Asian Journal of Information Technology 16 (8): 675-685, 2017
ISSN: 1682-3915

Data Driven Approach for Genetic Disorder Prediction by Aggregating Mutational Features

Sathyavikasini Kalimuthu and Vijaya Vijayakumar
Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

Abstract: In the recent genomic epoch, the recognition of the genetic diseases is paramount. It is a convoluted task to recognize a heritable disease that is certainly caused by genetic mutations. Identification of disease based on mutations in the gene sequences is an important and challenging task in the medical diagnosis of genetic disorders. This study addresses this problem by developing new model by extracting mutational features as discriminative descriptors for predicting the disease accurately. The disease gene sequences are mutated by espousing a technique like positional cloning on the reference cDNA sequence. A rare genetic disorder such as muscular dystrophy is taken as a sample for this research. This disease is a complicated neuromuscular ailment with a prominent social impact that impairs the working of the locomotive muscle tissues. The versatile causes of this disease bring about the requirement of new hereditary patterns that can diagnose patients using biological information. There are diverse significant forms of muscular dystrophy and it is imperative to identify the type of muscular dystrophy for proper medical diagnosis and medication. Hence, a data driven model is developed using pattern recognition techniques by aggregating the features related to all kinds of mutations for predicting the disease precisely. Results indicate that the SVM classifier is found to acquire the best accuracy of 90.5% for predicting muscular dystrophy.

Key words: Codon, classification, muscular dystrophy, positional cloning, pooled features, RSCU

INTRODUCTION

Monogenic or polygenic diseases ground the impairment of the normal structure or function of the organ. Mutilation in genetic characters due to single mutation in a specific gene is monogenic disease. Moreover, non-genetic mutations caused in multiple genes leads to polygenic diseases. Genetic disorders are caused by the deformities in the inherited genes and current encroachments in gene testing aids in diagnosing people at a risk of getting a disease in advance in a head of any indication of disease. An accurate gene test results in finding the disease-related gene mutation. The impact of the mutation on the gene sequence modifies the function of the gene. Substitution is an exchange of one base to another such as swapping a base from A-G. Mutations that show an impression on protein sequence include missense, nonsense, insertions, deletions, splicing and frame-shift mutations. Missense and non-sense are the non-synonymous single nucleotide variants where a single change in the gene alters the amino acid in the sequence (Ma *et al*., 2005; Zeng *et al*., 2014). Missense mutations are the substitution in a codon that encodes a different amino acid and alters the protein. Nonsense mutations are those where the protein attains to stop codon when a change occurs in the DNA sequence. Synonymous mutations are the silent mutations that the variant will not show amend in the amino acids. Silent mutations are a change in codon that encodes for the same amino acid and therefore, the translated protein is not modified (Kann, 2010). In detecting the type of disease it is necessary to consider the silent mutation as the changes can affect protein folding and function. Even though several codons encode for the same amino acid their frequency will vary and this is referred as codon bias. The increase in the number of the same nucleotides in a location is termed as duplications. Deletions are the mutations when a base or an exon is deleted from a sequence the mutations (Tranchevent *et al*., 2011). In the eukaryotic genes, the spliceosome catalyses the intervening sequences or introns which are spliced by the process of RNA splicing. Any change that occurs while splicing will lead to splicing mutation (Clancy, 2008).

Identification of genetic factors for complex diseases is a far more difficult task with the standard methods as it is difficult to analyze the data. The complex diseases provide a great deal of challenges to standard data

Corresponding Author: Sathyavikasini Kalimuthu, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India

analysis techniques. The apparent benefit of hereditary testing aids in identifying and understanding of risk for a certain disease. The traditional method of testing is time consuming and incurs cost over head. Therefore, it is essential to model and represent this knowledge in a computational form with minimal loss of biological context through a gene based approach. Disease-gene association needs to be designed to handle this type of data (Sathyavikashini and Vijaya, 2015a).

In this research, identification of disease is done with the help of sequence data and a data driven model is created using supervised learning techniques to infer the disease. For this purpose muscular dystrophy disease is considered, since, it is a monogenic disease caused by the mutations in the genes which are in charge of ordinary muscle function. Progressive muscle fatigue that impacts limb, axial and facial muscles is the fore most reason behind muscular dystrophy (Emery, 2002). Muscular dystrophy is believed as a genetic ailment run in a family, even if only one blood relation in the ancestor is affected. Rare forms of muscular dystrophy are duchenne, becker, emery-dreifuss, limb-girdle, facioscapulohumeral, myotonic, spinal, distal and charcot marie tooth disease. Duchenne Muscular Dystrophy (DMD) is the X-Linked and most common form of muscular dystrophy is caused by the mutations in the dystrophin gene located on the X chromosome. The absence of dystrophin gene occurs when a large number out frame deletions occur which is the major cause of DMD. Becker Muscular Dystrophy (BMD) is the X-Linked less defective mutations in the dystrophin gene display a much milder dystrophic phenotype in affected patients, known as Becker's muscular dystrophy. The mutations in the Emerin (EMD) and Lamin A/C (LMNA) genes cause Emery-Dreifuss Muscular Dystrophy (EMD). The defects in Limb-Girdle Muscular Dystrophy (LGMD) show a related distribution of muscle weakness that has an effect on both upper arms and legs. Charcot Marie Tooth disease (CMT) includes a number of disorders with an assortment of symptoms grounds damages in peripheral nerves. The disorder affects the peroneal muscle in the lower leg and hence the disease also is known as Hereditary Motor and Sensory Neuropathy (HMSN) and peroneal muscular atrophy (Uhmn *et al*., 2009). More than 30 forms of CMT are noticed and 30 genes are concerned, some may show severe brain malformations such as lissencephaly and hydrocephalus and hearing loss (Agnes *et al*., 2008).

Diagnosing muscular dystrophy is in progress with the help of muscle biopsy and DNA testing. The advantage of performing genetic testing over muscle biopsy is that in genetic testing, diagnosis is done with the blood sample to spot the alteration in the genes whereas the part of the tissue is required to perform the muscle biopsy. Gene therapy helps in knowing the exact mutation in the DMD gene and direct sequencing aids in identifying missense, nonsense, insertions, deletions and splicing mutations (Roberts *et al*., 1992). Laboratory methods such as Multiplex Ligation-dependent Probe Amplification (MLPA), PCR, Sanger's full gene sequencing is considered to be laborious, expensive, time-consuming and accuracy also cannot be attained (Chen *et al*., 2014; Bennett *et al*., 2009; Koenig *et al*., 1987). To overcome the challenges in laboratory methods, the process should be automated through the computational methods and disease should be identified efficiently. Machine learning techniques paved the way to predict the type of disease in some circumstances.

In the medical applications the capability of machine learning is well suited particularly on complex genomic and proteomic measurements. Models based on machine learning have been extensively used to analyze complex diseases such as diabetes (Ban *et al*., 2010) hepatitis rheumatoid arthritis (Briggs *et al*., 2010) schizophrenia (Nicodemus *et al*., 2010). However, not many studies have been carried out on variation of muscular dystrophy using machine learning algorithms. Also, the classification of this complex disease is done with the either protein data or micro array data as their inputs. Classification of Facio Scapulo Humeral muscular Dystrophy (FSHD) disease is done by monitoring of expression levels. Usually, microarray gene expression analysis is mainly focused to cancer diseases. In the study of Gonzalez *et al*. (2013) the research proposed an approach to classifying the types of Facio Scapulo Humeral muscular Dystrophy (FSHD). The microarray gene expression data from the DUX4 gene are taken into account for classification. A model was created using support vector machine to classify the types of FSHD.

The research by Mercuri and Muntoni (2013) developed a model using neural networks to identify whether the patient is affected from Limb Griddle Muscular Dystrophy (LGMD). The data based on the patient's family details are collected. The classification of disease status is made using the neural network and achieved an accuracy of 98%. The research by Noguchi *et al*. (2003) constructed a protein-protein interaction network to classify the sub-types of muscular dystrophy through machine learning techniques. Microarray gene expression datasets are analyzed and the protein data and their interaction data are collected and a network is constructed to classify the sub types. Multi class support vector machine is applied for the classification of 6 sub-types of muscular dystrophy. Noguchi *et al*. (2003) proposed a model to classify the

types of Human Leukocyte Antigen (HLA) gene into different functional groups by choosing the codon usage bias as input. In their research, they converted the gene sequence into 59 vector elements by calculating the RSCU values for the gene sequence. A model was created using support vector machine and achieved an accuracy rate of 99.3%. The researcher CM Nisha, Bhasker Pant and K.R. Pardasani proposed an approach based on codon usage pattern to classify the type of Hepatitis C Virus (HCV) that is the primary reason for the liver infection. To classify the sub class of its genotype a model was created using codon usage bias as input to multi class SVM (Nisha *et al*., 2012).

Kalari (2006) identified large mutations such as duplications and deletions through computational approach. A system speed was developed by utilizing the logical model tree method based on machine learning technique for the gene BRCA 1. High specificity was achieved with this technique. Wu *et al*. (2010) predicted the disease causing mutations through ensemble learning technique. The protein sequence dataset from swissprot database was used for classification. A comparative analysis was made between the traditional approach and ensemble approach and logit boost ensemble technique achieves high performance among all the methods compared.

Bioinformatic tools that are designed to assess the impact of genetic variation on splicing are NNS plice (Reese *et al*., 1997), maxentscan (Yeo and Burge, 2004) ESEF inder (Cartegni *et al*., 2003), spliceman (Lim and Fairbrother, 2012), skippy (Woolfe *et al*., 2010) and human splice finder (Desmet *et al*., 2009). Skippy is a web-based tool that defines exonic variants using the genomic features that modulate splicing. Single nucleotide variants relevant to splice-modulating genomic features variants are assessed and scored. Point mutations lay in the coding region show severe effects on gene function through disruption of splicing. Mutpred splice is a machine learning approach for identifying the coding region substitutions that disrupt pre-mRNA splicing. Disease causing splice altering variants, disease-causing splice neutral variants and polymorphic splice neutral variants are considered and discriminative descriptors are extracted from gene sequences. Supervised classification techniques such as random forest and SVM are employed for building models (Mort *et al*., 2014).

The classification of muscular dystrophy continues to evolve with the advances in understanding of their molecular genetics. Huge number of muscular dystrophy related faulty genes and proteins are identified but no successful treatments are known for many of its sub-types. The proportion of mutations in deletions, duplications and point mutations differs in each type of disease and the present methods cannot handle the entire mutational spectrum in a single platform. However, it is essential to look into the accurate mutation site and to predict the disease. HGMD-Human Gene Mutational Database is a cluster of the mutational information in genes coupled with the human inherited disease that is clinched from diverse research. Table 1 depicts the approximate number of mutational information for the muscular dystrophy disorder from HGMD database. The public version of HGMD is freely available to registered users from academic institutions/non-profit organizations.

Table 1: Proportion of mutation spectrum in HGMD for muscular dystrophy disease

Disease	Missense/Nonsense	Insertions/Duplications	Deletions	Splicing
DMD	460	295	826	150
BMD	70	63	283	70
EMD	95	14	50	14
CMT	284	36	56	40
LGMD	511	67	157	72

It is pioneered from the literatures that the disease identification problem can be modeled as pattern recognition task to identify the disease. As machine learning technique can automatically learn the model by taking intelligent hints from the data and predicts the output more accurately, it has been influenced in this research to extract and pool various discriminative features from the diseased gene sequences for building disease prediction models.

The primary focal point of this research is to build a disease identification model for diagnosing a genetic disease by extracting mutational features from the gene sequences. As muscular dystrophy is a heritable disorder caused by the mutations in the gene sequences this disease is considered for this research. Some of the diverse forms of this a genetic disorder is Duchenne Muscular Dystrophy (DMD), Becker's Muscular Dystrophy (BMD), Emery drefius Muscular Dystrophy (EMD) Limb Griddle Muscular Dystrophy (LGMD) Charcot Marie Tooth disease (CMT).

In the previous research by Sathyavikashini and Vijaya (2015b) features related to non synonymous mutations are considered and disease identification was done by extracting discriminative features from the cloned gene sequences. A model was developed based on pattern recognition techniques and high accuracy was attained from the decision tree classifier. In other research, Sathyavikashini and Vijaya (2016) silent mutational features were captured by calculating the RSCU values from the diseased gene sequences and an accuracy of 86% was attained using support vector machine and 90% of accuracy was achieved from Lib D3C classifier.

This stimulated to perform various autonomous implementations by increasing the dataset size to 1000 to classify the disease classification through various standard learning techniques based on all kinds of mutational features in order to predict the type of disease. The study is carried out to propose 5 different experiments by extracting diverse features pertaining to all kinds of mutations from 1000 mutated gene sequences. A cohesive approach is demonstrated based on computational intelligence technique to detect major 5 forms muscular dystrophy with diseased gene sequences as input. The pattern recognition algorithms such as decision tree, artificial neural network, naive bayes and support vector machine are utilized to train the model. Machine learning algorithms are data driven and are able to examine large amounts of data.

MATERIALS AND METHODS

Accurate prediction of genetic disorder is a complicated task as the pattern of the gene sequence varies for every individual. The key idea in this research is to pool out discriminative descriptors extracted from diseased gene sequences associated with all types of mutations and to provide an effective solution for predicting the type of disease. Multi-class classification is worked out through data modeling of gene sequences. The synthetic mutational gene sequences are created as the diseased gene sequences are not readily obtainable for this intricate disease. Positional cloning approach supports in generating disease gene sequences based on mutational information acquired from HGMD.

Data acquisition through positional cloning: Synthetic gene sequences are generated with the gene mutational information collected from the Human Gene Mutational Database (HGMD). The reference genes for the mutated genes are downloaded from NCBI. The mutation position and its location on the chromosome enable the synthesis of cloned gene sequences by employing the positional cloning approach.

The raw sequence obtained from HGMD is processed to form cDNA sequence and the nucleotide base alteration is done based on the mutational information. Using the traditional positional cloning approach the mutated sequences are generated and stored as fasta files. Consider the missense mutational information for the EMD phenotype from the emerin gene such as nucleotide change is 2 T>C which indicates in the position 2 the nucleotide changes from T-C alters the protein from met to thr. For example the cDNA sequence of EMD gene is:

ATGGACAACTACGCAGATCTTTCGGATACCGA...
↑

After the nucleotide change in the position 2:

ACGGACAACTACGCAGATCTTTCGGATACCGA...
↑

Sample output of cloning technique for gene sequence using positional information is shown in Fig. 1. Figure 1 depicts the generation of mutated gene sequence of the EMD phenotype from the emerin gene with the mutational information. The 7 types of mutations have been considered for generating mutated sequences. The

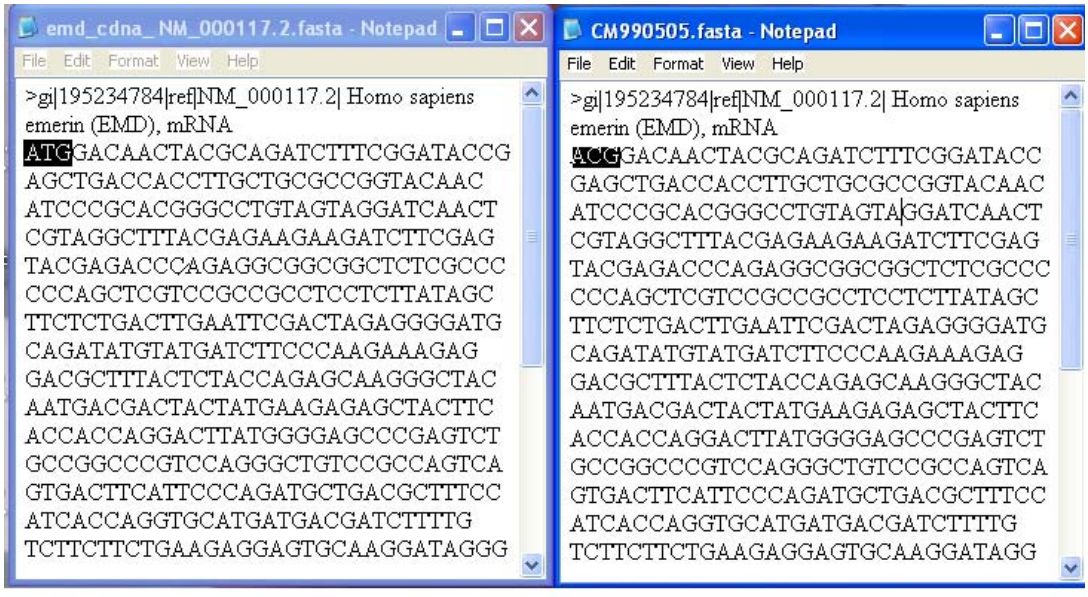

Fig. 1: Output of genrated mutated gene sequance

Table 2: Genes associated with different type of muscular dystrophy

Muscular dystrophy disease	Genes associated with the disease
Duchenne muscular dystrophy	Dystrophin
Becker's muscular dystrophy	Dystrophin
Emery-dreifuss muscular dystrophy	Emerin, LMNA/C
Limb griddle muscular dystrophy	ANO5, CAPN3, CAV3, DYSF, FKRP, FKTN, LMNA, MYOT, POMGNT1, POMT1,POMT2, SGCA, SGCB, S, GCD, SGCG, TCAP, TRIM32, TTN
Charcot marie tooth disease	AARS, AIFM1, BSCL2, DHTKD1, DNM2, DYNC1H1, EGR2, FGD4, FIG4, GARS, GDAP1, GJB1, HSPB1, HSPB8, INF2, KARS, KIF1B, LITAF, LMNA, LRSAM1, MED25, MFN2, MPZ, MTMR2, NDRG1, NEFL, PMP22, PRPS1, PRX, RAB7A, SBF2, SH3TC2, TRPV4, YARS

55 types of genes associated with the five types of neuromuscular disorder are studied. An analysis is made of fifty-five genes that are associated with five types of muscular dystrophy like DMD, BMD, EMD, LGMD and CMT. Table 2 summarizes genes associated with the disease. Several types of mutated sequences based on mutations like Missense, Nonsense, synonymous, Insertion/duplication, deletion mutations and splicing mutations are collected. For the purpose of this research in each category of muscular dystrophy disease, 200 synthetic mutated gene sequences are generated and a corpus comprising of 1000 sequences for all five categories of muscular dystrophy is developed.

Feature extraction: Change or mutation in the gene sequence, alters the structure of the sequence which implies the cause of disease. These structural changes are captured as features of mutational sequence to learn the prediction model. So far in the literature no attempt was made to build disease identification model by aggregating all kind of mutational descriptors and hence, it is significant to build this type of disease identification model. The mutational features to discriminate the disease are carefully examined and extracted. The 106 evocative features are cumulated and feature vectors are created for learning the disease prediction model.

Features of missense and nonsense mutations: The missense and nonsense mutational features are based on annotation, structure and alignment of the diseased gene sequences. The annotation features includes gene ID, gene symbol and chromosome number. Length of the sequence, alteration type, protein changed, reference allele, observed allele, mutation position, mutation start position, mutation end position, position of mutation in gene sequence amino acid change leads to stop codon, stop codon, position of start codon in cDNA sequence, position of stop codon in DNA sequence, the nucleotide composition of A, G, C, T, AT and GC component composition constitutes the structural features. The alignment features are edit distance scores, phred quality scores, substitution scores. These features are identified and defined as non synonymous mutational features.

Gene identifier and symbol of the gene is unique for every gene sequence. As many to one relationship occur between gene and the disease these features are captured. 55 genes are involved in five types of muscular dystrophy. Some form of MD was affected by the mutations in more than 20 genes and so, annotation descriptors are considered to differentiate the gene sequence.

The alteration type such as missense, nonsense, silent, deletion and duplications are encoded to numeric values from 1-5. The reference allele is the actual protein that is present in the cDNA sequence file and the observed allele is the protein observed after alteration. The length of the sequence plays an important role in examining the difference in length of the sequence. When the insertion or deletion mutation occurs the length of the sequence gets varied automatically.

The base composition A, C, G and T which is unique for every gene is calculated to count the number of occurrences of the four different nucleotides ("A", "C", "G" and "T") in the sequence. One of the most fundamental properties of a genome sequence is its AT and GC content. GC content is the fraction of the sequence that consists of Gs and Cs, i.e., the GC content is calculated as the percentage of the bases in the genome that are Gs or Cs that is:

$$\text{AT content} = \text{Number of As+Number of Ts} \times 100/\text{genome length}$$

$$\text{GC content} = \text{Number of Gs+Number of Cs} \times 100/\text{genome length}$$

The position of the Stop codon reveals the end of the coding part in the sequence. The position of the stop codon reveals the end of the coding part in the sequence. To find the position of start codon match pattern () function is used. Alignment scores are considered as the important feature for disease prediction. The global pair wise alignment based on edit distance is done with the mutated sequence against with the reference cDNA sequence and the alignment scores are calculated using edit distance scoring method. The phred quality measures

are calculated with the pattern quality and subject quality to examine the quality-based match and mismatch bit scores for DNA/RNA. The substitution scores are calculated by setting the error probability to 0.1.

Features of silent mutations: A codon is the triplet of nucleotides that code for a specific amino acid. Many to one relationship occur between the codon and amino acid. Many amino acids are coded by more than one codon because of the degeneracy of the genetic codes. A total number of codons in a DNA sequence counts to 64. Since, Amino acids methionine (ATG) and Tryptophan (TGG) posses only one codon, they are not included as their RSCU values are always equal to 1. The three stop codons (TGA, TAA, TAG) are also, not included. Accordingly, the number of codons considered here is 59. The differences in the frequency of occurrence of synonymous codons are referred as codon usage bias. The calculation of RSCU is done by dividing the number of times a particular codon observed relative to the number of times that the codon would be observed in the absence of any codon usage bias. The RSCU carries the value 1.00 if the codon usage bias of that particular codon is absent. If the codon is used less frequently than expected, the RSCU values tend to have the negative values. Following equation is used to calculate RSCU:

$$RSCU = \frac{X_{ij}}{(1/n_i \times S\{X_{ij};\ j = 1, n_i\})} \quad (1)$$

Where:

X_{ij} = Number of occurrences of the jth codon for the ith amino acid

n_i = Number of alternative codons for the ith amino acid

If the synonymous codons of an amino acid are used with equal frequencies then their RSCU values are 1. The RSCU values are derived for 59 codons from each mutated gene sequence and feature vectors are created.

Features of insertion/duplication, deletion mutations: The exonic and intronic features are considered from diverse gene families if extract the well defined descriptors related to insertion, deletion and duplication mutations in the mutated gene sequences. The extrinsic and intrinsic features more solely depend on the exons and introns that aids in identifying the disease affected by large insertions and deletions. Gene identifier, symbol of the gene, gene start position, gene end position, sequence length, number of exons inserted/deleted, exon and intron boundry, deletion type, exon type, alignment scores, conservation score, nucleotide composition values are identified as features of this kind of mutation.

In gross insertions and gross deletions, the numbers of exons inserted or deleted were noted cautiously as the count also, aids in deciding the type of the disease. Severe effect on the deletion of exons leads to DMD and mild deletion of exons will results in BMD. Location of the exons will be varied when a mutation occurs. The boundry of exons and introns were captured to identify the differences in the boundry between the normal and the diseased sequences. Deletion type is a contributive feature in identifying the type of the disease as in some diseases like BMD the sequence can be read after deletion and in some diseases like DMD the sequence cannot be read after deletion as it is outframe. Depending on location of the exon, the type may be initial, internal, terminal and single exons. The mutation in each type of exon has its own severity. The structure or the function of the sequence is identified by aligning the sequence with all organisms. University of California Santa Cruz (UCSC) genome browser is employed to calculate the conservation score.

Features of splicing mutations: The discriminative descriptors aids in diagnosing the identification of exonic single base substitutions that modulate splicing. exon number, variant exon number, exon boundary, intron boundary, sequence length, gene ID, gene symbol, chromosome number, splice site distance, phylop score, phastcons score, donor site score, acceptor site score, branch site score, ESR change, distance of alteration from 5 splice site, distance of alteration from 3' splice site, scoring splice site with PWM, flanking intron size, GC content, exon size, constitutive exon, exon type, coding region score. These are defined as contributive features of this splicing mutation.

Variant exon number gives the mutant exon's number in the target iso forms. The boundary of the affected exons and introns and the length of the sequence are captured using geneious pro tool. The other annotation descriptors are examined using genomic features in MATLAB. Features related to SNP are vital in identifying the disease and the mutations that disrupt splicing are the single nucleotide variants occur in both coding and non-coding regions. The distance of the substitution from the variant to the nearest splice site is identified and recorded as splice site distance. Phylop is an evolutionary conservation element that computed the base-wise sequence conservation score of single base substitution which is calculated based on multiple sequence alignment. Phastcons is a base wise conservation element examined from probability for substitution site, based on multiple alignments. Phylop and Phastcons scores are downloaded from the UCSC genome browser.

Acceptor site cut off score, branch site cut off score and donor site cut off score are calculated using ESE finder tool. Distance of alteration from 5' splice site, distance of alteration from 3' splice site are the distance between the variants and splice sites 3' and 5'. The regulatory sequences located within the exon and promoting exon inclusion are referred to as Exonic Splicing Regulatory (ESR) elements. ESR change identifies the change in the frequency of ESR elements with respect to single variants. To strengthen or repress the elements in the sequences Exonic Splicing Enhancers (ESE) and Exonic Splicing Silencers (ESS) is calculated using ESE finder tool. The ESR changes helps in recognizing the adjacent splice site. Counting the occurrences of nucleotides at each position within the 5' splice site is done using PWM-Position Weight Matrices that is calculated as log odds score.

The variation in the protein coding region makes a major impact on the gene and those exon-based descriptors focus mainly on the exons. Flanking intron size is the length of the base pairs of the up stream and down stream introns nearby the target exon. Constitutive exon is the boolean value that specifies whether the variant exon is present in every transcript. The score of the coding region was calculated using protein coding region calculator.

Data driven approach for machine learning algorithm: Data driven systems solves the problem by developing own models based on the examples and experiences. These methods develop intelligent systems that discover patterns from large datasets based on computational analysis that provides concrete theory and predictions. Four data driven supervised learning algorithms commonly used for classification task were used in this research of genetic disorder prediction. Decision tree, artificial neural network and support vector machine are the supervised learning algorithms employed to build disease identification model using MATLAB.

RESULTS AND DISCUSSION

Based on various mutational features five independent experiments were carried out in this study using decision tree, naive bayes, artificial neural network and support vector machine. Disease identification based on non synonymous mutational features disease identification based on-synonymous mutational features disease identification based on insertion/deletion and duplication mutational features disease identification based on splicing mutational features disease identification based on aggregated mutational features. The 10 fold cross validation is used to test the models and results are analysed.

Table 3: Training datasets

Types of Mutation	No. of Features	Dataset	Size of dataset
Non-synonymous	26	NSM	1000×26
Synonymous	59	SYM	1000×59
Insertion/duplication and deletion	25	IDM	1000×25
Splicing	24	SPM	1000×24
Aggregated	106	AGM	1000×106

Training dataset: The features extracted from each disease gene sequence forms a feature vector. Depending on the type of mutation the mutational features are varied and the size of the feature vector also varies here. Since, four kinds of mutations are taken into account, four exclusive datasets have been formed. Non-Synonymous (NSM), Synonymous (SYM) Insertion, Duplication and deletion (IDM) and Splicing Mutation (SPM) are the four datasets with different dimensions. By pooling all the mutational features, AGM (Aggregated mutational features) dataset is formed which is of dimension 106. Since, the corpus consists of 1000 diseased gene sequences all the above five datasets contains 1000 feature vectors. For each feature vector the class label is assigned a sequence number 1-5 according to the category of disease. Table 3 depicts the type of datasets with its dimensions.

The first experiment aims in predicting the disease using of NSM dataset. Point mutational features such as structural, annotation and alignment descriptors are considered to be the non-synonymous mutational features. The predictive performance of the disease classification shows that SVM classifier yielded a best accuracy of 84.9% and the results are tabulated in Table 4.

In the first experiment only non synonymous mutational features are taken into account to identify the disease where the silent mutational features are required to identify the disease that is caused due to synonymous mutations. Relative Synonymous Codon Usage (RSCU) values for 59 codons forms synonymous mutational features. The second experiment is conducted by learning SYM dataset using decision tree, naïve bayes, ANN and support vector machine. The predictive performance of the disease classification shows that decision tree classifier yielded a best accuracy of 86% and the results are tabulated in Table 5.

Insertions/duplications and deletions alter the structure of the sequence and throws a heavy impact and therefore in the third experiment imperative extrinsic and intrinsic descriptors are considered for learning the model

Table 4: Predictive performance of the classifiers (non-synonymous mutations)

Performance criteria	Decision tree classifier	Artificial neural network	Naive Bayes classifier	SVM
Correctly classified instances	805.000	793.000	698.000	849.000
Incorrectly classified instances	195.000	207.000	302.000	151.000
Prediction accuracy (%)	80.500	79.300	69.800	84.900
Precision	0.800	0.793	0.689	0.849
Recall	0.815	0.785	0.678	0.846
F1 score	80.900	78.800	69.900	85.100
Cohen's Kappa	0.802	0.793	0.692	0.841
Time taken to build the model (sec)	8.400	10.700	9.600	7.000

Table 5: Predictive performance of the classifiers (synonymous mutations)

Performance criteria	Decision tree classifier	Artificial neural network	Naive Bayes classifier	SVM
Correctly classified instance	860.000	833.000	840.000	846.000
Incorrectly classified instance	140.000	167.000	160.000	154.000
Prediction accuracy (%)	86.000	83.330	84.000	84.600
Precision	0.860	0.831	0.835	0.841
Recall	0.854	0.830	0.841	0.850
F1 score	85.600	83.100	83.300	84.800
Cohen's Kappa	0.860	0.810	0.830	0.840
Time taken to buildthe model (sec)	7.470	11.700	12.700	10.500

Table 6: Predictive performance of the classifiers (insertion/duplication, deletion mutations)

Performance criteria	Decision tree classifier	Artificial neural network	Naive Bayes classifier	SVM
Correctly classified instance	853.000	856.000	831.000	863.00
Incorrectly classified instance	147.000	144.000	169.000	137.00
Prediction accuracy (%)	85.300	85.600	83.100	86.30
Precision	0.853	0.860	0.831	0.863
Recall	0.850	0.800	0.830	0.87
F1 score	85.300	85.900	83.100	86.30
Cohen's Kappa	0.880	0.862	0.810	0.86
Time taken to build the model (sec)	8.700	9.600	11.70	7.60

Table 7: Predictive performance of the classifiers (splicing mutations)

Performance criteria	Decision tree classifier	Artificial neural network	Naive Bayes classifier	SVM
Correctly classified instances	849.000	835.000	813.000	867.000
Incorrectly classified instances	151.000	165.000	187.000	133.000
Prediction accuracy (%)	84.900	83.500	81.300	86.700
Precision	0.849	0.830	0.810	0.860
Recall	0.846	0.815	0.800	0.870
F1 score	85.100	82.900	79.900	86.700
Cohen's Kappa	0.841	0.810	0.802	0.867
Time taken to build the model (sec)	7.000	8.000	13.600	6.500

using supervised classification algorithms. This experiment was carried on IDD dataset and the predictive performance of the disease classification shows that SVM classifier yielded a best accuracy of 86.3% and the results are tabulated in Table 6.

The exons are formed by splicing out the introns during transcription and the mutations occurred while spicing should be considered to know the alteration after the splicing process. Hence, in the next consecutive experiment, exon, SNP and gene features are taken into account for building the model. The training was performed using SPM dataset and the predictive performance of the disease classification shows that SVM classifier attains an accuracy of 86.7%. The results are tabulated in Table 7.

In all the previous experiments, autonomous disease identification models were built based on the specific mutational features. But normally the type of mutation caused in the gene sequence may not be known explicitly and hence, all the mutational features are accumulated by eliminating the repetitive features without losing information to facilitate efficient learning for predicting the disease caused by any mutation. In this experiment AGM dataset is employed for training the decision tree, naive bayes, ANN and SVM models. The cross validation results of the classifiers are shown in Table 8 and illustrated in Fig. 2.

Feature selection: Feature selection or attribute subset selection look for the best descriptors for model construction. It aids in improving the accuracy and the learning time of the classifiers. The information gain selection attribute method is used here to select the subset of attributes and 73 highly ranked attributes are chosen. The experiment was carried out with selected subset of attributes and a model is built using the

Table 8: Predictive performance of the classifiers (pooled features)

Performance criteria	Decision tree classifier	Artificial neural network	Naive Bayes classifier	SVM
Correctly classified instance	847.000	829.000	823.000	872.000
Incorrectly classified instance	153.000	171.000	177.000	128.000
Prediction accuracy (%)	84.700	82.900	82.300	87.200
Precision	0.847	0.829	0.823	0.872
Recall	0.847	0.820	0.820	0.881
F1 score	84.100	82.100	82.100	87.200
Cohen's Kappa	0.847	0.830	0.830	0.870
Time taken to build the model (sec)	7.000	9.700	9.700	5.200

Table 9: Predictive performance of the classifiers (pooled features) after applying feature selection

Performance criteria	Decision tree classifier	Artificial neural network	Naive bayes classifier	SVM
Correctly classified instance	878.000	863.00	856.000	903.00
Incorrectly classified instance	122.000	137.00	144.000	97.00
Prediction accuracy (%)	87.800	86.30	85.600	90.30
Precision	0.870	0.86	0.853	0.90
Recall	0.871	0.87	0.850	0.91
F1 score	87.800	86.30	85.100	90.10
Cohen's Kappa	0.880	0.87	0.850	90.0 0
Time taken to build the model (sec)	5.200	4.70	6.100	4.000

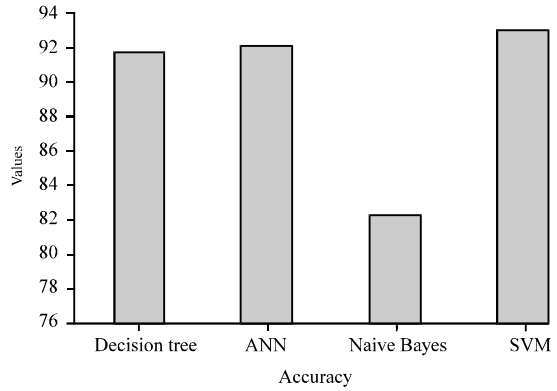

Fig. 2: Prediction accuracy of classifiers using pooled features

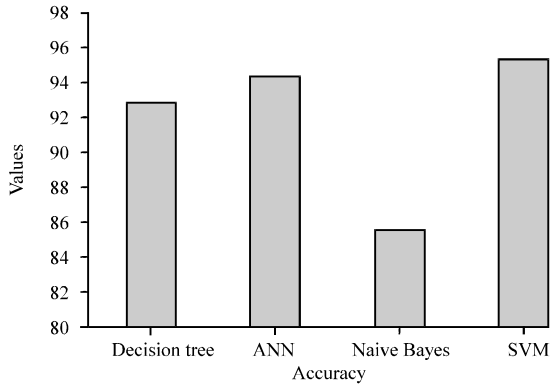

Fig. 3: Prediction accuracy of classifiers using pooled features

standard pattern recognition algorithms. The performance of SVM classifier observed better accuracy of 90.3%. The results of the classifiers are shown in Table 9 and illustrated in Fig. 3.

From the above experiments it was observed that the performance of the classifiers is high when training dataset contains summative features. The classification models built using non-synonymous mutational features produced an accuracy of about 84.9%. The classification models built using features related to synonymous mutations produced an accuracy of about 86%. About 86.3% accuracy was attained when insertion/duplication and deletion mutational features are taken into account. Disease prediction model reached an accuracy of 86.7% when splicing mutational features are considered. When all the mutational features are pooled together, the models showed an accuracy of about 87.2%.

Downsizing the features through feature selection expedites to improve the outcome and the prediction accuracy of the SVM classifier built using high ranked features was hoisted to 90.3%. Hence, it is observed that pooling the descriptors associated with all type of mutations produced an augmented trained model for meticulous disease prediction. In this research the mutation spectrum accompanies all types of muscular dystrophy diseases for modeling and therefore the task of full sequencing is eliminated. This approach generalizes the disease identification task as an automated practice which can be applied to identify any kind of genetic disease. Also, the prediction model is more effective and reliable, since, it is generated based on intelligent hints collected from mutated gene sequences.

CONCLUSION

This study demonstrates the modeling of disease identification research as the problem of learning multi-class classification system that can suits in

bioinformatics environment to identify the disease effectively. It describes the implementation of supervised learning approach for identifying the genetic disease based on the mutational features. Five different models were built to identify the disease based on diverse features associated to different kind of mutations. Currently, this problem has not been broadly studied in the literature and existing approaches are either restricted to a small number of classes due to computational issues or insufficient data. The fore most task in this research is to design the discriminative features from the mutated gene sequences and to build a data driven models for identifying the type of the genetic disorder. The proposed AGM Model is a generalized model which can identify any kind of disease effectively by aggregating all type of mutational features. The outcome of the experiments proves that the disease identification model is effectual when the collective features are used in learning. The experiments conducted on the diseased gene sequences and assessed with evaluation method on the model built, show that our method is valuable than existing disease identification procedures with respect to significant features.

REFERENCES

Agnes, J.A.M.D., K.M.S. Karen and E.S.M.D. Michael, 2008. Charcot-Marie-Tooth Neuropathies: Diagnosis and Management. Thieme Medical Publishers, Stuttgart, Germany.

Ban, H.J., J.Y. Heo, K.S. Oh and K.J. Park, 2010. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. BMC. Genet., 11: 26-26.

Bennett, R.R., H.E. Schneider, E. Estrella, S. Burgess and A.S. Cheng *et al*., 2009. Automated DNA mutation detection using universal conditions direct sequencing: Application to ten muscular dystrophy genes. BMC Genet., 10: 66-66.

Briggs, F.B.S., P.P. Ramsay, E. Madden, J.M. Norris and V.M. Holers *et al*., 2010. Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis. Genes Immune., 11: 199-208.

Cartegni, L., J. Wang, A. Zhu, M.Q. Zhang abd A.R. Krainer, 2003. ESEfinder: A web resource to identify exonic enhancers. Nucl. Acid Res., 31: 3568-3571.

Chen, C., H. Ma, F. Zhang, L. Chen and X. Xing *et al*., 2014. Screening of Duchenne Muscular Dystrophy (DMD) mutations and investigating its mutational mechanism in Chinese patients. PloS One, Vol. 9,

Clancy, S., 2008. RNA splicing: Introns, exons and spliceosome. Nat. Educ., 1: 31-31.

Desmet, F.O., D. Hamroun, M. Lalande, G.C. Beroud and M. Claustres *et al*., 2009. Human splicing finder: An online bioinformatics tool to predict splicing signals. Nucleic Acids Res., 37: e67-e67.

Emery, A.E., 2002. The muscular dystrophies. Lancet, 359: 687-695.

Goh, K.I. and I.G. Choi, 2012. Exploring the human diseasome: The human disease network. Briefings Funct. Genomics, 11: 533-542.

Gonzalez, N.F.F., L.A.B. Munoz and K.A.S. Colon, 2013. Effective classification and gene expression profiling for the facioscapulohumeral muscular dystrophy. PloS one, Vol. 8,

Kalari, K.R., 2006. Computational approach to identify deletions or duplications within a gene. Ph.D Thesis, University of Iowa, Iowa City, Iowa.

Kann, M.G., 2010. Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. Briefings Bioinf., 11: 96-110.

Koenig, M., E.P. Hoffman, C.J. Bertelson, A.P. Monaco and C. Feener *et al*., 1987. Complete cloning of the Duchenne Muscular Dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell, 50: 509-517.

Lim, K.H. and W.G. Fairbrother, 2012. Spliceman-a computational web server that predicts sequence variations in pre-mRNA splicing. Bioinf., 28: 1031-1032.

Ma, J., M.N. Nguyen, G.W. Pang and J.C. Rajapakse, 2005. Gene classification using codon usage and SVMS. Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'05), November 15, 2005, IEEE, Singapore, Asia, ISBN:0-7803-9387-2, pp: 1-8.

Mercuri, E. and F. Muntoni, 2013. Muscular dystrophies. Lancet, 381: 845-860.

Mort, M., T.S. Weiler, B. Li, E.V. Ball and D.N. Cooper *et al*., 2014. MutPred splice: Machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol., Vol. 15, 10.1186/gb-2014-15-1-r19

Nicodemus, K.K., J.H. Callicott, R.G. Higier, A. Luna and D.C. Nixon *et al*., 2010. Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: Biological validation with functional neuroimaging. Hum. Genet., 127: 441-452.

Nisha, C.M., B. Pant and K.R. Pardasani, 2012. SVM model for classification of genotypes of HCV using relative synonymous codon usage. J. Adv. Bioinf. Appl. Res., 3: 357-363.

Noguchi, S., T. Tsukahara, M. Fujita, R. Kurokawa and M. Tachikawa *et al*., 2003. cDNA microarray analysis of individual Duchenne muscular dystrophy patients. Hum. Mol. Genet., 12: 595-600.

Reese, M.G., F.H. Eeckman, D. Kulp and D. Haussler, 1997. Improved splice site detection in Genie. J. Comput. Boil., 4: 311-323.

Roberts, R.G., M.A.R.T.I.N. Bobrow and D.R. Bentley, 1992. Point mutations in the dystrophin gene. Proc. National Acad. Sci., 89: 2331-2335.

Sathyavikashini, K. and M.S. Vijaya, 2015a. Predicting muscular dystrophy through genetic testing: A study. Proceedings of the International Conference on Innovative Trends in Electronics Communication and Applications (ICIECA 2015), December 19-20, 2015, Indian Institute of Technology Madras, Chennai, India, pp: 64-71.

Sathyavikasini, K. and M.S. Vijaya, 2015b. Predicting muscular dystrophy with sequence based features for point mutations. Proceedings of the 2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), November 20-22, 2015, IEEE, Coimbatore, India, ISBN:978-1-4673-6734-9, pp: 235-240.

Sathyavikasini, K. and M.S. Vijaya, 2016. Muscular dystrophy disease classification using relative synonymous codon usage. Intl. J. Mach. Learn. Comput., 6: 139-144.

Tranchevent, L.C., F.B. Capdevila, D. Nitsch, B.D. Moor and P.D. Causmaecker *et al*., 2011. A guide to web tools to prioritize candidate genes. Briefings Bioinf., 12: 22-32.

Uhmn, S., D.H. Kim, Y.W. Ko, S. Cho and J. Cheong *et al*., 2009. A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis. Expert Syst., 26: 60-69.

Woolfe, A., J.C. Mullikin and L. Elnitski, 2010. Genomic features defining exonic variants that modulate splicing. Genome Boil., Vol. 11, 10.1186/gb-2010-11-2-r20.

Wu, J., W. Zhang and R. Jiang, 2010. Comparative study of ensemble learning approaches in the identification of disease mutations. Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics (BMEI) 2010, Vol. 6, October 16-18, 2010, IEEE, Beijing, China, ISBN:978-1-4244-6495-1, pp: 2306-2310.

Yeo, G. and C.B. Burge, 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol., 11: 377-394.

Zeng, S., J. Yang, B.H.Y. Chung, Y.L. Lau and W. Yang, 2014. EFIN: Predicting the functional impact of nonsynonymous single nucleotide polymorphisms in human genome. BMC Genomics, 15: 455-455.