European Journal of Biotechnology and Bioscience
Online ISSN: 2321-9122, Impact Factor: RJIF 5.44
www.biosciencejournals.com
Volume 5; Issue 3, May 2017; Page No. 28-33

Homology modeling and structural analysis of Env protein from Retroviridae

Jitender Singh, *Ashvinder Raina

Department of Experimental Medicine and Biotechnology, Post Graduate Institute of Medical Education and Research, Chandigarh, Punjab, India

Abstract
Retroviruses are viruses that are remarkable for their use of reverse transcription of viral RNA into DNA during replication. Members of this family include Human immunodeficiency virus (the virus that causes AIDS), feline leukemia, and several cancer-causing viruses. The genome of retroviridae is dimeric, unsegmented and contains a single molecule of linear. The genome is -RT and a positive-sense, single-stranded RNA. Minor species of non-genomic nucleic acid are also found in virions. The encapsidated nucleic acid is mainly of genomic origin but virions may also contain nucleic acid of host origin, including host RNA and fragments of host DNA believed to be incidental inclusions. The virions of a retroviridae consist of an envelope, a nucleocapsid and a nucleoid. The virus capsid is enveloped. The virions are spherical to pleomorphic and measure 80-100 nm in diameter. The surface projections are small or distinctive glycoprotein spikes that cover the surface evenly. The projections are densely dispersed and 8 nm long. The nucleoid is concentric or eccentric while the core is spherical. Here we present a computationally predicted structure of Env protein using homology modeling. Primary and secondary structure analysis suggested that Env is a hydrophilic protein containing a significant proportion of alpha helices, and sub cellular localization predictions suggested it is a cytoplasmic protein. The tertiary structure of protein Env was predicted by homology modeling. The results suggest a flexible structure which is also an important characteristic of active enzymes enabling them to bind various cofactors and substrates for proper functioning. Validation of 3D structure was done using Ramachandran plot. This predicted information will help in better understanding of mechanisms underlying Env protein. By using the information of 3d structure of this protein we can easily design drug against this protein which are fully responsible for diseases.

Keywords: homology modeling, retroviridae, structural analysis

1. Introduction
Retroviridae is a family of enveloped viruses that replicate in a host cell through the process of reverse transcription. A retrovirus is a single-stranded positive-sense RNA virus with a DNA intermediate and, as an obligate parasite, targets a host cell. Once inside the host cell cytoplasm, the virus uses its own reverse transcriptase enzyme to produce DNA from its RNA genome — the reverse of the usual pattern. This new DNA is then incorporated into the host cell genome by an integrase enzyme, at which point the retroviral DNA is referred to as a provirus. The host cell then treats the viral DNA as part of its own genome, translating and transcribing the viral genes along with the cell's own genes, producing the proteins required to assemble new copies of the virus. It is difficult to detect the virus until it has infected the host. At that point, the infection will persist indefinitely. In most viruses, DNA is transcribed into RNA, and then RNA is translated into protein. However, retroviruses function differently – their RNA is reverse-transcribed into DNA, which is integrated into the host cell's genome (when it becomes a provirus), and then undergoes the usual transcription and translational processes to express the genes carried by the virus. So, the information contained in a retroviral gene is used to generate the corresponding protein via the sequence: RNA → DNA → RNA → polypeptide. This extends the fundamental process identified by Francis Crick (one gene-one peptide) in which the sequence is: DNA → RNA → peptide (proteins are made of one or more polypeptide chain; e.g. haemoglobin is a four-chain peptide).Retroviruses are valuable research tools in molecular biology, and have been used successfully in gene delivery systems [1]. Retroviruses are a diverse family of enveloped RNA viruses that can be broadly categorized into two groups based on genome complexity: the simple retroviruses and the complex retroviruses. All retrovirus genomes contain three major open reading frames that encode the viral structural and enzymatic proteins: gag, pol and env. In addition to the three major genes, an additional domain, pro, codes for the viral protease that is also present in all retroviruses. Distinguishing them from simple retroviruses, complex retroviruses also encode a number of accessory proteins that carry out additional virus-specific functions. The proteins common to all retroviruses (Gag, Pol, Pro and Env) have the same function regardless of the specific virus. Env, the retroviral envelope protein, is the major viral protein present on the surface of retroviral particles. Env is translated as a polyprotein that is subsequently extensively post-translationally modified during trafficking through the biosynthetic pathway [2]. Env is cleaved by a furin-like protease in the Golgi into its two subunits: the surface unit (SU) protein and the transmembrane (TM) protein. These two proteins remain non-covalently associated in most retroviruses and further assemble into a homotrimeric complex that is the active form of Env [3]. The envelopes are typically derived from portions of the host cell membranes (phospholipids and proteins), but include some viral glycoproteins. They may help viruses avoid the host immune system. Glycoproteins on the surface of the envelope serve to identify and bind to receptor sites on the host's membrane. The viral envelope then fuses

with the host's membrane, allowing the capsid and viral genome to enter and infect the host. Env proteins play a role in association and entry of virions into the host cell. Possessing a functional copy of an env gene is what makes retroviruses distinct from retro-elements [4]. The ability of the retrovirus to bind to its target host cell using specific cell-surface receptors is given by the surface component (SU) of the Env protein, while the ability of the retrovirus to enter the cell via membrane fusion is imparted by the membrane-anchored trans-membrane component (TM). Thus it is the Env protein that enables the retrovirus to be infectious.

When retroviruses have integrated their own genome into the germ line, their genome is passed on to a following generation. These endogenous retroviruses (ERVs), contrasted with exogenous ones, now make up 5-8% of the human genome [5]. Most insertions have no known function and are often referred to as "junk DNA". However, many endogenous retroviruses play important roles in host biology, such as control of gene transcription, cell fusion during placental development in the course of the germination of an embryo, and resistance to exogenous retroviral infection. Endogenous retroviruses have also received special attention in the research of immunology-related pathologies, such as autoimmune diseases like multiple sclerosis, although endogenous retroviruses have not yet been proven to play any causal role in this class of disease [6]. While transcription was classically thought to occur only from DNA to RNA, reverse transcriptase transcribes RNA into DNA. The term "retro" in retrovirus refers to this reversal (making DNA from RNA) of the central dogma of molecular biology. Reverse transcriptase activity outside of retroviruses has been found in almost all eukaryotes, enabling the generation and insertion of new copies of retro-transposons into the host genome. These inserts are transcribed by enzymes of the host into new RNA molecules that enter the cytosol. Next, some of these RNA molecules are translated into viral proteins. For example, the gag gene is translated into molecules of the capsid protein, the pol gene is translated into molecules of reverse transcriptase, and the env gene is translated into molecules of the envelope protein. It is important to note that a retrovirus must "bring" its own reverse transcriptase in its capsid, otherwise it is unable to utilize the enzymes of the infected cell to carry out the task, due to the unusual nature of producing DNA from RNA.

2. Materials and methods

2.1 Sequence analysis and sub-cellular localization prediction

The amino acid sequence of Env protein was retrieved from the UNIPROT database using the primary accession number Q01281_9RETR. The sequence is 688 amino acids in length and was also cross checked in NCBI database (https://www.ncbi.nlm.nih.gov/) with accession number CAA41747.1.ProtParam [7] was used to predict the physiochemical properties. ProtParam computed the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropath city (GRAVY). Secondary structure predictions (helix, sheets, and coils) were done by using GOR4 [8]. Prediction of sub-cellular localization of Env was done by CELLO v.2.5 [9].

2.2 Domain Analysis of Env protein

Domains are the functional units of proteins. Env is a protein comprising of two domains. These domain sequences were retrieved from UNIPROT under accession IDQ01281_9RETR. Each domain sequence was then analyzed using Interproscan [10].

2.3 Homology modeling and loop optimization of Env protein

Homology modeling was used to determine the 3D structure of Env protein. As mentioned above Env consist of two domains. A BLASTP [11] search with default parameters were performed against the Protein Data Bank (PDB) to find suitable templates for homology modeling. PDB ID: 1AOL_A was identified as the best template based on sequence identity between query and template protein sequence. The alignment between template and query sequence was done using MODELLERv9.15 [12], which uses global dynamic programming with linear gap penalty function. This aligns the query and template sequences and the output is obtained in PIR format. The PIR format is used by MODELLER in the subsequent model-building stage. A three-dimensional structure was developed from sequence alignment between template and query sequence using MODELLERv9.15. Again 3D structure was cross checked using Phyre2 (www.sbg.bio.ic.ac.uk/~phyre/) and Swiss model (https://swissmodel.expasy.org/) online protein 3D structure modeling tools. Once the 3D model was generated using these three software tools, we selected 3D model which was common to all the three software. Further structural evaluation and stereo chemical analysis was performed using ProSA-webserver Z-scores https://prosa.services.came.sbg.ac.at/prosa.php) [13] and PROCHECK [14] Ramachandran plots by using https://services.mbi.ucla.edu. Furthermore Root Mean Squared Deviation (RMSD), superimposition of query and template structure and visualization of generated models was performed using PyMOL software.

2.4 Ligand binding sites prediction

After validation of 3D structure of env protein, further ligand binding sites were analyzed using Metapocket 2.0 (projects.biotec.tu-dresden.de/metapocket/)

3. Result and discussion

The present research study was focused on sequence and structural analysis of env protein by computational approach. We have taken primary sequence of env protein with accession no CAA41747.1 (Figure 1) ProtParam online web server was used to analyze different physiochemical properties from the amino acid sequence of this protein (Table 1). The env protein contains 688 amino acids, with a molecular weight of 75877.73Daltons and an isoelectric point of 8.49. An isoelectric point below 7 indicates a negatively charged protein corresponds to having more negatively charged residues. Although Expasy's Protparam computes the extinction coefficient for 276, 278, 279, 280 and 282 nm wavelengths, 280 nm is favored because proteins absorb light strongly there while other substances commonly in protein solutions do not. We observed that the extinction coefficient of env protein at 280 nm is 145300 M cm with respect to the concentration of Cys, Trp and Tyr. The high extinction

coefficient points towards the high concentration of Cys, Trp and Tyr. This extinction coefficient helps in the quantitative study of protein-protein interactions and ligand–protein interactions in the solution. The instability index provides an estimate of the stability of protein in a test tube. There are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable one. This method assigns a weight value of instability. Using these weight values it is possible to compute an instability index (II). 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 value for the env protein we found was to be 40.12; it means this protein is unstable. The aliphatic index (AI) which is defined as the relative volume of a protein occupied by aliphatic side chains (A, V, I and L) is regarded as a positive factor for the increase of thermal stability of globular proteins. Aliphatic index for this protein is 82.47.The Grand Average hydropathy (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. GRAVY indices of env protein we found was -0.225. This low range of value indicates the possibility of better interaction with water (Table- 1). Secondary structure analysis was performed using GOR4 tool and the protein was predicted to contain several helices and random coils (Figure 2A & B).The high percentage of helices in the structure makes the protein more flexible for folding, which might increase protein interactions. Subcellular localization is a key functional feature of a protein. Cellular functions are often localized in specific compartments; therefore, sub-cellular localization prediction of unknown proteins could be used to attain valuable information about their functions. Moreover, studying the sub-cellular localization of proteins is also helpful in understanding disease mechanisms which can have application in developing novel drugs. The consensus protein sub-cellular localization predictions suggest that Env protein is a cytoplasmic protein and had no transmembrane helices (Figure 3). Domain analysis of Env and homology modeling using Interproscan, predicted the two domains in Env gene. The two main domains predicted by Interproscan are F-MuLVreceptor-binding (IPR008981), TLV/ENV coat polyprotein (IPR018154). The F- MuLV receptor-binding domain forms part of the retroviral envelope glycoprotein in the murine leukemia virus. Envelope glycoproteins are synthesized as single chain precursors, which are subsequently cleaved into the surface subunit (SU) and the transmembrane subunit TM [15]. TLV/ENV coat polyprotein is a feature of enveloped viruses such as Human immunodeficiency virus 1, influenza virus, and Ebola virus sp express a surface glycoprotein that mediates both cell attachment and fusion of viral and cellular membranes. The env polyprotein contains two coat proteins which differ depending on the source [16]. Protein 3D structures can give an insight into how proteins interact and localize in their stable conformation. Homology modeling is one of the most common structure prediction methods in structural biology. Despite minimal modifications, one initial step that is common in all modeling tools and servers is to find the best matching template by performing a sequence homology search with BLASTP. Templates are experimentally determined 3D structures of proteins that share sequence similarity with the query sequence. The template sequence and the query protein sequence are aligned using pair-wise alignment algorithms [17]. A well defined alignment is very crucial for the prediction of a reliable 3D structure using homology modeling. Here we used the protocol of homology modeling to model env protein which was not modeled till now. A BLASTP search against the PDB database identified 1AOL_A as a best template as it shows maximum identity to query sequence. The identity between query and template was found to be 77.53%. 1AOL_A is an X-Ray diffraction model of a Murine Leukemia Virus Receptor-binding Domain. 3D structures were constructed using MODELLERv9.15, Phyre 2 and Swiss model Homology methods. Various models were obtained and the best model which was common by all the three tools was selected based on discrete optimized protein energy (Figure 4). The Z score is indicative of overall model quality and is used to check whether the input structure is within the range of scores typically found for native proteins of similar size. Z-scores of the query model was obtained from PROSA web (Figure 5A, B and C). The Z-Score of predicted model is -5.65. After modeling the 3D structure, we validated the 3D structure of the target protein using SAVES online software tool https://services.mbi.ucla.edu (took help of PROCHECK and Ramachandran plot-Figure and we found 100% of the residues had an averaged 3D-1D score >= 0.2, (Pass) and at least 80% of the amino acids have scored >= 0. 2 in the 3D/ 1D profile, thus passing the criteria of verifying 3D predicted structure. Finally, the Ramachandran plots were obtained for the homology model for quality assessment ((Figure 6). After verifying 3D structure, we checked the ligand binding sites on this env protein using metapocket online tool. We observed top three metapocket binding sites which are responsible for potential inhibitor target therapy (Figure 7A&B).

```
MDTRRPRQGSDHTPDKTIMESTTLSKPFKNQVNPWGPLIVLLILGGVNPVALGNSPHQVFNLSWEVTNGG
LETVWAITGNHPLWTWWPDLTPDLCMLALHGSYWGLEYRAPFSPPPGPPCCSGSNDSTSGCSRDCEEPLT
SYTPRCNTAWNRLKLSKVTHAHNEGFYVCPGPHRPRWARSCGGPESFYCASWGCETTGRASWKPSSSWDY
ITVSNNLTADQATPACKGNKWCNSLTIRFTSFGKQATSWVTGHWWGLRLYVSGHDPGLIFGIRLKITDLG
PRVPIGPNPVLSDRRPPSRPRPTRSPPSSNSTPTETPLTLPEPPPAGVENRLLNLVKGAYQALNLTSPDK
TQECWLCLVSGPPYYEGVAVLGTYSNHTSAPANCSVASQHKLTLSEVTGQGLCIGAVPKTLQVLCNTTQK
TSTGSYYLAAPTGTIWACSTGLTPCISTTILNLTTDYCVLVELWPRVTYHSPSYVYHQFERRAKYKREPV
SLTLALLLGGLTMGGIAAGVGTGTTALVATQQFQQLQAAMHDDLKEVEKSITNLEKSLTSLSEVVLQNRR
GLDLLFLKEGGLCAALKEECCFYADHTGLVRDSMAKLRERLSQRQKLFESQQGWFEGLFNRSPWFTTLIS
TIMGPLIILLLILLFGPCILNRLVQFIKDRISVVQALVLTQQYHQLKSIDPEEVESRE
```

Fig 1: Primary sequence of env protein

4. Conclusion

We used homology modeling tools/software to generate 3D structure of Env protein which is an important protein, from Retroviridae. We have also done functional analysis of the 3D env protein and also checked the functional domains. The predicted information is hoped to help in better understanding of the mechanisms and the production of novel therapeutic drugs for the treatment of long term diseases. Docking experiments may suggest potential lead compounds and medicinally significant conformations. Ultimately, the identification of pharmacologically active conformations via simulation will be great leap towards use of the new generation drugs or modern medicines.

Table 1: Physicochemical properties of env protein by Protparam tool

Protein name	env
Accession number	CAA41747.1
No. Of amino acid	688
Molecular weight	75877.73
pI (Isoelectric point)	8.49
+ Residues	53
- Residues	36
Atomic formula	$C_{3390}H_{5285}N_{927}O_{991}S_{31}$
Total no. Of atoms	10624
Ext. coefficient	146800
Instability index	40.12
Aliphatic index	82.47
GRAVY	-0.225

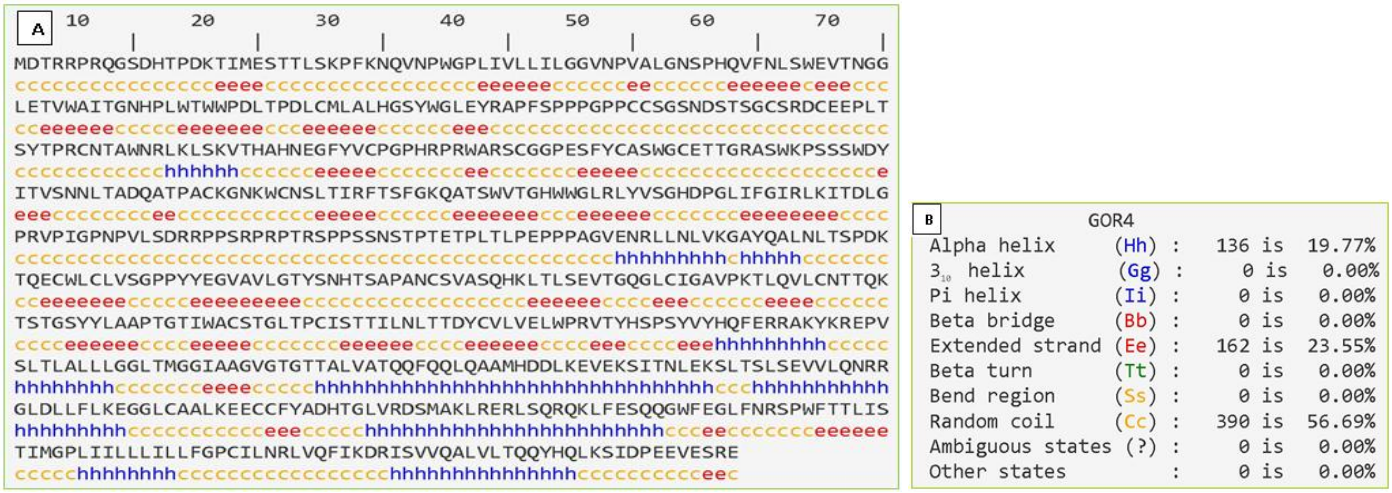


Fig 2(A&B): Secondary structure analysis of Env protein

LOCALIZATION	RELIABILITY	
OuterMembrane	0.333	
OuterMembrane	0.443	
Extracellular	0.373	
OuterMembrane	0.809	
OuterMembrane	0.645	
OuterMembrane	2.580	*
Extracellular	1.012	
Periplasmic	0.556	
InnerMembrane	0.539	
Cytoplasmic	0.313	

Fig 3: Sub cellular localization of env protein

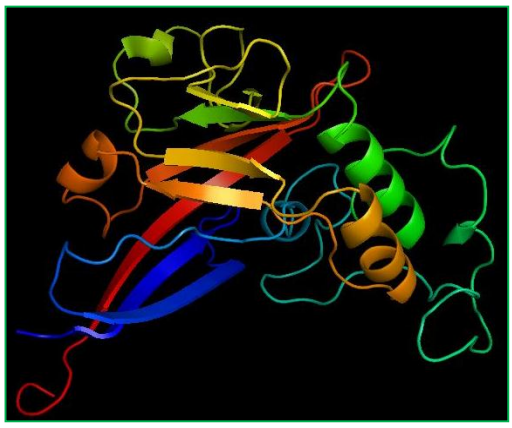


Fig 4: Predicted 3D structure of env protein

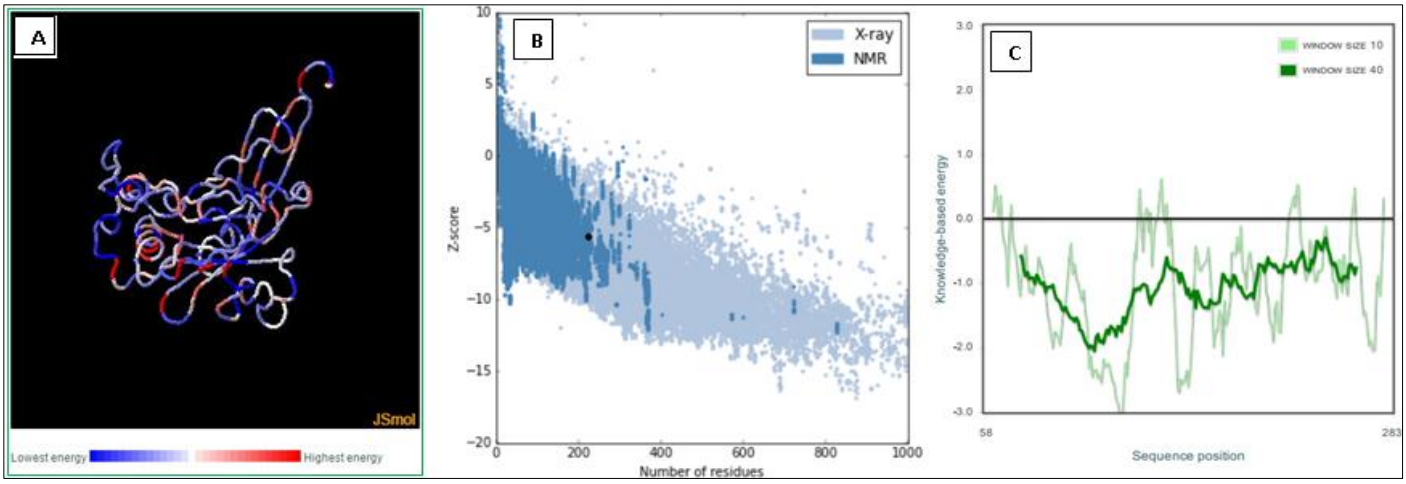
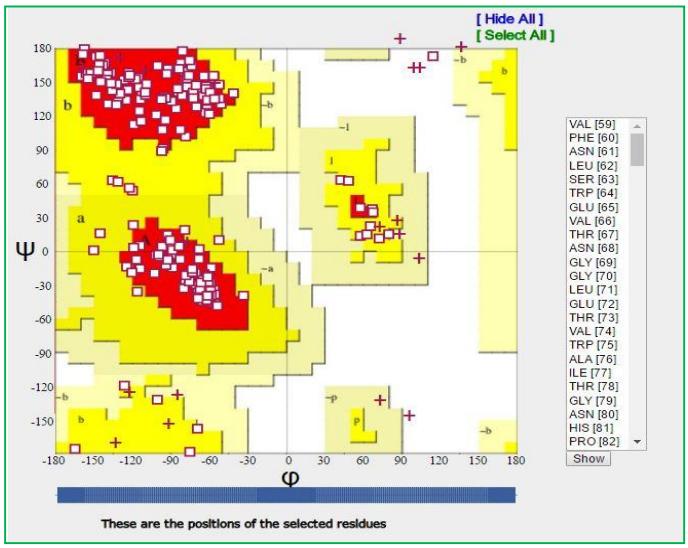
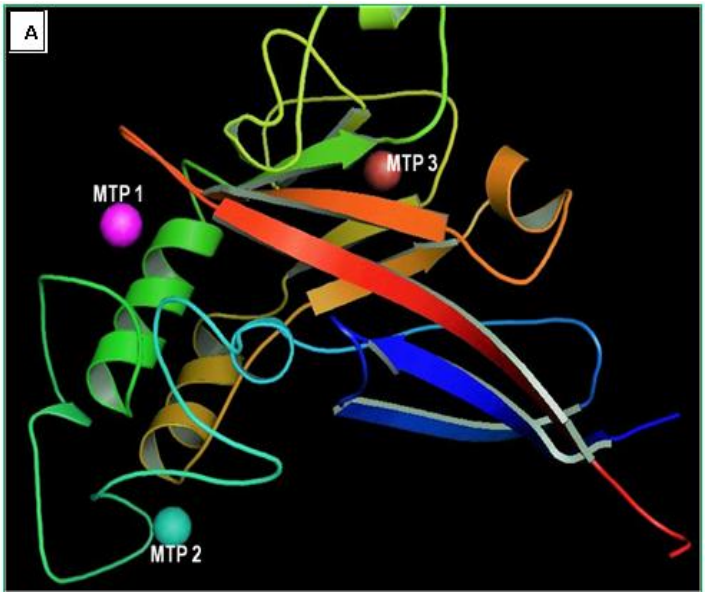


Fig 5 A, B and C: Z score model showing lowest and highest energy




```
+----------<<< P R O C H E C K    S U M M A R Y >>>----------+
|                                                            |
| model_01.pdb   1.5  A                          226 residues |
|                                                            |
*| Ramachandran plot:   84.7% core   14.2% allow   0.5% gener   0.5% disall |
|                                                            |
+| All Ramachandrans:    6 labelled residues (out of 224)    |
+| Chi1-chi2 plots:      1 labelled residues (out of 122)    |
|                                                            |
| Main-chain params:    6 better     0 inside     0 worse    |
| Side-chain params:    5 better     0 inside     0 worse    |
|                                                            |
*| Residue properties: Max.deviation:     7.1          Bad contacts:    0 |
*|                     Bond len/angle:    5.7    Morris et al class:  1  2  2 |
|                                                            |
| G-factors           Dihedrals:  -0.34  Covalent:  -0.13    Overall:  -0.24 |
|                                                            |
| M/c bond lengths: 99.5% within limits   0.5% highlighted   |
| M/c bond angles:  92.4% within limits   7.6% highlighted   |
*| Planar groups:    83.3% within limits  16.7% highlighted       2 off graph |
|                                                            |
+------------------------------------------------------------+
  + May be worth investigating further.  * Worth investigating further.
```

Fig 6: Ramachandran plot and procheck summary



B

```
HEADER binding site ID: 1
RESI   TYR_A^103^  PRO_A^266^  THR_A^143^  PRO_A^144^  ARG_A^145^
RESI   HIS_A^264^  VAL_A^261^  SER_A^262^  THR_A^140^  TYR_A^142^
RESI   LYS_A^154^  TRP_A^104^  LEU_A^259^  GLY_A^267^  PHE_A^187^
RESI   ARG_A^258^  ASP_A^265^  SER_A^102^  SER_A^141^  TYR_A^260^
RESI   GLY_A^263^  VAL_A^158^  GLU_A^164^
HEADER binding site ID: 2
RESI   ASN_A^125^  GLY_A^228^  GLY_A^123^  SER_A^124^  ASP_A^126^
RESI   LYS_A^227^  CYS_A^120^  PRO_A^119^  LYS_A^230^  PRO_A^224^
RESI   THR_A^128^  ALA_A^225^  ASN_A^229^  SER_A^122^  SER_A^127^
RESI   CYS_A^121^  CYS_A^131^
HEADER binding site ID: 3
RESI   THR_A^196^  THR_A^197^  SER_A^205^  SER_A^206^  ILE_A^211^
RESI   THR_A^212^  THR_A^240^  ASP_A^209^  SER_A^207^  PRO_A^204^
RESI   LYS_A^203^  GLY_A^198^  VAL_A^213^  SER_A^214^  ARG_A^199^
RESI   SER_A^201^  ALA_A^200^  ASP_A^89^   THR_A^236^  ARG_A^238^
```

Fig 7A and B: A) Env protein showing top three metapockets and B) binding site residues

5. References

1. Alice Telesnitsky. Retroviruses: Molecular Biology, Genomics and Pathogenesis Future Virol. 2010; 5(5):539-543.
2. Checkley MA, Luttge BG, Freed EO. HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation. J. Mol. Biol. 2011; 410:582-608.
3. Haffar OK, Dowbenko DJ, Berman PW. Topogenic analysis of the human immunodeficiency virus type 1 envelope glycoprotein, gp160, in microsomal membranes. J. Cell Biol. 1988; 107:1677-1687.
4. Kim FJ1, Battini JL, Manel N, Sitbon M. Emergence of vertebrate retroviruses and envelope capture. Virology. 2004; 318(1):183-91.
5. Belshaw R1, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A *et al.* Long-term reinfection of the human genome by endogenous retroviruses.Proc Natl AcadSci U S A. 2004; 101(14):4894-9.
6. Medstrand P1, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, Mager DL. Impact of transposable elements on the evolution of mammalian gene regulation. Cytogenet Genome Res. 2005; 110:342-52.
7. Wilkins MR1, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD *et al.* Protein identification and analysis tools in the ExPASy server. Methods Mol Biol. 1999; 112:531-52.
8. GOR secondary structure prediction method version IV, J. Garnier, J.-F. Gibrat, B. Robson, Methods in Enzymology, R.F. Doolittle Ed. 1996; 266:540-553.
9. Yu CS1, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins. 2006; 64(3):643-51.
10. Quevillon E1, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R *et al.* Inter Pro Scan: protein domains identifier. Nucleic Acids Res. 2005, 33. (Web Server issue):W116-20.
11. Altschul SF1, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389-402.
12. Sali A1, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993; 234(3):779-815.
13. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007, 35. (Web Server issue):W407-10.
14. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR. 1996; 8(4):477-86.
15. Fass D, Davey RA, Hamson CA, Kim PS, Cunningham JM, Berger JM. Structure of a murine leukemia virus receptor-binding glycoprotein at 2.0 angstrom resolution. Science. 1997; 277(5332):1662-6.
16. Fass D, Harrison SC, Kim PS. Retrovirus envelope domain at 1.7 angstrom resolution. Nat Struct Biol. 1996; 3(5):465-9.
17. Butt AM, Batool M, Tong Y. Homology modeling, comparative genomics and functional annotation of Mycoplasma genitalium hypothetical protein MG_237.Bioinformation. 2011; 7(6):299-303.