11] ZhengGuo, Tianwen Zhang, Xia Li, Qi Wang, JianzhenXu, Hui Yu, Jing Zhu, Haiyun Wang, Chenguang Wang, Eric J Topol, Qing Wang and ShaoqiRao, “Towards precise classification of cancers based on robust gene functional expression profiles”, BMC Bioinformatics 2005.

[12] Julie A. Hawkins, Colin E. Hughes, Robert W. Scotland, “Primary Homology Assessment, Characters and Character States”, Cladistics 13, 1997.

[13] P.Vivekanandan, R. Nedunchezhian, “A New Incremental Genetic Algorithm Based Classification Model To Mine Data With Concept Drift”, JATIT, 2010.

[14] Paivi Onkamo1,HannuToivonen, “A survey of data mining methods for linkage disequilibrium mapping”, Human Genomics, 2006.

[15] S.Vidhya, S.Karthikeyan, “A Security Based Data Mining Approach In Data Grid”, Journal Of Computing, 2010.

[16] Mahmood Akhtar, “Comparison of Gene and Exon Prediction Techniques for Detection of Short Coding Regions”, IJIT 2005.

[17] Shreyas Sen, Seetharam Narasimhan, Amit Konar, “Biological Data Mining for Genomic Clustering Using Unsupervised Neural Learning”, Engineering Letters, 2007.

[18] Dominik Grimm, “Data Mining in Bioinformatics Day 6: Classification in Next Generation Sequencing Data Analysis”, 2013.

[19] NavinChatlani, John J. Soraghan, “Local Binary Patterns For 1-D Signal Processing”, EUSIPCO-2010.

[20] JianfengRen, “Noise-Resistant Local Binary Pattern With an Embedded Error-Correction Mechanism”, IEEE Transactions On Image Processing, 2013.

[21] M. Ben Haj Hmida, Y. Slimani, “High performance Grid computing for detecting gene-gene interactions in genome-wide association studies”, geonomics 2000.

[22] Jiang, D., Pei, J. and Zhang, A. . DHC: A Density-based Hierarchical Clustering Method for Time- series Gene Expression Data. In Proceeding of BIBE2003: 3rd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, March 10-12 2003.

[23] Edward B. Suh, “Parallel Computing Methods for Analyzing Gene Expression Relationships”, SPIE 2001.

[24] http://www.ncbi.nlm.nih.gov/nuccore/M17262.1