International Journal of Biology Research
www.biologyjournal.in
ISSN: 2455-6548
Received: 18-06-2021, Accepted: 04-07-2021, Published 20-07-2021
Volume 6, Issue 3, 2021, Page No. 24-29

Homolgy modelling of hypothetical protein (Apolipoprotein L6)

Sushma Kumari, Nayan Mehta, Varinder Kumar, Jasmine Dua
Department of Bioinformatics, GGDSD College Chandigarh, India

Abstract
Apolipoprotein L(Apo L) belongs to the high density lipoprotein family that plays a central role in cholesterol transport.There are six apo L genes located very close to each other on chromosome 22q12 in humans. Apolipoprotein L6 is a protein that in humans is encoded by the APOL6 gene. It is found that there is a number of proteins related to this family whose structure is unknown till date. I performed homology modelling which is a technique of choice when experimental structure data are not available but three-dimensional coordinates are needed. Chromosome Locus 22q12 is a confirmed high-susceptibility locus for schizophrenia and close to the region associated with velocardiofacial syndrome that includes symptoms of schizophrenia. Using bioinformatics tools and softwares, a homology based 3-D model of hypothetical protein sequence is generated.

Keywords: apolipoprotein, homology modelling, hypothetical protein

Introduction

Apolipoproteins are proteins that bind lipids (oil-soluble substances such as fat and cholesterol) to form lipoproteins. They transport the lipids through the lymphatic and circulatory systems. Apolipoproteins also serve as enzyme cofactors, receptor ligands, and lipid transfer carriers that regulate the metabolism of lipoproteins and their uptake in tissues.
Apolipoprotein L –Also known as APOLVI; APOL-VI. Apolipoprotein L6 is a protein that in humans is encoded by the *APOL6* gene. Human proteins containing this domain are:
APOL1; APOL2; APOL3; APOL4; APOL5; APOL6; APOLD1;
This gene is a member of the apolipoprotein L gene family. The encoded protein is found in the cytoplasm, where it may affect the movement of lipids or allow the binding of lipids to organelles.

Synthesis and Regulation

Apolipoprotein synthesis in the intestine is regulated principally by the fat content of the diet. Apolipoprotein synthesis in the liver is controlled by a host of factors, including dietary composition, hormones (insulin, glucagon, thyroxin, estrogens, and rogens), alcohol intake, and various drugs (statins, niacin, and fibric acids). Apo B is an integral apoprotein whereas the others are peripheral apoproteins. This gene is a member of the apolipoprotein L gene family. The encoded protein is found in the cytoplasm, where it may affect the movement of lipids or allow the binding of lipids to organelles.

Methodology and Observation

- The hypothetical protein was taken in FASTA format from NCBI. Here the protein of interest is apolipoprotein l6, a hypothetical protein whose structure have not yet been discovered. The basic information about this particular protein is as below:

Protein **(**Recommended name) -Apolipoprotein L6

1. Alternative name (s)- Apolipoprotein L-VI Short name - ApoL-VI
2. Gene name - APOL6
3. ORF Names - UNQ3095/PRO2134
4. Biological process – Lipid transport

The value ‘Evidence at protein level’ indicates that there is clear experimental evidence for the existence of the protein. The criteria include Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies.

Function: It affect the movement of lipids in the cytoplasm or allow the binding of lipids to organelles.

Tissue Specificity

Highly expressed in the uterus, fetal brain and spinal cord, also detected in heart, liver, lung, colon, spleen, thymus, prostate, placenta, adrenal gland, salivary and mammary gland.
The hypothetical protein was searched in Homolo Gene section of NCBI. Homolo Gene is a system for automated detection of homologs (similarity attributable to descent from a common ancestor) among the annotated genes of several completely sequenced eukaryotic genomes. The Homolo Gene processing consists of the protein analysis from the input organisms. Sequences are compared using blastp, then matched up and put into groups, using a taxonomic tree built from sequence similarity, where closer related organisms are matched up first, and then further organisms are added to the tree.
Here the search result of hypothetical apolipoprotein L6 indicated that the protein has gene homologs in following organisms:

- *P. troglodytes*
- *M. mulatta*
- *C. lupus*
- *B. taurus*
- *M. musculus*

The pairwise alignment identity scores of the hypothetical protein with other homologous organisms’ sequence is given in figure. It shows that the protein has maximum percentage identity with *P. troglodytes* with 97.4 (protein) and 98.9 (DNA).
The very basic steps for predicting the structure of a protein from its sequence needs the complete information about its chemical and physical properties and other related parameters. Here, I have used ProtParam, a tool ExPASy server. ProtParam computes various physico-chemical properties that can be deduced from a protein sequence**.**
The parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw. The amino acid and atomic compositions are self-explanatory. All the other parameters will be explained below:

Extinction coefficients
The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. The ex. coefficient for the query protein sequence = 44920 M^{-1} cm^{-1}
The absorbance (optical density) -Abs 0.1% (=1 g/l) 1.178

Instability index (II)
The instability index provides an estimate of the stability of your protein in a test tube. A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable. The instability index (II) = 38.70

Aliphatic index
The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine):
Aliphatic index: 86.82

GRAVY (Grand Average of Hydropathy)
The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence.
Grand average of hydropathicity (GRAVY): -0.401
Next I searched for the conserved domains of homologous sequences of hypothetical protein. Domains are often identified as recurring (sequence or structure) units, which may exist in various contexts. In molecular evolution such domains may have been utilized as building blocks, and may have been recombined in different arrangements to modulate protein function. Conserved domains contain conserved sequence patterns or motifs, which allow for their detection in polypeptide sequences. The CDD (Conserved Domain Database) is such a database to find out the conserved domains. The output shows the colored amino acids one letter code and some blank spaces in multiple sequence alignment.
For the prediction of protein localization sites, I used PSORT which is a computer program for the prediction of protein localization sites in cells. It receives the information of an amino acid sequence and its source origin, *e.g.*, Gram-negative bacteria, as inputs. Then, it analyzed the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reported the possibility for the input protein to be localized at each candidate site with additional information.

- **Prediction of Membrane Topology**

PSORT uses Hartmann *et al*.'s method called "MTOP" for the prediction of membrane topology, which assumes that the overall topology is determined from the net charge difference of both sides of 15 residues flanking the most N-terminal transmembrane segment.

I (middle): 92 Charge difference (C-N): -1.5

Recognition of Signal Sequence
PSORT first predicts the presence of signal sequences by McGeoch's method. It considers the N-terminal basically-charged region (CR) and the central hydrophobic region (UR) of signal sequences.
Length of UR: 6
Peak Value of UR: 0.03
Net Charge of CR: -3
Discriminant Score: -18.20
PSORT applies von Heijne's method of signal sequence recognition. It is a weight- matrix method and incorporates the information of consensus pattern around the cleavage sites (the (-3,-1)-rule) and thus it can be used to detect uncleavable signal sequences.

Signal Score (-3.5): -10.7
Possible cleavage site: 35

Recognition of Transmembrane Segments
PSORT is programmed to detect potential transmembrane segments. It attempts to identify the most probable transmembrane segment from the average hydrophobicity value of 17-residue segments, if any. It predicts whether the segment is a transmembrane segment (INTEGRAL) or not (PERIPHERAL) comparing the discriminant score.
INTEGRAL Likelihood = -7.11 Transmembrane 324 - 340 (322 - 342)
INTEGRAL Likelihood = -3.35 Transmembrane 85 - 101 (79 - 102)

PERIPHERAL Likelihood = 3.76
modified ALOM score: 1.52

Recognition of Mitochondrial Proteins
PSORT employs a simple method to recognize mitochondrial targeting signals using the discriminant analysis from values of partial amino acid composition.
negative (-6.56)

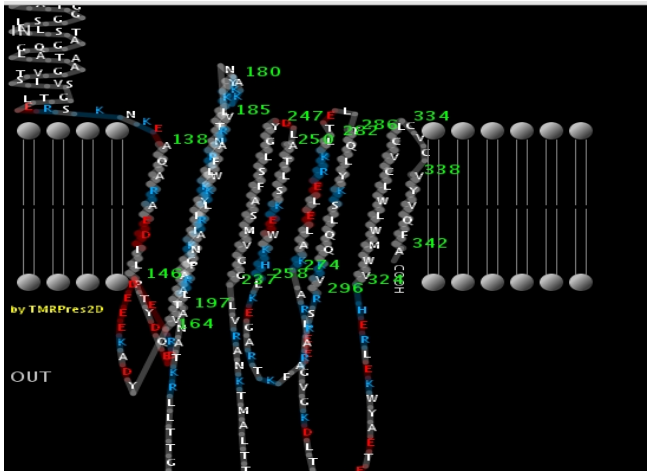

Fig 1: Showing Transmembrane Regions of Sequence (PRED TMBB)

Results and Discussion
Performance of the combined prediction
The prediction quality of RHYTHM improved compared to our previous analysis. This is due to the enlarged data set of helical membrane proteins and the combination of the matrix prediction method with the prediction from evolutionary conservation. The average AUC-values (from a leave-one-out cross validation) for the prediction of helix–helix contacts are 0.72 for channels [as in our previous analysis and 0.68 for membrane–coils, respectively. The corresponding values for the prediction of helix–membrane contacts are 0.75 and 0.73. Best predictions were obtained for helix–helix contacts of the translocon channel (PDB-entry: 1rh5, AUC-value: 0.78) and for helix–membrane contacts of the ABC-transporter protein (PDB-entry: 2qi9, AUC-value: 0.86).
Here the predicted result showed that the residues from amino acids 138 -342 lie in transmembrane region.

Secondary Structure Prediction
Alpha helix (Hh): 208 is 60.64%
3_{10} helix (Gg): 0 is 0.00%
Pi helix (Ii): 0 is 0.00%
Beta bridge (Bb): 0 is 0.00%
Extended strand (Ee): 39 is 11.37%
Beta turn (Tt): 0 is 0.00%
Bend region (Ss): 0 is 0.00%
Random coil (Cc): 96 is 27.99%
Ambigous states (?): 0 is 0.00%
Other states: 0 is 0.00%

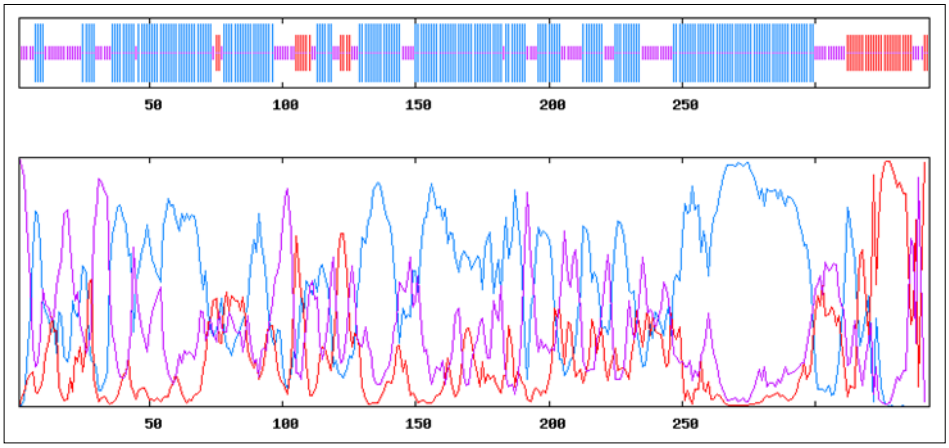

Fig 2: GOR predicted result in graphical format

The protein sequence, in FASTA format was pasted in Swiss model workspace under "automated mode" where the server predicted the tertiary structure of protein from the sequence.

Building a homology model comprises four main steps:

- Identification of structural template(s),
- Alignment of target sequence
- Template structure(s),
- Model building, and model quality evaluation.

These steps can be repeated until a satisfying modelling result is achieved. Each of the four steps requires specialized software and access to up-to-date protein sequence and structure database that is promised by Swiss model workspace.

Fig 3: Automated Modelling of hypothetical protein by Swiss Model Workspace

Template Searching:

The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The critical first step in homology MODELLING is the identification of the best template structure, if indeed any are available.

Here the protein sequence is taken in FASTA format and then blastp was preformed nd templates structures were searched in PDB. The BLASTP search returned the following five templates as follows:

Out of those, the upper four templates 3CTP A,31ZB M, 2XSJ M, 1SIH M were taken into consideration and processed with MODELLER and PHYRE2.

Tertiary Structure Prediction:

Phyre2 uses the alignment of hidden Markov models via HHsearch[1]to significantly improve accuracy of alignment and detection rate.Phyre2 also incorporates a new *ab initio* folding simulation called Poing[2] to model regions of your proteins with no detectable homology to known structures.

Secondary Structure and disorder prediction:

The prediction is 3-state: either α-helix, β-strand or coil. Green

helices represent α-helices, blue arrows indicate β-strands and faint lines indicate coil. The 'SS confidence' line indicates the confidence in the prediction, with red being high confidence and blue low confidence. However the two middle helices are associated with orange, yellow and green indicating a weak prediction. The 'Disorder' line contains the prediction of disordered regions in your protein and such regions are indicated by question marks (?)

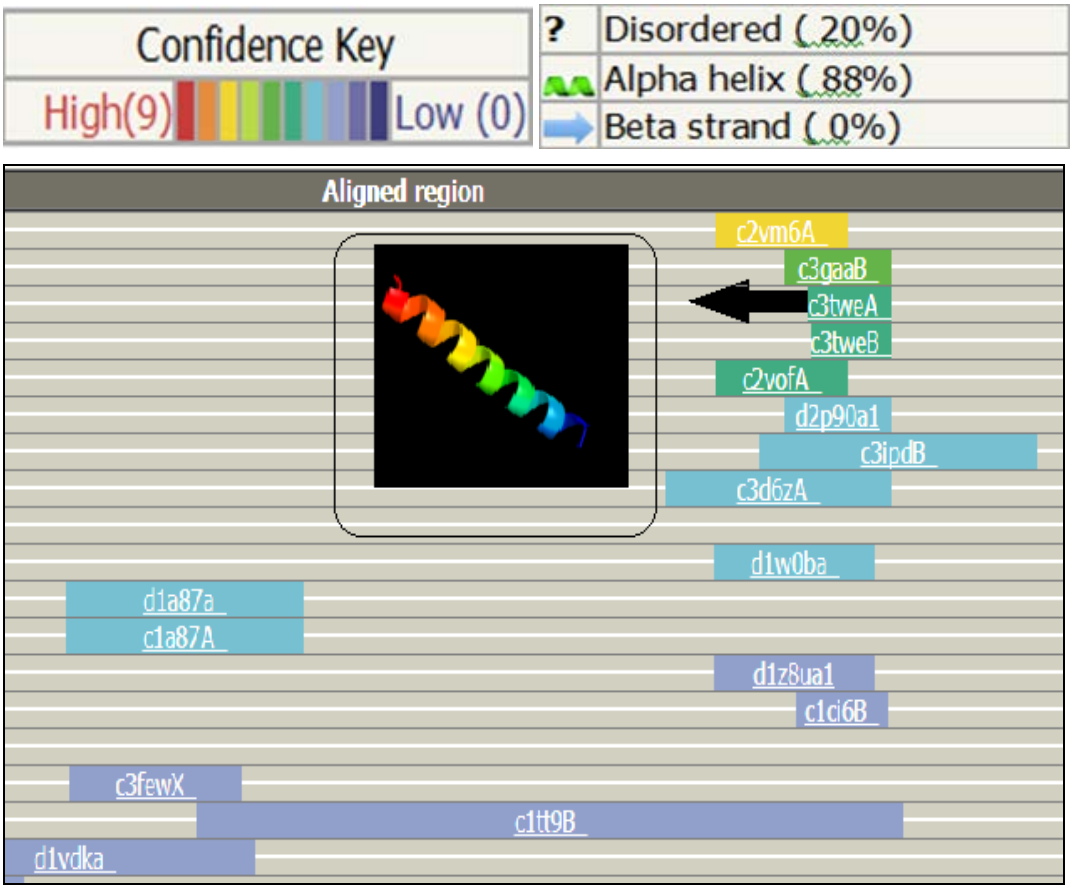

Fig 4: showing homologous domains in sequenc

The domain analysis section illustrates where along sequence matches have been found, colour-coded by confidence. These codes are from SCOP as indicated by the initial '**d**' standing for **d**omain. It was found that the sequence has 4 domains. The initial '**c**' indicates this protein is a whole **c**hain taken from the PDB. Here I have shown the PDB identifier **c3tweA** and chain identifier A. Here top six templates were shown with details. The topmost template is the only one to show maximum similarity of 71%.

Template: Coc2vm6A_nf Confidence: 71% % i.d: 11

Template Information
PDB header: immune system Chain: A: PDB Molecule: bcl-2-related protein a1; PDB Title: human bcl2-a1 in complex with bim-bh3 peptide

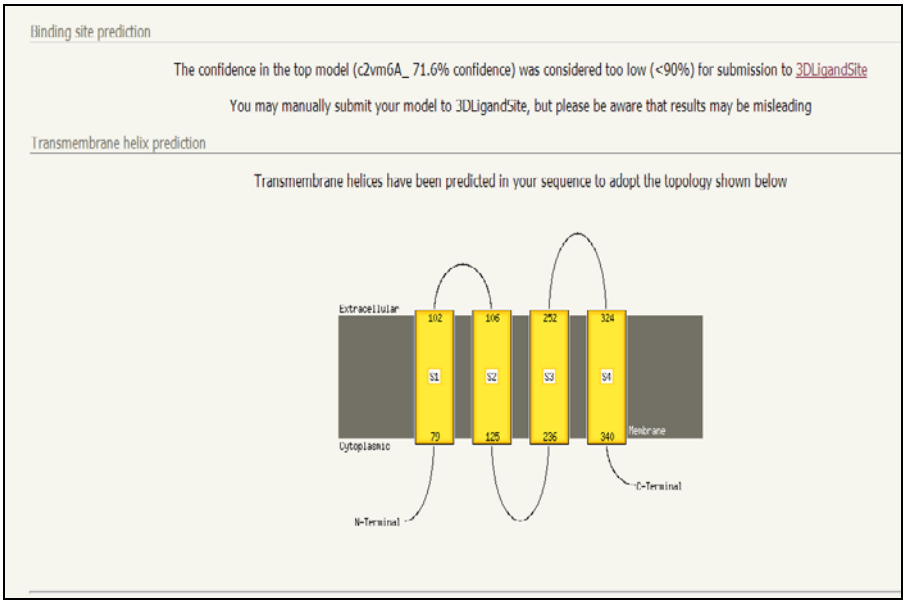

Fig 5: Transmembrane helix prediction of protein sequence

Transmembrane Helix Prediction

The sequence and the set of homologues detected by PSI-Blast are processed by a Support Vector Machine to

a. determine whether your sequence is likely to contain transmembrane helices
b. Predict their topology in the membrane.

The extracellular and cytoplasmic sides of the membrane are labelled and the beginning and end of each transmembrane helix illustrated with a number indicating the residue index.

Here the transmembrane regions were found in lying between Amino acid Range: 79-102, 106 -125
236-252, 324–340

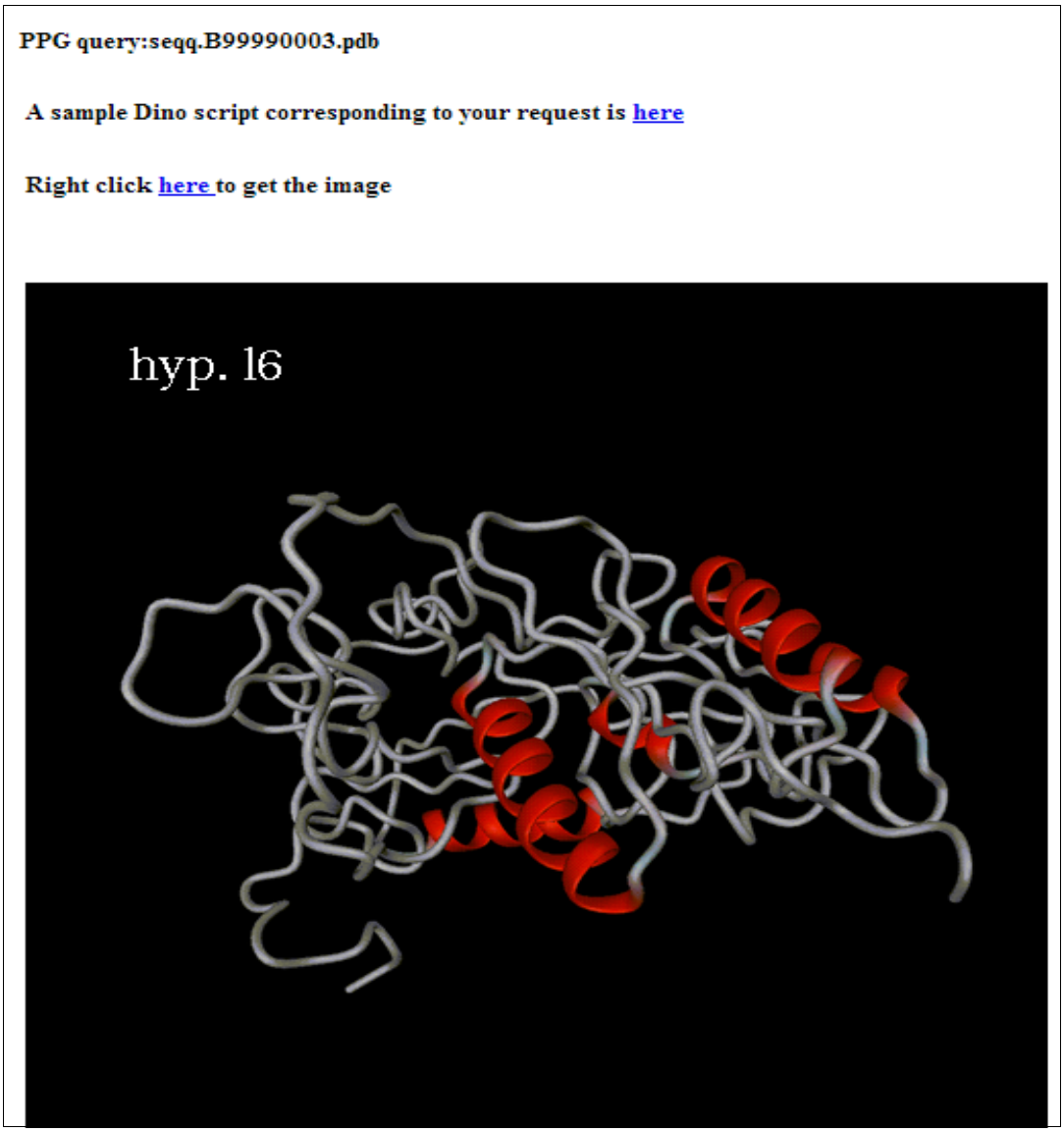

Fig 6: modeled structure of hypothetical protein by protein picture generator (PPG)

The hypothetical protein sequence was processed with MODELLER which is a standalone program. T MODELLER calculates a 3D model of the target completely automatically, using its automodel class. MODELLER implements comparative protein structure MODELLING by satisfaction of spatial restraints.

Four models (PDB files) had generated using MODELLER. All four structures has studied for their various values like Native energy, Z scores, % sequence identity, Compactness etc. Out of four, the seqq.B99990003.pdb most stable structure. Other details are given below:

Chain identifier: _
% sequence identity: 10.000000
Sequence length: 343
Compactness: 0.286940
Native energy (pair): 568.703756
Native energy (surface): 48.461046
Native energy (combined): 16.610775
Z score (pair): -0.823259
Z score (surface): -2.047491
Z score (combined): -1.792999
GA341 score: 0.011575

Structure Validation

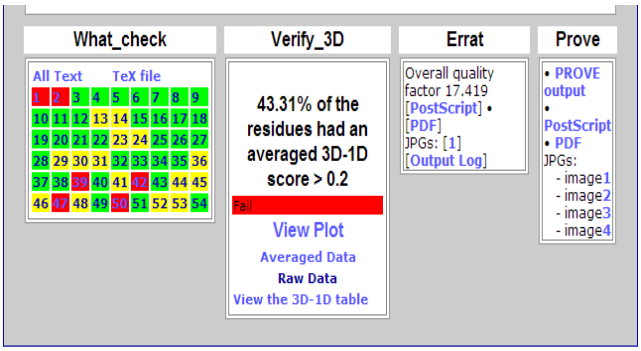

Fig 7: SAVES server validation result summary

The red, yellow and light yellow regions represent the favored, allowed, and "generously allowed" regions as defined by Pro Check.
The very final step in homology MODELLING is to validate the modelled protein structure. Here, it is done with using SAVES server which checks the validity of modelled structure

Ramachandran Plot

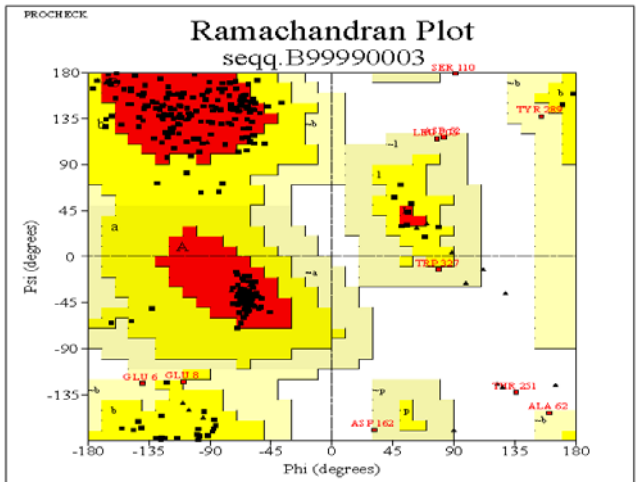

Fig 8: interactive Ramachandran plot

The Ramachandran plot shows the phi-psi torsion angles for all residues in the structure. Glycine residues are separately identified by triangles as these are not restricted to the regions of the plot appropriate to the other side chain types.
The coloring/shading on the plot represents the different regions. The darkest areas (here shown in *red*) correspond to the "core" regions representing the most favorable combinations of phi-psi values.

Ramachandran Plot Regions
Plot Statistics:
Residues in most favoured regions [A,B,L] 249 79.8%
Residues in additional allowed regions [a,b,l,p] 53 17.0%
Residues in generously allowed regions [~a,~b,~l,~p] 8 2.6%
Residues in disallowed regions 2 0.6%

Number of non-glycine and non-proline residues 312 100.0%
Number of end-residues (excl. Gly and Pro) 2
Number of glycine residues (shown as triangles) 23
Number of proline residues 6

Conclusion
Here the hypothetical protein Apolipoprotein L (Apo L) is modeled on the basis of homology. Apo L belongs to the high density lipoprotein family that plays a central role in cholesterol transport. It is found that the site of the gene confirmed high-susceptibility locus for schizophrenia and close to the region associated with velocardiofacial syndrome that includes symptoms of schizophrenia.
Finally, here I am concluding this project here by the statement that in its most elementary form, homology modelling involves calculating the structure of Apo L protein for which only the sequence is known using its alignment with a homologous protein for which the structure is known. The process starts with the detection of a suitable template; an alignment is produced; insertions, deletions and residue substitutions are performed; the model is optimized. At last, a Ramachandran plot is also drawn of the modeled protein.

Abbreviations
APO (Apolipoprotein), CASP (Critical Assessment of Structure Prediction),
SCOP (Structural Classification of Protein), SAVES (Structure analysis and verification server)

References

1. The human apolipoprotein L gene cluster identification classification and sites of distribution N M Butlin D J Lomthaisong K Lowry PJ Genomics,2001:74:71-78.
2. Predicting Coiled Coils from Protein Sequences, Lupas A Van Dyke M and Stock, J Science,1991:252:1162-1164.
3. "Porter a new accurate server for protein secondary structure prediction". Bioinformatics G Pollastri, A McLysaght, 2005:21:(8):1719-1720.
4. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier Sapay N, Guermeur Y Deleage GBMC Bioinformatics, 2006:16;7(1):255.
5. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997:1:25(17):3389-3402.
   Altschul SF Madden TL Schaffer AA Zhang JZhang Z *et al*.
6. "Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence", *Proteins*,2002:49:154-166.
7. 3DLigandSite: predicting ligand-binding sites using similar structures Wass MN *et al*. Structural Bioinformatics Group Centre for Bioinformatics Imperial College London, London SW7 2AZ, UK
8. GOR V server for protein secondary structure prediction *Bioinformatics* Sen T Z Jernigan R L Garnier, J Kloczkowski A,2005:21:(11):2787-2788>
9. Apolipoproteins in the brain implications for neurological and psychiatric disorders David A Elliott *et al*, 2011.
10. Gene expression analysis in schizophrenia Reproducible up-regulation of several members of the apolipoprotein L family located in a high-susceptibility locus for schizophrenia on chromosome 22-Michael L Mimmack *et al*, 2002.
11. Apolipoprotein L6 Induced in Atherosclerotic Lesions Promotes Apoptosis and Blocks Beclin 1-dependent Autophagy in Atherosclerotic Cells Siqin Zhaorigetu *et al*, 2011.
12. Human Apolipoprotein L1 (ApoL1) in Cancer and Chronic Kidney Disease Chien-An A Hu *et al*, 2012.
13. Family-based association analysis of 42 hereditary prostate cancer families identifies the Apolipoprotein L3 region on chromosome 22q12 as a risk locus Bo Johanneson, *et aL*, 2010.
14. The apolipoprotein L family of programmed cell death and Immunity genes rapidly evolved in primates at discrete sites of host–pathogen interactions Eric E Smith.
15. Kabsch W, Sander C A dictionary of protein secondary structurepolymers,1983:22:2577-2637.