In- silico analysis of hematopoeitic stem cell antigen CD34 to identify its functional implications

A Major Project thesis

Submitted in partial fulfillment of the requirements
For the degree of

Master of Technology

In Bioinformatics

Submitted By

Unnati Goel

Enrolment No. - 2K11/BIO/20

Under the supervision of

Dr. Vimal Kishor Singh

Delhi Technological University
Shahbad Daulatpur, Bawana Road
New Delhi- 110042
July 2013

CERTIFICATE

This is to certify that the M. Tech. dissertation entitled **“To find out molecular structure of CD34 gene and protein and to explore it as an adhesion molecule”**, submitted by **UNNATI GOEL (2K11/BIO/20)** in partial fulfillment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honoring of any other degree.

Date:

Dr. Vimal Kishor Singh
(Project Mentor)
Department of Bio-Technology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)

ACKNOWLEDGEMENT

I take this opportunity to express a deep sense of gratitude towards my guide Dr. Vimal Kishor Singh, for providing excellent guidance, encouragement and inspiration throughout the project work. Without his invaluable guidance, this work would never have been a successful one. I would also like to thank all my colleagues, lab staff as well as all the faculties of Biotechnology for their valuable suggestions and helpful discussions.

Unnati Goel
Delhi Technological University
July 2013

Contents

Page No.

ABSTRACT **6**

ACKNOWLEDGEMENTS **3**

LIST OF FIGURES **5**

1. **INTRODUCTION** **8**

2. **REVIEW OF LITERATURE** **16**

   2.1 Application of stem cells in treating various lung disorders

   2.2 Stem cell markers

   2.3 Separation of stem cells

3. **METHODOLOGY** **20**

   1. To find out molecular structure of CD34 protein 22
   2. To find out molecular structure of CD34 gene 22
   3. To prove that CD34 is an adhesion molecule 23

4. **RESULTS AND DISCUSSIONS** **24**

5. **CONCLUSIONS AND FUTURE PERSPECTIVES** **67**

6. **REFERENCES** **68**

List of Figures –

Page No.

1. Classification of stem cells on basis of potency 8
2. Classification of stem cells on basis of source 9
3. Classification of adult stem cells 10
4. Development of human Embryonic stem cells 11
5. Separation of stem cells 15

To find out molecular structure of CD34 gene and protein and to explore it as an adhesion molecule

Unnati Goel
Delhi Technological University, India

ABSTRACT

Stem cells nowadays, are a broad area of research. These have the capacity to divide and differentiate and form various organs on basis of their properties. Furthermore, they show clinical importance also in treating disorders like, SCID, COPD, and Emphysema which all have been described in this work. This requires association of some proteins which help in identification of the target sites and also have therapeutic value and are known as stem cell markers. Various types of stem cell markers and their roles have also been described in the thesis work. So, the work was to investigate the properties of these stem cell markers which help them in disease identification and treatment, particularly focusing attention on CD34 marker which has its role in homing as well as cell adhesion. Evaluation of its cell adhesion property was done by performing BLAST with similar molecules which are for their adhesion properties. The results proved the adhesion property of this molecule and to confirm this, another search was performed using CLUSTALW for multiple sequence alignment and to construct a phylogenetic tree to evaluate distance homology between them. It showed, close homology to N-CAM and PECAM-1 which are adhesion molecules. So, it was finally proved that CD34 is also an adhesion molecule which is a major finding in stem cell therapy used for lung regeneration.

Keywords- SCID, Emphysema, Therapeutic, Cell adhesion, Homology.

CHAPTER-1

INTRODUCTION

Stem cells are the cells required for tissue regeneration and repair based on their characteristics. It includes development of damaged organs of the body as well but for different organs potential is different and this requires involvement of markers to identify which organ to be repaired so that stem cell will reach to that particular site only which will have receptor for that particular marker. There are various stem cell markers which have been dealt with in detail later in this dissertation. In this particular phenomenon major characteristic that plays role is homing capacity of the stem cell. Homing is the process whereby cells migrate to the organ of their origin. By homing, transplanted hematopoetic stem cells are able to travel to and engraft or establish residence in the bone marrow. Various chemokines and receptors are involved in the homing of hematopoietic stem cells.

This particular study involves in silico analysis of the properties of various markers on stem cell which is to recognize the particular organ, that how it attaches to that organ. Here, cell adhesion plays role but it was yet to be confirmed which has been confirmed in this work using various softwares and tools like PSI- BLAST, CLUSTAL W after which guide tree will tell us the relationship between the concerned molecule and other adhesion molecules.

PSI- BLAST program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which summarizes significant features present in these sequences. A query against the protein database is then run using this profile, and a larger group of proteins is found. This larger group is used to construct another profile, and the process is repeated. By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.

CLUSTAL W program accepts a wide range of input formats, including NBRF/PIR, FASTA, EMBL/Swiss-Prot, Clustal, GCC/MSF, GCG9 RSF, and GDE. The output format can be one or many of the following: Clustal, NBRF/PIR, GCG/MSF, PHYLIP, GDE, or NEXUS.

The adhesion property of these molecules helps in attaching them to the target site for which they are specific. In this study CD34 has been used as a model as it has been found that CD34 is most effective in lung regeneration therapy based on its adhesion properties. CD34 is also a stem cell marker where CD stands for cluster of differentiation, is a protocol used for the identification and investigation of cell surface molecules providing targets for immunophenotyping of cells. Physiologically, CD molecules can act in numerous ways, often acting as receptors or ligands (the molecule that activates a receptor) important to the cell. A signal cascade is usually initiated, altering the behavior of the cell.

CHAPTER- 2

REVIEW OF LITERATURE

Stem cells are the starting cell type for every cell in the body. When developing in the uterus, every cell started out as a stem cell and then developed into a specific cell type such as lung, heart, or tissue cell. Yet during this differentiation process, stem cells also retain the ability to make copies of them in an undifferentiated state. Thus they are considered “immortal” for the life of an organism, and can even be grown in culture indefinitely. Depending on the type of stem cells, some possess the ability to differentiate into one or many cell types, some stem cells can even develop into any cell type. Due to this ability (pluripotency), they are of clinical importance in regenerative therapies for all sorts of diseases such as Alzheimer’s, Parkinson’s, heart disease and many others.

2.1 Stem Cell Classification by Potency

All stem cells can be classified under four different categories that describe their potency, or, the extent into which they can differentiate. These four categories are totipotent, pluripotent, multipotent, and unipotent. A totipotent stem cell can give rise to all the cell types in the body including the entire embryo and placenta; a fertilized egg cell is the only cell that is considered totipotent.

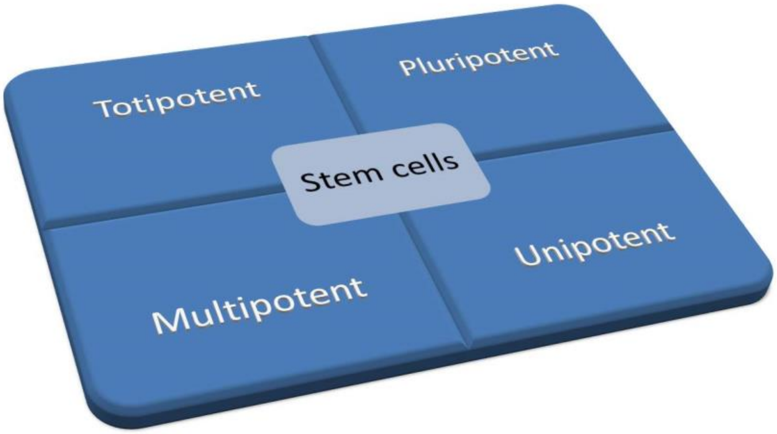


Figure 1 Classification of stem cells on basis of potency (totipotent- the cells have potential to develop into complete organism, pluripotent- ability to develop into an organ (ES cells),

multipotent- ability to develop into more than one kind of organs (HSCs), unipotent- ability to form whole organism).

2.2 Stem Cell Classification by Source

Stem cells can also be classified by their source. There are two different types of stem cells based on source. The first and least controversial are adult stem cells and one of the most common being hematopoietic stem cells. Hematopoietic stem cells are the precursors of mature red and white blood cells that have the ability to replace bone morrow upon its destruction as well as produce mature blood cells. Other adult stem cell types can be found in the brain, blood, cornea, retina, heart, intestines and several other areas. Some advancement has already been made in this field, such as the ability to perform adult stem cell replacement, through bone marrow transplantation, as a treatment for blood cancers and other blood disorders. Thus far, umbilical cord blood stem cells have been used for stem cell transplantation to reconstitute blood cell formation in patients that have been exposed to radiation, or given drugs for cancer or leukemia. Also, in some genetic diseases, a transplantation of umbilical cord blood cells can give them a new system that can form healthy blood cells.

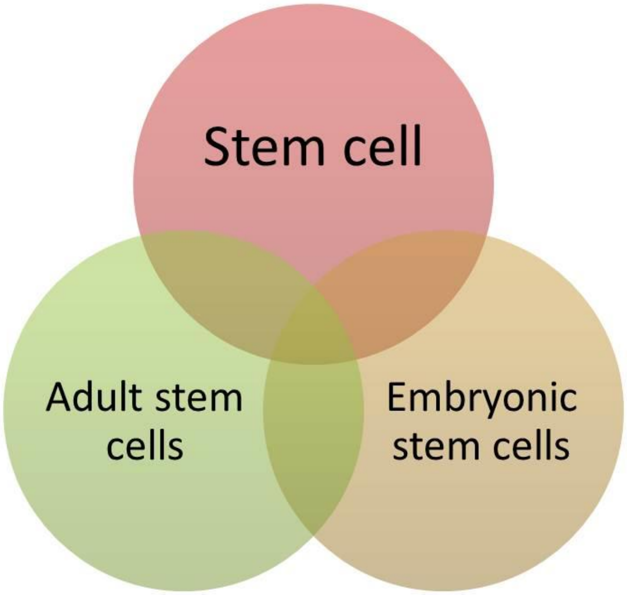


Figure 2 Classification on the basis of source- Adult stem cells and embryonic stem cells

Adult stem cells can be further classified into various kinds of stem cells as follows-

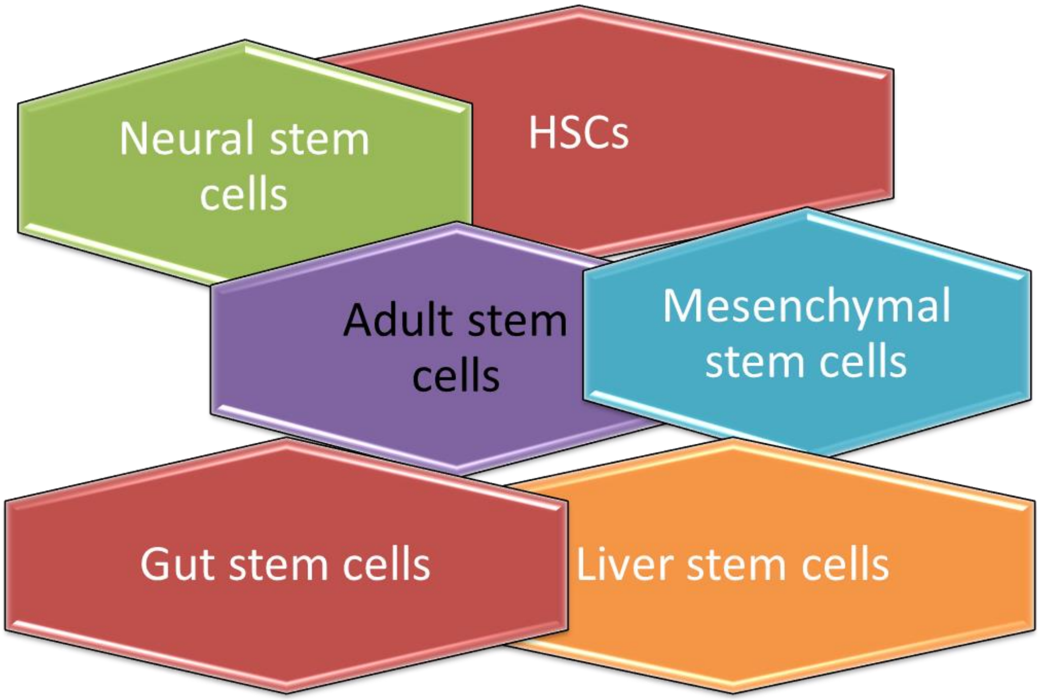


Figure 3 Classification of Adult Stem Cells- firstly it can be classified into hematopoetic stem cells, secondly comes mesenchymal stem cells, thirdly gut stem cells, fourthly liver stem cells, fifthly neural stem cells and so on).

Mesenchymal stem cells form bone, muscle, fat, and cartilage. Mesenchymal stem cells are also involved with repair in bone and cartilage. Once these cells divide, their progeny become committed to one function that is characteristic of a specific tissue type (e.g. cartilage). [Kaplan A I (1991)]

Another type of adult stem cell is the neural stem cell; these can grow from adult brain tissue in culture media. The term 'neural stem cell' is used to describe the cells that (i) can generate neural tissue or are derived from the nervous system, (ii) have some capacity for self-renewal, and (iii) can give rise to cells other than themselves through asymmetric cell division. [Fred H. Gage (2000)]

Research in the field of adult stem cells has given us much insight into what could be done with these cell lines, such as controlling and protecting vital organs from inflammatory and destructive autoimmune reactions, [Laar J M et al. (2005)] treatment of cancers [Weiss M L et al. (2003)], and their possible use in the treatment of several other debilitating disorders.

Limitations of using Adult Stem cells over Embryonic Stem cells-

- They are limited in their usage.
- They have already transformed into a specific type of cell.
- They are only found in one or very few areas and therefore could not be used to create any type of cell.
- ES cells are pluripotent and therefore have the ability to become any cell type.

Limitations of ES cells-

- It has to terminate potential human life in order to obtain the cells.
- Time taking process.

Embryonic Stem Cells-

Embryonic stem (ES) cells are pluripotent stem cells derived from the inner cell mass of a blastocyst-stage embryo. A blastocyst-stage embryo is a human embryo that has not been implanted into a uterus about five days after being fertilized in vitro. A blastocyst has two parts. The first part is the trophoblast, or outer cell ring. Inside the outer cell ring is a mass of about 30 cells referred to as the inner cell mass. The inner cell mass is made up of pluripotent stem cells.

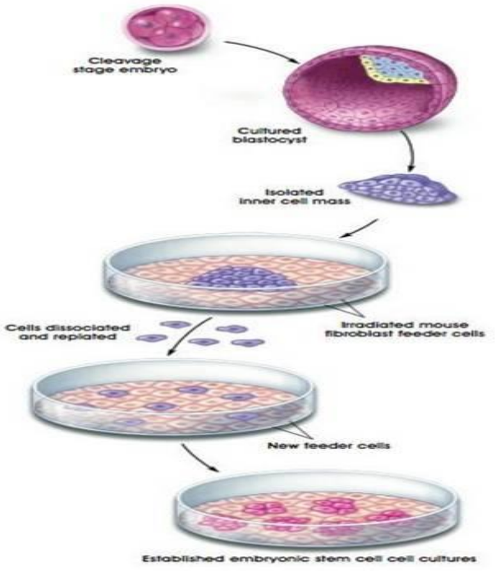

Cleavage stage embryo
Cultured blastocyst
Isolated inner cell mass
Cells dissociated and replated
Irradiated mouse fibroblast feeder cells
New feeder cells
Established embryonic stem cell cell cultures

Figure 4 Showing the development of human embryonic stem cells. 1. Cleavage of 8-celled stage blastocyst, 2&3. Culture of the blastocyst and isolation of inner cell mass, 4. Spread of inner cell mass over irradiated mouse fibroblast cells feeder layer, 5. Dissociation and

replacement of cells, 6. Establishment of ES cell line (*Image Courtesy- Techniques for Generating Embryonic Stem Cell Cultures. (© 2001 Terese Winslow, Caitlin Duckwall)*).

Fundamental Properties of an Embryonic Stem cell-

1. Capable of undergoing an unlimited number of symmetrical divisions without differentiating (long-term self-renewal).
2. Exhibit and maintain a stable, full (diploid), normal complement of chromosomes (karyotype).
3. Pluripotent ES cells can give rise to differentiated cell types that are derived from all three primary germ layers of the embryo (endoderm, mesoderm, and ectoderm).
4. Capable of integrating into all fetal tissues during development.
5. Clonogenic i.e. a single ES cell can give rise to a colony of genetically identical cells, or clones, which have the same properties as the original cell.
6. **Expresses the transcription factor Oct-4, which** can be induced to continue proliferating or to differentiate.
7. **Lacks the G1 checkpoint in the cell cycle.** ES cells spend most of their time in the S phase of the cell cycle, during which they synthesize DNA. Unlike differentiated somatic cells, ES cells do not require any external stimulus to initiate DNA replication.
8. Do not show X inactivation. In every somatic cell of a female mammal, one of the two X chromosomes becomes permanently inactivated. X inactivation does not occur in undifferentiated ES cells.

2.3 Stem cell applications

2.1Application of stem cells in treating various disorders-

1. <u>**Lung Disorders**</u>**-**

Incurable lung diseases pose a major challenge to medical science. Cystic fibrosis, asthma, pulmonary fibrosis, cancer, hyaline membrane disease and emphysema are examples of diseases that may be eventually conquered by stem cell therapies. However, due to its geometrical complexity it is difficult to deliver therapeutic agents to the diseased regions of the lung where they may affect cures. Two of the diseases have their treatments using stem cells discovered so far. These are-

a) **Interstitial Lung Disease**- (ILD)

Interstitial lung disease is a term used for a particular type of interstitium inflammation of the lungs. The interstitium is the tissue, surrounds and separates the tiny air sacs (alveolae) in the lungs. Interstitial lung disease involves an inflammation of this supportive tissue between the air sacs. The scarring (fibrosis) begins in Interstitium. This damage of lung tissue occurs due to known or unknown reason. ILD may be called as interstitial pulmonary fibrosis or pulmonary fibrosis.

The lung is a complex organ with limited regenerative capacity and lung injuries are leading causes of morbidity and mortality worldwide. It is believed that stem cells act through paracrine mechanism. Injured lung is thought to produce several chemokines, including hyaluronan, osteopontin, stromal-derived factor 1α, and secondary lymphoid chemokine, which interact with several receptors present on stem cells that stimulate proliferation of cells and migrate to sites of injury.

Regeneration of tissue by stem cells from endogenous, exogenous, and even genetically modified cells is a promising new therapy under evaluation for ILD, not responding to conventional therapy. Different studies are conducted to support that bone marrow progenitor cells contribute to repair and remodeling of lung in animal models of progressive pulmonary fibrosis. Recent studies have demonstrated paracrine effects of administered cells, including stimulation of angiogenesis and modulation of local inflammatory and immune responses in model mouse lung disease. Also some recent studies demonstrate that mesenchymal stem cells (MSCs) can modulate local inflammatory and immune responses in mouse lung disease models.

b) **Pulmonary Hypertension**-

Pulmonary hypertension generally results from constriction or stiffening of the pulmonary arteries that supply blood to the lungs, making it difficult for the heart to pump blood forward through the lungs. This stress on the heart leads to enlargement of the right heart and eventually fluid can build up in the liver and other tissues, such as in the legs. Pulmonary hypertension is divided into two main categories; **1)** primary pulmonary hypertension (not caused by any other disease or condition); Primary pulmonary hypertension has no identifiable underlying cause and is also referred to as idiopathic pulmonary hypertension. And **2)** secondary pulmonary hypertension (caused by another underlying condition).

It is believed that Endothelial Progenitor Cells (EPCs) are involved in endothelial homeostasis, as well as physiological and pathological angiogenesis. It helps in replace the damaged blood vessels in the lungs of the pulmonary hypertension patients. Even patients no longer needs to be supplemented with oxygen or considered for a lung transplant. EPC-based therapies in patients with pulmonary hypertension show benefit of this approach, thus revealing EPCs as potential therapeutic targets.

The clear advantage of EPCs would be their natural homing to the site of angiogenesis, which would target them to the site of muscle regeneration. It is hypothesized that neovascularization in the lung could increase the volume of the vascular bed in the pulmonary circulation and thus reduce the development of pulmonary hypertension (PH). These suggest that EPC transplantation can be a new therapeutic option for PPH.

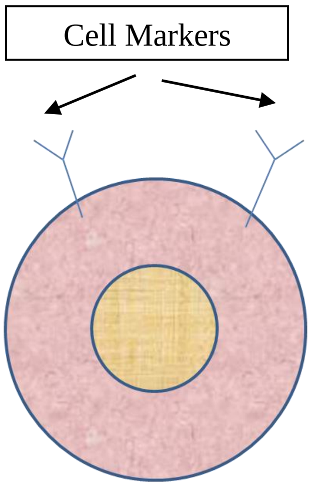
c) **Emphysema**-

Emphysema is a long-term lung disease. In people with emphysema, the tissues necessary to support the shape and function of the lungs are destroyed. It is included in a group of diseases called chronic obstructive pulmonary disease or COPD. Emphysema is called an obstructive lung disease because the destruction of lung tissue around smaller sacs, called alveoli, makes these air sacs unable to hold their functional shape upon exhalation. Emphysema is most often caused by tobacco smoking and long-term exposure to air pollution. With the discovery of multipotent lung stem cells in 2011, a new treatment option may soon become available. Scientists injected human lung stem cells into mice with damaged lungs. The stem cells formed human bronchioles, alveoli, and pulmonary vessels integrated structurally and functionally with the damaged mouse organ. The May 2011 report in the New England Journal of Medicine concluded that human lung stem cells "have the undemonstrated potential to promote tissue restoration in patients with lung disease".

2.3 **Stem Cell Markers**-

Stem cell markers are genes and their protein products used by scientists to isolate and identify stem cells. Stem cells can also be identified by functional assays. Below is a list of genes/protein products that can be used to identify various types of stem cells.

- **c-Kit**
- **Oct4** (ATGCAAAT) POU Family Protein
- **CD34**
- **CD44**
- **CD133**
- **Nestin**



Stem Cell

c-kit- Mast/stem cell growth factor receptor (SCFR), also known as proto-oncogene c-Kit or tyrosine-protein kinase Kit or CD117, is a protein that in humans is encoded by the KIT gene.[Andre C et al. (1997)] CD117 is an important cell surface marker used to identify certain types of hematopoietic (blood) progenitors in the bone marrow. To be specific, hematopoietic stem cells (HSC), multipotent progenitors (MPP), and common myeloid progenitors (CMP) express high levels of CD117.

Function- CD117 is a cytokine receptor expressed on the surface of hematopoietic stem cells as well as other cell types.[Edling C E et al. (2007)] CD117 is a receptor tyrosine kinase type III, which binds to stem cell factor (a substance that causes certain types of cells to grow), also known as "steel factor" or "c-kit ligand". Activating mutations in this gene are associated with gastrointestinal stromal tumors, testicular seminoma, mast cell disease, melanoma, acute myeloid leukemia, while inactivating mutations are associated with the genetic defect piebaldism.

CD34- CD34 molecule is a cluster of differentiation molecule present on certain cells within the human body. It is a cell surface glycoprotein and functions as a cell-cell adhesion factor. It may also mediate the attachment of stem cells to bone marrow extracellular matrix or directly to stromal cells. [Simmons D L et al. (1992)] The CD34 protein is a member of a family of single-pass transmembrane sialomucin proteins that show expression on early hematopoietic and vascular-associated tissue. CD34 is also an important adhesion molecule and is required for T cells to enter lymph nodes. [Nielsen J S et al. (2008)] It is expressed on lymph node endothelia whereas the L-selectin to which it binds is on the T cell. [Berg E L et al. (1998); SSuzawa K et al. (2007)].The presence of the surface antigen CD34 on hematopoeitic stem cells is presently used as a selective marker for enumeration of these populations for hematopoeitic disorders. [Burt KE. (1999)]. Efforts are being done to unravel the structural and biological significance of this molecule have shown that it is a highly glycosylated, type I integral protein. Other studies have implicated CD34 in regulation of the adhesive properties of hematopoeitic progenitors and in cytokine- induced differentiation. However, the role of CD34 in stem cell function remains to be elucidated.

The predicted CD34 structure consists of a large extracellular domains (278 residues), followed by a stretch of hydrophobic amino acids within the transmembrane region (~22 residues) and a small cytoplasmic tail (73 residues). [Civin CL et al. (1984), Lanza F, L Healy et al. (2001)].

Oct-4- Oct-4 (octamer-binding transcription factor 4) also known as **POU5F1** (POU domain, class 5, transcription factor 1) is a protein that in humans is encoded by the *POU5F1* gene. [Takeda et al. (1992)] This protein is critically involved in the self-renewal of undifferentiated embryonic stem cells. Oct-4 transcription factor is initially active as a maternal factor in the oocyte but remains active in embryos throughout the preimplantation period. Oct-4 expression is associated with an undifferentiated phenotype and tumors. Mouse embryos that are Oct-4-deficient or have low expression levels of Oct-4 fail to form the inner cell mass, lose pluripotency and differentiate into trophoectoderm. [Looijenga et al. (2003)] Therefore, the level of Oct-4 expression in mice is vital for regulating pluripotency and early cell differentiation since one of its main functions is to keep the embryo from differentiating.

CD44- CD44 antigen is a cell-surface glycoprotein involved in cell–cell interactions, cell adhesion and migration. In humans, the CD44 antigen is encoded by the CD44 gene on Chromosome 11. [Oxley SM et al. (1994)] CD44 is a receptor for hyaluronic acid and can also interact with other ligands, such as osteopontin, collagens, and matrix metalloproteinases (MMPs). CD44 function is controlled by its posttranslational modifications. One critical modification involves discrete sialofucosylations rendering the selectin-binding glycoform of CD44 called HCELL (for Hematopoietic Cell E-selectin/L-selectin Ligand). The HCELL glycoform was originally discovered on human hematopoietic stem cells and leukemic blasts, [Sackstein et al. (2000a, 2008b; Dimitroff et al. (2001); Hanley W D et al. (2005)] and was subsequently identified on cancer cells. [Burdick et al. (2006); Hanley W D et al. (2006); Napier S L et al. (2007); Thomas S N et al. (2008)].

CD133- CD133 is a glycoprotein also known in humans and rodents as Prominin 1 (PROM1). [Corbeil D et al. (2001)] Currently the function of CD133 is unknown. It is a member of pentaspan transmembrane glycoproteins (5-transmembrane, 5-TM), which specifically localize to cellular protrusions. Recent studies in brain tumors have identified a CD133+ cell population thought to be a cancer stem cell population, which is rare, undergoes self-renewal and differentiation, and can propagate tumors when injected into immune-compromised mice. [Singh SK et al., (2003); Hemmati H D et al., (2003); Galli R et al., (2004)]

2.4 **Separation of Stem Cells**-

- Cells in suspension are tagged with fluorescent markers specific for undifferentiated stem cell.
- Labeled cells are sent under pressure through a small nozzle and pass through an electric field.
- A cell generates a negative charge if it fluoresces and a positive charge if it does not.

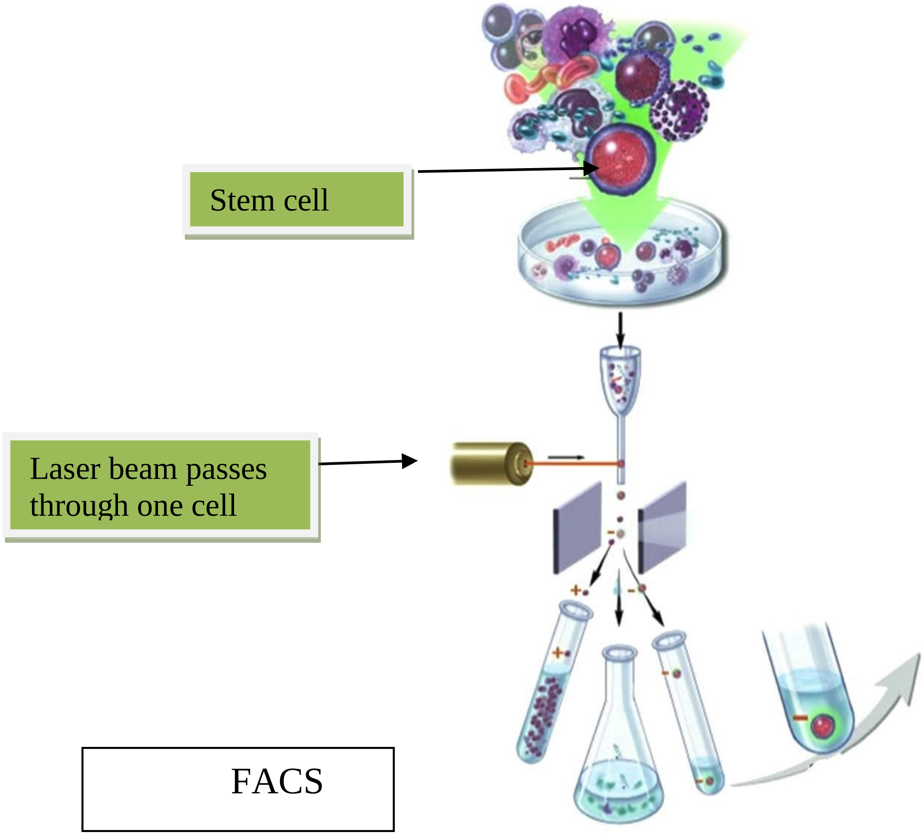


Figure 5 Separation of Stem Cells through Fluorescence Activated Cell Sorting- Mixture of labeled stem cells pass through funnel, laser beam passes through one cell at a time, charge separation takes place hence positively charged non-fluorescent stem cells collected in a separate centrifuge tube and negatively charged fluorescent stem cells in another tube. (Fig. Ref- © 2001 Terese Winslow, Lydia Kibiuk, Caitlin Duckwall)

2.5 Role of human lung stem cell in tissue regeneration-

Using lung tissue from surgical samples, researchers identified and isolated the human lung stem cell and tested the functionality of the stem cell both in vitro and in vivo. Once the stem cell was isolated, researchers demonstrated in vitro that the cell was capable of dividing both into new stem cells and also into cells that would grow into various types of lung tissue. Next, researchers injected the stem cell into mice with damaged lungs. The injected stem cells differentiated into new bronchioles, alveoli and pulmonary vessel cells which not only formed new lung tissue, but also integrated structurally to the existing lung tissue in the mice.

The researchers define this cell as truly "stem" because it fulfills the three categories necessary to fall under stem cell categorization: first, the cell renews itself; second, it forms into many different types of lung cells; and third, it is transmissible, meaning that after a mouse was injected with the stem cells and responded by generating new tissue, researchers were then able to isolate the stem cell in the treated mouse, and use that cell in a new mouse with the same results. These are the critical first steps in developing clinical treatments for those with lung disease for which no therapies exist. Further research is needed, in this field.

CHAPTER- 3

METHODOLOGY

A. OBJECTIVE- To find out molecular structure of CD34 protein.

Procedure-

a. Analysis of amino acid composition of CD34 protein using ANNIE.

- Amino acid composition was found out using ANNIE. (An integrated de novo protein sequence annotation tool).

Significance of finding this amino acid composition of a protein-

When we know the amino acid composition we can predict secondary and tertiary structure of a protein knowing the occurrence of a particular amino acid residue in a particular structure as it is already known since a long time that which amino acids contribute to which particular structure.

b. Secondary Structure prediction of CD34 protein.

There are a number of tools available for predicting secondary structure of a protein as mentioned below-

- **AGADIR**- which predicts helical content.
- **BCM PSSP**- This provides a rich set of programs for protein secondary structure determination. For example-
  - **Coils** - prediction of coiled coil regions
  - **nnPredict** - uses a 2 layer neural network
  - **PSSP / SSP** - segment-oriented prediction
  - **PSSP / NNSSP** - nearest-neighbor prediction
  - **SAPS** - statistical analysis of protein sequences
  - **TMpred** - transmembrane region and orientation prediction
  - **PHDsec** - profile network method
  - **PSA** - for single domain globular proteins
  - **SOPM** - self optimized prediction method
  - **SSPRED** - with residue exchange statistics

- **Swiss-Model** - from alignment to crystallographic data

- **PSIpred-** PSIpred - Prediction of secondary structure from multiple sequences
  MEMSAT 2 - Prediction of transmembrane topology from multiple sequences
  GenTHREADER - Fast and reliable protein fold recognition

- **HMMTOP-** HMMTOP is an automatic server for predicting topology of transmembrane proteins. The method is based on the hypothesis that topology is determined by the maximum divergence of the amino acid distributions of the various structural parts in membrane proteins.
- **Predict Protein Server (Maxhom/PHD) -** PredictProtein is an automatic service for the prediction of aspects of protein structure. You send an amino acid sequence and PredictProtein returns a multiple sequence alignment, and a prediction of the secondary structure, residue solvent accessibility, and helical transmembrane regions.
- **TMpred-** (prediction of transmembrane regions and orientation) this program tries to find putative transmembrane domains in proteins and also speculates on the possible orientation of these segments. For its scoring, it uses a combination of multiple weight-matrices that have been extracted from a statistical analysis of TMbase, a collection of all annotated transmembrane proteins present in SwissProt.
- **SignalP- (**predicts signal peptides of secretory proteins) SignalP predicts signal peptides of secretory proteins. For cleaved signal peptides, the precise location of the cleavage site in the amino acid sequence is predicted. The prediction is optimized for three different types of organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.
- **SOSUI- (secondary structure prediction of membrane proteins)** The SOSUI system is a useful tool for secondary structure prediction of membrane proteins from a protein sequence. The basic idea of prediction in this system is based on the physicochemical properties of amino acid sequences such as hydrophobicity and charges. The system deals with three types of prediction: discrimination of membrane proteins from soluble one, prediction of existence of transmembrane helices and determination of transmembrane

helical regions. The accuracy of this system, discrimination of membrane proteins, existence of transmembrane helices and transmembrane helical regions, are about 99%, 96% and 85% respectively.

- **Prediction of helical content in amino acid sequence of CD34** (using SOSUI)

  Average of hydrophobicity calculated by the server was -0.3205.

  Two transmembrane alpha helices were predicted-

Rest of the secondary structure information retrieved at Mobyle@pasteur portal using GOR method.

- Coiled regions found.
- Consists of helix, beta strands and turns also.
- ***Individual percentage of these secondary structures is as follows-**

Helix= 24.9, beta strand= 32.8, turns= 17.4, Coil= 29.4

- **To find out motifs present in the sequence- using MOTIF Scan**

B. OBJECTIVE- To find out molecular structure of CD34 gene.

<u>Procedure</u>-

a. To find out nucleic acid composition of CD34. (using word count at Mobyle@pasteur portal)

Located on chromosome no. 1q32 consists of 24801 nucleotides. Composition is as follows-

T	7044
A	6405
C	5709
G	5643

b. To predict secondary structure of the nucleotide sequence.

c. To find motifs in a DNA sequence.

C. **OBJECTIVE**- To generated and validate 3D model of CD34 protein

- 3 methods involved-
  - Threading/ fold recognition
  - Comparative modeling
  - Ab initio prediction

These all utilize different tools for prediction which have been used for producing the output here.

D. **OBJECTIVE-** To prove that CD34 is an adhesion molecule.

<u>Procedure</u>-

Perform PSI-BLAST of CD34 with adhesion molecules known.

- Divide the entire stretch of CD34 containing 373 amino acids into 3 fragments of following sizes- 1-259, 260-300, 301-373.
- Compute alignment scores individually for each fragment.

a. **CD34 with CADHERIN** (CADHERIN is 140 amino acid containing adhesion molecule having accession no.- BAA21568.1
   For fragment1 (1-259 amino acids)

b. CD34 with PECAM-1(PECAM-1 is platelet endothelial cellular adhesion molecule or CD31 having accession no. AAB28645.1
   For fragment 1 (1-259 amino acids)

c. **CD34 with N-CAM** (N-CAM is Neural cell adhesion molecule or CD56 HAVING ACCESSION NO. AAB31836.1)
   For fragment 1 (1-259 amino acids)

d. **CD34 with PODOCALYXIN** (PODOCALYXIN is also an adhesion molecule which is a sialoglycoprotein and also a member of CD34 family of transmembrane proteins having accession no. NP_001018121.1)
   For fragment 1 (1-259 amino acids)

CHAPTER 4

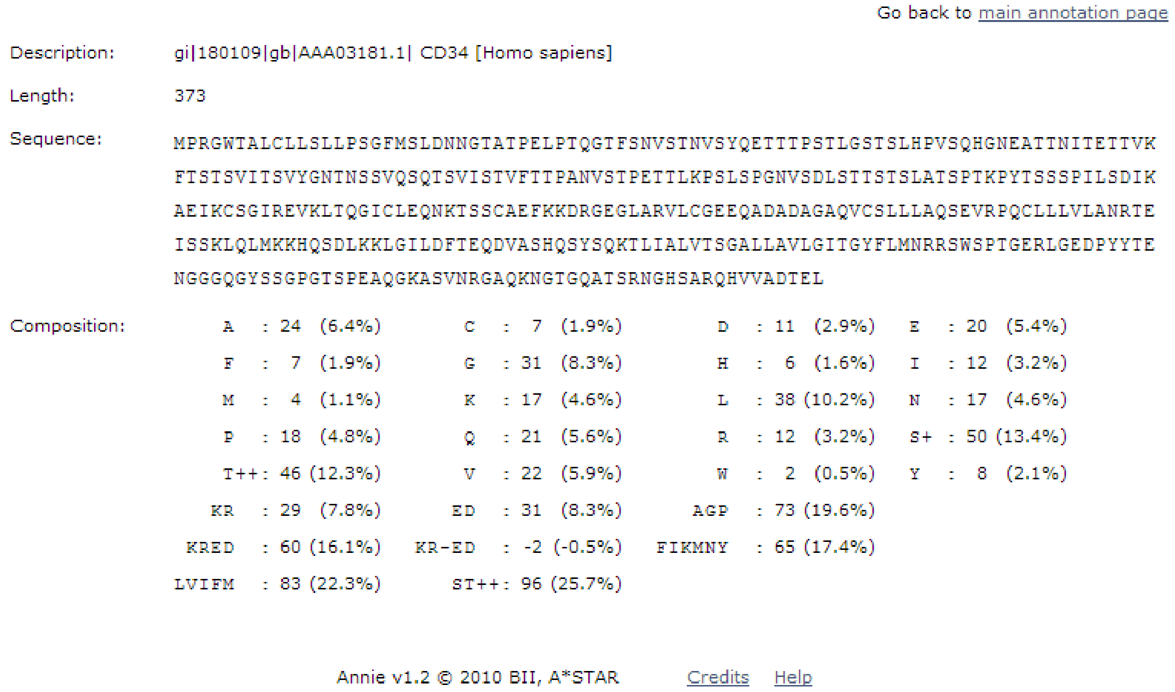
RESULTS AND DISCUSSIONS

A. OBJECTIVE- To find out molecular structure of CD34 protein.

FASTA sequence of CD34 protein, 373 amino acids, linear, Accession no. M81104.1

```
>gi|180109|gb|AAA03181.1| CD34 [Homo sapiens]
MPRGWTALCLLSLLPSGFMSLDNNGTATPELPTQGTFSNVSTNVSYQETTTPSTLGSTSLHPVSQHGNEA
TTNITETTVKFTSTSVITSVYGNTNSSVQSQTSVISTVFTTPANVSTPETTLKPSLSPGNVSDLSTTSTS
LATSPTKPYTSSSPILSDIKAEIKCSGIREVKLTQGICLEQNKTSSCAEFKKDRGEGLARVLCGEEQADA
DAGAQVCSLLLAQSEVRPQCLLLVLANRTEISSKLQLMKKHQSDLKKLGILDFTEQDVASHQSYSQKTLI
ALVTSGALLAVLGITGYFLMNRRSWSPTGERLGEDPYYTENGGGQGYSSGPGTSPEAQGKASVNRGAQKN
GTGQATSRNGHSARQHVVADTEL
```

- Amino acid composition was found out using ANNIE. (An integrated de novo protein sequence annotation tool).



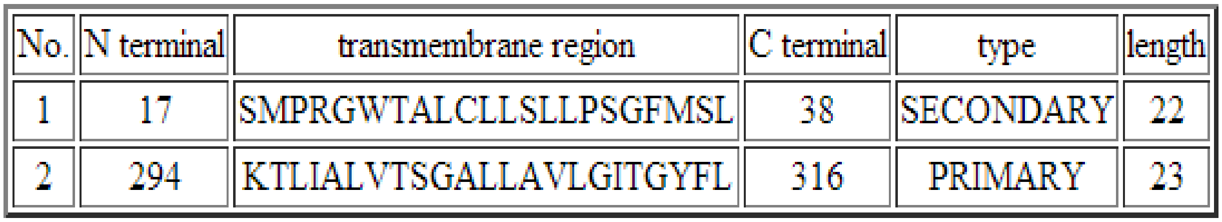
```
Description:    gi|180109|gb|AAA03181.1| CD34 [Homo sapiens]

Length:         373

Sequence:       MPRGWTALCLLSLLPSGFMSLDNNGTATPELPTQGTFSNVSTNVSYQETTTPSTLGSTSLHPVSQHGNEATTNITETTVK
                FTSTSVITSVYGNTNSSVQSQTSVISTVFTTPANVSTPETTLKPSLSPGNVSDLSTTSTSLATSPTKPYTSSSPILSDIK
                AEIKCSGIREVKLTQGICLEQNKTSSCAEFKKDRGEGLARVLCGEEQADADAGAQVCSLLLAQSEVRPQCLLLVLANRTE
                ISSKLQLMKKHQSDLKKLGILDFTEQDVASHQSYSQKTLIALVTSGALLAVLGITGYFLMNRRSWSPTGERLGEDPYYTE
                NGGGQGYSSGPGTSPEAQGKASVNRGAQKNGTGQATSRNGHSARQHVVADTEL

Composition:        A   : 24  (6.4%)        C   :  7  (1.9%)         D   : 11  (2.9%)   E   : 20  (5.4%)
                    F   :  7  (1.9%)        G   : 31  (8.3%)         H   :  6  (1.6%)   I   : 12  (3.2%)
                    M   :  4  (1.1%)        K   : 17  (4.6%)         L   : 38 (10.2%)   N   : 17  (4.6%)
                    P   : 18  (4.8%)        Q   : 21  (5.6%)         R   : 12  (3.2%)   S+  : 50 (13.4%)
                    T++: 46 (12.3%)         V   : 22  (5.9%)         W   :  2  (0.5%)   Y   :  8  (2.1%)
                   KR   : 29  (7.8%)       ED   : 31  (8.3%)       AGP   : 73 (19.6%)
                 KRED   : 60 (16.1%)    KR-ED   : -2 (-0.5%)    FIKMNY   : 65 (17.4%)
                LVIFM   : 83 (22.3%)      ST++: 96 (25.7%)
```

This amino acid composition helps us to predict N and O glycosylation sites of a protein by observing which amino acid residues are occurring frequently and repeatedly. For example- Threonine, Serine and Aspartate are occurring repeatedly here and these contribute to random coil regions and this also gives idea about the stability of the molecule if similar types of amino acid residues will be near molecule will be unstable due to repulsion.

- **Prediction of helical content in amino acid sequence of CD34** (using SOSUI)

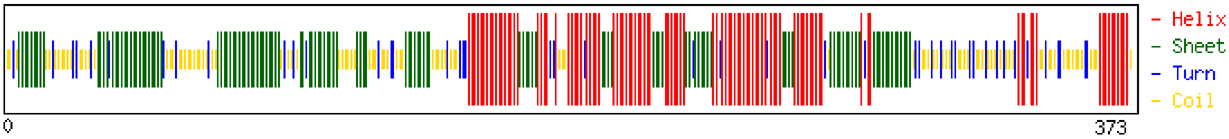
  Average of hydrophobicity calculated by the server was -0.3205.

  Two transmembrane alpha helices were predicted-



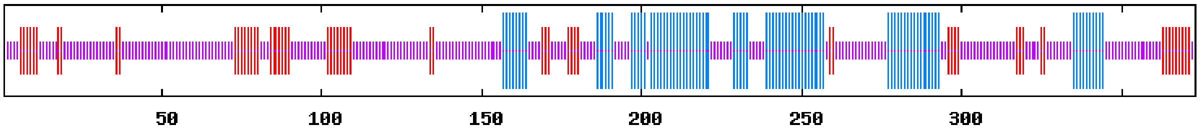
No.	N terminal	transmembrane region	C terminal	type	length
1	17	SMPRGWTALCLLSLLPSGFMSL	38	SECONDARY	22
2	294	KTLIALVTSGALLAVLGITGYFL	316	PRIMARY	23

Prediction of Helical Content, strand and turn content using CFSSP (Chou and Fasman secondary structure prediction server) at ExPAsy.



Helix= 46.6, Strand= 41, Turns= 16.1

All types of secondary structures found out using GOR method available at ExPAsy.



GOR4-

Alpha Helix	89 is 23.86%
3_{10} Helix	0 is 0.00%
Pi Helix	0 is 0.00%
Beta bridge	0 is 0.00%
Extended strand	62 is 16.62%
Beta Turn	0 is 0.00%
Bend Region	0 is 0.00%
Random Coil	222 is 59.52%
Ambiguous States	0 is 0.00%
Other States	0 is 0.00%

Protein structure prediction using SCRATCH Protein Predictor at ExPAsy-

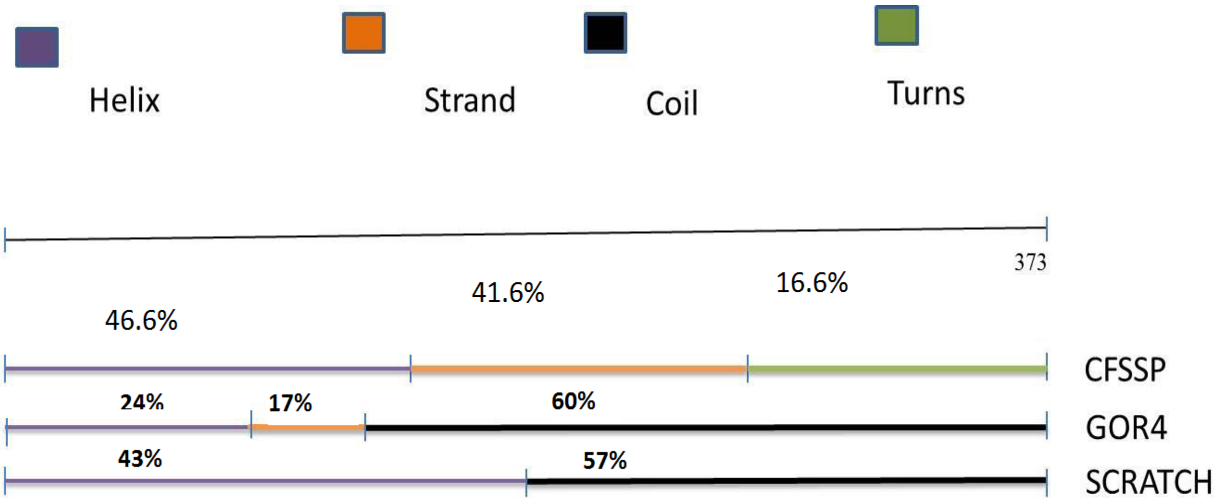
```
Amino Acids:
MPRGWTALCLLSLLPSGFMSLDNNGTATPELPTQGTFSNVSTNVSYQETTTPSTLGSTSLHPVSQHGNEATTNITET
TVKFTSTSVITSVYGNTNSSVQSQTSVISTVFTTPANVSTPETTLKPSLSPGNVSDLSTTSTSLATSPTKPYTSSSP
ILSDIKAEIKCSGIREVKLTQGICLEQNKTSSCAEFKKDRGEGLARVLCGEEQADADAGAQVCSLLLAQSEVRPQCL
LLVLANRTEISSKLQLMKKHQSDLKKLGILDFTEQDVASHQSYSQKTLIALVTSGALLAVLGITGYFLMNRRSWSPT
GERLGEDPYYTENGGGQGYSSGPGTSPEAQGKASVNRGAQKNGTGQATSRNGHSARQHVVADTEL

Predicted Secondary Structure:
CCCHHHHHHHHHHCCCCCEECCCCCCCCCCCCCCCCCCCCCCCCCCCECCCCCCCCCCCCCCCCCCCCCCCCCEEEE
EEEEECCEEEECCCCCCCCCCCCCCEEEEEEEECCCCCCCCCEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CHHHCCCEEEECCCEEEEEHCCEEEECCCCCCHHHHCCCCCCCHHHHECCHHHCCCCCCHHHHHHHHHHHCCCHHHH
HHHHHCHHHHHHHHHHHHHHHHHHHHCCCHHCCHHHHCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHECCCCCCCC
CCCCCCCCCEECCCCCCCCCCCCCCCCHCCCCEEECCCCCCCCCCCCCCCCCCCCCCEEEECCCC

Predicted Secondary Structure (8 Classes):
CCCHHHHHHHHHHCTTTCEEECTTSCCCCCCCCCCCCCCCCCCCCEEECCCCCCCSCEEEEEECTTSCEEEEEEEEE
EEEEEEEEEEEECCTCCCCEEEEEEEEEEEEEECCCCCCCCEEEECCCCCTTCCEECCCCCCEECCCTCCCCCCSCH
HHHHHHHEEEETTHEEEEEHHEEEEEHCCCCCHHHHHHHHTTHHHHEEEEHHHHCCHHTTHHHHHHHHHHHHCHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHTCEECCHHHHHHHHHHCHHHHHHHHHHHHHHHHHHHHHHHHEECCCCCTT
SCECTCCCEEEETTCCCCCCCSCTCCHHHTTCEEEETCCCTTCCCEEECTTSCCCEEEEEECCCC
```

Of all tools used, conclusion comes out to be that the protein consists mostly of random coil regions and alpha helices.

```
Predicted Domains:
Domain 1: 1 - 136
Domain 2: 137 - 373
```

Predicted domains came out to be in the above regions.

Summary-



All the three methods gave near about similar results with major percentage of being alpha helix and random coil present in CD34 structure.

- To find out motifs present in the sequence**-** using MOTIF Scan

Motifs	**Range**
ASN_GLYCOSYLATION	24-27 39-42 43-46 73-76 95-98 114-117 130-133 182-185 237-240 350-353
CK2_PHOSPHO_SITE	45-48 116-119 186-189 264-267 333-336
MYRISTYL	35-40 67-72 92-97 213-218 286-291 322-327 346-351 353-358
PKC_PHOSPHO_SITE	78-80 121-123 242-244 275-277 356-358 362-364
TYR_PHOSPHO_SITE	311-318
CD34_antigen	178-373

B. OBJECTIVE- To find out molecular structure of CD34 gene

- **To find out nucleic acid composition of CD34. (using word count at Mobyle@pasteur portal)**

  Located on chromosome no. 1q32 consists of 24801 nucleotides. Composition is as follows-

T	7044
A	6405
C	5709
G	5643

- To predict secondary structure of the nucleotide sequence.

  1. **Inverted repeats using Einverted at Mobyle@pasteur portal**.

4475atttatttatttatttatttgagaagaagtttctctcttgttgctcaggctggagtacaatggtgtgatctcggatcaccgcaacct
ctgcctcccaggttcaagcagttctccttcctcagcctcctaagtagctgggattacaggcatgcgccaccatgcccagctaa-
ttttgtatttttagtagagacgggttttctctatattggtcaggctggtctcgaactcctgacttcagatgatccacccacatcggcctc
ccaaaatgttgggattacaggcgtgagccatggtgtctggcc**4776**
6206taaggaaagaaagaaaaaaagaactctgtcagagtgagacaacgggtccgacctcacgtcaccgtgctagaaccgac
cgacgttggtgacggagggtccaagttctgtaagaggacggagtcggagggttcatcgactctgatgtccgtacacggtggtac
gggccgattaaaaacataaaaatcatctctgccccaaatcggtataaccggttcgaccagagtttgaggactggagtccactaga
cggacggagccggagggcttcacgaccctaacgtccgtactcggtggcacgcaccgg**5904**

2. To find motifs in a DNA sequence

Motif	Position	Strand score	Sequence
SRF	5768 to 5780	+6.99	Ttccatatcaggg
TATA	5796 to 5810	+9.23	Ctataaagggccaga
TATA	5831 to 5845	-7.57	Ccatatatggcccct
SRF	5834 to 5846	+7.78	Ggccatatatggt
SRF	5835 to 5847	-10.5	Aaccatatatggc
GATA	5867 to 5879	-7.63	Gtaagataagaca
LSF	5914 to 5928	+6.26	Ggtggctcatgcctg
ERE	5985 to 5998	+6.45	Agaccagcttggcc
Mef-2	6022 to 6033	-7.03	Ttgtatttttag
Mef-2	6031 to 6042	-6	Ggctaatttttg
Myc	6052 to 6061	-7.98	Ggcatgtgcc
LSF	6118 to 6132	+6.66	Agtggttgcagccag
LSF	6119 to 6133	-7.33	Gctggctgcaaccac
LSF	6144 to 6158	-6.79	Gctggagtgcagtgg
SRF	20328 to 20340	+9.72	Agccatgtaaggc
CCAAT	20375 to 20390	-9.14	Ctgaaccaatcaacag

C. **OBJECTIVE-** To predict 3D structure of CD34 protein.

- 3 methods used-
  - Threading
  - Comparative modeling
  - Ab- initio method

Threading- **Protein threading**, also known as **fold recognition**, is a method of protein modeling (i.e. computational protein structure prediction) which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure. It differs from the homology modeling method of structure prediction as it (protein threading) is used for proteins which do not have their homologous protein structures deposited in the Protein Data Bank (PDB), whereas homology modeling is used for those proteins which do. Threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the protein which one wishes to model.

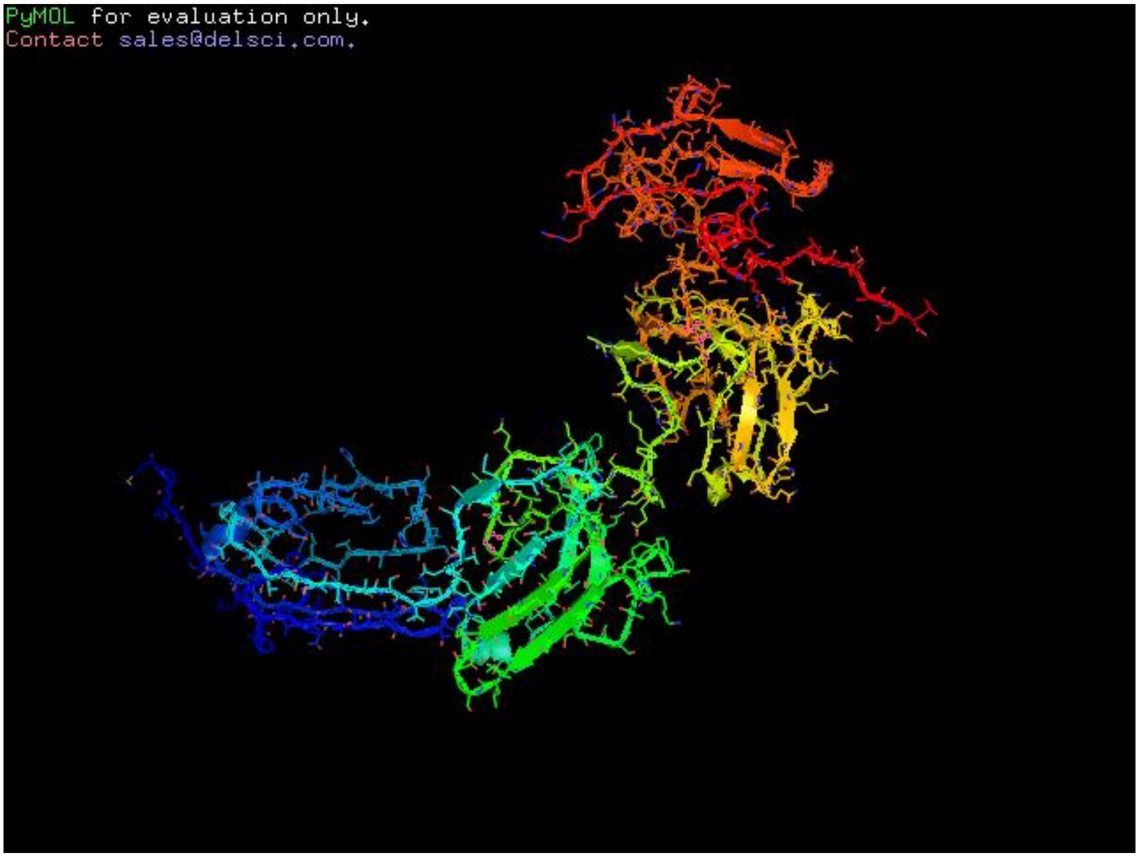
The prediction is made by "threading" (i.e. placing, aligning) each amino acid in the target sequence to a position in the template structure, and evaluating how well the target fits the template. After the best-fit template is selected, the structural model of the sequence is built based on the alignment with the chosen template. Protein threading is based on two basic observations: that the number of different folds in nature is fairly small (approximately 1300); and that 90% of the new structures submitted to the PDB in the past three years have similar structural folds to ones already in the PDB.

Tool for threading used was **MUSTER**- (Cited by- S. Wu, Y. Zhang. (2008)

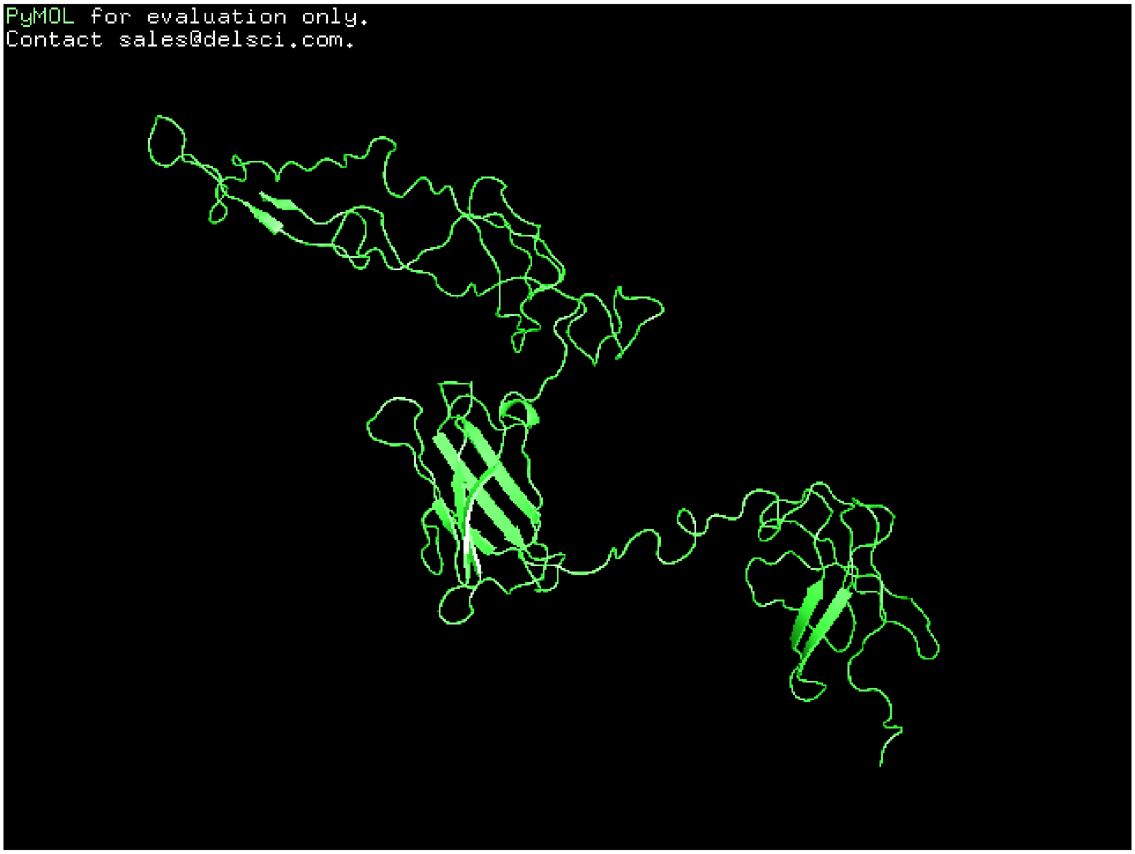
MUSTER is a MUlti-Source ThreadER program, which considers six different sources: (1) sequence-derived profiles; (2) secondary structures; (3) structured-derived profiles; (4) solvent accessibility; (5) torsion angles (psi and phi angles); (6) hydrophobic scoring matrix. The optimized threading is found by global dynamic programming.

Rank	Template	Align_length	Coverage	Zscore	Seq_id	Type	Target template alignments	3D models	Full length models
1	2owcA	362	0.970	5.232	0.088	Bad	Alignment_1	Threading_1	Model1
2	3chnS	373	1	5.141	0.099	Bad	Alignment_2	Threading_2	Model_2
3	1zlgA	362	0.970	5.139	0.135	Bad	Alignment_3	Threading_3	Model_3
4	1gwsA	368	0.986	5.065	0.062	Bad	Alignment_4	Threading_4	Model_4
5	4acqA	373	1	5.028	0.080	Bad	Alignment_5	Threading_5	Model_5
6	3ucpA	372	0.997	4.963	0.108	Bad	Alignment_6	Threading_6	Model_6
7	4akfA	365	0.978	4.910	0.115	Bad	Alignment_7	Threading_7	Model_7
8	2e84A	369	0.989	4.909	0.087	Bad	Alignment_8	Threading_8	Model_8
9	4jrfA	370	0.991	4.866	0.122	Bad	Alignment_9	Threading_9	Model_9
10	4ak1A	371	0.994	4.884	0.108	Bad	Alignment_10	Threading_10	Model_10

Model 1 as viewed in PYMOL-

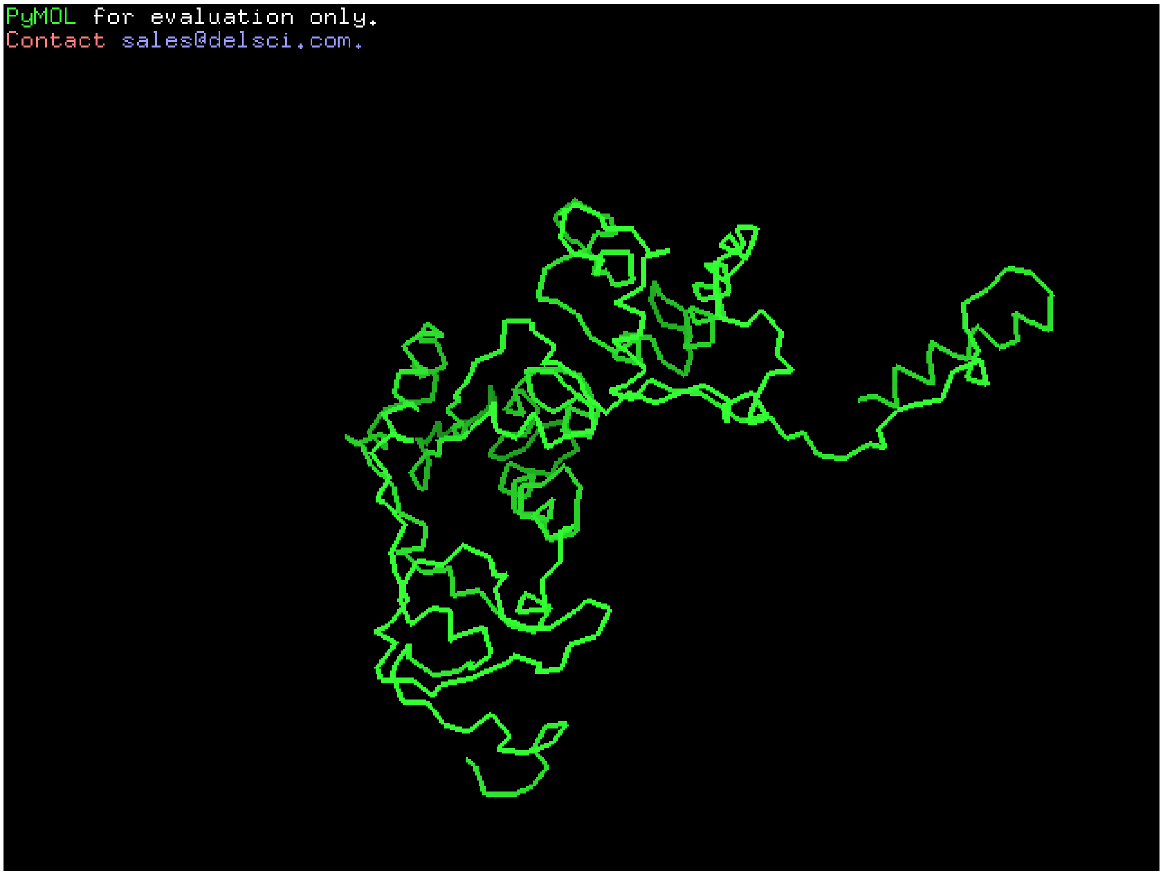


Model 2-

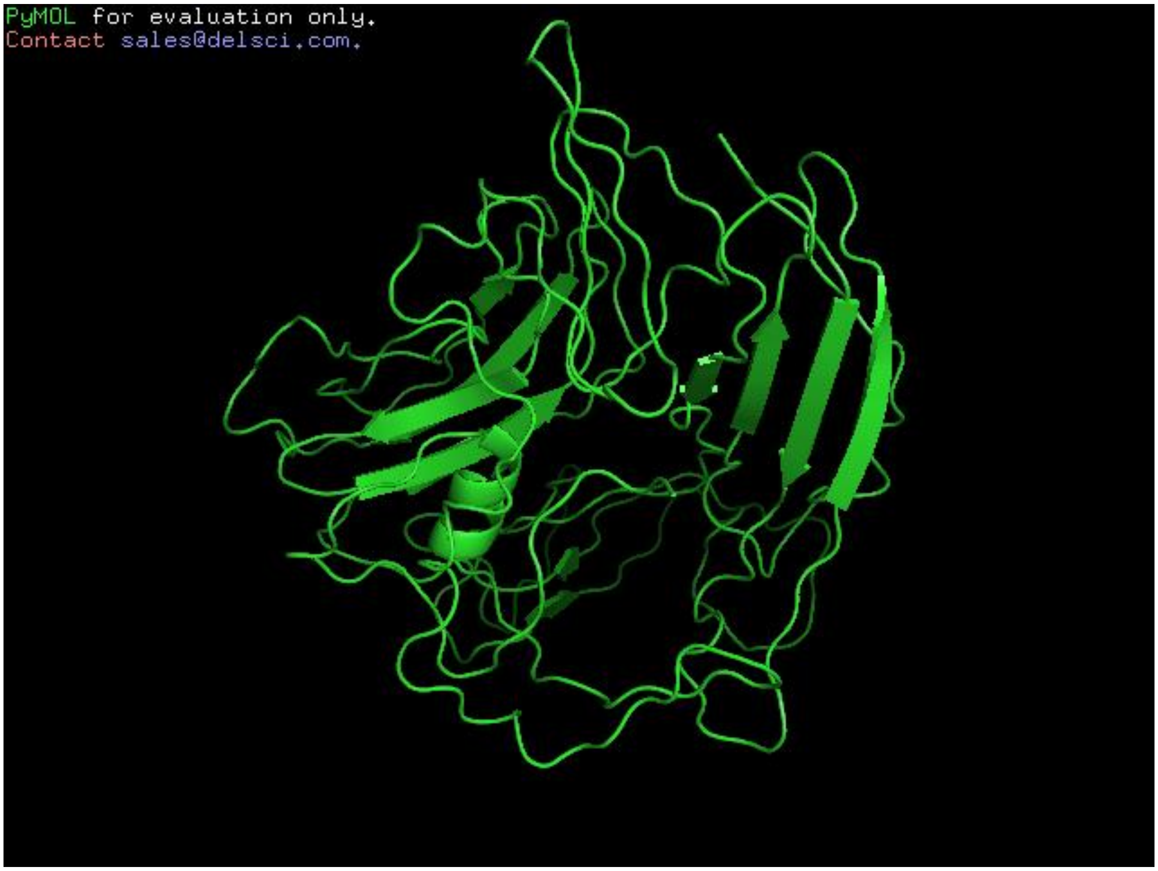


PyMOL for evaluation only.
Contact sales@delsci.com.

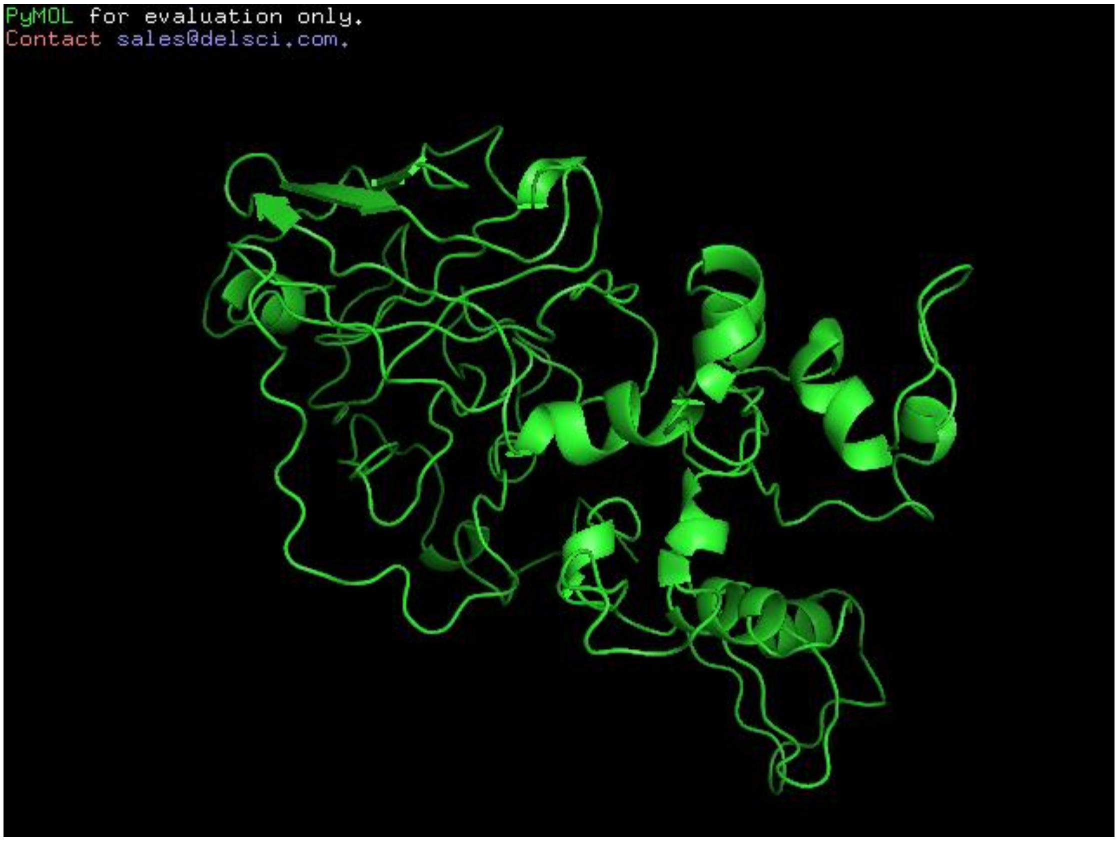
Model 3-



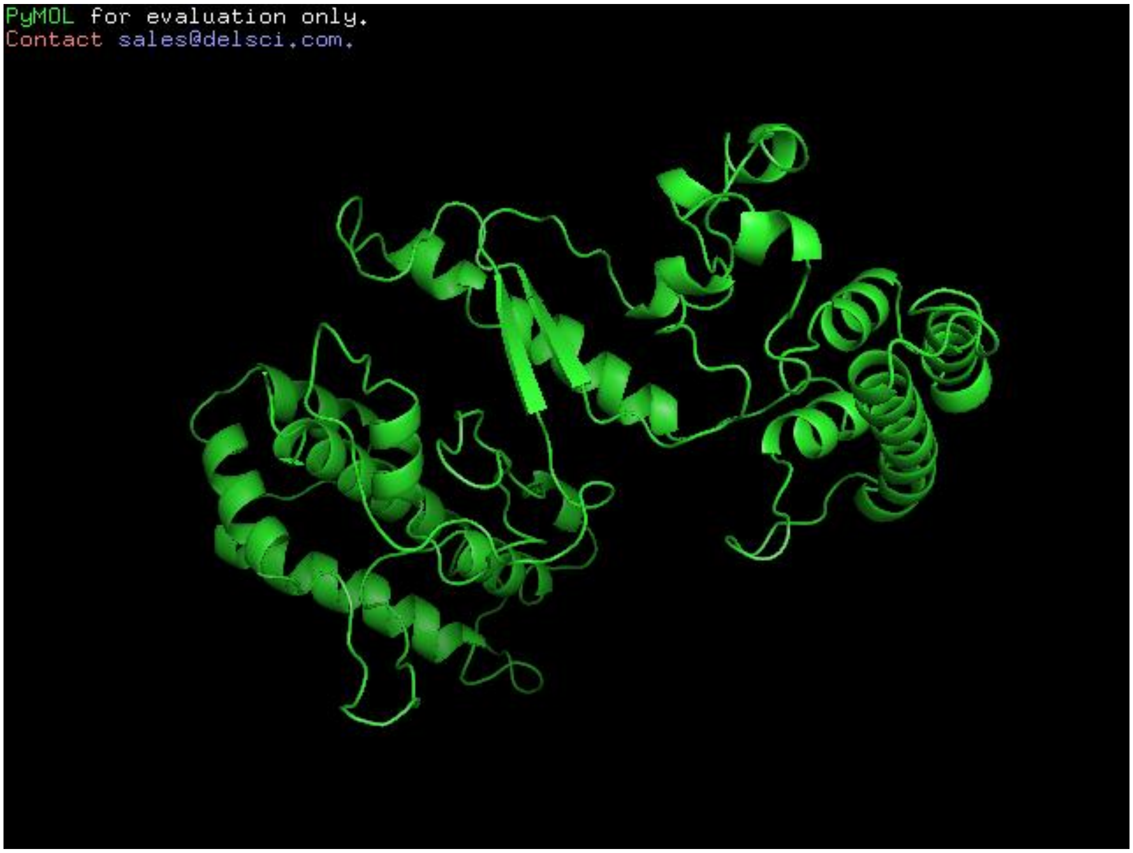
Model 4-



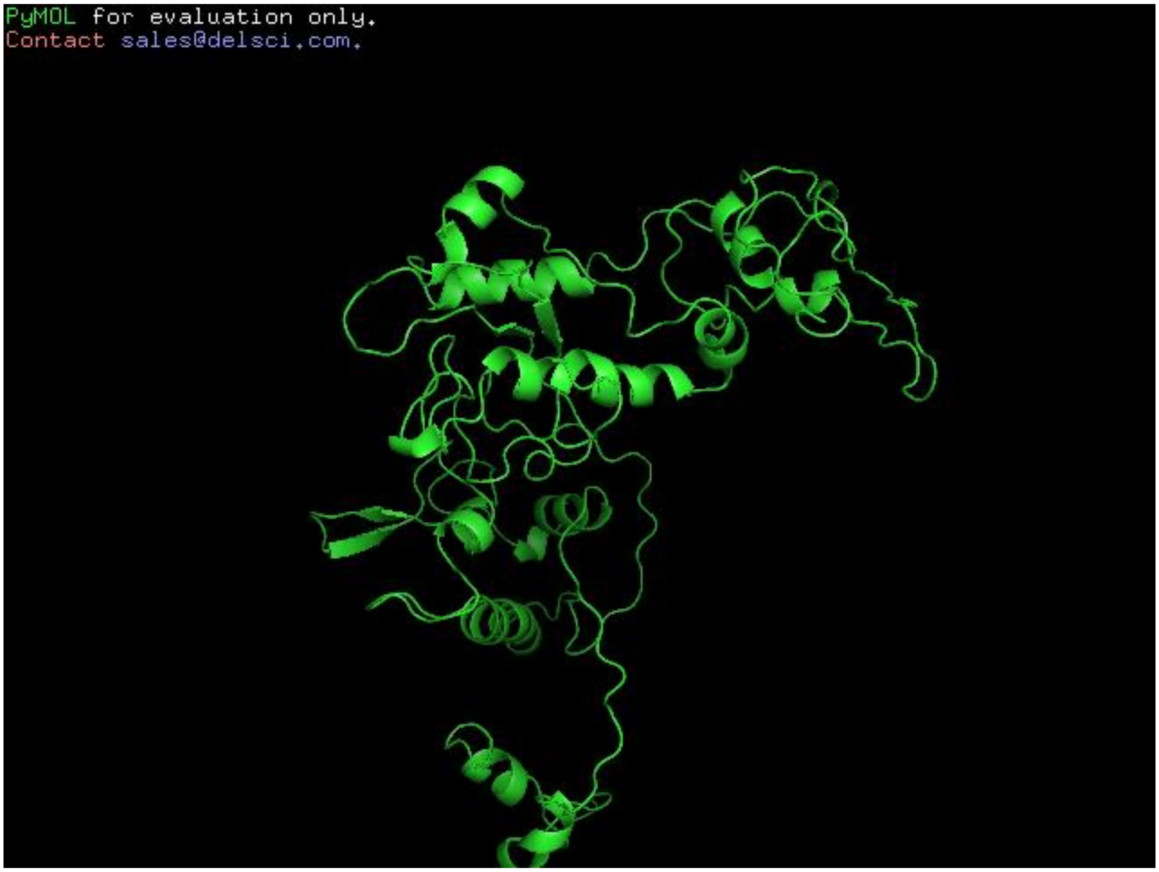
Model 5-



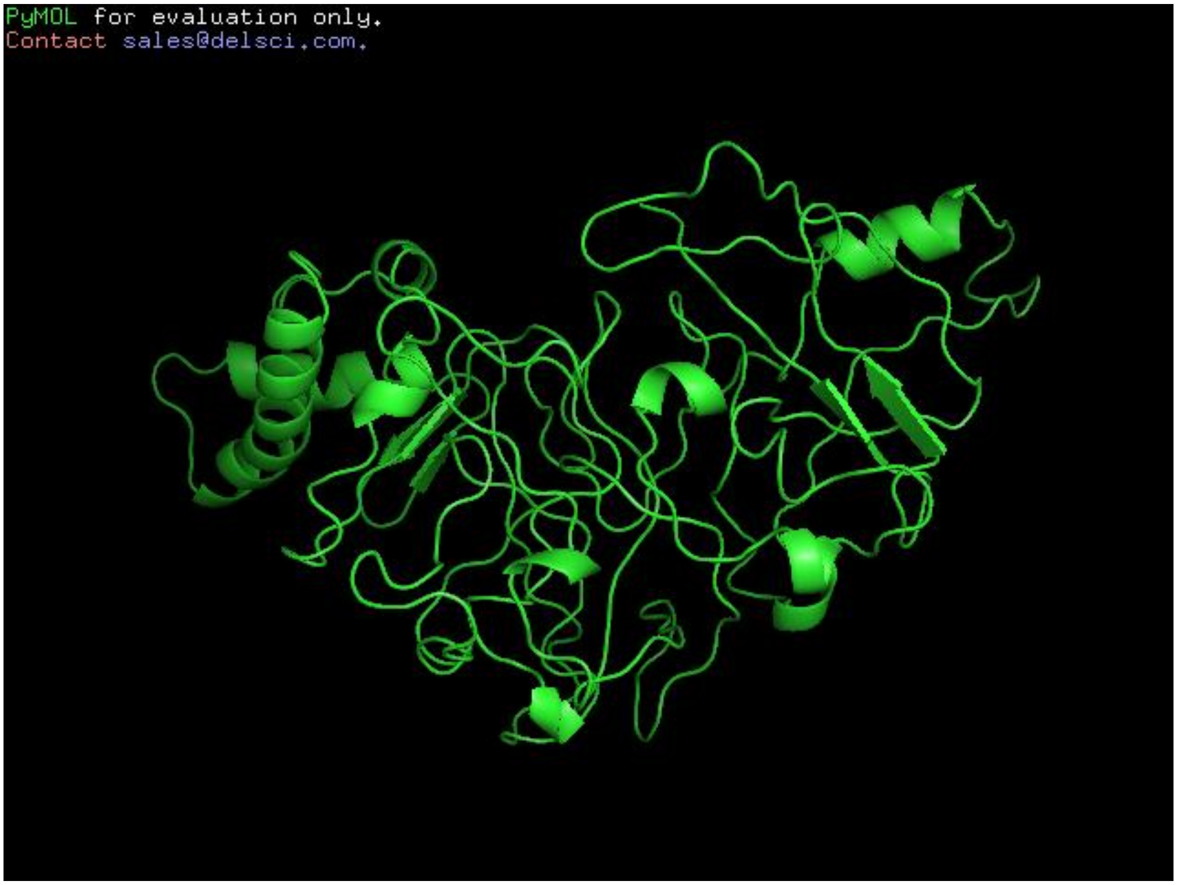
Model 6-



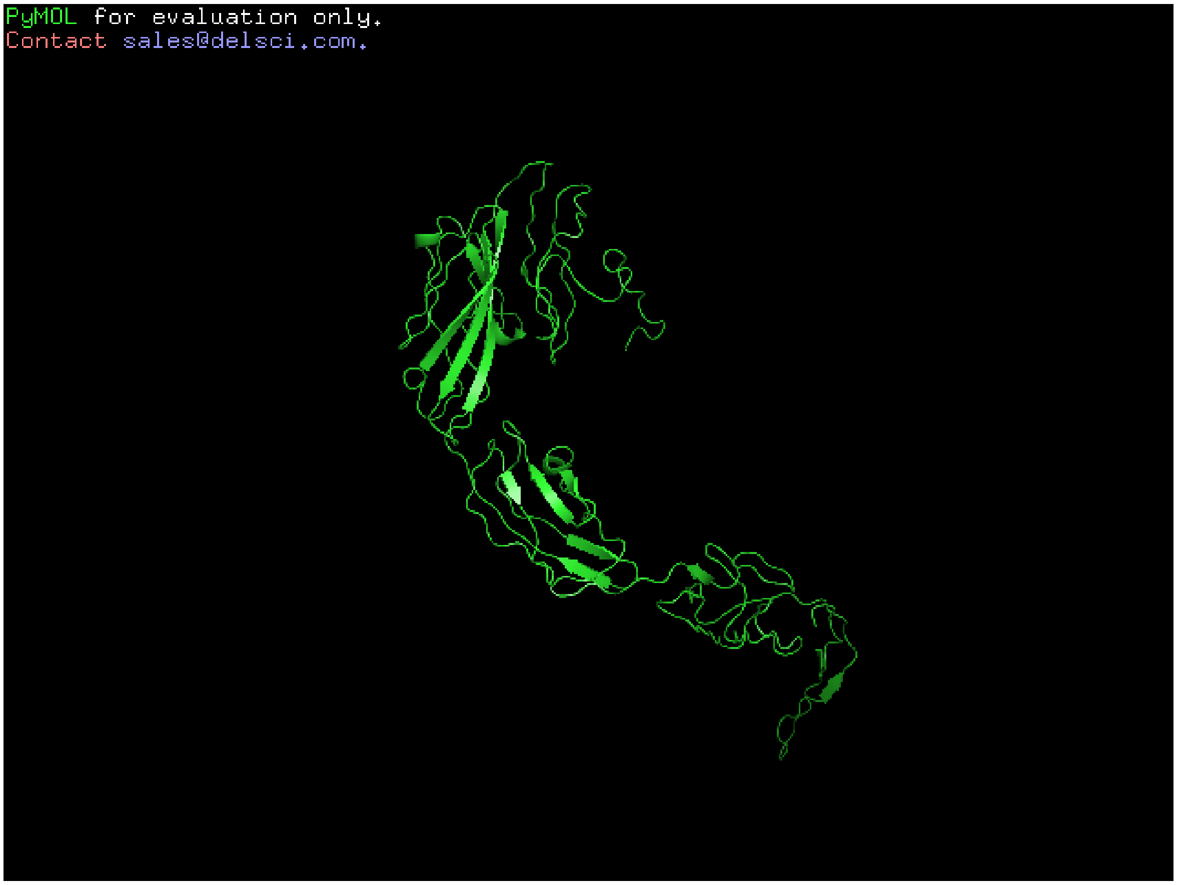
Model 7-



Model 8-



Model 9-

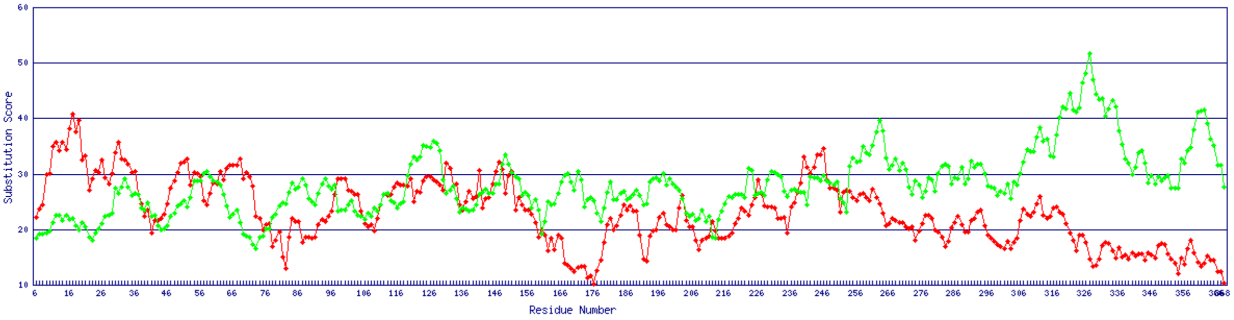


These models have been generated by MUSTER or Multi Source Threader which uses secondary structure information by predicting the same by itself.

Validation of 3D models- using Harmony Structure Validation Server (Cited by- Dr. R. Sowdhamini et al. ns@mbu.iisc.ernet.in)

Model 1-

Detection of local error-



Green= query sequence

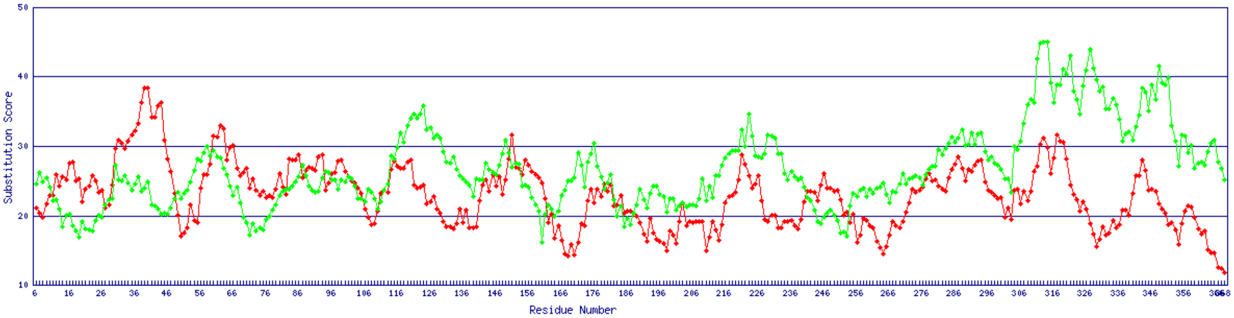
Red= reverse sequence

Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1465.00012	9462.47461

Model 2-

Detection of local error-



Green= query sequence

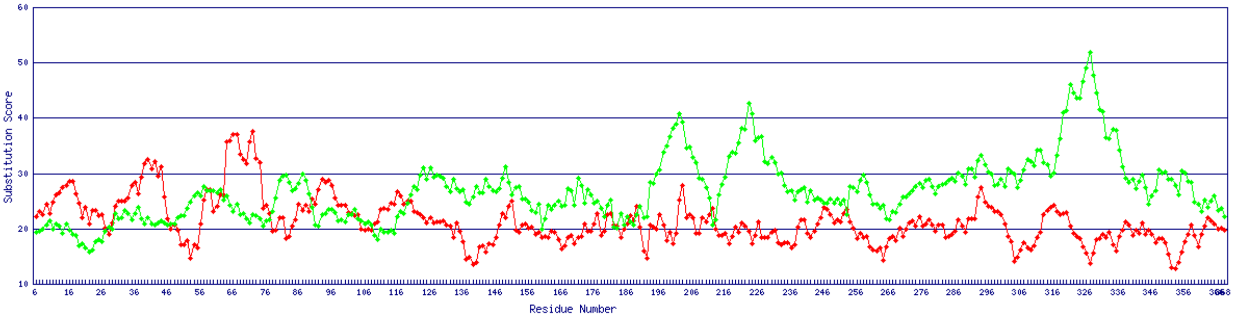
Red= reverse sequence

Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1374.59949	9018.4248

Model 3-

Detection of Local error-

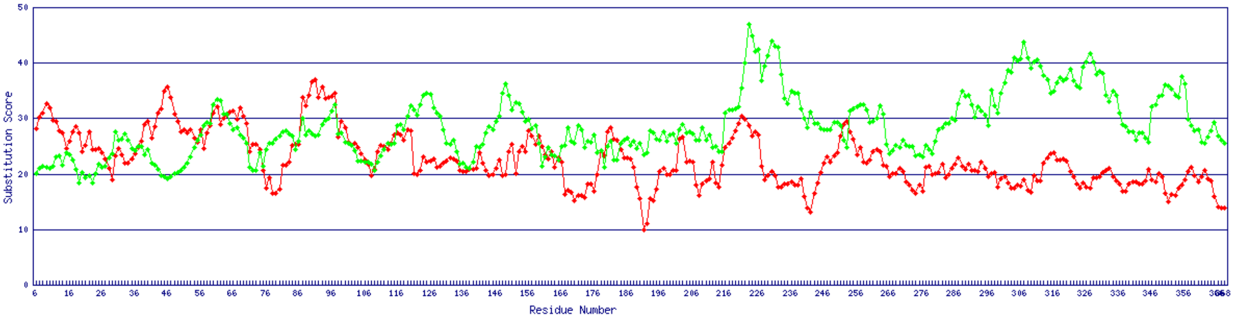


Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1314.39954	9086.55762

Model 4-

Detection of local error-

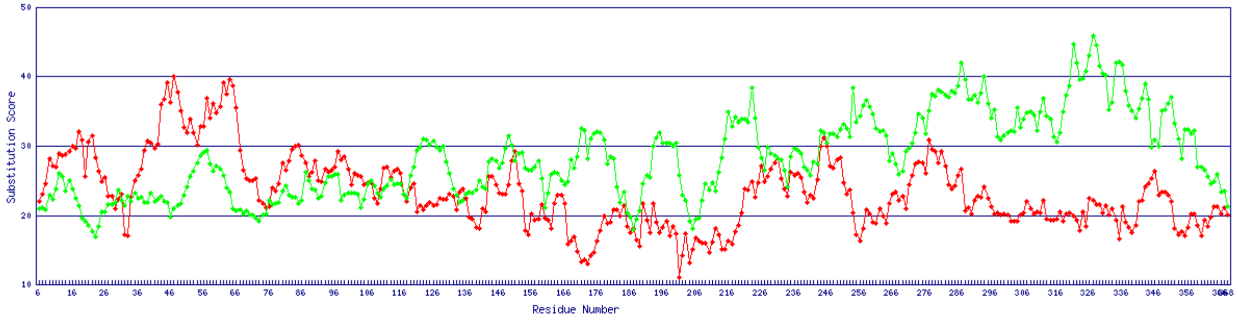


Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1458.7002	9656.08789

Model 5-

Detection of local error-

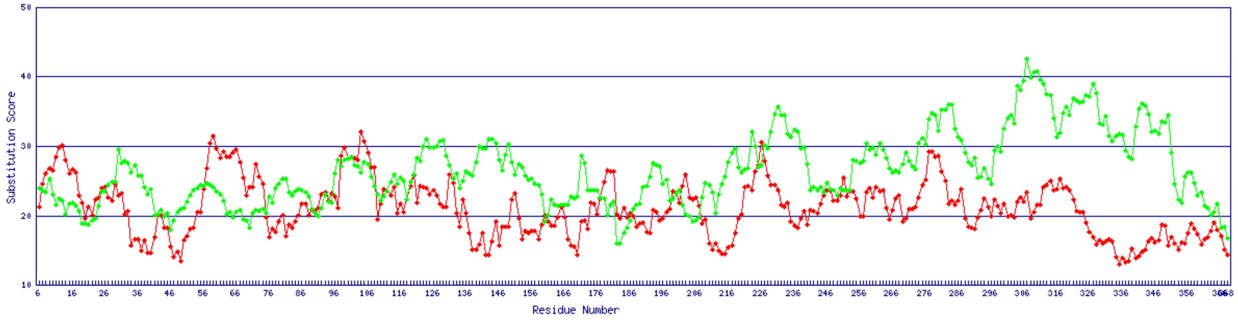


Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1450.40076	9520.90233

Model 6-

Detection of local error-

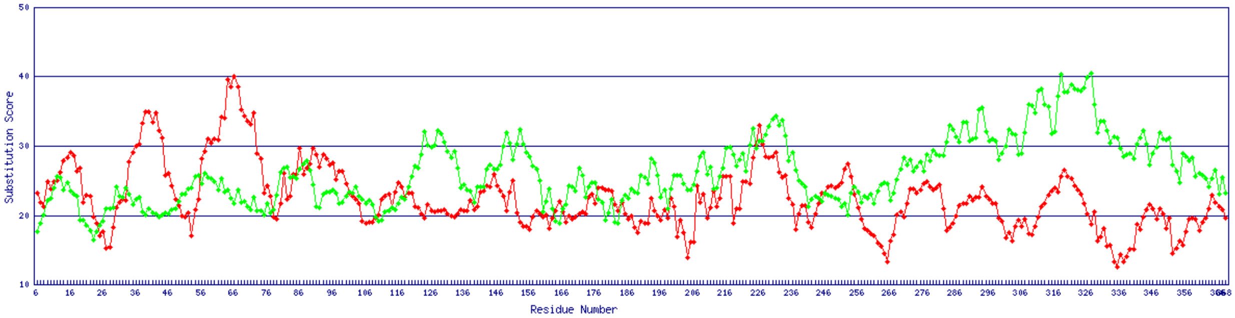


Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1327.89978	8943.4502

Model 7-

Detection of local error-

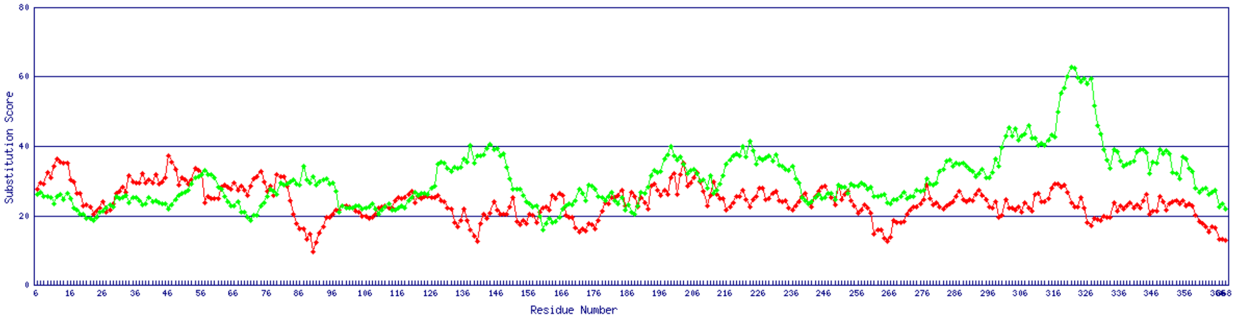


Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1332.7002	8800.14551

Model 8-

Detection of local error-

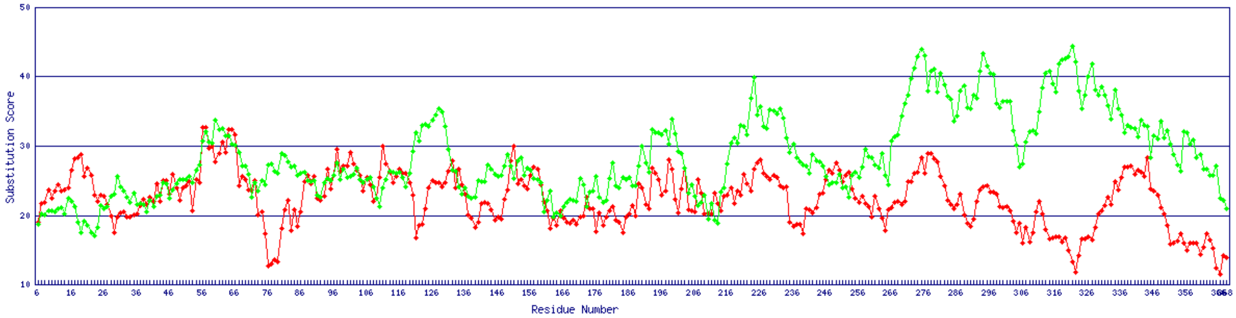


Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1585.79993	10226.4795

Model 9-

Detection of local error-



Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1474.49963	9602.375

Harmony is a server to assess the compatibility of an amino acid sequence with a proposed three-dimensional structure. Structural descriptors such as backbone conformation, solvent accessibility and hydrogen bonding are used to characterize the structural environment of each residue position. Propensity and Substitution values are used together to predict the occurrence of an amino acid at each position in the sequence on the basis of the local structural environment.

Top model generated using Phyre2 server-

Top template information-

PDB header- Nuclear Protein

Chain A: PDB molecule microprocessor complex subunit dgcr8

Confidence and coverage= 3.4% and 4%

15 residues have been modeled with 4% confidence.

Final model generated-

PyMOL for evaluation only.
Contact sales@delsci.com.

Model dimensions (Å): **X**:15.957 **Y**:25.304 **Z**:21.677

COMPARATIVE MODELING-

The process of building a comparative model is conceptually straightforward. First, an alignment is performed between the sequence for which the structure has been determined by experimental methods (the parent) with the sequence to be modeled (the target). This sequence alignment is used to construct an initial model (sometimes referred to as a framework or template) by copying over some main chain and side chain coordinates from the parent structure based on the equivalent residue in the sequence alignment. Side chains must be built for residues in the target that does not correspond to an identity in the alignment, and for residues where the side chain conformation is thought to vary in the target relative to the parent structure. Main chains must be built in the case of insertions, regions surrounding a deletion, and in other regions of suspected main chain variation.

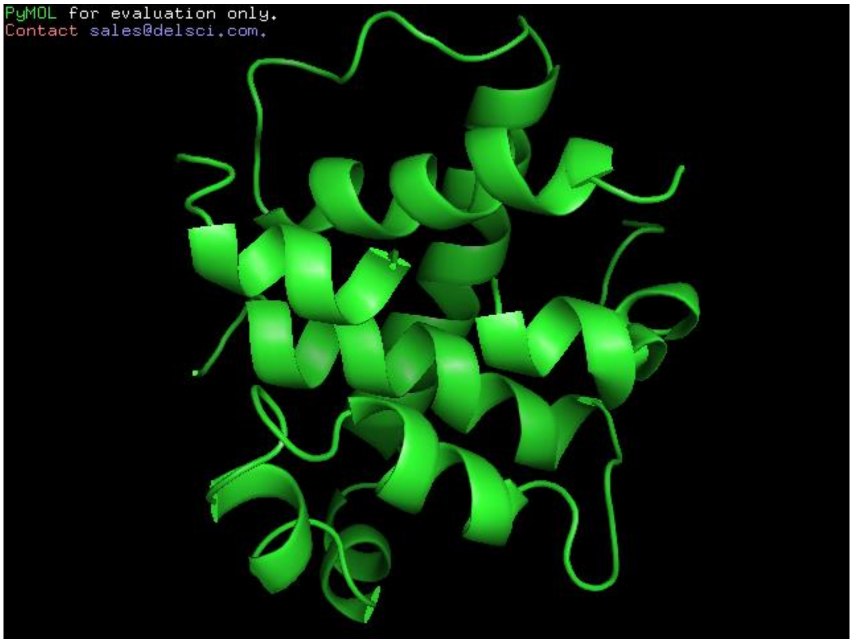
It can be done using 3D Jigsaw protein comparative modeling server, PSIpred and CPH model server, ATOME2, FUGUE.

In this work FUGUE has been used for comparative modeling using FASTA sequence.

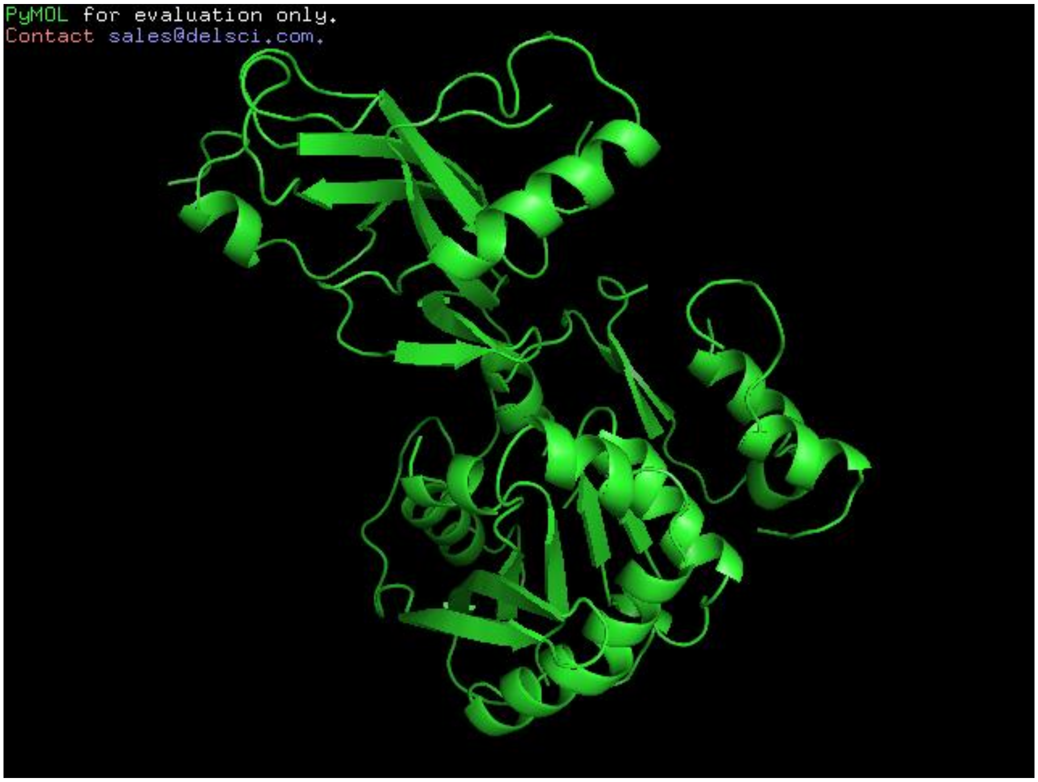
Output-

Profile hit	PLEN	RAWS	RVN	Z- score	ZORI	AL
Hs2aboa	131	-268	38	2.82	4.16	02
Hs2b20a	384	-458	96	2.52	5.16	00
Hs2od0a	100	-324	14	2.26	5.03	02
Hs1w8xp	79	-279	26	2.13	3.60	02
Hs3doha	374	-464	94	2.02	4.87	00
Hs1m06j	24	-328	8	2.00	3.34	02
Hsd1gba_1	105	-298	51	1.96	4.57	02
Hsd1bia_1	63	-309	29	1.96	4.49	02
Hs2rf4b	61	-303	5	1.93	4.13	02
Hs4gdkc	34	-337	10	1.92	3.88	02

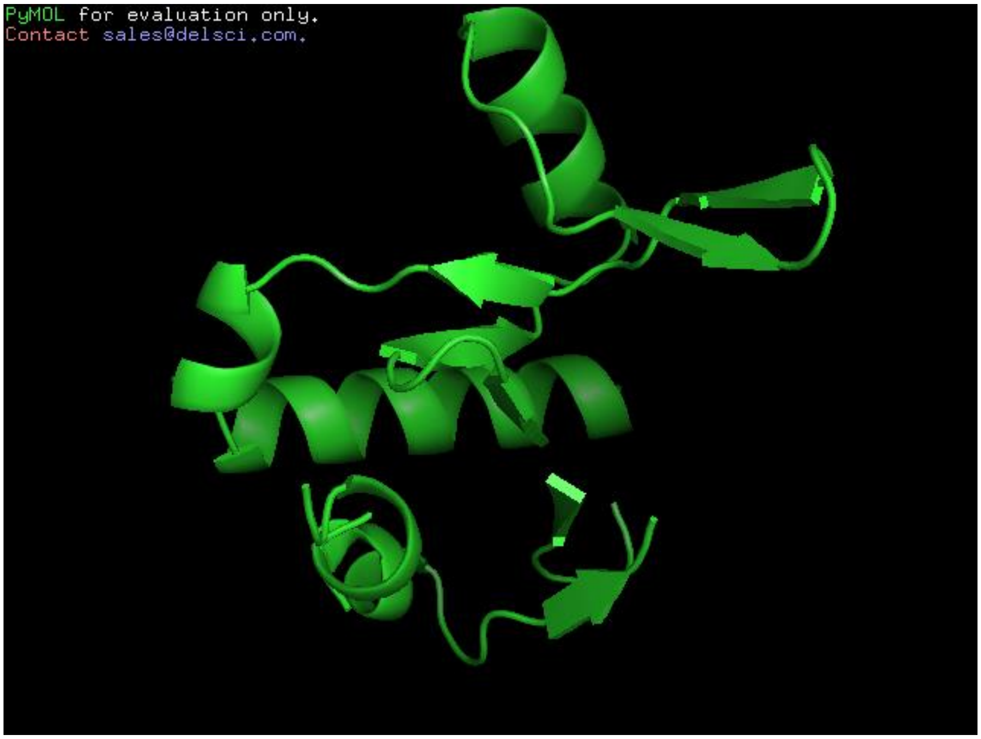
Model of structural homolog 1 generated on above score-



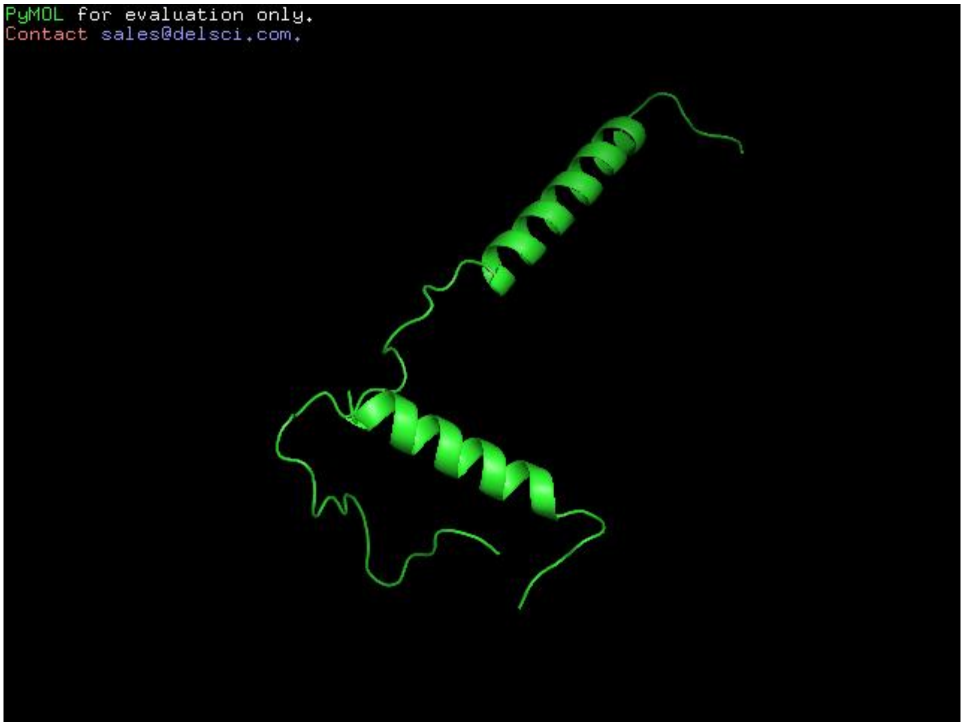
Model of structural homolog 2 generated on the above score-



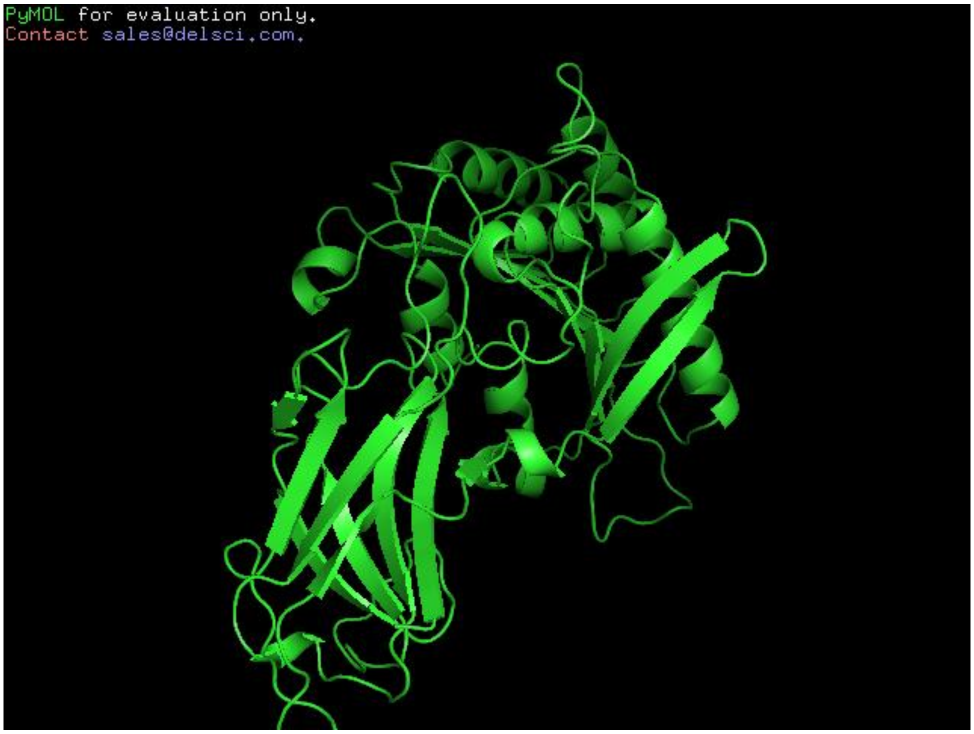
Model of structural homolog 3 generated on the above score-



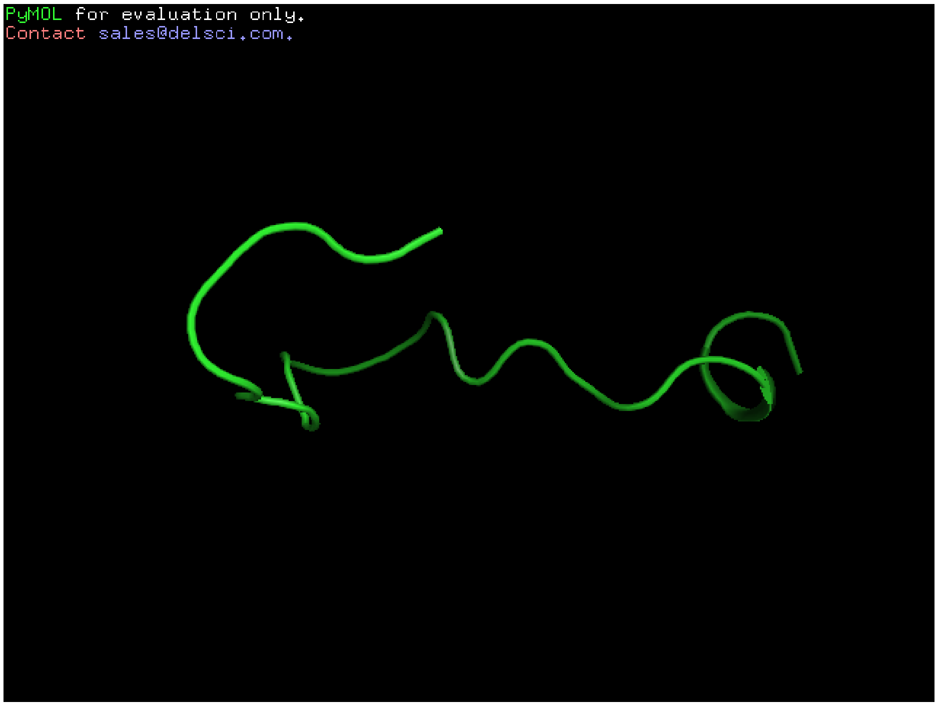
Model of structural homolog 4 generated on the above score-



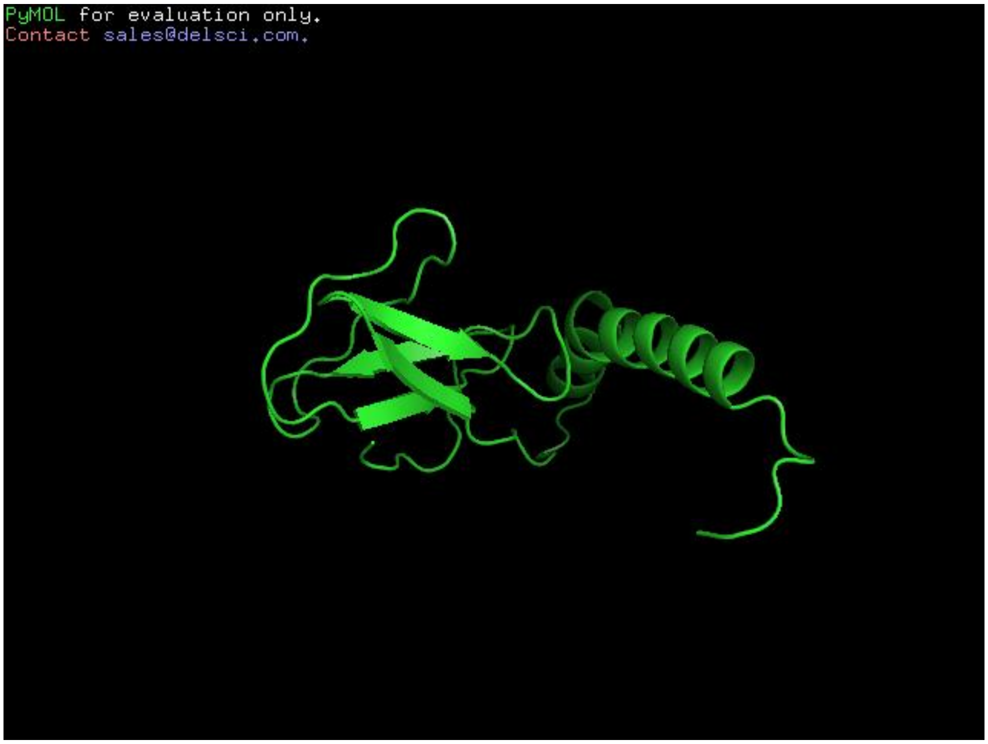
Model for structural homolog 5 generated on the above score-



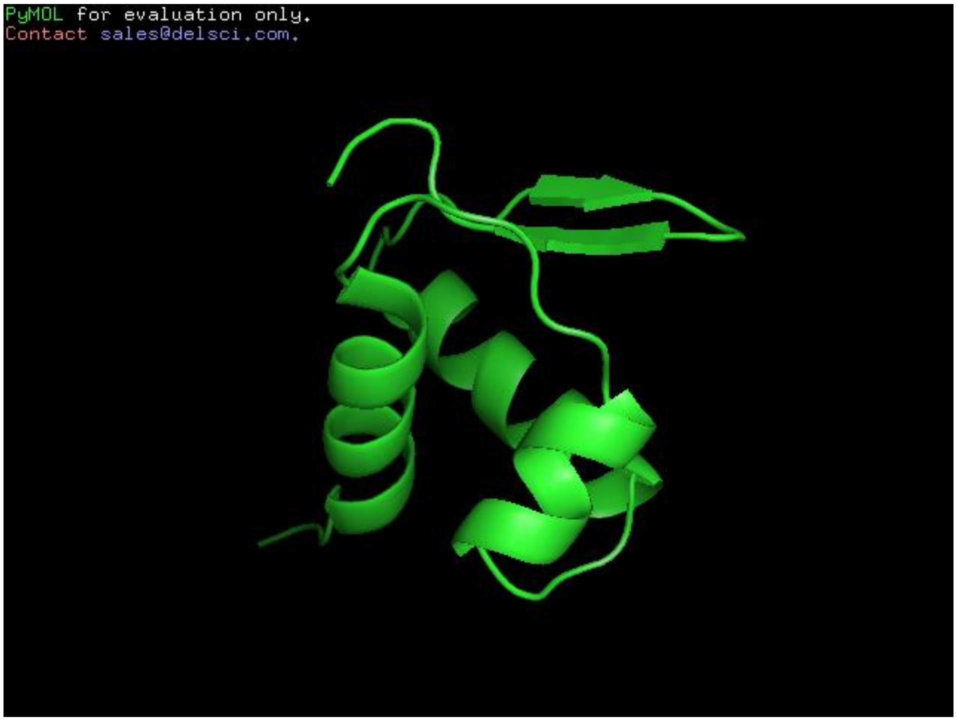
Model for structural homolog 6 generated on the above score-



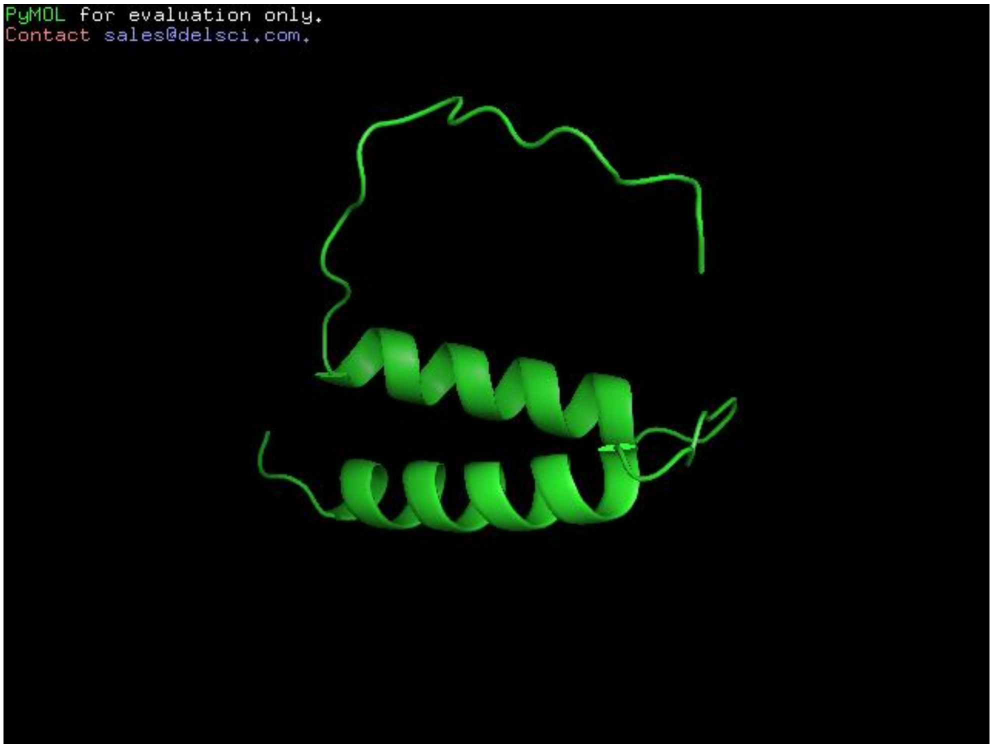
Model for structural homolog 7 generated on the above score-



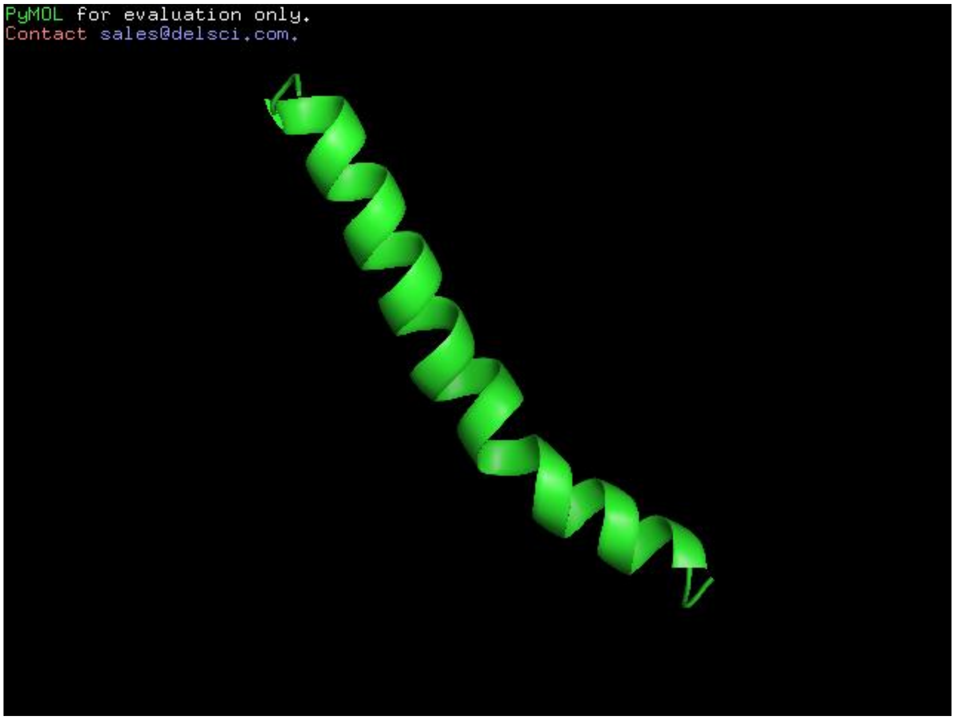
Model for structural homolog 8 generated on the above score-



Model for structural homolog 9 generated on the above score-



Model of structural homolog 10 generated on the above score-



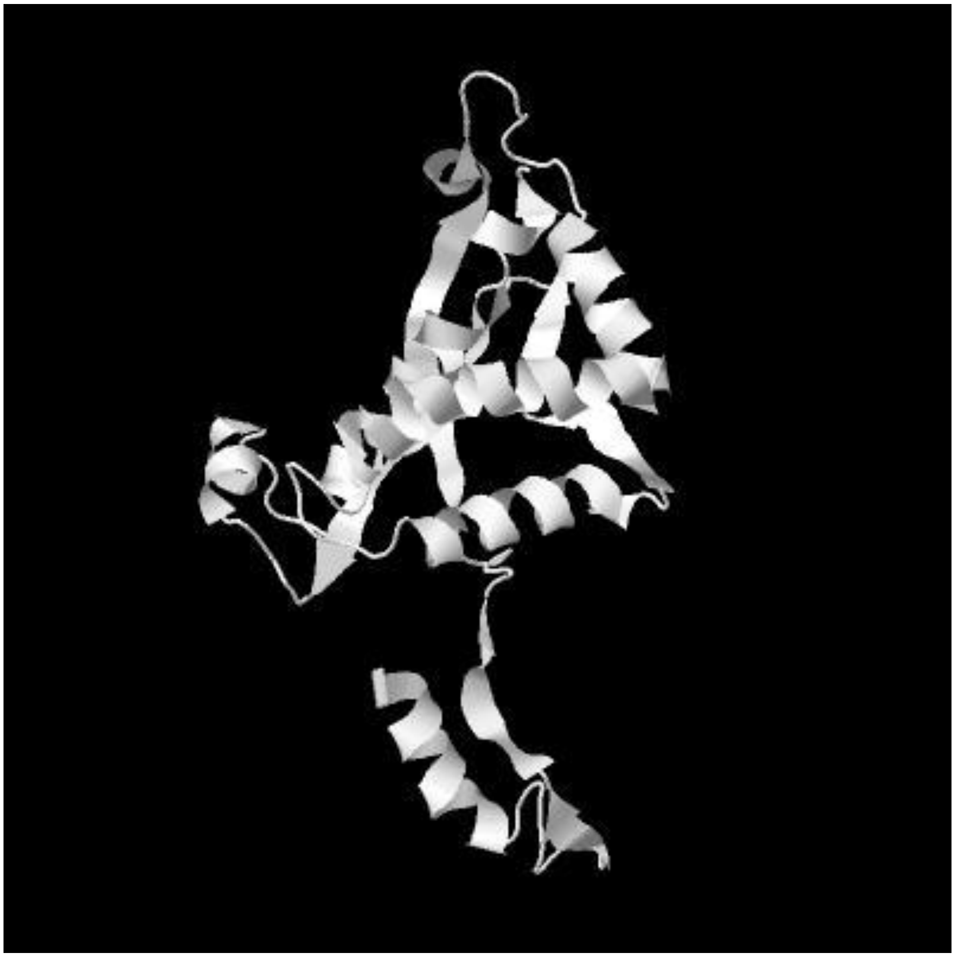
Comparative modeling using CPH model server-

On profile- profile alignment

Score= 31.0 bits

Identity= 42.6%

Model built-



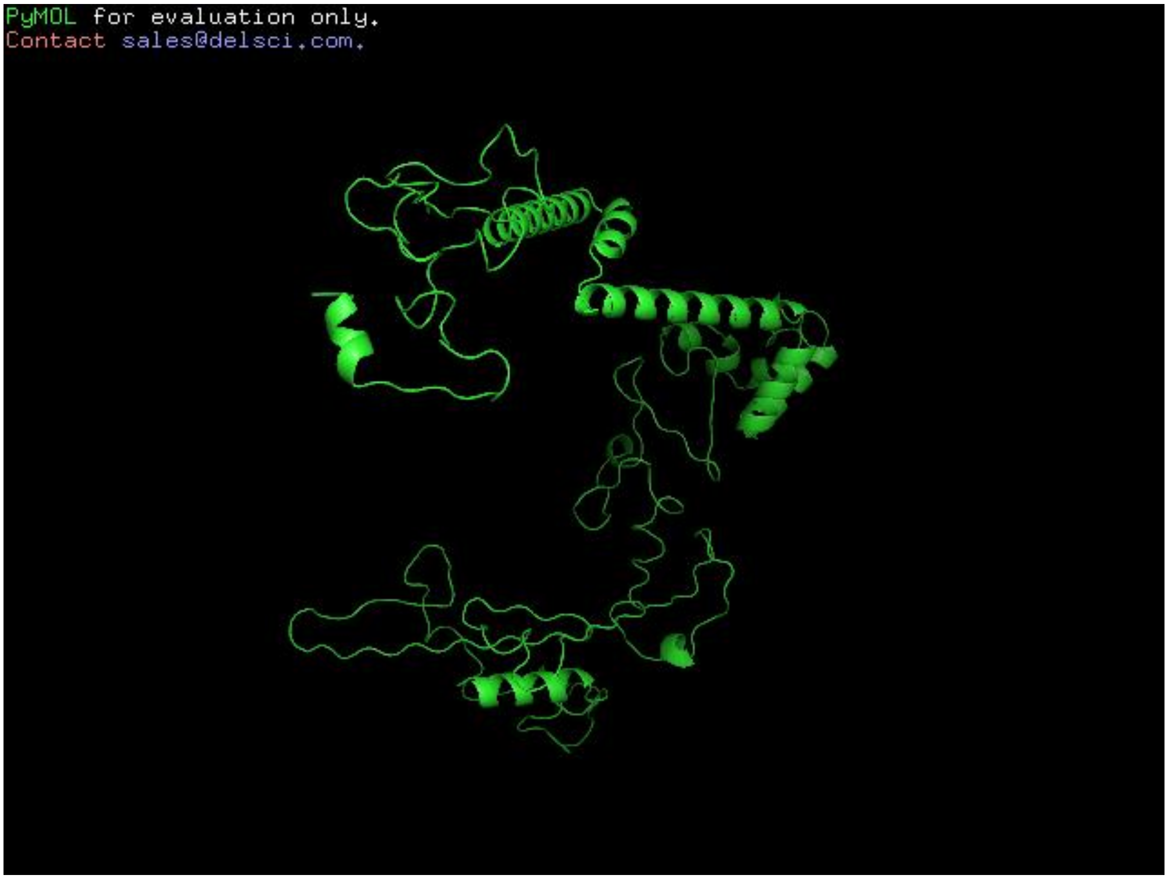
Ab- Initio Prediction of Protein structure-

Some methods for Ab initio prediction include Molecular Dynamics (MD) simulations of proteins and protein-substrate complexes provide a detailed and dynamic picture of the nature of inter-atomic interactions with regards to protein structure and function; Monte Carlo (MC) simulations that do not use forces but rather compare energies, via the use of Boltzmann probabilities; Genetic Algorithms which tries to improve on the sampling and the convergence of MC approaches, and exhaustive and semi-exhaustive lattice-based studies which are based on using a crude/approximate fold representation (such as two residues per lattice point) and then exploring all or large amounts of conformational space given the crude representation.

This can be done by a number of tools like 3D Jigsaw, Scratch etc.

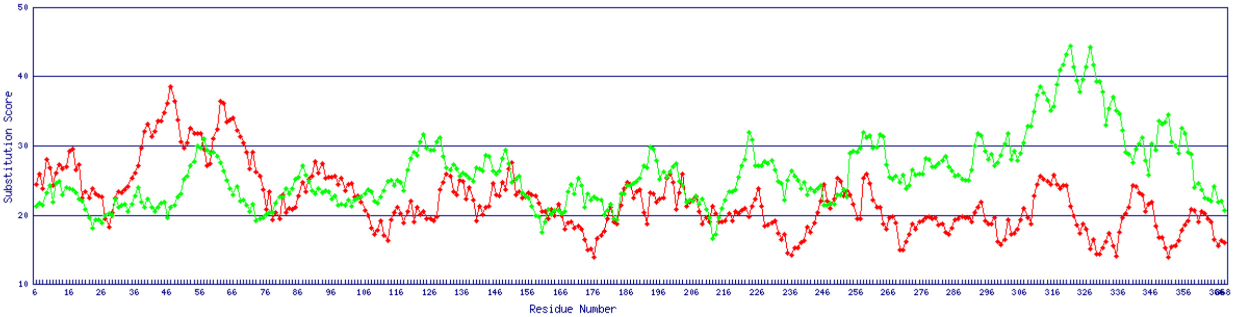
In this work Scratch protein predictor has been used.

Model built using this tool-



Validation of this model using HARMONY-

Detection of local error-



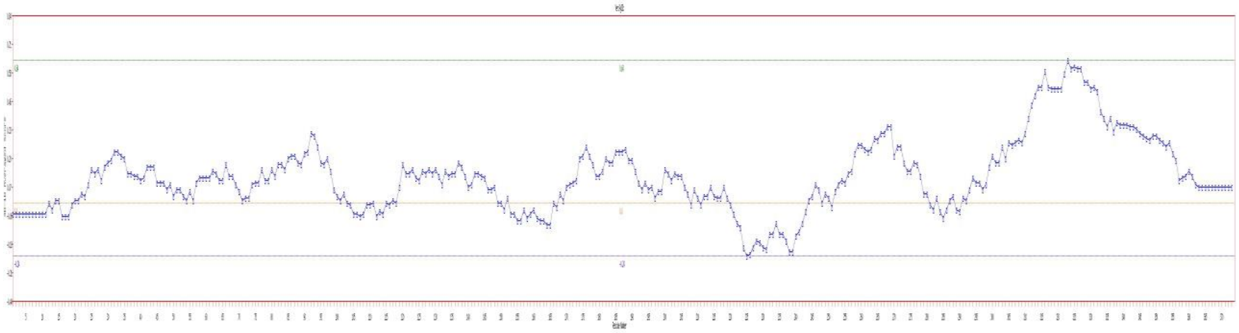
Green – query sequence

Red – reverse sequence

Total propensity and substitution score-

No of Residues	Propensity Score	Substitution Score
368	1342.09961	8771.15332

Validation using **Verify3D server-**



In the above plot, the vertical axis represents the average 3D-1D profile score for residues in a 21-residue sliding window, the center of which is at the sequence position indicated by the horizontal axis. Scores for the first 9 and the final 9 sequence positions have no meaning. A window length of 21 residues strikes a useful balance between smoothing fluctuations and localizing the error.

3D JIGSAW tool- Pfamily hit found is **Q4R4G9_MACFA which is a brain cDNA** clone and is similar to CD34 antigen. No confident structural templates of above protein have been found in PDB. It is an **unnamed protein product**. Its function has been reported in literature to be in cell adhesion and when **aligned by ClustalO it showed 35% similarity to CD34** on which basis CD34 can be proved to be an adhesion molecule.

Output-

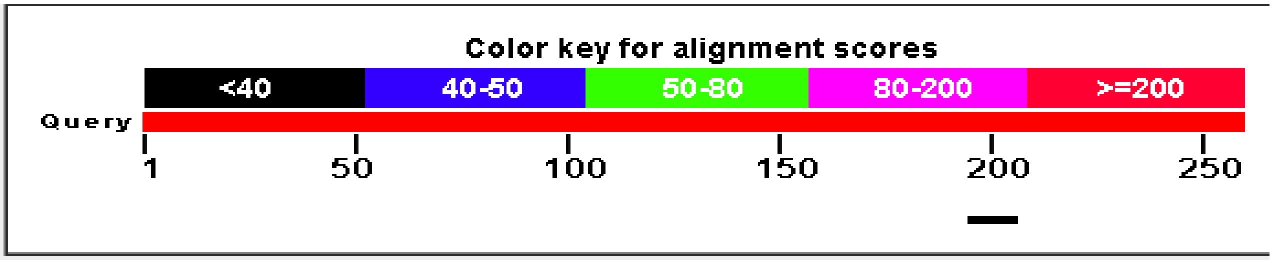
```
1 ------------MPRGWTALCLLSLLPSGFMSLDNNGTATPELPTQGTFSNVSTNVSYQE 48 gi|180109|gb|AAA03181.1|
1 MLVRRGARAGPGMPRGWTALCLLSLLPSGFTSANDTSTVTPKSSTQGTFSTVSTNVSYQE 60 Q4R4G9 Q4R4G9_MACFA
*: ******** ************************* ******.************:**
49 TTTPSTLGSTSLHPVSQHGNEATTNITETTVKFTSTSVITSVYGNTNSSVQSQTSVISTV 108 gi|180109|gb|AAA03181.1|
61 TAIPSTLGSTSPHPVSQHGNEATTNITETTVKFTSTSGITSVYGTTNSSVQSQTSVITTV 120 Q4R4G9 Q4R4G9_MACFA
******:******** **********************.***** **
109 FTTPANVSTPETTLKPSLSPGNVSDLSTTSTSLATSPTKPYTSSSPILSDIKAEIKCSGI 168 gi|180109|gb|AAA03181.1|
121 FTTPANISTPETTLKSSLSPGNVSDLSTTSTSLATSPTDPYTSSPPIP------------ 168 Q4R4G9 Q4R4G9_MACFA
169 REVKLTQGICLEQNKTSSCAEFKKDRGEGLARVLCGEEQADADAGAQVCSLLLAQSEVRP 228 gi|180109|gb|AAA03181.1|
169 ------------------------------------------------------------ 168 Q4R4G9 Q4R4G9_MACFA
229 QCLLLVLANRTEISSKLQLMKKHQSDLKKLGILDFTEQDVASHQSYSQKTLIALVTSGAL 288 gi|180109|gb|AAA03181.1|
169 ------------------------------------------------------------ 168 Q4R4G9 Q4R4G9_MACFA
289 LAVLGITGYFLMNRRSWSPTGERLGEDPYYTENGGGQGYSSGPGTSPEAQGKASVNRGAQ 348 gi|180109|gb|AAA03181.1|
169 ------------------------------------------------------------ 168 Q4R4G9 Q4R4G9_MACFA
349 KNGTGQATSRNGHSARQHVVADTEL 373 gi|180109|gb|AAA03181.1|
169 ------------------------- 168 Q4R4G9 Q4R4G9_MACFA
```

Identical positions	134
Identity	34.805%
Similar positions	12

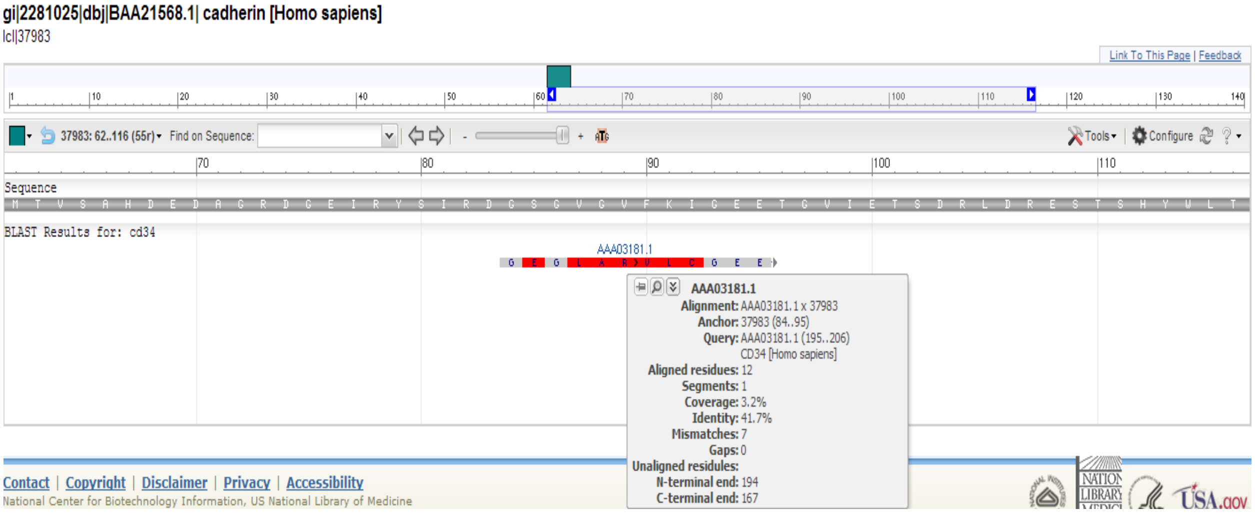
D. OBJECTIVE- To prove that CD34 is an adhesion molecule.

- **Perform PSI-BLAST of CD34 with adhesion molecules known**

a. CD34 with CADHERIN (CADHERIN is 140 amino acid containing adhesion molecule having accession no.- BAA21568.1



The alignment score was found to be around 80-100 which is quite significant but still not appropriate to prove CD34 as an adhesion molecule.


gi|2281025|dbj|BAA21568.1| cadherin [Homo sapiens]
lcl|37983

Link To This Page | Feedback
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
37983: 62..116 (55r)
Find on Sequence:
Tools
Configure
70
80
90
100
110
Sequence
M T V S A H D E D A G R D G E I R Y S I R D G S G V G V F K I G E E T G V I E T S D R L D R E S T S H Y V L T
BLAST Results for: cd34
AAA03181.1
G E G L A R V L C G E E
Contact | Copyright | Disclaimer | Privacy | Accessibility
National Center for Biotechnology Information, US National Library of Medicine
USA.gov

AAA03181.1
Alignment: AAA03181.1 x 37983
Anchor: 37983 (84..95)
Query: AAA03181.1 (195..206)
CD34 [Homo sapiens]
Aligned residues: 12
Segments: 1
Coverage: 3.2%
Identity: 41.7%
Mismatches: 7
Gaps: 0
Unaligned residues:
N-terminal end: 194
C-terminal end: 167

For fragment 2 (260-300 amino acids) - for this particular fragment no significant similarity was found.

For fragment 3 (301-373 amino acids) - for this particular fragment no significant similarity was found.

Hence, CD34 cannot be proved to be an adhesion molecule.

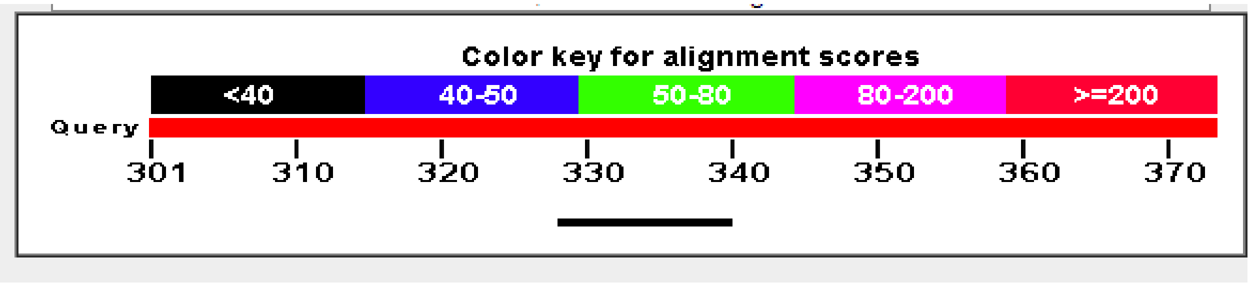
- **CD34 with PECAM-1(PECAM-1 is platelet endothelial cellular adhesion molecule or CD31 having accession no. AAB28645.1**

For fragment 1 (1-259 amino acids)

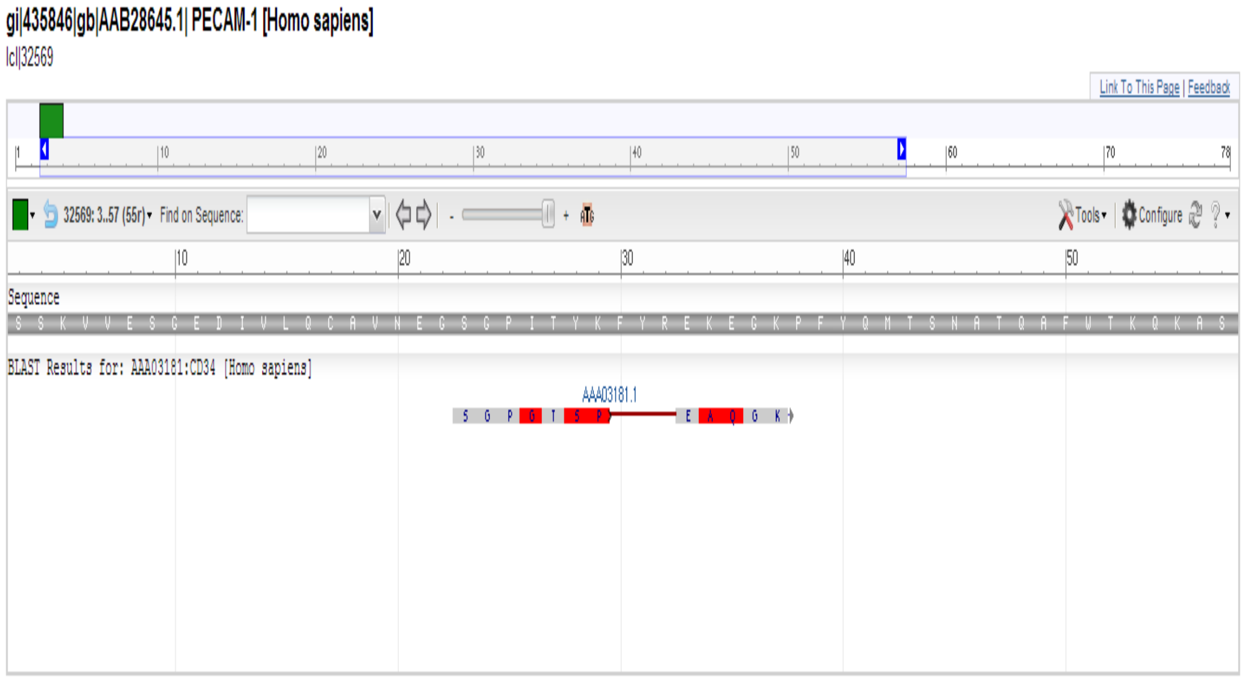
No significant similarity found in this range.

For fragment 2 (260-300 amino acids)

No significant similarity found in this range.



Alignment score was found to be around 58 which is quite significant to prove that CD34 is an adhesion molecule.



gi|435846|gb|AAB28645.1| PECAM-1 [Homo sapiens]

lcl|32569

32569: 3..57 (55r)

Find on Sequence:

Sequence

S S K V V E S G E D I V L Q C A V N E G S G P I T Y K F Y R E K E G K P F Y Q M T S N A T Q A F V T K Q K A S

BLAST Results for: AAA03181:CD34 [Homo sapiens]

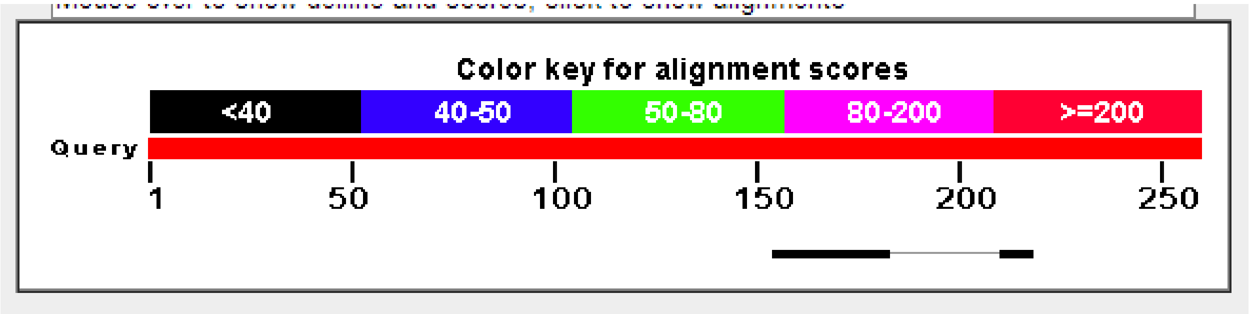
AAA03181.1

S G P G T S P E A Q G K

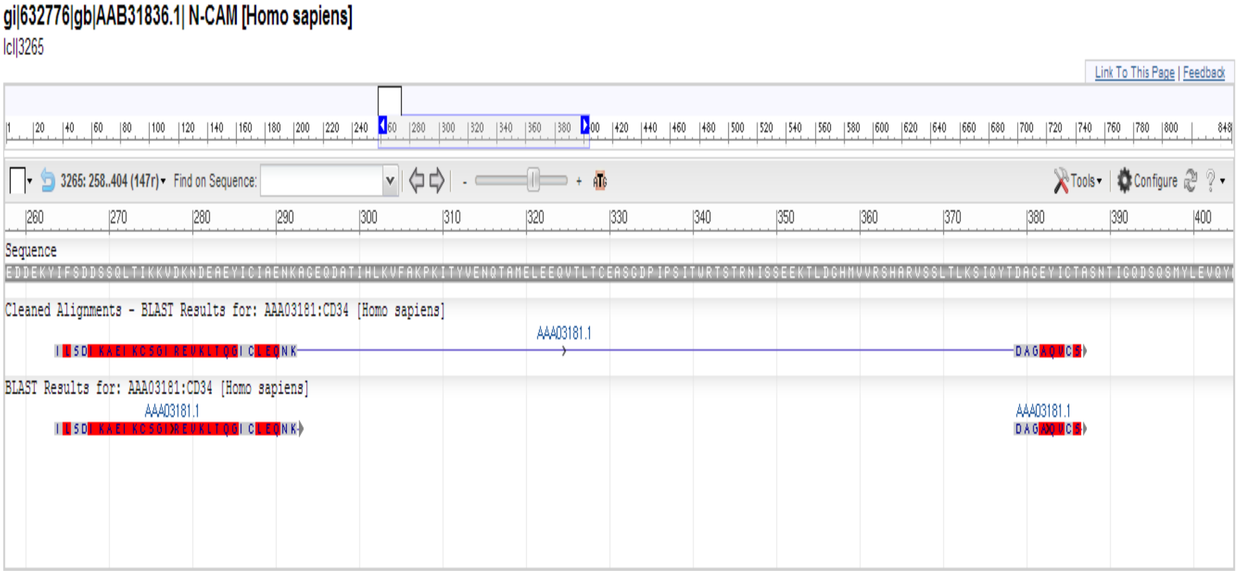
Hence, CD34 can be proved to be an adhesion molecule.

- ➢ **CD34 with N-CAM** (N-CAM is Neural cell adhesion molecule or CD56 HAVING ACCESSION NO. AAB31836.1)

  For fragment 1 (1-259 amino acids)


Mouse over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
50
100
150
200
250

Alignment score was found to be around 110 for residue range 155-183 & >200 for residue range 211-218 which is very good score to prove CD34 as an adhesion molecule.



Sequence

EDDEKYIFSDDSSQLTIKKVDKNDEAEYICIAENKAGEQDATIHLKVFAKPKITYVENQTAMELEEQVTLTCEASGDPIPSITWRTSTRNISSEEKTLDGHMVVRSHARVSSLTLKSIQYTDAGEYICTASNTIGQDSQSMYLEVQY

Cleaned Alignments - BLAST Results for: AAA03181:CD34 [Homo sapiens]

ILSDIKAEIKCSGIREVKLTQGICLEQNK DAGAQVCS

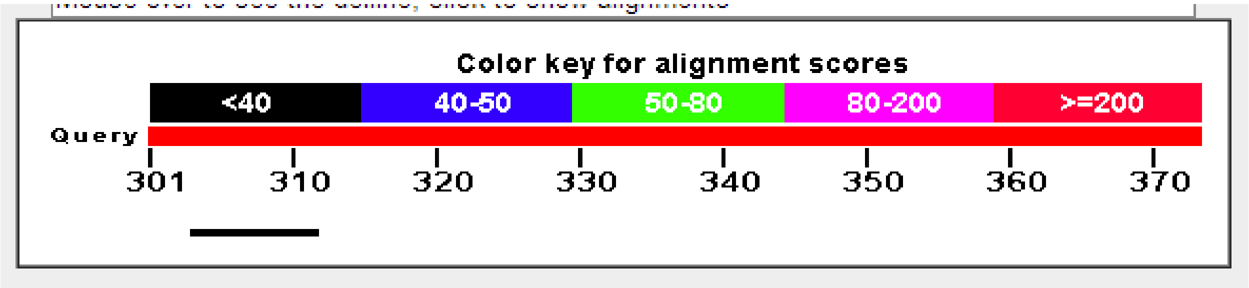
BLAST Results for: AAA03181:CD34 [Homo sapiens]

ILSDIKAEIKCSGIREVKLTQGICLEQNK DAGAQVCS

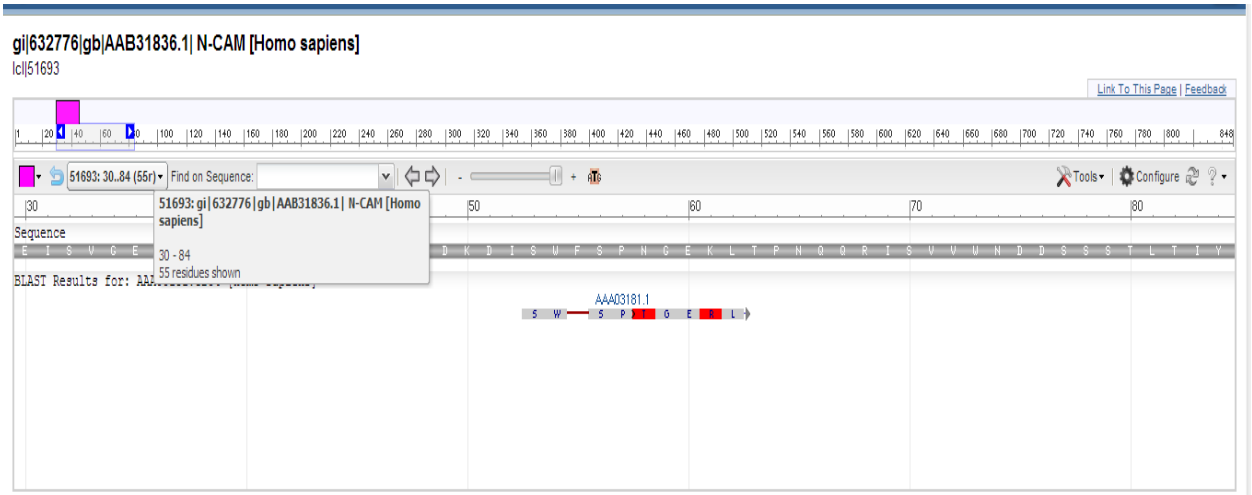
For fragment 2 (260-300 amino acids)

No significant similarity found in this range

For fragment 3 (301-373 amino acids)


Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
301
310
320
330
340
350
360
370

Alignment score was found to be <40 which is insignificant.



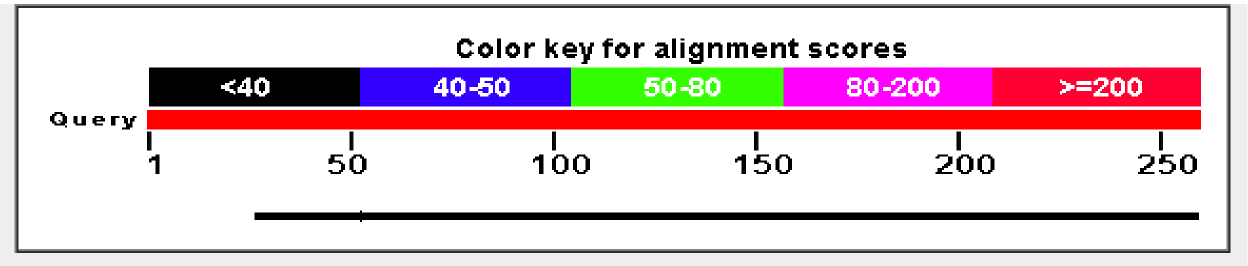
gi|632776|gb|AAB31836.1| N-CAM [Homo sapiens]
lcl|51693

51693: gi|632776|gb|AAB31836.1| N-CAM [Homo sapiens]
30 - 84
55 residues shown

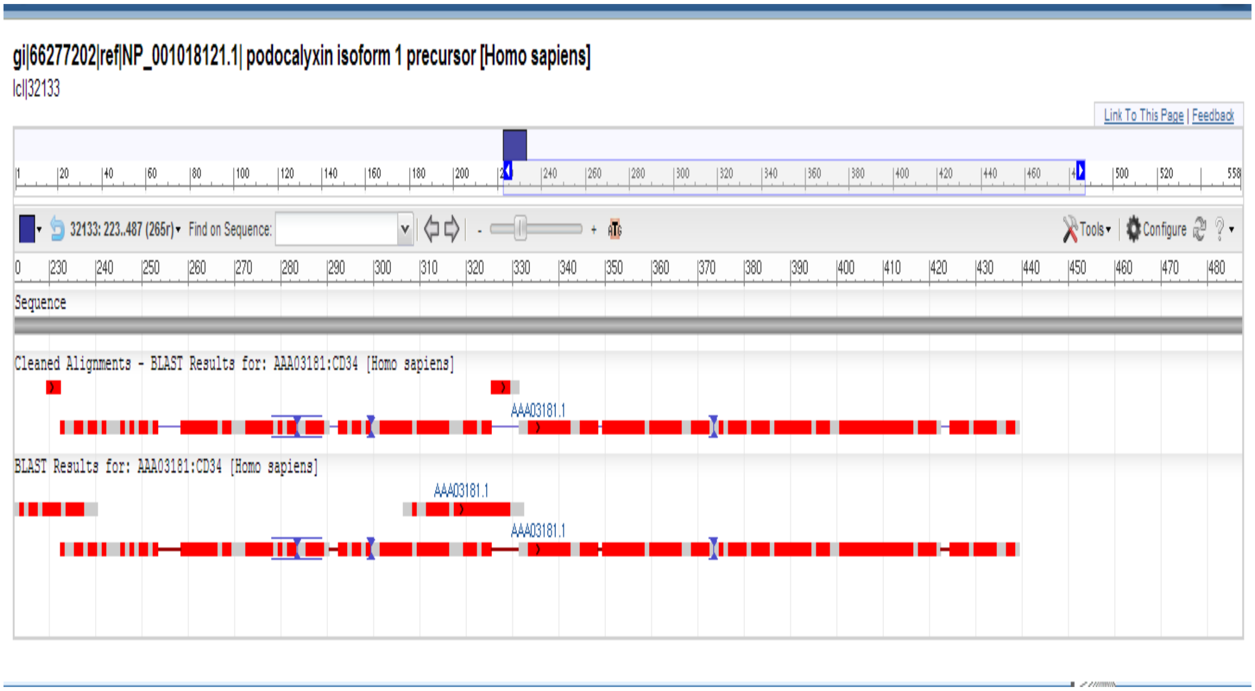
Hence, still CD34 can be proved to be an adhesion molecule.

- **CD34 with PODOCALYXIN** (PODOCALYXIN is also an adhesion molecule which is a sialoglycoprotein and also a member of CD34 family of transmembrane proteins having accession no. NP_001018121.1)

For fragment 1 (1-259 amino acids)



Alignment score was found to be <40 from residue 33-53 & 40-300 for residues ranging from 54-259.



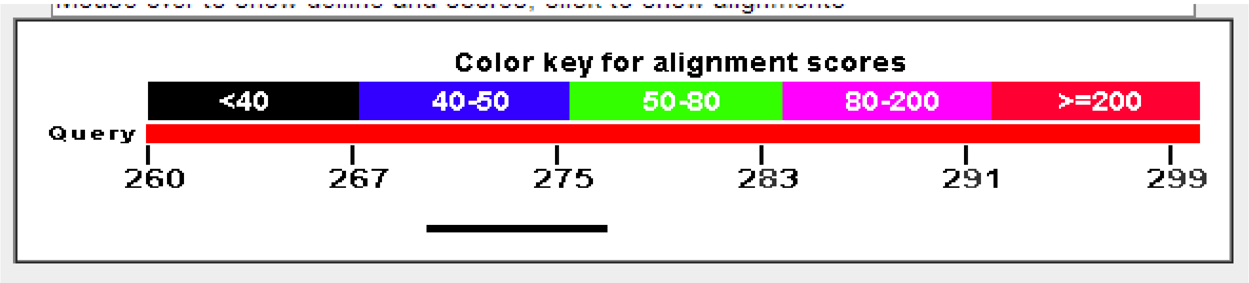
gi|66277202|ref|NP_001018121.1| podocalyxin isoform 1 precursor [Homo sapiens]

lcl|32133

Cleaned Alignments - BLAST Results for: AAA03181:CD34 [Homo sapiens]

BLAST Results for: AAA03181:CD34 [Homo sapiens]

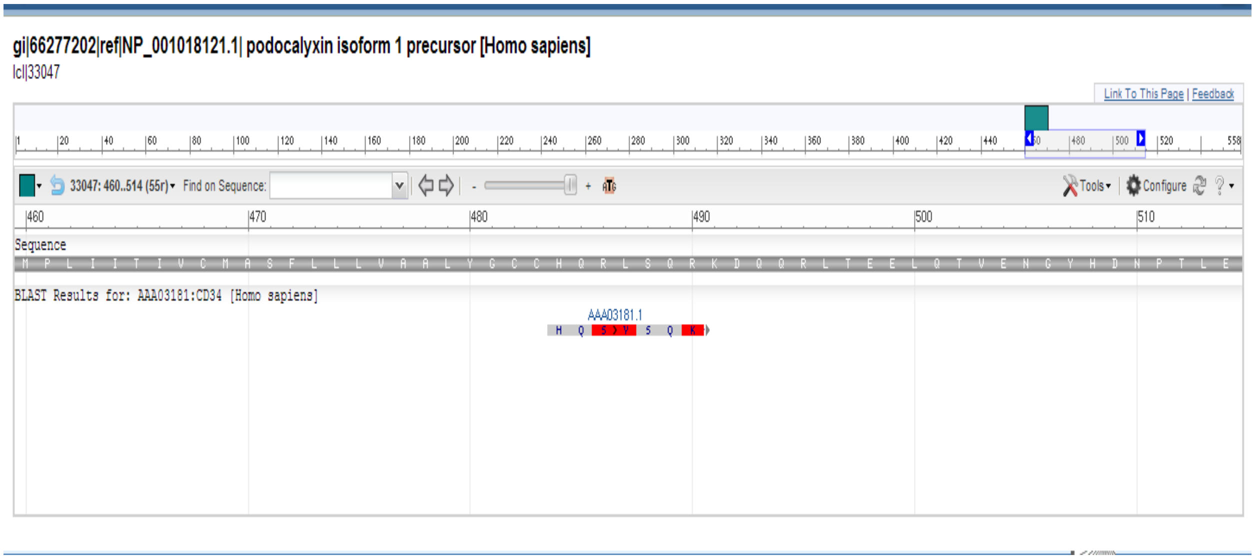
For fragment 2 (260-300 amino acids)



Color key for alignment scores

<40	40-50	50-80	80-200	>=200

Alignment score was found to be from 43-60 for residues ranging from 271-277.



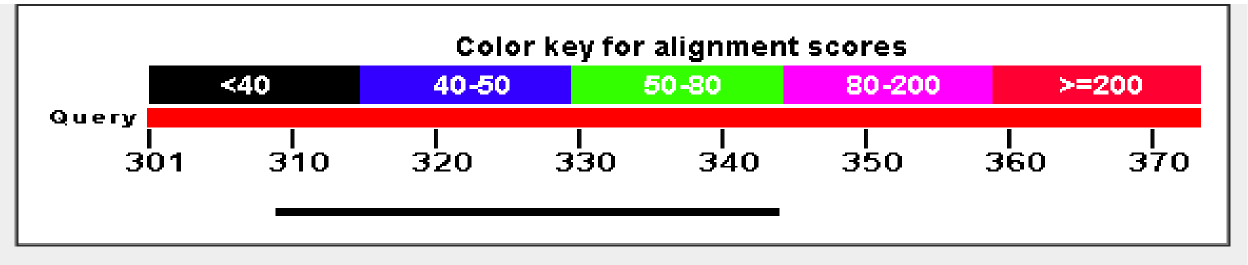
gi|66277202|ref|NP_001018121.1| podocalyxin isoform 1 precursor [Homo sapiens]

lcl|33047

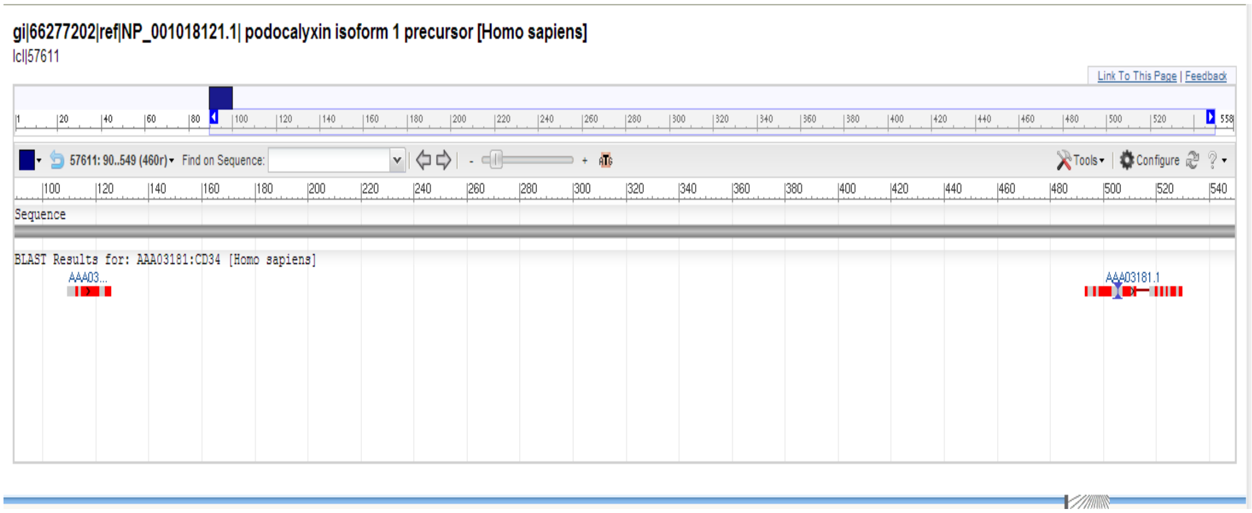
BLAST Results for: AAA03181:CD34 [Homo sapiens]

Link To This Page | Feedback
33047: 460..514 (55r)
Find on Sequence:
Tools
Configure
460
470
480
490
500
510
Sequence
M P L I I T I V C M A S F L L L V A A L Y G C C H Q R L S Q R K D Q Q R L T E E L Q T V E N G Y H D N P T L E
AAA03181.1
H Q S V S Q K

For fragment 3 (301-373 amino acids)



Alignment score was found to be ~50 for residues ranging from 310-344.


gi|66277202|ref|NP_001018121.1| podocalyxin isoform 1 precursor [Homo sapiens]
lcl|57611
Link To This Page | Feedback
57611: 90..549 (460r)
Find on Sequence:
Tools
Configure
Sequence
BLAST Results for: AAA03181:CD34 [Homo sapiens]
AAA03...
AAA03181.1

Observing all the three fragments it can be said that CD34 is an adhesion molecule.

E. Perform Multiple Sequence alignment of aligned sequences with CD34

Output file-

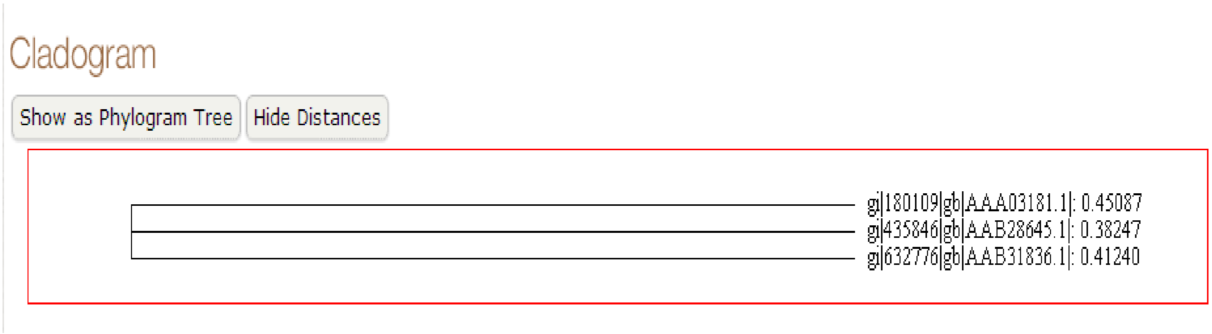
```
Sequences (1:2) Aligned. Score: 16.6667
Sequences (1:3) Aligned. Score: 13.6729
Sequences (2:3) Aligned. Score: 20.5128
```

Sequence1= CD34

Sequence2= PECAM-1

Sequence3= N-CAM

Calculated guide tree (Cladogram tree)



Cladogram

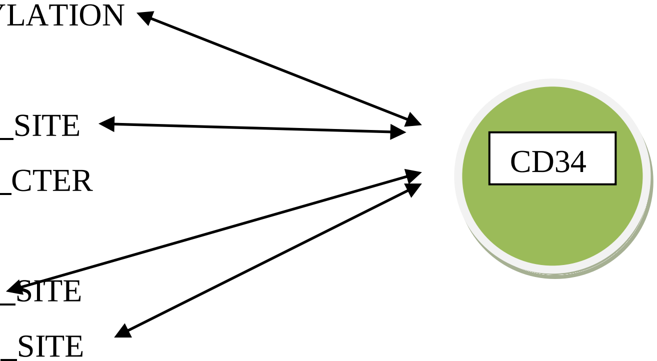
Show as Phylogram Tree | Hide Distances

gi|180109|gb|AAA03181.1|: 0.45087
gi|435846|gb|AAB28645.1|: 0.38247
gi|632776|gb|AAB31836.1|: 0.41240

F. Comparison of motifs found in N-CAM and PECAM-1 with that of CD34 using MOTIF Scan

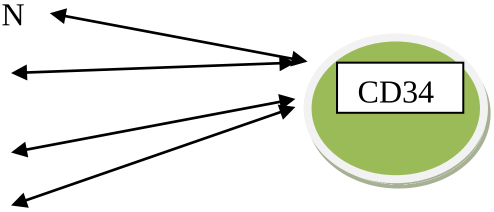
Following motifs were found in N-CAM

ASN_GLYCOSYLATION
ATP_GTP_A
CK2_PHOSPHO_SITE
MICROBODIES_CTER
MYRISTYL
PKC_PHOSPHO_SITE
TYR_PHOSPHO_SITE
Ig domain



Following motifs were found in PECAM-1

ASN_GLYCOSYLATION
CK2_PHOSPHO_SITE
PKC_PHOSPHO_SITE
TYR_PHOSPHO_SITE



CONCLUSIONS AND FUTURE PERSPECTIVES

The CD34 protein is a member of a family of single-pass transmembrane sialomucin proteins that show expression on early hematopoietic and vascular-associated tissue. CD34 is also an important adhesion molecule and is required for T cells to enter lymph nodes. In- silico analysis was done to study the structural aspects of CD34 for which secondary structure details were retrieved via tools which are implicated for predicting secondary structure of a protein molecule. CFSSP, GOR4 and SCRATCH were the tools used in this work which gave near about similar results i.e. alpha helix 45% on an average and random coil >50%. The next step was to figure out the similarity of CD34 amongst other different cell adhesion molecules, for which PSI BLAST was performed of CD34 along with molecules such as N- CAM (stands for neural cell adhesion molecule, is a hemophilic binding glycoprotein expressed on the surface of neurons, glia, skeletal muscle and natural killer cells. NCAM has been implicated as having a role in cell–cell adhesion), PECAM- 1(is a protein that in human is encoded by the PECAM1 gene found on chromosome 17) and Cadherin (are a class of type-1 transmembrane proteins and play important roles in cell adhesion). CD34 showed 42 and 58% similarity respectively to both N- CAM and PECAM- 1. Further to confirm the results multiple sequence alignment was performed on these three and the aligned sequences were used to build a phylogenetic tree which showed close homology between these three molecules. Till date confirmed 3D structure of CD34 is not available however, preliminary reports have shown that it consists of extracellular domains, cytoplasmic tail and consistent coiled regions as well as alpha helices. In this work, methods used for 3D structure prediction were threading, comparative modeling and ab initio. Comparative modeling and ab initio tools for example, 3D JIGSAW, CPH model server produced templates resembling the structure of CD34 for example, **Q4R4G9_MACFA** which is an unknown protein product of 168 amino acids and found to be similar to CD34 antigen. It is a brain cDNA clone from macaque monkey, when ClustalO was performed on these two it showed 35% similarity. The predicted CD34 structure consisted of a large extracellular domains (278 residues), followed by a stretch of hydrophobic amino acids within the transmembrane region (~22 residues) and a small cytoplasmic tail (73 residues). Based on the above results, it can be concluded that CD34 behaves as an adhesion molecule or cell adhesion can be one of the functions of CD34 whose functions are still unclear.

REFERENCES

Andre C, Hampe A, Lachaume P, Martin E, Wang XP, Manus V, Hu WX, Galibert F (January 1997). "Sequence analysis of two genomic regions containing the KIT and the FMS receptor tyrosine kinase genes". *Genomics* **39** (2): 216–26.

Berg EL, Mullowney AT, Andrew DP, Goldberg JE, Butcher EC (February 1998). "Complexity and differential expression of carbohydrate epitopes associated with L-selectin recognition of high endothelial venules". *Am. J. Pathol.* **152** (2): 469–77.

Burdick MM, Chu JT, Godar S, Sackstein R (May 2006). "HCELL is the major E- and L-selectin ligand expressed on LS174T colon carcinoma cells". *J. Biol. Chem.* **281** (20): 13899–905.

Corbeil D, Fargeas C, Huttner W (2001). "Rat prominin, like its mouse and human orthologues, is a pentaspan membrane glycoprotein". *Biochem Biophys Res Commun* **285** (4): 939–44.

Dimitroff CJ, Lee JY, Rafii S, Fuhlbrigge RC, Sackstein R (June 2001). "Cd44 Is a Major E-Selectin Ligand on Human Hematopoietic Progenitor Cells". *J. Cell Biol.* **153**(6): 1277–86.

Edling CE, Hallberg B (2007). "C-Kit--a hematopoietic cell essential receptor tyrosine kinase". *Int. J. Biochem. Cell Biol.* **39** (11): 1995–8.

Fred H. Gage. Mammalian Neural Stem Cells. Science 25 February 2000: 287 (5457), 1433-1438.

Furness SG, McNagny K (2006). "Beyond mere markers: functions for CD34 family of sialomucin in hematopoiesis". *Immunol. Res.* **34** (1): 13–32.

Galli R, Binda E, Orfanelli U, Cipelletti B, Gritti A, De Vitis S, Fiocco R, Foroni C, Dimeco F, Vescovi A (2004). "Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma". *Cancer Res* **64** (19): 7011–7021.

Hanley WD, Burdick MM, Konstantopoulos K, Sackstein R (July 2005). "CD44 on LS174T colon carcinoma cells possesses E-selectin ligand activity". *Cancer Res.* **65**(13): 5812–7.

Hanley WD, Napier SL, Burdick MM, Schnaar RL, Sackstein R, Konstantopoulos K. (Dec 2005). "Variant isoforms of CD44 are P- and L-selectin ligands on colon carcinoma cells". *FASEB J* **20** (2): 337–9.

Hemmati HD, Nakano I, Lazareff JA, Masterman-Smith M, Geschwind DH, Bronner-Fraser M, Kornblum HI (2003). "Cancerous stem cells can arise from pediatric brain tumors". *Proc Natl Acad Sci U S A*; **100** (25): 15178–15183.

Kaplan AI. Mesenchymal Stem Cells. J Orthop Res. 1991 Sep; 9(5):641-50.

Laar JM, Farge D, Tyndall A. Autologous Stem cell Transplantation International Scleroderma (ASTIS) trial: hope on the horizon for patients with severe systemic sclerosis. Ann Rheum Dis 2005; 64:1515.

Looijenga LH, Stoop H, de Leeuw HP, *et al.* (2003). "POU5F1 (OCT3/4) identifies cells with pluripotent potential in human germ cell tumors". *Cancer Res.* **63** (9): 2244–50.

Napier SL, Healy ZR, Schnaar RL, Konstantopoulos K (Feb 2007). "Selectin ligand expression regulates the initial vascular interactions of colon carcinoma cells: the roles of CD44v and alternative sialofucosylated selectin ligands". *J Biol Chem.* **282** (6): 3433–41.

Nielsen JS, McNagny KM (2008). "Novel functions of the CD34 family". *J of Cell Science* **121** (Pt 22): 3682–3692.

Oxley SM, Sackstein R (November 1994). "Detection of an L-selectin ligand on a hematopoietic progenitor cell line". *Blood* **84** (10): 3299–306.

Sackstein R, Dimitroff CJ (October 2000). "A hematopoietic cell L-selectin ligand that is distinct from PSGL-1 and displays N-glycan-dependent binding activity". *Blood* **96** (8): 2765–74.

Sackstein R, Merzaban JS, Cain DW, Dagia NM, Spencer JA, Lin CP, Wohlgemuth R (February 2008). "Ex vivo glycan engineering of CD44 programs human multipotent mesenchymal stromal cell trafficking to bone". *Nat. Med.* **14** (2): 181–7.

Simmons DL, Satterthwaite AB, Tenen DG, Seed B (1 January 1992). "Molecular cloning of a cDNA encoding CD34, a sialomucin of human hematopoietic stem cells". *J. Immunol.***148** (1): 267–71.

Singh SK, Clarke ID, Terasaki M, Bonn VE, Hawkins C, Squire J, Dirks PB (2003). "Identification of a cancer stem cell in human brain tumors". *Cancer Res* **63** (1): 5821–5828.

SSuzawa K, Kobayashi M, Sakai Y, Hoshino H, Watanabe M, Harada O, Ohtani H, Fukuda M, Nakayama J (July 2007). "Preferential induction of peripheral lymph node addressin on high endothelial venule-like vessels in the active phase of ulcerative colitis". *Am. J. Gastroenterol.* **102** (7): 1499–509.

S. Wu, Y. Zhang. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Structure, Function, and Bioinformatics 2008; 72: 547-556

Takeda J, Seino S, Bell GI (September 1992). "Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues". *Nucleic Acids Res.* **20** (17): 4613–20.

Thomas SN, Zhu F, Schnaar RL, Alves CS, Konstantopoulos K (Jun 2008)."Carcinoembryonic Antigen and CD44 Variant Isoforms Cooperate to Mediate Colon Carcinoma Cell Adhesion to E- and L-selectin in Shear Flow". *J Biol Chem.* **283** (23): 15647–55.

Weiss ML, Mitchell KE, Hix JE et al. Transplantation of porcine umbilical cord matrix cells into the rat brain. Exp.Neurol. 2003; 182:288-299.