ABSTRACT

As the growth of data increases, storage and analysis becomes incredible, this in turn increases the processing time and cost efficiency. Data mining and machine learning techniques aids in extraction and prediction of effective information from the data. In the current genomic era, a large collection of genes has been cataloged in the human genome. This biological data paves the way for challenges in computing. High-throughput genotyping and sequencing techniques are rapidly and inexpensively provides large amounts of human genetic variation data. Prediction of complex diseases is an important issue in bioinformatics research. Identifying trait diseases through DNA analysis is a prime task in diagnosing an ailment. Huge number of sequence variations is found in large set of genes. Genetic disorders are caused by the deformities in the inherited genes and an accurate gene test facilitates in finding the disease. Identification of genetic features for complex diseases is a far more difficult task with the standard methods. Disease identification model is essential to represent this knowledge in a computational form with minimal loss of biological context through a gene based approach.

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 and 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. Nonsense mutations are those where the protein attains to stop codon when a change occurs in the DNA sequence. Silent mutations are changes in codon that encodes for the same amino acid and therefore the protein is not altered. 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. Deletions are the mutations when a base or an exon is deleted from a sequence the mutations. There may be single or gross insertions and duplications.

Muscular dystrophy is a group of hereditary progressive muscle disorders caused by mutations in genes that encode for proteins that are necessary for regular muscle function. Muscular dystrophy is an association of muscle diseases that worsen the musculoskeletal system and hamper the locomotive performance. Muscular Dystrophy is a multi-system disorder

exhibiting indications in quite a lot of organ systems in human body. The subset of Muscular dystrophy subtypes hit distinct sets of muscles in various onset ages, severity and patterns of inheritance. There are about 30 major forms in muscular dystrophy and a better understanding is needed to predict this genetic disease. The mutation in the genes causes most of these disorders. There are currently no effective treatments to halt the muscle breakdown in muscular dystrophies.

However, finding muscular dystrophy disease comparing with other hereditary traits seems very complicated. Lack of specific features to guide diagnosis and the wide range of phenotypes that are possible with many genetic forms combine to make it a challenge to predict the correct genetic cause from the laboratory reports and history. Therefore, muscular dystrophy disease prediction needs to be made systematic to find the type of disease accurately for proper diagnosis and treatment. The focus of this research is to develop a new model for predicting the muscular dystrophy disease accurately with the gene sequences based on the computational techniques.

The main aim of this research work titled “Identification of Rare Genetic Muscular Dystrophy from Gene Sequences and Mutation Based Features through Shallow and Deep Learning” is to propose models for predicting the major five forms of muscular dystrophy disease automatically from cloned gene sequences through self extracted and hand crafted features based on shallow and deep learning. Duchenne Muscular Dystrophy (DMD), Becker Muscular Dystrophy (BMD), Emery-dreifuss Muscular Dystrophy (EMD), Limb-Girdle Muscular Dystrophy (LGMD) and Charcot Marie Tooth disease (CMT) are five major categories of muscular dystrophy considered for this study.

The core objectives of this research work are as follows

- To generate synthetic gene sequences by adopting positional cloning approach
- To identify and capture discriminative descriptors from the diseased gene sequences related to mutations like missense, nonsense, silent, insertion, deletion, duplication and splicing
- To build autonomous disease identification models based on different kinds of mutational features using supervised pattern classification techniques
- To build muscular dystrophy disease identification model using ensemble learning approach
- To define two mapping schemes namely nucleotide mapping and codon mapping schemes to encode the diseased gene sequences for deep learning

- To build Deep Neural Network classifiers with nucleotide mapping and codon mapping for predicting the category of diseased gene sequences in tensorflow environment

The thesis explains a novel and never before tried methodology in disease identification wherein the muscular dystrophy disease identification problem is modeled as a classification task. Muscular dystrophy disease classification is done based on both hand crafted mutational features and self extracted features through shallow and deep learning.

The key point of this research is to pinpoint discriminative descriptors and to provide an efficient machine learning solution for predicting the type of muscular dystrophy disease. Multi-class classification is worked out through data modeling of gene sequences. The availability of diseased gene sequences is a real challenge for this intricate disease, which stimulates the need for the generation of synthetic mutational gene sequences. The cloned gene sequences are synthesized based on the mutation position and its location on the chromosome by employing the positional cloning approach. The information about the position of mutations in the gene sequences is available in HGMD (Human Gene Mutation Database). It is a collection of data on germ-line mutations in genes with their human hereditary disease which are grasped from various literatures. The reference genes are identified from OMIM (Online Mendelian Inheritance in Man) database and its corresponding reference gene sequences are downloaded from NCBI (National Center for Biotechnology Information). The positional change of the nucleotide is done in cDNA sequence against the reference gene sequence and the new mutated gene sequences for five forms of muscular dystrophy diseases are generated. For the purpose of this research work, in each category of muscular dystrophy disease, 200 synthetic mutated gene sequences are generated and a corpus comprising of 1000 sequences is developed.

In this work, the discriminative features associated with different kind of mutations are defined and extracted to capture the effect of mutations in the gene sequences to construct the feature vectors/datasets. In the first case, the features related to missense and nonsense (Non-synonymous) mutations are considered. Annotation, structure and alignment features have been extracted from the corpus of sequences and the dataset NSM with dimension 26 is generated.

In the second case, the silent mutational features are taken into account as it is required to identify the disease that is caused due to synonymous mutations. 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 descriptors

for specifying different gene families for all types of diseases. Codon usage bias helps in identifying Silent mutations and hence 59 RSCU (Relative Synonymous Codon Usage Bias) values have been determined from the same corpus and the dataset SYM with 1000 feature vectors is created.

Prediction of muscular dystrophy disease using features related to insertion and deletion mutations was done in third experiment. The extrinsic and intrinsic features more solely depend on the exons and introns that enable to identify the disease affected by large insertions and deletions. Twenty three such exonic and intronic descriptors related to Insertion/Duplication, deletion were extracted from the gene sequences and dataset IDM is prepared.

In the next case, the mutations occurred while spicing is considered to know the alteration after the splicing process as the exons are formed by splicing out the introns during transcription. Exon, Single Nucleotide Polymorphism (SNP) and gene features are taken into account. Position of the spliced introns and exons are carefully examined and twenty four such features are identified and extracted to capture the variations due to splicing in the mutated gene sequences. The dataset SPM of size 1000 instances with dimension 24 is generated.

Autonomous data driven models have been built based on the above mutational feature sets using supervised classification algorithms such as Decision tree, Artificial neural network, Naïve bayes and Support Vector Machines. The predictive performance of the disease classification models are evaluated using 10-Fold cross validation and analyzed using various metrics like predictive accuracy, precision, recall, F-measure and time taken to build the model.

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. Information gain feature selection method is employed to select high ranked effective features and the dataset AGM is generated. Data driven models are built using above mentioned supervised classification algorithms and the predictive performance of the models have been analysed.

In machine learning, the hybrid approach has been an ongoing research area for gaining best performance for classification or prediction problems over a single learning approach. Ensemble models have been developed using LibD3C classifiers by learning the above five independent datasets. The performance of these ensemble models are evaluated in the similar fashion and the results are compared with standard pattern recognition algorithms.

In deep learning approach, two mapping schemes such as nucleotide mapping and codon mapping were proposed by considering the diseased gene sequence as a sequence of categorical values. The trick in deep learning approach to represent biological data is converting disease gene sequence into 1-D representation by encoding. Deep neural network was employed to build the muscular dystrophy disease identification model. Deep models were built with the self extraction of features using deep neural network coded through jupyter notebook in TensorFlow environment. Self taught learning attempts to automatically learn good features or representations based on training data. Deep neural network classifier was implemented based on scikit flow with 1000 gene sequences. A,T,G,C nucleotide values were hardcoded using nucleotide and codon mapping schemes. The disease gene sequences were converted into numpy array based on the hardcoded values and directly fed into tensorflow. Tensorflow linear classifier and Tensorflow Deep neural network were employed and their parameters were tuned to attain good results.

Exhaustive experimentations carried out on disease gene sequences ascertain that the classification modeling through shallow and deep learning is effective for predicting the type of neuro muscular disorder muscular dystrophy. In this work the mutation spectrum accompanies all types of muscular dystrophy diseases for modeling and therefore the task of full sequencing is eliminated. It is concluded that shallow and deep learning techniques are suitable to predict muscular dystrophy disease when any types of mutational features are utilized for building the models. These approaches generalize the disease identification task as an automated practice, which can be applied to identify any kind of genetic disease. These approaches for disease identification exceedingly simplify the traditional disease identification problem and the prediction model is more effective, reliable since it is generated based on intelligent hints collected from mutated gene sequences. The promising results obtained from these approaches will facilitate the scientist to identify the hereditary traits from the features extracted from gene sequences for global genetic disease prediction.