Muscular Dystrophy Disease Classification Using Relative Synonymous Codon Usage

K. Sathyavikasini and M. S. Vijaya

***Abstract*—Genetic diseases are predictable by the mutations in the gene sequences. Predicting a disease based on mutations is an important and challenging task in the medical diagnosis of genetic disorders such as Muscular dystrophy. Currently, this problem is handled for non-synonymous single nucleotide variants (SNVs) that capture only missense and nonsense mutations. Silent mutations do not result in changes in the encoded protein, but appear in the variation of codon usage pattern that results in disease. Hence, a new computational model is proposed for recognizing the disease using synonymous codon usage. The model adopts codon usage bias pattern as a feature vector by calculating the Relative Synonymous Codon Usage (RSCU) values from the mutated gene sequences. This paper addresses the problem by formulating it as multi-classification trained with feature vectors of fifty-nine RSCU frequency values from the mutated gene sequences. The outcome of trained model reports that the prediction accuracy of 86% in multi-class SVM with the RBF kernel.**

***Index Terms*—Codon, codon usage bias, positional cloning, RSCU, silent mutations.**

I. Introduction

A muscular dystrophy is a group of continuous muscle disorders caused by mutations in genes that encode for proteins that are necessary for regular muscle function [1]. Currently there is no cure for muscular dystrophy and the affected patients go down their life in earlier stages. The disease should be diagnosed early and effectively to understand and gain the life hope and quality of life of patients. Various types of muscular dystrophy are Duchenne, Becker, Emery-dreifuss, Limb-Girdle muscular dystrophy, Facioscapulohumeral, Myotonic and Congenital Muscular Dystrophy.

Duchenne Muscular Dystrophy (DMD) and Becker Muscular Dystrophy (BMD) are caused by the mutations in the dystrophin gene located on the X chromosome. Dystrophin is the massive human gene that is 2.5mb long and encompasses of 79 exons. The absence of dystrophin gene occurs when a large number of exons are deleted, which is the major cause of DMD. When the effect of mutations is less in the dystrophin gene then the disease is known as Becker's muscular dystrophy [2]. In this type of disease patients will have milder dystrophic phenotype.

The mutations in the Emerin (EMD) and Lamin A/C (LMNA) genes cause the third category of muscular dystrophy disease called as Emery-dreifuss muscular dystrophy (EMD). In this case the patients can be affected in their childhood or in the early asolescent years with muscle contractures.

Limb-Girdle Muscular Dystrophy (LGMD) can be seen in both boys and girls. Nearly 18 genes involved in the mutation of LGMD. The defects in 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. CMT causes mild and also severe muscle degeneration, which are dependent on its mutation. More than 30 forms of CMT are noticed and 30 genes are concerned.

Muscular dystrophy can be diagnosed with the results of muscle biopsy, electromyography, electrocardiography and DNA analysis. Muscle biopsy and DNA testing are widely used tests to predict muscular dystrophies. DNA or Genetic testing is an initial step tested on a blood sample to spot the alteration in the genes so as to help in the diagnosis of muscular dystrophy without performing a muscle biopsy. Carrier mothers, those who may be at risk of passing this disease on to their children are identified by genetic testing and preventive measures can be provided.

Amend in the genetic code that causes a permanent change in the DNA sequence is termed as mutation. DNA mutations perceptibly root to genetic diseases. Single character change in a gene makes an impact on the gene which in turn changes the function of the gene. Substitution is an exchange of one base to another, such as swapping a base from A to G. There are six types of mutations generally occur. They are Missense, Nonsense, Silent, Insertions, Deletions and Frame shift mutations.

Missense mutations are the substitution in a codon that encodes a different amino acid and cause a small change in the protein [3]. Nonsense mutations are those where the protein attains to stop codon when a change occurs in the DNA sequence.

Insertions are the mutations where a new base is added into the sequence that alters the function of a gene. An increase in the number of the same nucleotides in a location is termed as duplications. There may be single or gross insertions and duplications.

Deletions are the mutations when a base or an exon is deleted from a sequence the mutations. Frame shift mutations cause changes in the reading frame, [4] pointing to an induction of irrelevant amino acids into the protein, generally followed by a stop codon [5].

Silent mutations are changes in codon that encodes for the same amino acid and therefore the protein is not altered [6]. The information from the genes transfers the nucleic acid to

Manuscript received December 18, 2015; revised March 5, 2016.

The authors are with the PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India (e-mail: mail2sathyavikashini@gmail.com, msvijaya@psgrkc.com).

doi: 10.18178/ijmlc.2016.6.2.588

proteins in the form of codons. During the process of translation, the synonymous codons have different frequencies [7] which are referred as codon usage bias that is dynamic. The functionality of the gene depends on the codon usage bias [8]. Also, the protein tertiary structure is related to the codon bias pattern. When a silent mutation occurs the protein remains unchanged, but the variation occurs in the bias pattern. Therefore, the codon usage is important in mutation studies and in molecular evolution.

II. Synonymous Mutation

The silent mutation is a kind of point mutation which changes the codon usage pattern. The translated protein in the amino acid sequence is not modified with the synonymous codon changes. Modern investigations show that the silent mutation changes can affect protein folding and function. More than 50 diseases are correlated with this kind of point mutation. Silent mutation alters the secondary structure of mRNA and hence the stability of the mRNA will be reconstructed. Even though several codons encode for the same amino acid the frequency will differ and this is referred as codon bias.

Only very few research has been carried out on gene sequences based on RSCU (Relative Synonymous Codon Usage) to predict or classify either type of gene or virus or diseases. The authors in [8] 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 work 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 percent.

The authors C.M.Nisha, Bhasker Pant, and K. R. Pardasani proposed a new approach based on codon usage pattern to classify the type of Hepatitis C virus (HCV) that are the primary reason for the liver infection.To classify the subclass of its genotype a model was created using codon usage bias as input to multi class SVM [9].

The authors in [10] analyze synonymous codon variations in the protein coding genes of begomoviruses that cause severe disease in major crop plants. Fourteen codons were determined as translational optimal ones according to the comparison of codon usage patterns between highly and lowly expressed genes.

The above literature survey motivates that the classification of disease can also be carried out by modeling silent mutations with RSCU values. Hence, it is proposed in this paper to demonstrate the application of RSCU in disease gene sequences to model the silent mutations for predicting the type of muscular dystrophy diseases.

III. Muscular Dystrophy Disease Identification

The gene sequences and its pattern vary in every human. Also the pattern gets altered when mutations occur in the chromosome. Therefore the accurate prediction of the disease is a highly complicated and challenging task. The principal focus of this research is to provide an efficient machine learning solution for predicting the type of muscular dystrophy disease with the silent mutations. Multi-class classification is formulated through data modeling of gene sequences. The mutational gene sequences are generated as the synonymous mutated gene sequences are not explicitly available for this complicated disease. Five types of muscular dystrophy namely DMD, BMD, EMD, LGMD amd CMT have been considered for building the disease prediction model.

A. Disease Identification Model

The development of Muscular dystrophy disease Identification model comprises of five phases such as mutational gene sequence generation, feature extraction, RSCU calculation, building the model and classification. The framework of the proposed model is illustrated in Fig. 1.

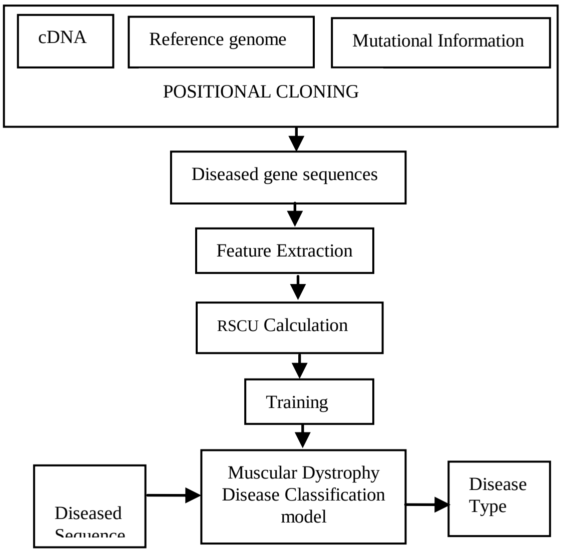


Fig. 1. Disease identification model.

B. Generation of Mutational Gene Sequences

Mutated gene sequences are generated through positional cloning based on the mutation and its location on the chromosome. The information on the position of mutations in the gene sequences is available in HGMD (Human Gene Mutation Database) [11] is a core collection of data on germ-line mutations in genes coupled with the human inherited disease which are grasped from various literatures. The public version of HGMD[1] is freely available to registered users from academic institutions/non-profit organizations.

The positional change of the nucleotide is done in cDNA sequence against the reference gene sequence and the new mutated gene sequences for muscular dystrophy are generated through R script. The cDNA sequence and the reference sequence are first stored as text files. Using the Stringreplace() function from the stringi library the required position is to be altered is identified and replaced with the nucleotide specified in the nucleotide change column of HGMD database.Five types of mutations have been considered for generated for generating mutated sequences. Using the traditional positional cloning approach the mutated sequences are generated and stored as fasta files.

[1] http://www.hgmd.org

Consider the silent mutational information for the DMD phenotype from the Dystrophin gene such as nucleotide change is 8628 G>A which indicates in the position 8628 the nucleotide changes from G to A and the protein is not altered. The codon change CAG-CAA reflects to the same protein Glutamine (Glu).

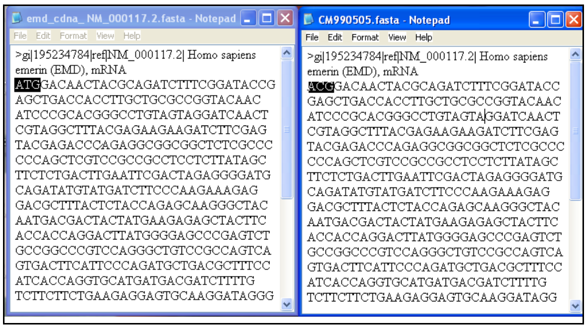
The cDNA sequence of EMD gene is

ACAGAGCAGCCTTTGGAAGGACTAGA

After the nucleotide change in the position 2

ACAGAGCAACCTTTGGAAGGACTAGA

The generation of a mutated gene sequence in Dystrophin gene with the mutational information is shown in Fig. 2.


emd_cdna_ NM_000117.2.fasta - Notepad
File Edit Format View Help
CM990505.fasta - Notepad
File Edit Format View Help

```
>gi|195234784|ref|NM_000117.2| Homo sapiens
emerin (EMD), mRNA
ATGGACAACTACGCAGATCTTTCGGATACCG
AGCTGACCACCTTGCTGCGCCGGTACAAC
ATCCCGCACGGGCCTGTAGTAGGATCAACT
CGTAGGCTTTACGAGAAGAAGATCTTCGAG
TACGAGACCCAGAGGCGGCGGCTCTCGCCC
CCCAGCTCGTCCGCCGCCTCCTCTTATAGC
TTCTCTGACTTGAATTCGACTAGAGGGGATG
CAGATATGTATGATCTTCCCAAGAAAGAG
GACGCTTTACTCTACCAGAGCAAGGGCTAC
AATGACGACTACTATGAAGAGAGCTACTTC
ACCACCAGGACTTATGGGGAGCCCGAGTCT
GCCGGCCCGTCCAGGGCTGTCCGCCAGTCA
GTGACTTCATTCCCAGATGCTGACGCTTTCC
ATCACCAGGTGCATGATGACGATCTTTTG
TCTTCTTCTGAAGAGGAGTGCAAGGATAGGG
```

```
>gi|195234784|ref|NM_000117.2| Homo sapiens
emerin (EMD), mRNA
ACGGACAACTACGCAGATCTTTCGGATACC
GAGCTGACCACCTTGCTGCGCCGGTACAAC
ATCCCGCACGGGCCTGTAGTAGGATCAACT
CGTAGGCTTTACGAGAAGAAGATCTTCGAG
TACGAGACCCAGAGGCGGCGGCTCTCGCCC
CCCAGCTCGTCCGCCGCCTCCTCTTATAGC
TTCTCTGACTTGAATTCGACTAGAGGGGATG
CAGATATGTATGATCTTCCCAAGAAAGAG
GACGCTTTACTCTACCAGAGCAAGGGCTAC
AATGACGACTACTATGAAGAGAGCTACTTC
ACCACCAGGACTTATGGGGAGCCCGAGTCT
GCCGGCCCGTCCAGGGCTGTCCGCCAGTCA
GTGACTTCATTCCCAGATGCTGACGCTTTCC
ATCACCAGGTGCATGATGACGATCTTTTG
TCTTCTTCTGAAGAGGAGTGCAAGGATAGG
```

Fig. 2. Output of generated mutated sequences.

There are about fifty five genes evacuated with five types of muscular dystrophy. Table I summarizes genes associated with the muscular dystrophy disease.

TABLE I: 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

For each phenotype, 30 mutated gene sequences are generated and dataset comprises of 150 mutated gene sequences combining all forms of muscular dystrophy is developed.

C. Feature Extraction and Training

Feature extraction plays an important role in machine learning for improving the classification effectiveness, computational efficiency or both. In this research work, the codon usage patterns are considered as the contributing features for representing the mutated gene sequences. Since codon usage patterns are diverse in different gene families, this feature input is a well-chosen discriptors for specifying different gene families for all types of diseases.

A codon is the triplet of nucleotides that code for a specific amino acid. Many to one relationship occurs 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 methionine (ATG) and tryptophan (TGG) have only one corresponding codon, they are not counted and are eliminated from the analysis 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 is 59. Therefore, irrespective of the size, the DNA sequence is converted to a feature vector of 59 elements.

The RSCU features are extracted from mutated gene sequences through R script that is created using seqinr() package downloaded from www.CRAN.org.

D. Calculation of RSCU

The differences in the frequency of occurrence of synonymous codons are referred as codon usage bias. The formula for calculating RSCU can be explained as, the number of times a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias [8].

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 formula is used to calculate RSCU.

$$\text{RSCU} = X_{ij} / (1/n_i *\text{S} \{X_{ij}; j=1, n_i\})$$

where X_{ij} is the number of occurrences of the j^{th} codon for the i^{th} amino acid, and n_i is the number of alternative codons for the i^{th} amino acid.

If the synonymous codons of an amino acid are used with equal frequencies, then their RSCU values are 1.

In this manner the RSCU values are derived for 59 codons from each mutated gene sequence which forms a feature vector for classification task. Since the corpus consists of 150 sequences of 5 types of Muscular dystrophy diseases, a training set with 150 feature vectors has been created and for each feature vector the class label is assigned from 1 to 5 indicating the five types of muscular dystrophy diseases.

IV. MACHINE LEARNING SCHEMES

Four standard supervised pattern learning algorithms such as Decision Tree learning, Na ïve bayes classifier, artificial neural network and Support vector machine have been used for learning the muscular dystrophy disease type prediction model.

A. Decision Tree Learning

Classification in the Decision tree learning is performed by sorting the instances down the tree from the root to a leaf node, which provides the classification of the instance [12]. A binary tree like structure is generated as the outcome of the decision tree, where the choice between the alternatives is represented in each branch node and each leaf node represents a classification or decision. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. A Decision Tree model contains rules to predict the target variable [12].

TABLE II: RSCU Values for 59 Codons

Codon	Value	Codon	Value	Codon	Value
AAA	1.05	CCC	0.97	GGC	0.92
AAC	0.812	CCG	0.12	GGG	0.75
AAG	0.948	CCT	1.64	GGT	0.64
AAT	1.18	CGA	0.87	GTA	0.81
ACA	1.52	CGC	0.54	GTC	0.93
ACC	0.76	CGG	0.66	GTG	1.40
ACG	0.24	CGT	0.63	GTT	0.85
ACT	1.48	CTA	0.73	TAC	0.61
AGA	1.84	CTC	0.87	TAT	1.38
AGC	0.99	CTG	1.41	TCA	1.23
AGG	1.42	CTT	1.03	TCC	0.91
AGT	1.36	GAA	1.22	TCG	0.14
ATA	0.52	GAC	0.81	TCT	1.33
ATC	1.10	GAG	0.77	TGC	1.16
ATT	1.36	GAT	1.18	TGT	0.833
CAA	0.87	GCA	1.23	TTA	0.71
CAC	0.86	GCC	1.18	TTC	0.64
CAG	1.13	GCG	0.15	TTG	1.23
CAT	1.14	GCT	1.42	TTT	1.63
CCA	1.25	GGA	1.67		

B. Na ïve Bayes Classifier

A Naive Bayesian model is simple to construct, but useful for very large datasets with no difficult iterative parameter estimation. The Naive Bayes classifier (NB) can be applied in the applications such as natural language processing, information retrieval [13]. A Naive Bayes probabilistic classifier is based on the Bayes' theorem. Na ïve Bayes classifiers assume that the effect of a variable value in a given class is independent of the values of other variables. The Naive-Bayes activator figures out the conditional probabilities of the classes given the instance and picks the class with the top posterior value [14].

C. Artificial Neural Network

An artificial neuron network (ANN) is based on the structure and function of the biological network that forms a computational model. The structure of the ANN will be based on the information flows through the network. A technical neural network consists of simple processing units, the neurons, and directed, weighted connections between those neurons. Data are transferred between neurons via connections with the connecting weight being either excitatory or inhibitory. The performance of ANN will depend upon the trained parameters and the data set relevant to the training. The foremost advantage of this algorithm is that it can be easily implemented in parallel architectures. The Deep Neural Networks are an interesting kind of ANNs that works with a giant number of units and layers that achieves a high level of learning with low supervision.

D. Support Vector Machine

Support Vector Machine a new approach to supervised pattern classification which has been successfully applied to a wide range of pattern recognition problems. SVM is most suitable for working accurately and efficiently with high dimensionality feature spaces. SVM is based on strong mathematical foundations and results in a simple yet very powerful algorithm [15]-[17].

The standard SVM algorithm builds a binary classifier. A simple way to build a binary classifier is to construct a hyperplane separating class members from non-members in the input space. The system automatically identifies a subset of informative points called support vectors and uses them to represent the separating hyperplane which is sparsely a linear combination of these points.

The simplest model of SVM called Maximal Margin classifier, constructs a linear separator (an optimal hyperplane) given by $w^Tx - \gamma = 0$ between two classes of examples. The free parameters are a vector of weights **w** which is orthogonal to the hyperplane and a threshold value γ. These parameters are obtained by solving the following optimization problem using Lagrangian duality

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} \quad D_{ii}(\mathbf{w}^T\mathbf{x}_i - \gamma) \geq 1, i = 1, \ldots, l. \end{aligned}$$

where D_{ii} corresponds to class labels, which assumes value +1 and –1. The instances with non null weights are called support vectors.

When the number of classes is more than two, then the problem is called multiclass SVM. There are two types of approaches for multiclass SVM. In the first method called indirect method, several binary SVM's are constructed and the classifier's output are combined for finding the final class. In the second method called direct method, a single optimization formulation is considered. The formulation of one of the direct methods called Crammer and Singer Method [18] is

Minimize

$$\frac{1}{2}\sum_{k=1}^{N}\mathbf{w}_k^T\mathbf{w}_k + C\sum_{i=1}^{n}\xi_i$$

subject to the constraints

$$\mathbf{w}_{k_i}^T\phi(\mathbf{x}_i) - \mathbf{w}_k^T\phi(\mathbf{x}_i) \geq e_k^i - \xi^i \quad , \forall k \neq k_i$$

where k_i is the class to which the training data xi belong,

$$e_k^i = 1 - c_k^i$$

$$c_k^i = \begin{cases} 1 \text{ if } k_i = k \\ 0 \text{ if } k_i \neq k \end{cases}$$

The decision function for a new input data x_i is given by

$$\hat{d}_j = \arg \max_k \left\{ f_k(\mathbf{x}_j) \right\}$$

where

$$f_k(\mathbf{x}_j) = \mathbf{w}_k^T \phi(\mathbf{x}_j) - \gamma_k$$

V. Experiment and Results

Muscular dystrophy disease identification model is developed using R, an open source software environment for statistical computing, a universe of analysis. The sequence analysis packages are downloaded from Bioconductor the biological software repository of R. Five types of muscular dystrophy disease are taken into account for implementing the model and thus the disease identification problem becomes multi classification.

In this experiment, the mutated gene sequences of muscular dystrophy are generated through positional cloning. The diseases are identified from the silent mutations occurred in the gene sequences by calculating the RSCU features for the synonymous codon usage. The training data set with instances related to five categories of muscular dystrophy that is Duchenne muscular dystrophy, Becker's muscular dystrophy, Emery-Dreifuss, Limb-girdle muscular dystrophy and Congenital muscular dystrophy has been developed as described in Section III.

Evaluating the generalization power of the classifiers and to estimate their predictive capabilities for unknown samples, a standard 10- fold cross-validation technique is used to split the data randomly and repeatedly into training and test sets.

The standard supervised pattern learning techniques, namely Naïve Bayes Classifier, Decision tree induction, artificial neural network and Support vector machine (SVM) have been used to learn and build the classifiers. Independent trained models have been used for predicting the type of disease. The performance of trained models is evaluated using 10-fold cross validation and measured in terms of classification accuracy. The prediction accuracy is defined as the ratio of the number of correctly classified instances in the test dataset and the total number of test cases. In this work, Radial basis kernel is employed for the multi-class SVM classification with the cost value of 1 and gamma value 0.016. The number of support vectors created by this model is 90. The results of the experiments are summarized in Table III and Table IV. The prediction accuracy is illustrated in Fig. 3.

TABLE III: Predictive Performance of the Classifiers

	Classifiers			
Evaluation criteria	NB	ANN	Decision Tree	SVM
Kappa Statistic	0.784	0.776	0.786	0.822
Correctly classified instances	126	125	127	128
Incorrectly classified instances	24	25	23	21
Prediction accuracy	84%	83.33%	84.6%	86%

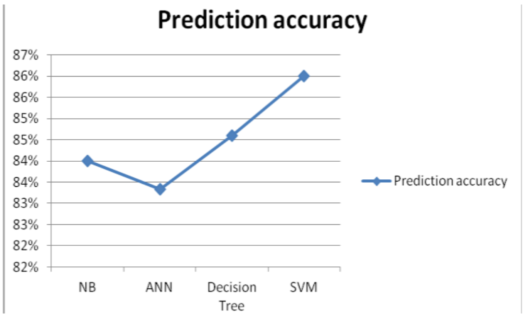


Fig. 3. Prediction accuracy.

TABLE IV: Predictive Performance of the Classifiers

Class	Classifier	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
1	ANN	1	0.80	0.6	0.80	0.22	0.22	0.37	0.9
	NB	0.23	0.78	0.23	0.22	0.05	0.05	0.22	0.5
	DT	1	0.8	0.6	1	0.22	0.22	0.37	0.9
	SVM	1	0.78	0.56	1	0.22	0.22	0.4	0.89
2	ANN	0.33	1	1	0.85	0.22	0.07	0.07	0.66
	NB	0.3	0.8	0.3	0.22	0.06	0.06	0.22	0.55
	DT	0.33	1	1	0.84	0.22	0.07	0.07	0.6
	SVM	0.23	1	1	0.82	0.22	0.05	0.05	0.61
3	ANN	0.93	0.98	0.93	0.78	0.21	0.2	0.21	0.95
	NB	0.21	0.78	0.2	0.21	0.04	0.04	0.21	0.49
	DT	0.93	0.98	0.93	0.98	0.21	0.2	0.21	0.95
	SVM	1	1	1	1	0.21	0.21	0.21	1
4	ANN	1	0.98	0.92	0.80	0.19	0.18	0.20	0.99
	NB	0.16	0.8	0.16	0.18	0.02	0.02	0.18	0.48
	DT	1	0.98	0.92	1	0.18	0.18	0.20	0.99
	SVM	1	1	1	1	0.18	0.19	0.19	1
5	ANN	0.9	1	1	0.83	0.16	0.14	0.14	0.95
	NB	0.19	0.85	0.19	0.15	0.03	0.03	0.15	0.52
	DT	0.9	1	1	0.98	0.15	0.14	0.14	0.95
	SVM	1	1	1	1	0.15	0.15	0.15	1

From the above results, it is perceived that the kappa statistic and prediction accuracy are high for SVM than other algorithms. The sensitivity and specificity measure for all classes is prominent in SVM with RBF kernel when compared with other learning techniques. SVM gives an elevated score value for the positive predictive value and negative predictive value. The proportion value of the particular disease is given by the prevalence that gives a stabilized value in SVM for all the five classes. The sensitivity measure or recall depends on the prevalence and where the specificity is independent of prevalence. The detection rate and detection prevalence also depend on the prevalence measure which is also stabilized. Overall, the balance accuracy measure is also eminent in SVM when measured with other algorithms.

As this experiment is the computational based approach, the cost incurred in laboratory is reduced by capturing the attributes of the synonymous mutations as feature vectors. The model creates an automatic system for the disease prediction and it creates an error free report and it is more reliable and the output will be more effective. The proposed model aids in classifying type of muscular dystrophy in mutated gene sequences by capturing RSCU features of silent mutations.

VI. Conclusion

This research work demonstrates the development of muscular dystrophy disease prediction model using mutated gene sequences and Codon usage bias. The RSCU values are extracted as features and a model is built by employing supervised machine learning algorithms such as NB, ANN, Decision Tree and SVM. The performance of the learning methods was evaluated based on their predictive accuracy. The results indicate that the SVM outperforms in prediction than Decision tree learning, Naïve Bayes methods and artificial neural network. As the nature of the application demands more accurate prediction, it is found that the SVM is better to identify the type of muscular dystrophy disease in DNA analysis with features of silent mutations. This model can also be applied in investigating the changes in protein folding and function. The work can be further extended by adding more sequences and repeating the experiment with other techniques.

References

[1] L. Fajkusova, Z. LukasIb, M. Tvrdoakova, V. Kuhrova, J. Haajekb, and J. Fajkusc, “Novel dystrophin mutations revealed by analysis of dystrophin mRNA: Alternative splicing suppresses the phenotypic effect of a nonsense mutation,” *Neuromuscular Disorders*, vol. 11, 2001.

[2] E. E. Zubrzycka-Gaarn, D. E. Bulman *et al*., “The Duchenne muscular dystrophy gene product is localized in sarcolemma of human skeletal muscle,” *Nature*, vol. 333, no. 6172, pp. 466-469, 1988.

[3] M. G. Kann, “Advances in translational bioinformatics: computational approaches for the hunting of disease genes,” *Briefings in Bioinformatics*, vol. 11, pp. 96–110, 2009.

[4] L.-C. Tranchevent *et al.*, “A guide to web tools to prioritize candidate genes,” *Briefings in Bioinformatics*, vol. 12, pp. 22–32, 2010.

[5] K. N. North and K. J. Jones, “Diagnosing childhood muscular dystrophies,” *Journal of Paediatrics and Child Health*.

[6] M. Koenig, E. P. Hoffman, C. J. Bertelson *et al*., “Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene innormal and affected individuals,” *Cell*, vol. 50, pp. 509-517, 1987.

[7] D. Charif, J. Thioulouse, J. R. Lobry, and G. Perrière, “Online synonymous codon usage analyses with the ade4 and seqinR packages,” *Bioinformatics Oxford Journal*, vol. 21, no. 4, pp. 545-547, 2005.

[8] J. M. Ma, M. N. Nguyen, G. W. L. Pang, and J. C. Rajapakse, “Gene Classification using Codon Usage and SVMs,” IEEE, 2005.

[9] C. M. Nisha, B. Pant, and K. R. Pardasani, “SVM model for classification of genotypes of HCV using relative synonymous codon usage,” *Journal of Advanced Bioinformatics Applications and Research*, vol. 3, issue 3, pp. 357-363, 2012.

[10] X.-Z. Xu *et al*., “Analysis of synonymous codon usage and evolution of begomoviruses,” *Journal of Zhejiang University Science B*, vol. 9, no. 9, pp. 667-674, 2008.

[11] P. D. Stenson, M. Mort, E. V. Ball *et al.*, “The human gene mutation database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine,” July 2013.

[12] Decision tree learning. [Online]. Available: www.cs.princeton.edu/courses/archive/spring07/cos424/papers/mitchell-dectrees

[13] F. C. Peng, *Augmenting Naive Bayes Classifiers with Statistical Language Models*, University of Massachusetts, Amherst, 2003.

[14] I. H. Witten and E. Frank, *Data Mining — Practical Machine Learning Tools and Techniques*, 2nd Edition, Elsevier, 2005.

[15] T. Joachims, B. Schölkopf, C. Burges, and A. Smola, “Making large-scale SVM learning practical,” *Advances in Kernel Methods - Support Vector Learning*, MIT Press, Cambridge, MA, USA, 1999.

[16] S.-T. John and N. Cristianini, “Support vector machines and other kernel-based learning methods,” Cambridge University Press, UK, 2000.

[17] V. N. Vapnik, *Statistical Learning Theory*, J. Wiley & Sons, Inc., 1998.

[18] K. P. Soman, R. Loganathan, and V. Ajay, *Machine Learning with SVM and Other Kernel Methods*, PHI, India, 2009.

K. Sathyavikasini pursuing her doctoral program from PSGR Krishnammal College for women, Coimbatore. data mining is her research area. She had participated in various conferences, workshops and seminars. She received the Best Out Going Student Award in her PG degree and she is a proficiency rank holder. Her areas of interest include bioinformatics, data mining, machine learning, R programming, big data and deep learning.