Identification of functional SNPs in VEGF gene and in silico analysis of damaging SNPs based on data procured from dbSNP database

Brijesh Dabhi and Kinnari N Mistry*

Ashok and Rita Patel Institute of Integrated Study and Research in Biotechnology and Allied Sciences (ARIBAS)
Charutar Vidyamand (CVM) (Affiliated to Sardar Patel University), New Vallabh Vidhyanagar, Anand 388 121, India

Received 20 October 2014; revised 17 May 2015; accepted 12 June 2015

VEGF (vascular endothelial growth factor) gene is known to cause angiogenesis and participate in many diseases like prostate cancer and diabetic retinopathy. Several single nucleotide polymorphisms (SNPs) have been described in the VEGF gene, some of which have been reported to be associated with differential expression of VEGF in vitro. In the present study, polymorphism of VEGF gene was studied in relation to damaging mutations. Of total SNPs in VEGF gene, 38 were non-synonymous SNPs (nsSNPs), 47 were synonymous SNPs, 66 were in non-coding regions, which comprised of 13 SNPs in 5'UTR region and 53 SNPs were in 3'UTR region. The rest were in the intronic region. SIFT (Sorting Intolerant from Tolerant) analysis showed that L163P, H200W, P312T, R288W and G365V mutations were damaging, whereas PolyPhen predicted P163L and H200W nsSNPs as damaging. Among the predicted nsSNPs, rs4645843, rs1800620 were identified as deleterious and damaging by the SIFT and PolyPhen programs. Protein structural analysis with these amino acid variants was performed by using I-Mutant, Swiss PDB viewer, ANOLEA (Atomic Non-Local Environment Assessment), MUSTER (MUlti-Sources ThreadER) and NOMAD-Ref servers to check their molecular dynamics and energy minimization calculations. This in silico analysis suggested that mutations in VEGF, such as, G365V, P312T and R288W, could directly or indirectly destabilize the amino acid interactions and hydrogen bond networks, thus explaining the functional deviations of protein to some extent. Thus screening for VEGF G365V, P312T and R288W variants in the population-based study may be useful to check disease susceptibility.

Keywords: in silico analysis, mutation, SNP, VEGF

Introduction

Vascular endothelial growth factor (VEGF), a mitogen that promotes vascular endothelial cell proliferation and angiogenesis, is a 45 kDa glycoprotein secreted in the vascular wall by endothelial and smooth muscle cells. Although VEGF has been considered relatively specific for endothelial cells, it also influences monocyte activation and migration, and vascular smooth muscle cell migration. Several studies have investigated the association of VEGF gene polymorphisms with diseases in which angiogenesis plays a major role in pathogenesis, such as, diabetic retinopathy, renal cell carcinoma, acute renal allograft rejection, prostate cancer and malignant melanoma. Several single nucleotide polymorphisms (SNPs) have been described in the VEGF gene, some of which have been reported to be associated with differential expression of VEGF in vitro. Two of these SNPs (positions 22578 & 21154) are located in the VEGF promoter and one SNP (position 2634) is located in the 59 untranslated region of the gene (position +405 after transcription initiation site).

Since VEGF gene plays an important role in diverse range of human diseases, its functional genomics based on mutation analysis is conceived to give key leads in disease diagnosis and therapy. The harmful mutations for VEGF gene have not been predicted to date in silico. To explore the possible associations between genetic mutation and phenotypic variation different genomics tools were used for prioritization of high-risk non-synonymous mutations in coding regions that are likely to have an effect on the structure and function of VEGF. Different algorithms like PupaSuite, SIFT (Sorting Intolerant from Tolerant) and PolyPhen were used for prioritization of high-risk non-synonymous single nucleotide polymorphisms (nsSNPs) in coding regions that are likely to have an effect on the function and structure of the protein. The frequent mutations were analyzed in the present study, in...
which 3D model structures of the mutant proteins were compared with the native protein structure. We further examined the native and mutant protein structures for solvent accessibility and secondary structure analyses. Our in silico study further suggests the presence of additional deleterious mutations in VEGF gene that may affect the structure and function of proteins with apparent roles in many diseases.

Materials and Methods

Datasets
The SNPs and their related protein sequence for VEGF gene were obtained from the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/) for computational analysis.

Prediction of Tolerated and Deleterious SNPs using SIFT
SIFT, version 2 is a program that can identify if an amino acid substitution influences a protein function as well as a phenotypic change. The SIFT tool estimates the consequence of an amino acid substitution on protein function by considering the sequence homology and physical properties of amino acids. It has been reported that SIFT can distinguish between functionally neutral and deleterious amino acid changes in mutagenesis studies and on human polymorphisms (http://blocks.fhcrc.org/sift/SIFT.html)\(^1\). SIFT analysis was based on algorithms to find homologous sequences using database SWISS-PORT version 51.3 and TrEMBL 34.3, by selecting median conservation sequence score 3.00. The cutoff value in SIFT program is tolerance index of \( \geq 0.05 \). Higher the tolerance index, less functional impact a particular amino acid substitution is likely to have.

Prediction of Functional Modification of Coding nsSNPs
The PolyPhen (Phenotyping Polymorphism) server characterizes protein structural modifications by using the rs id of SNPs as input. PolyPhen software version 2.0.9 automatically predicts the consequence of an amino acid change on the structure and function of a protein on specific empirical rules on the sequence. PolyPhen input option are protein sequence, attainment number P68871 or database ID/accession number combined with sequence position with amino acid variants AA1 and AA2. We submitted the query in the form of protein sequence with mutational position and two amino acid variants. Sequence based characterization of the substitution site, profile analysis of homologous sequences and mapping of substitution site to a known protein 3-dimensional (3D) structures are the parameters taken into account by PolyPhen server to calculate the score. It calculates position-specific independent counts (PSIC) scores for each of the two variants, and then computes the PSIC scores difference(PSIC SD) between them. Higher the PSIC SD, higher is the functional impact a particular amino acid substitution is likely to have.

PROVEAN (Protein Variation Effect Analyzer) Prediction of nsSNP
It is a software tool, which predicts whether an amino acid substitution or indel has an impact on the biological function of a protein. Variant with a score equal to or above \(-2.5\) are considered to be deleterious, whereas below \(-2.5\) score are considered to be neutral\(^14,15\). In put for this tool includes protein sequence in FASTA format and mutations to be analyzed.

nsSNPAnalyzer
nsSNPAnalyzer is a tool to predict whether a non-synonymous single nucleotide polymorphism (nsSNP) has a phenotypic effect. nsSNPAnalyzer also provides additional useful information about the SNP to facilitate the interpretation of results, e.g., structural environment and multiple sequence alignment. nsSNPAnalyzer uses information contained in the multiple sequence alignment and information contained in the 3D protein structure to make predictions. It uses a machine learning method called Random Forest to classify the nsSNPs. It is customized using a curated SNP dataset prepared from the SwissProt database. nsSNPAnalyzer calculates three types of information from the user's input: 1) The structural environment of the SNP, including the solvent accessibility, environmental polarity and secondary structure\(^16\); 2) the normalized probability of the substitution in the multiple sequence alignment\(^17\); and 3) the similarity and dissimilarity between the original amino acid and mutated amino acid.

The minimum input includes a protein sequence in FASTA format and a substitution file denoting the SNP identities to be analyzed. Substitution file format: A substitution is denoted as X#Y, where X is the original amino acid in one letter, # is the position of the substitution (starting from 1) and Y is the mutated amino acid in one letter. Multiple substitutions should be separated by new line characters.
Modeling nsSNP Locations on Protein Structure and Their RMSD Difference

Structural information has been extensively used for studying the effects of nsSNPs. We used the database dbSNP\(^{18}\) for identifying the nsSNPs onto the structure by using limits option. We confirmed the mutation positions and the mutation residues by this database. The mutation was performed by using SWISSPDB viewer and energy minimization for 3D structures was performed by NOMAD-Ref server. This server use Gromacs as default force field for energy minimization based on the methods of steepest descent, conjugate gradient and L-BFGS methods\(^{19}\).

I-Mutant and FOLD-X

I-Mutant (version 2.0) (http://folding.uib.es/i-mutant/i-mutant2.0.html) is a neural network based tool for the routine analysis of protein stability and alterations by taking into account the single-site mutations. I-Mutant also provides the scores for free energy alterations, calculated with the FOLD-X energy based web server. By assimilating the FOLD-X estimations with those of I-Mutant, the 93% precision can be achieved. The FASTA sequence of protein retrieved from NCBI is used as an input to predict the mutational effect on protein stability. The FOLD-X tool provides the comparison between wild type and mutant models in the form of van-der Waals clashes, which greatly influence the energy decomposition. Sometimes the mutations cause strain in the original native structure and sometimes reduce the strain\(^{20}\).

MUSTER

MUSTER is a MUlti-Source ThreadER program, which considers six different sources: (1) Sequence-derived profiles, (2) secondary structures, (3) structured-derived profiles, (4) solvent accessibility, (5) torsion angles (psi and phi angles), and (6) hydrophobic scoring matrix. MUSTER (v1.0) (http://zhanglab.ccmb.med.umich.edu/MUSTER/) is a valuable threading tool for protein structure prediction. Muster provides the Z-score and complete full-length models by using MODELLER v8.2. If the calculated Z-score is greater than 7.5, the corresponding template is considered good otherwise designated as bad. Sequence-derived profiles, secondary structures, structured-derived profiles, solvent accessibility, backbone torsion angles and hydrophobic scoring matrix are the six different sources used by MUSTER\(^{21}\).

I-TASSER

I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) creates the full length protein models by excising continuous fragments from threading alignments and further reconstructs them by using replica exchanged Monte Carlo simulations. During the simulation, decoys (low temperature replicas) are clustered by SPICKER and the full-length atomic models are generated by selecting the top five cluster centroids. The number of decoys at the unit space of SPICKER clusters defines the cluster density. The quality of predicted structure is estimated by I-TASSER in the form of confidence score (C-score). The C-score lies in the range of –5 to 2, with higher values depicting the high confidence for predicted model. The C-score has a correlation with the TM-score and RMSD (root-mean-square deviation). If the native structure is known, the TM-score and RMSD are used to measure the accuracy of predicted structure, else these are used to predict the quality of the modeling prediction by calculating the distance between two predicted models (by considering the one model as a native on the basis of C-score). The TM-score is a recently proposed scale to solve the local error problem of RMSD. A TM-score >0.5 highlights a model of correct topology and a TM-score < 0.17 indicates a random similarity\(^{22}\).

Results

SNP Dataset from dbSNP

The VEGF gene investigated in this work was retrieved from dbSNP database. It contained a total of 454 SNPs. Of which, 38 were nsSNPs, 47 were synonymous SNPs, and 66 were in non-coding regions, which comprises of 13 SNPs in 5’UTR region and 53 SNPs were in 3’UTR region (Fig. 1). The rest were in the intronic region. We selected only nsSNPs for our investigation.

![Fig. 1 — A graphical representation of distribution of nsSNP, intronic SNP, 5’UTR and 3’ UTR SNPs in VEGF gene (based on the dbSNP database).](image-url)
Analysis of Deleterious and Damaged nsSNPs
An amino acid substitution carries the potential impact on protein structure and function. In this study, the impact of VEGF variants on its structure and function was determined by using various computational tools.

**SIFT Analysis**
SIFT scores were classified as damaging (0.00-0.05), possibly damaging (0.051-0.10), borderline (0.101-0.20) or tolerant (0.201-1.00). Of the 38 screened nsSNPs, L163P, H200W, P312T and R288W showed the score of 0.00, whereas remaining G365V showed 0.01 score (Table 1). Of the total 38 nsSNPs screened by SIFT, five nsSNPs (rs 28940297, rs 28940298, rs 62401172, rs 76869573 & rs 114262569) were found to be deleterious.

**Prediction of Protein Structural and Functional Modifications**
rs ids of nsSNPs along with protein sequence of VEGF were submitted to PolyPhen server. PolyPhen score differences (PSIC SD) were assigned as probably damaging (2.00 or more), possibly damaging (1.40-1.90), potentially damaging (1.20-1.50) and benign (0.00-0.90). PolyPhen predicted two nsSNPs (rs 28940297 & rs 28940298) as potentially damaging out of all 38 nsSNPs.

**PROVEAN Prediction of nsSNP**
We wanted to validate the results obtained by SIFT and PolyPhen, so we submitted only five nsSNP to PROVEAN, which were found to be damaging by SIFT. PROVEAN predicted three nsSNP (rs 28940298, rs 76869573 & rs 114262569) as deleterious, whereas two ns SNPs (rs 28940297 & rs 62401172) were found to be neutral.

<table>
<thead>
<tr>
<th>SNPs id</th>
<th>AA change</th>
<th>SIFT tolerance index</th>
<th>PSIC SD</th>
<th>PROVEAN score</th>
<th>PROVEAN prediction (cut off -2.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs 28940297</td>
<td>P163L</td>
<td>0.00</td>
<td>1.117</td>
<td>-0.864</td>
<td>Neutral</td>
</tr>
<tr>
<td>rs 28940298</td>
<td>H200W</td>
<td>0.00</td>
<td>1.426</td>
<td>-4.590</td>
<td>Deleterious</td>
</tr>
<tr>
<td>rs 62401172</td>
<td>G365V</td>
<td>0.01</td>
<td>Not found</td>
<td>-1.556</td>
<td>Neutral</td>
</tr>
<tr>
<td>rs 76869573</td>
<td>P312T</td>
<td>0.00</td>
<td>Not found</td>
<td>-5.382</td>
<td>Deleterious</td>
</tr>
<tr>
<td>rs 114262569</td>
<td>R288W</td>
<td>0.00</td>
<td>Not found</td>
<td>-6.804</td>
<td>Deleterious</td>
</tr>
</tbody>
</table>

**nsSNPAnalyzer**
All five amino acid variants (P163L, H200W, G365V, P312T & R288W) were submitted to nsSNPAnalyzer tool along with protein sequence of TNF-α in FASTA format. We could not find any significant result for the P84L variant. The output of final result of nsSNPAnalyzer tool is shown in Table 2.

**Modelling of Mutant Structure**

**Prediction of Protein Structural Stability**
The five mutations (P163L, R200W, G365V, P312T & R288W) in VEGF gene were selected on the basis of SIFT score. These variants were given to I-Mutant web server to predict the DDG stability and reliability index (RI) upon mutation (Table 3).

**Models Generated by MUSTER**
We found 10 different template of VEGF gene by MUSTER, which gave Z-score based on alignment diversity. The optimized threading was found by global dynamic programming, which is given in Table 4.

Mapping the deleterious nsSNPs into the protein structure information was obtained from Single Amino Acid Polymorphism database (SAAPdb) and dbSNP. The available structure for VEGF gene has the PDB id 1VPF. According to SIFT and PROVEAN, mainly three mutations occurred for 1VPF at 3 SNP ids, namely, rs 62401172, rs 76869573 and rs 114262569, which were mapped by CPH model server23. Then, energy minimizations were carried out by the NOMAD-Ref server for the native and the 3 mutant type proteins (G365V, P312T & R288W) (Fig. 2). The total energy for the native type was –4549.106 kJ/mol, while mutant type

<table>
<thead>
<tr>
<th>SNP</th>
<th>Phenotype</th>
<th>Environment</th>
<th>AreaBuried</th>
<th>FracPolar</th>
<th>Secondstr</th>
</tr>
</thead>
<tbody>
<tr>
<td>G365V</td>
<td>Disease</td>
<td>EC</td>
<td>0.033</td>
<td>0.875</td>
<td>C</td>
</tr>
<tr>
<td>P312T</td>
<td>Disease</td>
<td>ES</td>
<td>0.145</td>
<td>0.698</td>
<td>S</td>
</tr>
<tr>
<td>R288W</td>
<td>Disease</td>
<td>B3S</td>
<td>0.573</td>
<td>0.531</td>
<td>S</td>
</tr>
</tbody>
</table>
Table 3—Stability of VEGF protein due to mutation predicted by I-mutant

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Wild type</th>
<th>New</th>
<th>pH</th>
<th>Temperature (°C)</th>
<th>Stability</th>
<th>DDG (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P163L</td>
<td>P</td>
<td>L</td>
<td>7.0</td>
<td>25</td>
<td>Increase</td>
<td>0.68</td>
</tr>
<tr>
<td>H200W</td>
<td>H</td>
<td>W</td>
<td>7.0</td>
<td>25</td>
<td>Increase</td>
<td>0.13</td>
</tr>
<tr>
<td>G365V</td>
<td>G</td>
<td>V</td>
<td>7.0</td>
<td>25</td>
<td>Increase</td>
<td>0.28</td>
</tr>
<tr>
<td>P312T</td>
<td>P</td>
<td>T</td>
<td>7.0</td>
<td>25</td>
<td>Increase</td>
<td>0.01</td>
</tr>
<tr>
<td>R288W</td>
<td>R</td>
<td>W</td>
<td>7.0</td>
<td>25</td>
<td>Decrease</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

Where, “Wild type” is the amino acid in native protein, “New” is mutant amino acid and DDG is the stability (DDG<0: Decrease stability, DDG>0: Increase stability).

Table 4—Different templates predicted by MUSTER program by multiple sequence alignment

<table>
<thead>
<tr>
<th>Rank</th>
<th>Template</th>
<th>Align_length</th>
<th>Coverage</th>
<th>Zscore</th>
<th>Seq_id</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3chnS</td>
<td>397</td>
<td>0.963</td>
<td>7.441</td>
<td>0.108</td>
<td>Bad</td>
</tr>
<tr>
<td>2</td>
<td>1w0sA</td>
<td>406</td>
<td>0.985</td>
<td>7.44</td>
<td>0.091</td>
<td>Bad</td>
</tr>
<tr>
<td>3</td>
<td>1w0rA</td>
<td>368</td>
<td>0.893</td>
<td>7.045</td>
<td>0.111</td>
<td>Bad</td>
</tr>
<tr>
<td>4</td>
<td>3ucpA</td>
<td>404</td>
<td>0.98</td>
<td>6.943</td>
<td>0.101</td>
<td>Bad</td>
</tr>
<tr>
<td>5</td>
<td>1zlgA</td>
<td>399</td>
<td>0.968</td>
<td>6.5</td>
<td>0.083</td>
<td>Bad</td>
</tr>
<tr>
<td>6</td>
<td>1yo8A</td>
<td>371</td>
<td>0.9</td>
<td>6.406</td>
<td>0.035</td>
<td>Bad</td>
</tr>
<tr>
<td>7</td>
<td>4acqA</td>
<td>410</td>
<td>0.995</td>
<td>6.289</td>
<td>0.093</td>
<td>Bad</td>
</tr>
<tr>
<td>8</td>
<td>3o94A</td>
<td>399</td>
<td>0.968</td>
<td>6.177</td>
<td>0.123</td>
<td>Bad</td>
</tr>
<tr>
<td>9</td>
<td>2ahxB</td>
<td>397</td>
<td>0.963</td>
<td>6.17</td>
<td>0.093</td>
<td>Bad</td>
</tr>
<tr>
<td>10</td>
<td>4b2nA</td>
<td>392</td>
<td>0.951</td>
<td>6.136</td>
<td>0.079</td>
<td>Bad</td>
</tr>
</tbody>
</table>

If Z-score >7.5, the corresponding template is a ‘Good’ template. Otherwise, it is a ‘Bad’ template.

Discussion

The SNPs are hypothesized to play an important role in several human diseases. Approx, more than 4 million unique human SNPs have now been reported by a number of SNP databases. About 2% of the all known single nucleotide variants associated with monogenic disease are nsSNPs in protein-coding regions (SNPs that alter a single amino acid in a protein molecule). As a result, it is anticipated that this class of SNPs are related to complex inherited disease traits. In silico analysis using powerful software tools can facilitate predicting the phenotypic effect of nsSNPs on the physico-chemical properties of the concerned proteins. Such information is critical for genotype-phenotype correlations and also to understand disease biology.

Given the fact that nsSNPs in critical cellular genes, such as, VEGF, participates in vascularization and angiogenesis, it is believed to play an important role in disease predisposition also. Therefore, an effort was made to identify SNPs that can modify the structure, function and expression of the VEGF gene. As described in the methodology section, a structure G365V, P312T and R288W were found to be –4549.106, –4597.938 and -4325.349 kJ/mol, respectively.
computational approach was undertaken to study systematic analysis of SNPs to predict the benign mutations by using the SIFT, PolyPhen and PROVEAN servers. According to the SIFT results, only 5 mutations, viz., P163L, H200W, G365V, P312T and R288W were found to be deleterious by PolyPhen, whereas three mutations H200W, P312T and R288W were found to be damaging by PROVEAN. We confirmed these results by submitting deleterious variants found by SIFT, PolyPhen and PROVEAN in nsSNPAnalyzer server. We found that among five variants only three variants, viz., G365V, P312T and R288W, may have disease phenotype. The classical molecular dynamics approach was also applied for studying native and fetal mutations using simulations in explicit solvent and examining the differences in dynamics and stability of the VEGF protein by I-Mutant server.

and disrupting the proper folding of VEGF protein. The change in arginine to tryptophan at 288 introduces aromatic amino acid in place of basic amino acid, which can disturb the electrostatic interactions with other molecules. R288W showed decrease stability so it could be a potential variant to be involved in disease, in which low level of VEGF protein can elucidate disease. P163L, H200W, G365V and P312T variants may not disturb the bonding and electrostatic interaction with nearby residue. Hence these variants were not found to decrease stability of VEGF protein. Mutation at specified position was performed by CPH model and SWISSPDB viewer to get modeled structure. Then, energy minimizations were performed by NOMAD-Ref server for both the native structure and mutant modeled structure. Energy minimization value of mutant models is different from wild types, which denotes that nsSNPs associated with this mutant structure was highly affected in sequence level (SIFT, PolyPhen & PROVEAN) and structural level.

Several studies have investigated the role of VEGF polymorphisms as genetic determinant for susceptibility and outcome of breast, prostate, NSCL and colorectal cancer. Several polymorphisms have been reported within the promoter (-2578C>A, -2489C>T, -1498C>T & -1154G>A), 5’-UTR (-634G>C & -7C>T) and 3’-UTR (936C>T & 1612G>A) for the VEGF gene. The variant allele for -1154G>A and 936C>T results in lower VEGF expression, whereas the variant allele for -1498C>T and -7C>T results in increased concentrations of VEGF Mrna. VEGF variants found in this study have not been reported earlier so they need to be validated to check its significance. Our results indicate that the procedure of computational algorithms offer an alternative strategy to choose SNPs targets by considering the role of SNPs on the functional attributes or molecular phenotype of a protein. In silico approaches based on these investigations will surely assist in our understanding of the inheritance of complex human diseases.

Acknowledgement

We are thankful to NCBI databank server for providing data for SNPs for our analysis. We would also like to acknowledge CVM and ARIBAS for providing platform for this work.

References