



Received on 07 February, 2017; received in revised form, 12 April, 2017; accepted, 26 April, 2017; published 01 September, 2017

## CHARACTERIZATION OF ENDOGLIN GENE USING *IN SILICO* TOOLS

V.G. Vidhya <sup>\*1</sup> and Anusha Bhaskar <sup>2</sup>

Department of Biotechnology <sup>1</sup>, Faculty of Science and Humanities, SRM University, Kattankulathur - 603 203, Kancheepuram District, Tamil Nadu, India.

ABHI Diagnostic and Research Lab <sup>2</sup>, Perambalur - 621212, Tamil Nadu, India.

### Keywords:

Diabetic retinopathy,  
PDR, Endoglin, Neoangiogenesis

### Correspondence to Author:

**V. G. Vidhya**

Assistant Professor,  
Department of Biotechnology,  
Faculty of Science and Humanities,  
SRM University, Kattankulathur -  
603203, Kancheepuram District,  
Tamil Nadu, India.

**E-mail:** vidhyavg@gmail.com

**ABSTRACT:** Diabetic retinopathy (DR) is a micro vascular disorder affecting the retina of the eye. Many genes have been identified for its pathogenesis. Endoglin (ENG) is one of the candidate gene involved in DR. Several literature studies helped us understand that the *in vitro* and *in vivo* studies of these candidate genes had demonstrated the role of a new gene called endoglin (ENG) capable of inhibiting angiogenesis. In this present study we characterized the physicochemical and functional properties of endoglin gene using bioinformatics tools as this had not been studied previously. The three dimensional structure of endoglin was also computed using homology modelling and the reliability of the model was assessed by predicting its secondary structure along with validation. Additionally, using experimental evidence score, phylogenetic analysis and protein-protein network was framed between endoglin and its interacting neighbouring partners. These results are innovative and relevant enough to start prospective studies that will allow us to establish the relative strength of the prediction of diabetic retinopathy according to the endoglin levels presented by the patient.

**INTRODUCTION:** Diabetic retinopathy, a leading cause of blindness in the developing countries is a prototypical micro-vascular disorder associated with micro-aneurysms, intraretinal hemorrhages, capillary non - perfusion, intraretinal micro-vascular abnormalities and neo-vascularization. Global population based data indicate that it is the fifth most common cause of blindness in the world <sup>1</sup>. The duration of diabetes and glycemic control is the two most important factors in the development of retinopathy <sup>2</sup>.

However, these factors alone do not explain the occurrence of retinopathy. It may be absent in some patients with poor glycaemic control even over a long period of time, while others may develop retinopathy in a relatively short period despite good glycaemic control.

This raises the possibility of a genetic predisposition to retinopathy. Supportive evidence for a genetic role of retinopathy derives from twin, family and transracial studies demonstrating the importance of inherited factors in the aetiology of diabetes and its complications <sup>3 - 5</sup>. Diabetic retinopathy is primarily classified into Background diabetic retinopathy; Non-Proliferative diabetic retinopathy (NPDR) and Proliferative diabetic retinopathy (PDR) <sup>6</sup>. Visual impairment in diabetic retinopathy also occurs due to diabetic macular edema (DME) which is defined as retinal

QUICK RESPONSE CODE	DOI: 10.13040/IJPSR.0975-8232.8(9).3837-42
	Article can be accessed online on: <a href="http://www.ijpsr.com">www.ijpsr.com</a>
DOI link: <a href="http://dx.doi.org/10.13040/IJPSR.0975-8232.8(9).3837-42">http://dx.doi.org/10.13040/IJPSR.0975-8232.8(9).3837-42</a>	

thickening / hard exudates which are due to increased permeability of retinal vessels<sup>7,8</sup>.

### Causes of Diabetic Retinopathy:

- Glycosylated proteins and free radicals
- Lack of oxygen to the retina
- Hyperglycemia
- Duration of diabetes
- Hypertension
- Gender
- Elevated serum lipids
- Pregnancy
- Alcohol
- Anemia
- Obesity

Candidate genes for diabetic retinopathy have also been selected from the proposed pathogenic pathways that include the polyol pathway, AGEs, the renin - angiotensin system, growth factors (VEGF, GH and IGF-1), oxidative damage and ROS, PKC, GLUT1, PPAR $\gamma$ , extracellular matrix homeostasis and tissue matrix metalloproteinases, inflammation, thrombogenesis, apolipoprotein A and Vitamin D also contribute to the identification of susceptible genes.

Endoglin (ENG), also known as CD105, is a disulphide-linked homodimeric transmembrane glycoprotein of 180 kDa that consists of an extracellular domain (561 amino acid residues) and a cytoplasmic region (serine/threonine-rich 47 amino acid residues)<sup>9-10</sup>. The extracellular domain comprises a zona pellucida (ZP) domain, with an arginine - glycine - aspartic acid (RGD) binding motif<sup>11-12</sup>, while several potential phosphorylation sites are located in the intracellular region<sup>13</sup>. ENG gene is located in chromosome 9 (9q33 - q34.1) and has 14 exons and 22 distinct introns which are associated with human vascular endothelium<sup>14-15</sup>. Mutations in this gene cause hereditary hemorrhagic telangiectasia, also known as Osler-Rendu-Weber syndrome 1, an autosomal dominant multisystemic vascular dysplasia. Spliced transcript variants encoding different iso forms have been reported for this gene.

### MATERIALS AND METHODS:

**Primary Structure Prediction:** Physicochemical properties like molecular weight, theoretical pI, %

composition of amino acids, extinction coefficient<sup>16</sup>, estimated half-life, instability index<sup>17</sup>, aliphatic index, Grand average of hydropathicity (GRAVY)<sup>18</sup> of linear sequence of ENG gene was calculated using ExPASy PROTPARM server.

**Secondary Structure Prediction:** Secondary structure of protein sequence was predicted using GOR IV and SOPMA. GOR IV method is based on information theory and was developed by J. Garnier, D. Osguthrope and B. Robson<sup>19</sup>. SOPMA (Self-optimized prediction method with alignment) is an improvement of SOPM method. These methods are based on the homologue method of Levin *et al.* This self-optimized prediction method builds sub databases of protein sequences with known secondary structures; each of the proteins in a sub database is then subjected to secondary structure prediction based on sequence similarity. The information from the sub databases is then used to generate a prediction on the query sequence.

**Protein Motif Identification:** Finger PRINT Scan was used to identify the motif of the target protein. This tool identified the closest matching PRINTS sequence and motif fingerprints in a protein sequence.

**Trans-Membrane Prediction:** TMPred relies on a database of trans-membrane proteins called TMbase<sup>20</sup>. TMbase which is derived from uniprot, contains additional information on each sequence regarding the number of trans-membrane domains they possess, the location of these domains, and the nature of the flanking sequences. TMPred uses this information in conjunction with several weight matrices in making its predictions.

The sequence in one letter code is pasted into the query sequence box, and the user can specify the minimum and maximum lengths of the hydrophobic part of the trans-membrane helix to be used in the analysis. The output has four sections: a list of possible transmembrane helices, a table of correspondences, suggested models for trans-membrane topology and a graphic representation of the same results.

**Homology Modelling:** The most robust of the structure prediction techniques is homology modelling. If sequence similarity search of target

sequence is more than 60 % then homology modelling can be useful for 3D structure prediction of the target protein. Homology modelling was done using a template sequence whose structure has been solved by either X-ray diffraction or NMR technology. Swiss Model is an online tool for 3D structure prediction based on homology modelling<sup>21</sup>. Swiss model is used to determine whether a sequence can be modelled at all; when a sequence is submitted.

It first compares the target protein with the crystallographic database in pdb and modelling is attempted only if there is a homolog. The template structures are selected if there is at least 25 % sequence identity in a region more than 20 residues long. If this approach finds one or more appropriate entries in pdb, atomic models are built and energy minimization is performed to generate the best model. Those results can be resubmitted to Swiss model using its optimizing mode, which allows for alteration of the proposed structure based on other knowledge, such as biochemical information. The modelled structure was estimated by assessing the QMEAN score and Z-score and was visualized using Rasmol software.

**Phylogenetic Analysis:** The phylogenetic analysis of endoglin gene was done to identify the number of species that share common structural and functional features. The multiple sequence alignment was done using Clustal Omega with default parameters. The output was analyzed for sequences that are aligned for the complete length, scores, alignment, conserved residues, substitutes and semi conserved substituted residue patterns.

The phylogenetic tree was constructed based on the bootstrap Neighbour Joining (NJ) method<sup>22 - 24</sup>. The stability of the internal nodes was assessed by bootstrap analysis with 100 replicates.

## RESULTS AND DISCUSSION:

**Sequence Retrieval:** The endoglin protein sequence was retrieved from our curated diaretinopathy database<sup>25</sup>. NCBI's BLAST tool for protein sequence was used to search similar sequences. The results retrieved 9 sequences from 8 different species with blast hits ranging from 62 - 96 %. All the sequences showed E-value of 0.00 (Table 1).

**TABLE 1: BLAST RESULT OF THE ENG PROTEIN**

Accession no.	Organism	Blast hit (%)	E-value
EHH62423.1	<i>Macaca fascicularis</i>	89%	0.0
JAV38373.1	<i>Castor Canadensis</i>	76 %	0.0
NP_001010968.1	<i>Rattus norvegicus</i>	73%	0.0
CAA54917.1	<i>Mus musculus</i>	72 %	0.0
NP_999196.1	<i>Sus scrofa</i>	69%	0.0
NP_035708.2	<i>Mus musculus</i>	62 %	0.0
NP_001182613.1	<i>Homo sapiens</i>	62 %	0.0
EHH24226.1	<i>Macaca mulatta</i>	94 %	0.0
NP_001126409.1	<i>Pongo abelii</i>	96 %	0.0

**Primary Structure Prediction:** Primary structure of ENG protein was predicted and its physiochemical properties were analyzed by using ExPASy's Protparam server. Result showed that endoglin protein has 658 amino acid residues and the estimated molecular weight is 70578.0. iso-electric point (pI) is the pH at which the surface of the protein is covered with charge but net charge of protein is zero. The calculated iso-electric point (pI) was computed to be 6.14. The computed value is less than 7 which indicates that the protein is acidic. The maximum number of amino acid present in the sequence was found to be leucine (13.1%) and the least was that of tryptophan (1.1%). The total number of positively charged residues (Asp + Glu) was 54 and the total number of negatively charged residues (Arg + Lys) was 46.

The very high aliphatic index (95.29) indicates that this protein is stable for a wide range of temperature range while the instability index (50.29) provides the estimate of the stability of protein in a test tube. Endoglin protein is found to be unstable because of the low instability index. The GRAVY value is low 0.076 indicating better interaction of the protein with water.

**Functional Site Prediction:** The domain search was done by conserved domain search on BLAST site and it showed the presence of single domain - the zona pellucida superfamily (Fig. 1).

**Secondary Structure Prediction:** The secondary structure is composed of alpha helix and random

coil and the secondary structure is predicted using GOR IV and SOPMA.

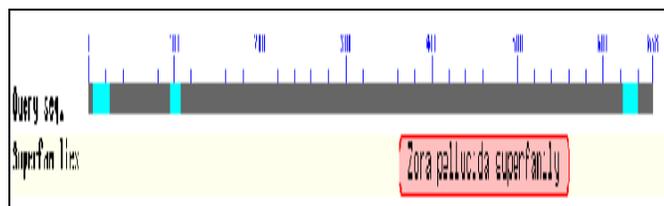


FIG. 1: GRAPHICAL REPRESENTATION OF CONSERVED DOMAIN IN ENG PROTEIN

Table 2 presents the comparative analysis of GOR IV and SOPMA from which it is clear that random coil is predominantly present when the structure was predicted both by SOPMA and GOR followed by alpha helix and extended strand. The secondary structure prediction was done and random coil was

found to be frequency (54.71 %) followed by alpha helix (24.77 %) and random coil was found to be least frequent (20.52 %). This is graphically represented in Fig. 2.

TABLE 2: SECONDARY STRUCTURE OF ENG BY SOPMA AND GOR

Secondary Structure	SOPMA	GOR
Alpha helix	20.36%	24.77%
3 <sub>10</sub> helix	0.00%	0.00%
Pi helix	0.00%	0.00%
Beta bridge	0.00%	0.00%
Extended strand	26.14%	20.52%
Beta turn	5.47%	0.00%
Bend region	0.00%	0.00%
Random coil	48.02%	54.71%
Ambiguous states	0.00%	0.00%
Other states	0.00%	0.00%
Sequence length	658	658

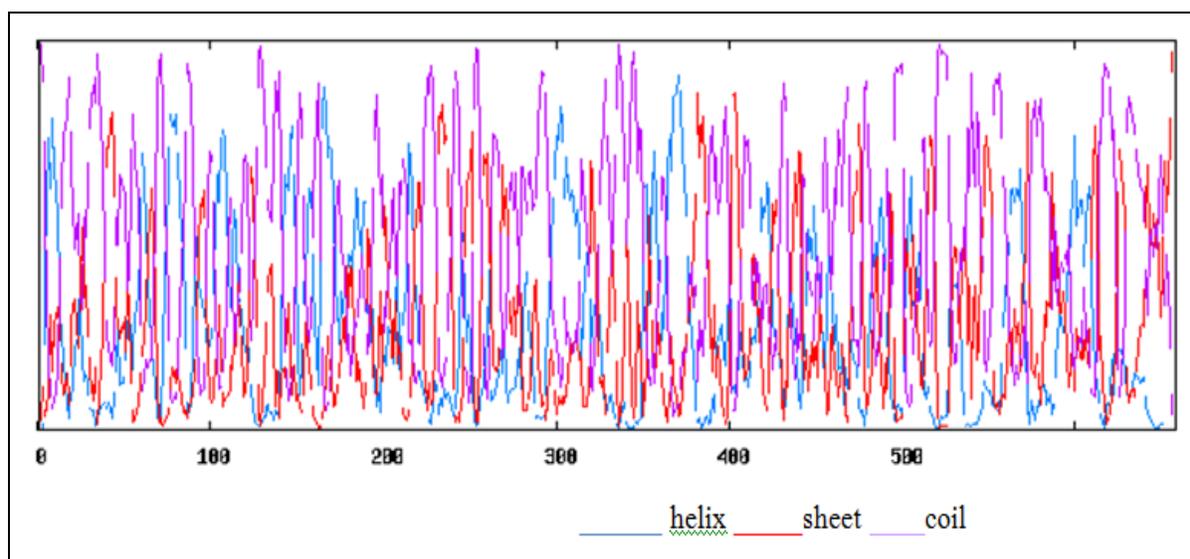


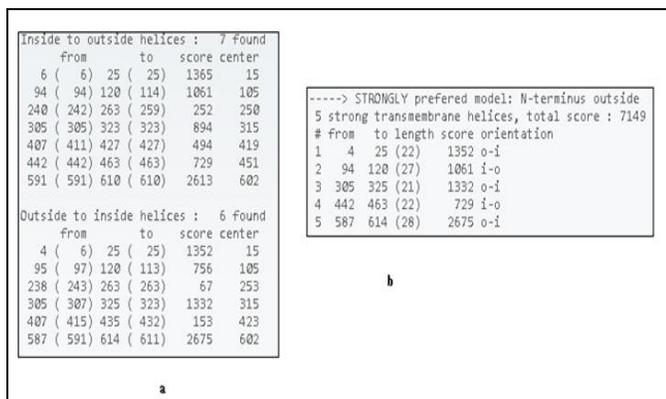
FIG. 2: GRAPHICAL REPRESENTATION OF SECONDARY ELEMENTS IN SORD PROTEIN

**Protein Motif Identification:** The FINGER Print scan of the sequence showed 10 fingerprints having 2 motifs for each fingerprint in the sequence (Table 3)

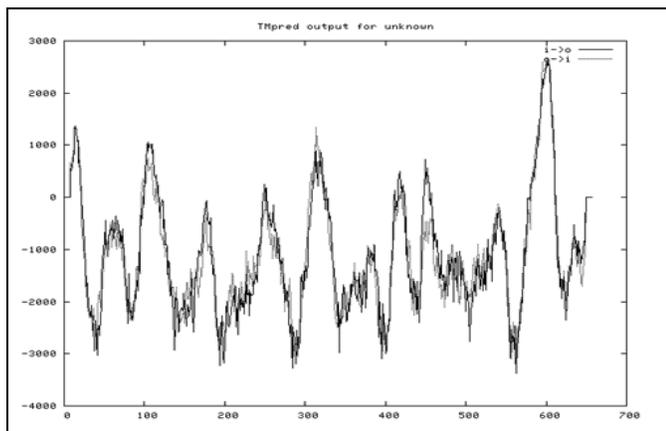
TABLE 3: FINGER PRINT SCAN RESULT OF ENG PROTEIN

Finger Print	No. of Motifs
Lipocalinimr	2
Annexin	2
Integrinb	2
Bradykinb1r	2
GPR153	2
Salspvbprot	2
Nmdareceptor	2
Fmrfamidep	2
Januskinase3	2
Ligninase	2

**Transmembrane Prediction:** The TMPred program makes a prediction of membrane spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase a database of naturally occurring trans-membrane proteins. The prediction is made using a combination of several weight-matrices for scoring for the endoglin protein. The results predicted 7 helices from inside to outside and 6 helices from outside to inside. As we know that scores above 500 are considered significant only one prominent trans-membrane region was identified for the protein. However according to TMPred the strongly preferred model for trans-membrane topology showed total score of 7149 with 5 strong TM helices (Fig. 3) and (Fig. 4).

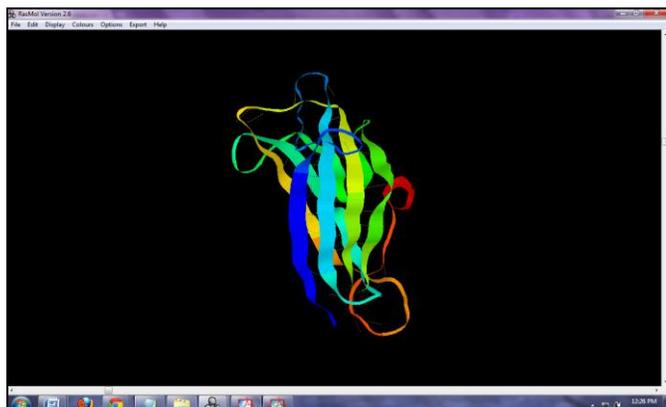


**FIG. 3: POSSIBLE TRANSMEMBRANE HELICES PREDICTED USING TMPRED**

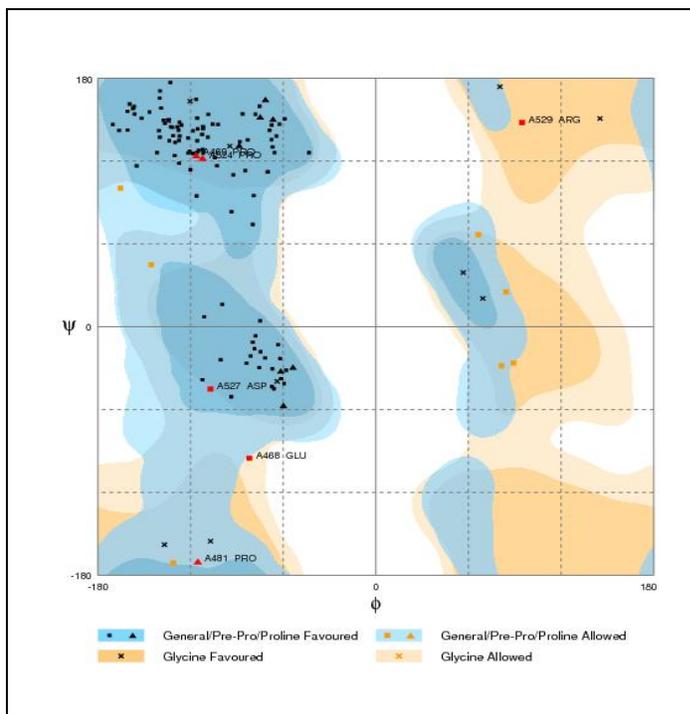


**FIG. 4: GRAPHICAL REPRESENTATION OF TM HELICES BY TMPRED**

**Homology Modelling:** The tertiary structure was modelled using Swiss model by using the templates from BLAST similarity and PDB Sum. 3qw9.pdb was used as template to model the structure (Fig. 5). The modelled structure showed 72 H bonds, 1 helices, 11 strands and 16 turns. The modelled structure was validated using SAVES tool and Ramachandran plot was plotted using RAMPAGE (Fig. 6).



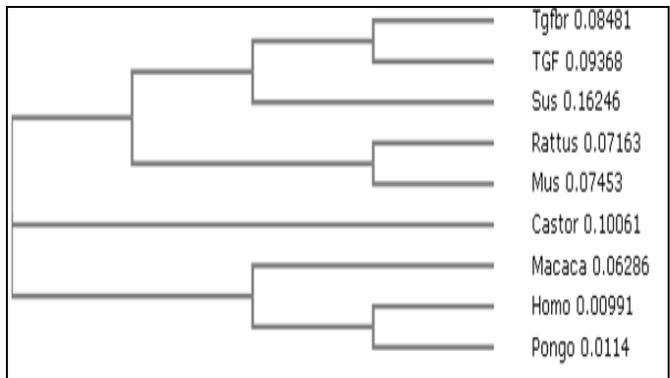
**FIG. 5: THREE DIMENSIONAL STRUCTURE OF ENDOGLIN PROTEIN**



**FIG. 6: GRAPHICAL REPRESENTATION OF RAMACHANDRAN PLOT BY RAMPAGE**

Number of residues in favoured region (~98.0% expected): 118 (90.1%)  
 Number of residues in allowed region (~2.0% expected): 7 (5.3%)  
 Number of residues in outlier region: 6 (4.6%)

**Phylogenetic Analysis:** The input for multiple sequence alignment was taken from 9 species whose blast hit was between 70 - 98 %. The clustal omega tool was run with default parameters such as a value of 10 for gap open penalty and a value of 0.05 and 0.05 for extending a gap and separating a gap respectively. The phylogenetic tree was drawn using NJ plot (Fig. 7).



**FIG. 7: PHYLOGENETIC ANALYSIS OF ENG PROTEIN USING NJ PLOT**

The results revealed that the endoglin protein of Homo sapiens is closely related to *Pongo abelii* and *Macaca fascicularis* with distance value of 0.0114

and 0.06286 respectively and distantly related to *Sus scrofa* with distance value of 0.16246.

**CONCLUSION:** Endoglin is a homodimeric membrane glycoprotein primarily associated with human vascular endothelium. The regulation through the polyol pathway is thought to affect the accumulation of endoglin that is associated with diabetes mellitus and its complications. Further research involving development of appropriate strategies for studying this protein could be significance in preventing diabetic complications. In the present study the sequence and structural analysis of endoglin gene was done by various computational tools. These results are innovative and relevant enough to start prospective studies that will allow us to establish the relative strength of the prediction of diabetic retinopathy according to the endoglin levels presented by the patient.

**ACKNOWLEDGEMENT:** Nil.

**CONFLICT OF INTEREST:** We declare that we have no conflict of interest.

## REFERENCES:

- Martin MN and Michael WU: Diabetic retinopathy - ocular complications of diabetes mellitus. *World J Diabetes*. 2015; 6(3): 489-499.
- Ahsan H: Diabetic retinopathy--biomolecules and multiple pathophysiology. *Diabetes Metab Syndr*. 2015; 9(1): 51-4.
- Klaassen I, Van Noorden C J and Schlingemann RO: Molecular basis of the inner blood-retinal barrier and its breakdown in diabetic macular edema and other pathological conditions. *Prog Retin Eye Res*. 2013; 34: 19-48.
- Nittala MG, Keane PA, Zhang K, and Srinivas RS: Risk factors for proliferative diabetic retinopathy in a Latino American Population. *Retina*. 2014; 34(8): 1594-1599.
- Kumar V, Sharma N and Bhalla TC: *In silico* Analysis of  $\beta$  Galactosidases Primary and Secondary Structure in relation to Temperature Adaptation. *Journal of Amino Acids* 2014.
- Vidhya GK and Anusha B: Diaretinopathy database – A gene database for diabetic retinopathy. *Bioinformation*. 2014; 10(4): 235 – 40.
- Jethra G, Sharma R, Singh P and Choudhary S: *In-silico* analysis of fenugreek (*Trigonella foenumgraecum*) protein. *International Journal of Advances in Science Engineering and Technology*. 2015; 3(3): 66-67.
- Mihasan M: Basic protein structure prediction for the biologist: A review. *Arch. Biol. Sci., Belgrade*. 2010; 62: 857-871.

- Prajapat R and Bhattacharya I: *In Silico* Structure Analysis of Type 2 Diabetes Associated Cysteine Protease Calpain-10 (CAPN10). *Advances in Diabetes and Metabolism*. 2016; 4(2): 32 – 43.
- Picos-Cárdenas VJ, Sáinz-González E, Miliar-García A, Romero-Zazueta A, Quintero-Osuna R, Leal-Ugarte E, Peralta-Leal V and Meza-Espinoza JP: Calpain-10 gene polymorphisms and risk of type 2 diabetes mellitus in Mexican mestizos. *Genet Mol Res*. 2015; 14(1): 2205 - 2215.
- Prajapat R, Marwal A and Gaur RK: Recognition of Errors in the Refinement and validation of three-dimensional structures of AC1 proteins of begomovirus strains by using ProSA-Web. *J. of Viruses* 2014.
- Hertel JK, Johansson S, Midthjell K, Nygard O, Njolstad PR and Molven A: Type 2 diabetes genes –Present status and data from Norwegian studies. *Norsk Epidemiologi* 2013; 23(1): 9-22.
- Xian G, Chao Z, Aisa Y, Ying S, Xuewei Y, Aosiman N, Yaqun G and Shuhua X: A comparative analysis of genetic diversity of candidate genes associated with type 2 diabetes in worldwide populations. *Yi Chuan*. 2016; 38(6): 543-559.
- Wheeler-Jones CP, Clarkin CE, Farrar CE, Dhadda P, Chagastelles P, Nardi N and Jones PM: Endoglin (CD105) is not a specific selection marker for endothelial cells in human islets of Langerhans. *Diabetologia*. 2013; 56(1): 222-224.
- Dorajoo R, Liu J and Boehm BO: Genetics of Type 2 Diabetes and Clinical Utility. *Genes (Basel)*. 2015; 6(2): 372-384.
- Kalman M and Ben-Tal N: Quality assessment of protein model structures using evolutionary conservation. *Bioinformatics*. 2010; 26: 1299-1307.
- Narayana Swamy A, Valasala H and Kamma S: *In silico* Evaluation of Nonsynonymous Single Nucleotide Polymorphisms in the ADIPOQ Gene Associated with Diabetes, Obesity, and Inflammation. *Avicenna J Med Biotechnol*. 2015; 7(3): 121-127.
- Chu H, Wang M, Zhong D, Shi D, Ma L, Tong N and Zhang Z: AdipoQ polymorphisms are associated with type 2 diabetes mellitus: a meta-analysis study. *Diabetes Metab Res Rev*. 2013; 29(7): 532-45.
- Garnier J, Osguthrope DJ AND Robson B: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*. 1978; 120(1) 97 – 120.
- Singh N, Upadhyay S, Jaiswar A and Mishra N: *In silico* Analysis of Protein. *J Bioinform Genomics, Proteomics*. 2016; 1(2): 1007.
- Ertugrul F and Ibrahim K: *In silico* sequence analysis and homology modeling of predicted beta-amylase 7-like protein in *Brachypodium distachyon* L. *J Bio Sci Biotech*. 2014; 3: 61-67.
- Wulandari D, Rachmadi L and Sudiro TM: Phylogenetic analysis and predicted functional effect of protein mutations of E6 and E7 HPV16 strains isolated in Indonesia. *Med J Indones*. 2015; 24(4): 197-205.
- Mirzaei K, Bahramnejad B, Shamsifard MH and Zamani W: *In Silico* Identification, Phylogenetic and Bioinformatic Analysis of Argonaute Genes in Plants. *Int J Genomics*. 2014.
- Mahmood N and Moosa MM: *In silico* analysis of the NBS protein family in *Ectocarpus siliculosus*. *Indian Journal of Biotechnology*. 2013; 12: 98-102.
- <http://diaretinopathydatabase.com/>

### How to cite this article:

Vidhya VG and Bhaskar A: Characterization of endoglin gene using *in silico* tools. *Int J Pharm Sci Res* 2017; 8(9): 3837-42. doi: 10.13040/IJPSR.0975-8232.8(9).3837-42.

All © 2013 are reserved by International Journal of Pharmaceutical Sciences and Research. This Journal licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

This article can be downloaded to **ANDROID OS** based mobile. Scan QR Code using Code/Bar Scanner from your mobile. (Scanners are available on Google Playstore)