Cystic fibrosis (CF; MIM#219700) is a life-threatening autosomal recessive disorder commonly seen in the populations of European descendants [1]. The disorder is caused due to mutations arising in the gene that encodes for the cystic fibrosis trans membrane conductance regulator (CFTR). The CFTR gene is located on the long arm of chromosome-7, region q31 [2]. The CFTR protein is a member of the ABC-transporter family of proteins and it is located in the apical membrane of epithelial cells. CFTR protein comprises of two membrane spanning domains (MSD1 and MSD2) that form the chloride ion channel; two nucleotide-binding domains (NBD1 and NBD2) that bind and hydrolyzes adenosine triphosphate (ATP); and a regulatory (R) domain involved in the cAMP-activated transport of chloride, bicarbonate and glutathione [3].

The major disease attributes of loss of CFTR function includes increased chloride concentrations in sweat [4]; low chloride conductance of airway epithelium [5]. It has been observed that the major disease-causing mutations of CFTR occur in the sequence that codes for the first NBD1. Understanding gene variations could provide essential insights into the role of these variations that could influence the severity of the disease and symptom progression. The availability of human whole genome sequence [6, 7] made it possible to investigate the effects of missense or non-synonymous mutations on the structure and functions of a gene and related protein [9, 10]. These investigations involving bioinformatics tools could provide critical information regarding the deleterious or non-deleterious nature of missense mutations and aid in developing potential therapeutics to eliminate or reduce disease condition. In this study, we report those nsSNPs that were identified either as deleterious or damaging to the overall functioning of the CFTR protein.

This study was carried out using some of the regularly used computational methods to analyze CFTR gene variants. A comparative analysis was also performed to look into those nsSNPs found to be deleterious or damaging that were common in all the bioinformatics tools tested. The required dataset of CFTR gene variants (nsSNPs) and their related protein sequences from Homo sapiens was retrieved from the dbSNP for computational analysis [11]. Each SNP carries unique ID and reference ID (rsIDs). Details about the SNP and the amino acid changes in its structures including positions and their corresponding accessions IDs were obtained by hitting on each rsIDs button. The dataset was subjected to at least three bioinformatics tools viz. SIFT, PolyPhen-2 and I-Mutant2 analyses.

Objective: The major objective of the study was to carry out comparative bioinformatics analyses to identify different nsSNPs that were predicted to be deleterious or damaging to the structure and functions of CFTR protein causing cystic fibrosis.

Methods: The CFTR gene variants (nsSNPs) and their related protein sequences from Homo sapiens were subjected to computational analyses using the following bioinformatics tools (a) SIFT: a sequence-homology based prediction tool that can be used to distinguish between the intolerant from tolerant SNP changes. (b) PolyPhen2: a structure and sequence-based physical and comparison tool to study the impact of amino acid substitution on the structure and function of human proteins and (c) I-Mutant2: to predict the protein stability changes arising due to single point mutations.

Results: SIFT, PolyPhen2, and I-Mutant2 analyses indicated that 21 out of 108 nsSNPs were identified to be common that were strongly predicted to be deleterious and damaging for CFTR protein in cystic fibrosis conditions. Most of the substitutions in the CFTR protein contained the amino acids valine followed by cysteine and proline respectively. Homology modeling carried out to determine if any of these nsSNPs had a role in changing the conformation of CFTR protein drastically. Homology modeling of selected nsSNP variants indicated that these substitutions, however, did not change the overall CFTR protein structure but predicted to cause severe damaging changes to the phenotypes of CFTR protein. Results indicated that multiple bioinformatics tools are needed to predict the effect of substitutions and these prediction tools need to be analyzed more into detail and common determination factors are required to predict a nsSNP to be deleterious or damaging to the overall functioning of the CFTR protein.

Conclusion: Multiple bioinformatics tools are in fact the need of the hour to establish if a strong relationship between nsSNPs that could alter the protein stability and cause a deleterious or damaging phenotypic change to the individual with cystic fibrosis involving the CFTR protein.

Keywords: Cystic fibrosis, CFTR protein, SIFT, PolyPhen2, I-Mutant2, Homology model
SIFT is a sequence-homology based prediction tool that can be used to
distinguish between the intolerant from tolerant SNP changes. This
tool can predict if an amino acid substitution can lead to phenotypic
changes in the protein [12]. The notion behind this method is that the
evolution of the protein is correlated with its function indicates that
phenotypic value with other methods. SIFT and PolyPhen2 are the two tools
in this range are more confidently predicted to be deleterious. The overall
predictions based on scores are (a) 0.0-0.15: Variants with scores in
this range are predicted to be benign (b) 0.15 to 1.0: Variants with
scores in this range are also predicted to be damaging. One important
observation would be that PolyPhen-2 and SIFT scores fall in the same
range, 0.0 to 1.0, but with quite opposite implications. A CFTR variant
with a PolyPhen-2 score of 0.0 is predicted to be benign whereas a CFTR variant with a SIFT score of
1.0 is predicted to be benign.

I-Mutant v2.0 is a Support Vector Machine (SVM) based tool to predict
the protein stability changes arising due to single point mutations [14].
The initiations were done either by using protein structure or more
precisely from the protein sequence. The output values are calculated
as free energy changes represented by $\Delta G$. A positive $\Delta G$ value
indicates the protein exhibiting a higher stability and vice versa. Also,
the results can be interpreted in terms of Reliability Index on a scale of
0-9. A high RI score indicates the protein to be highly stable while a
less RI score indicates that the protein is relatively less stable towards
AA substitutions. RI scores on a scale 0-9 were calculated. In this
study, RI values of >5 were regarded as highly unstable while RI scores
of ≤5 were regarded as highly stable.

Finally, randomly selected deleterious or damaging nsSNPs were used
to prepare 3-D models using SWISS-MODEL software. SWISS-
MODELL is at present the most accurate method to generate reliable
three-dimensional protein structure models and is routinely used in
many practical applications [15]. Homology (or comparative)
modeling methods make use of experimental protein structures
(‘templates’) to build models for evolutionarily related proteins
(‘targets’). Template search was carried out using BLAST (Basic
Local Alignment Search Tool) [16] against the SWISS-MODEL
template library. The target sequence was searched with BLAST (https://blast.ncbi.nlm.nih.gov/blast)
against the primary amino acid sequence contained in the SMTL. The template’s
quality for each identified template has been predicted from features of
the target-template alignment. The templates with the highest
quality have then been selected for model building. Models are built
based on the target-template alignment using ProMod-II.

Coordinates which are conserved between the target and the
template are copied from the template to the model. Insertions and
deletions are remodeled using a fragment library. Side chains are then
rebuilt. Finally, the geometry of the resulting model is regularized by using a force field. In case loop modelling with
ProMod-II does not give satisfactory results, an alternative model is
built with MODELLER.

A total of 108 nonsynonymous single nucleotide polymorphisms
(nsSNPs) in the CFTR protein were identified and manually
retrieved from dbSNP for analyses. Using bioinformatics tools, a
comparative approach was carried out to identify deleterious
substitutions in the nsSNPs and the influence of these substitutions
on the stability of the CFTR protein. The selected nsSNPs were
individually subjected to the following tools viz. SIFT, PolyPhen2 and
I-Mutant2 and the results were analyzed for further analysis
( Supplementary table 1). The outcomes of the analyses were
summarized by the following means to determine if a nsSNP would
be deleterious or probably damaging or benign. (A) The SIFT scores
were represented as tolerance index (TI) values. TI values of ≤0.05
are predicted by the algorithm to be intolerant (weak) or deleterious
amino acid substitution, whereas TI scores ≤0.5 are predicted to
be deleterious (strong). Higher tolerance index indicates that the protein
encounters a less functional impact towards a substitution. (B) PolyPhen2 analyses are based on Position-Specific Independent
Counts (PSIC) 5D scores: 0.00-0.15: Benign; 0.16-1.00: Possibly
damaging (PoD); 0.85-1.0: Confidently predicted to be damaging
(CPD). (C) I-Mutant result analysis was carried out to indicate the
protein stability changes based on Reliability Index (RI) scores. A
high RI score indicates the protein to be highly stable while a less RI
score indicates that the protein is relatively less stable towards AA
substitutions. RI scores (scale 0-9) (RI values of ≤5 were regarded as
stable; 6-9 were regarded as highly stable) were predicted to be
not found. Selected nsSNPs that were predicted to be deleterious,
damaging and destabilizing the CFTR protein by all the three
methods were randomly selected and were used to build homology
models based on the normal functioning CFTR protein to observe for
any structural changes in the CFTR protein.

SIFT analysis of the selected 108 nsSNPs identified 49 variants to
possess least tolerant (deleterious) for substitutions. Data could not be
generated for 36 variants that were indicated as not found. Remaining
variants were predicted to be highly stable. PolyPhen2 analysis predicted a total of 91 substitutions to be highly damaging.
The remaining variants seem to be either benign or least damaging
on the CFTR protein organization. I-Mutant analysis resulted in
a diversified data compared to SIFT and PolyPhen. Some of the
substitutions that were predicted to be highly deleterious or
damaging in SIFT or PolyPhen2 seem to possess stable protein
structures with higher RI scores in I-Mutant analysis. This may be
due to the fact that not all the substitutions necessarily bring about
changes in the protein architecture resulting in a decrease in the
stability of the protein. However, we have identified 21 AA
substitutions to be potentially damaging or deleterious in all the
three bioinformatics tools analyzed. The 21 AA substitutions found
to be common in all the three analyses were F508C, D1270H, G551D,
D649V, G1707P, D1249E and E92K respectively. Three of the 21
substitutions had valine (nonpolar and hydrophobic) in place of
other amino acids; three substitutions had cysteine (-SH containing
amino acid); two substitutions had proline (nonpolar, amino acid);
two substitutions had lysine (positively charged); two substitutions had
tryptophan (aromatic amino acid); one each of histidine (positively
charged), aspartic acid (negatively charged), asparagine (polar
amino acid), serine (polar amino acid), arginine (positively charged);
tyrosine (aromatic amino acid) and phenylalanine (aromatic amino
acid) respectively. The details of these 21 nsSNPs are summarized in
table 1. A few of these 21 variants were randomly selected for
developing homology models based on the native CFTR protein
domain and structure. SIFT, PolyPhen, PupaSuite, FASTSNP, ASA View, DSSP and
SRate tools were used to identify the deleterious nsSNPs that are
likely to affect the function and structure of the protein and showed
the hitSNPs which are in the haplotype blocks using HAPAP analysis.
Based on an evolutionary perspective SNPs identified using SIFT
tool were indicated that 17 nsSNPs (44%) were found to be deleterious.
PolyPhen server identified 26 nsSNPs (66%) may disrupting protein
function and structure. The Pupa Suite tool predicted the phenotypic
effect of SNPs on the structure and function of the affected protein
[17]. However, comparative analyses of the data obtained from
nsSNPs in CFTR gene by involving different bioinformatics tools might
provide a better understanding of the consequences of mutations in
CFTR gene. In this study we have included a comparative approach

for the identification of nsSNPs using SIFT, PolyPhen2 and I-Mutant which was not included in earlier studies. Our comparative study using SIFT, PolyPhen2 and I-Mutant identified 21 nsSNPs could be deleterious or damaging to the protein structure which could result in severe disease manifestation related to cystic fibrosis. Homology modeling structures developed from selected SNPs showed that these substitutions did not have any impact on the 3-D structure of the CFTR protein indicating that these substitutions might be playing a role in the functional aspects of the CFTR protein (fig. 1). In vivo assays are necessary to determine the deleterious or damaging effects of the CFTR protein in suitable host models to clinically determine the importance of these AA substitutions. The present study is of clinical importance and can be useful for developing clinical measures for the management of cystic fibrosis.

Table 1: List of 21 AA substitutions predicted to be deleterious or damaging to the CFTR protein. The overall prediction obtained from all the three prediction tools is summarized

<table>
<thead>
<tr>
<th>rsID</th>
<th>AA change</th>
<th>SIFT (Tolerance Index)</th>
<th>PolyPhen2 (PSIC SD)</th>
<th>I-Mutant (RI)</th>
<th>Overall prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1800093</td>
<td>F508C</td>
<td>0.00</td>
<td>1.000</td>
<td>3</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs11971167</td>
<td>D1270H</td>
<td>0.01</td>
<td>1.000</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs75527207</td>
<td>G551D</td>
<td>0.00</td>
<td>1.000</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs74503330</td>
<td>S1251N</td>
<td>0.00</td>
<td>0.993</td>
<td>5</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs75961395</td>
<td>G65E</td>
<td>0.01</td>
<td>0.995</td>
<td>2</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs78194216</td>
<td>R1066C</td>
<td>0.00</td>
<td>1.000</td>
<td>4</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs79282516</td>
<td>G480C</td>
<td>0.03</td>
<td>1.000</td>
<td>3</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs80034486</td>
<td>N1303K</td>
<td>0.00</td>
<td>1.000</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs80825262</td>
<td>G178R</td>
<td>0.00</td>
<td>1.000</td>
<td>4</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs113993958</td>
<td>D110Y</td>
<td>0.04</td>
<td>1.000</td>
<td>4</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909009</td>
<td>G458V</td>
<td>0.00</td>
<td>1.000</td>
<td>5</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909011</td>
<td>R334W</td>
<td>0.02</td>
<td>1.000</td>
<td>3</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909013</td>
<td>G551S</td>
<td>0.00</td>
<td>0.999</td>
<td>4</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909017</td>
<td>S492F</td>
<td>0.02</td>
<td>0.993</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909020</td>
<td>A1067P</td>
<td>0.00</td>
<td>1.000</td>
<td>4</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909021</td>
<td>A349V</td>
<td>0.02</td>
<td>0.907</td>
<td>3</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909033</td>
<td>D648V</td>
<td>0.03</td>
<td>0.001</td>
<td>3</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909037</td>
<td>Q1071P</td>
<td>0.03</td>
<td>0.999</td>
<td>5</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909040</td>
<td>G1249E</td>
<td>0.00</td>
<td>1.000</td>
<td>1</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs113999123</td>
<td>R334W</td>
<td>0.02</td>
<td>1.000</td>
<td>3</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>rs121909027</td>
<td>E92K</td>
<td>0.01</td>
<td>1.000</td>
<td>2</td>
<td>Probably damaging</td>
</tr>
</tbody>
</table>

Fig. 1: Selected homology models of native and nsSNP regions in CFTR protein

CONCLUSION

Several bioinformatics tools provide useful information regarding the effects of nsSNPs or missense mutations on the stability and structure of a protein under study. In this study, we analyzed a total of 108 nsSNPs and subjected to analyses using SIFT, PolyPhen2 and I-Mutant2 bioinformatics tools. We identified a total of 21 nsSNPs common to all the three methods that have a deleterious or damaging effect on the CFTR protein. This information could provide a researcher to explore these variants in in vivo experiments to
understand the consequences of the nsSNPs in individuals affected with cystic fibrosis.

ACKNOWLEDGEMENT

The authors thank the management, K L University for providing the necessary facilities to carry out this work.

CONFLICTS OF INTERESTS

There is no conflict of interest among the authors.

REFERENCES


How to cite this article