Original Research Article

Insilco Study on the Structural Characterization and Inhibitor Detection for Super Small Secreted Glycoprotein of Reston Ebolavirus

Abstract

Super small secreted glycoprotein (ssGP), a recognized as a virulent protein which plays an important role during the Ebola viral infection. This study entails *in silico* structural and finding inhibitor compound for the secreted glycoprotein of Reston ebolavirus (strain Philippines-96). Initially, the physical and stereochemical properties, the secondary and the tertiary structure of the protein were predicted. Later, as the predicted model was evaluated to be a reliable structure, binding pockets were predicted and known binding ligands to similar proteins were identified. Analog compounds to known ligands were collected and docked against ssGP. Compound having least binding energy was identified and recommended as potent inhibitor towards ssGP. *In vitro* and *in vivo* analysis has to be done for the selected ligand from this study for designing effective drugs against Reston ebolavirus.

Keywords: EHF, Reston ebolavirus, Small secreted glycoprotein (ssGP), docking, SHFV

Introduction

Ebola hemorrhagic fever (EHF) or Ebola virus disease (EVD) is caused by the Ebola virus. It came to prominence in the year 1976, near the Ebola River present in the todays Democratic Republic of the Congo (DRC). 2014 Ebola epidemic in West Africa has gained attention globally. The period of incubation for EHF is first 3 weeks and common symptoms include chills, fever, myalgia, and malaise, followed by the onset of symptoms like multi-organ stress and subsequent failure. This fatal disease infected approx. 28,500 people and killed around eleven thousand people in recent years, mostly in West and Central Africa [1]. Since then, EHF has been listed as an endemic disease.

Humans and other primates are infected with the Ebola virus, which is an enveloped negative-stranded RNA virus. The Ebola virus is a member of the Filoviridae family [2]. Fruit bats, which are members of the Pteropodidae family, are primarily thought to be Ebola virus natural hosts [3]. The disease is transmitted in two ways: enzootic and epizootic. Humans contract the disease from wild animals, and it spreads among the human population

through human-to-human contact. Several human infections were investigated among personnel who came into contact with Ebola Reston-infected monkeys or pigs.

Reston ebola virus is one of five Ebola virus species [4], with Reston Ebola virus strain Philippines-96 being the subject of this research. Reston virus is non-pathogenic to humans but dangerous to pigs and monkeys, hence it is classified as a level-4 organism;[5,6] the monkey's coinfection with Simian hemorrhagic fever virus (SHFV) [7] increased the sense of its fatality. Clinical testing on a sample acquired from domestic pigs in the Philippines region yielded positive results [8]. This means that the Reston ebolavirus has a greater chance of infecting humans than previously thought.

The virus has a 19-kilobyte genome and non-segmented negative-sense (NNS) RNA with seven genes. The viral proteins NP, VP35, VP30, and L wrap around the centre of the Ebola virus, which contains a helical single-stranded RNA genome. The nucleocapsid is covered by viral glycoprotein (GP) spikes on the outer viral envelope, and the viral proteins VP40 and VP24 are found between the nucleocapsid and the envelope [9].

The EBOV glycoprotein is a enveloped viral protein and a key component of immunogenicity; it is produced by the GP gene. GP gene also produces two other forms: soluble GP (sGP) and small soluble GP (ssGP) via a co-transcriptional editing mechanism [10, 11]. It's been suggested that sGP and ssGP play a key role in infecting the host [10, 12]. With only 5 percent of GP gene transcript is expected to be specific for ssGP mRNA, after ebolavirus infection ssGP is the second nonstructural glycoprotein generated through RNA editing [13]. ssGP shares a 295 N-terminal aminoalkanoic acid sequence with GP1,2 and sGP, with the C-terminal regions varying [10]. A particular frame shift occurs when two adenosine residues are co-transcribed, or when one adenosine is removed, resulting in a diminutive transcript that encodes ssGP [10]. Three primary transcripts were obtained by Co-transcriptional editing at the GP gene editing site , which are then translated into a variety of glycoprotein products, including pre-sGP, pre-GP, and pre-ssGP. Pre-ssGP, like other glycoproteins, they do not undergo post-translational cleavage but is O-glycosylated.

Since the function of ssGP is still not confirmed and from prior studies as it was noted that ssGP is involved in the pathogenicity of the EVD, this research covers structural and functional characterization of ssGP present in Reston Ebola Virus strain Philippines-96.

2. Methodology

Sequence retrieval and analysis:

The sequence of the ssGP protein was retrieved from the UNIPROTKB database with ID P0C770. It consists of 332 amino acids with a molecular weight of 37495.69 Da. From the database annotation it is known that the structure of this protein is not predicted experimentally and thereby not in structure database. Hence, steps are carried out to predict the 3D structural model of the protein using *in silico* tools.

First, to compute Physico-chemical properties of the ssGP protein, Protparam tool was used. (http://www.expasy.org/tools/protparam)

Secondary and Tertiary structure determination

SOPMA (Self-Optimized Prediction Method with Alignment) is a tool for predicting the secondary structure of ssGP protein based on the primary sequence.

SWISS-MODEL Workspace is a Web-based modelling expert system for homology modelling of three-dimensional protein structures. This tool takes the sequence of ssGP and searches the PDB database for a template structure, then, based on a sequence alignment between the target protein and the template structure, it generates a three-dimensional model for the target protein. The reliability of the resultant models is estimated using model quality evaluation techniques. (https://swissmodel.expasy.org/)

Analysis of modeled Tertiary structure

ProSA is a popular tool for checking 3D models of protein structures for mistakes. It connects a model's z-score to the z-scores calculated from all experimental structures in the PDB. The Z-score represents the departure of the structure's total energy from an energy distribution obtained from random conformations. Also, the model quality is evaluated by plotting energies across amino acid sequence positions. The modeled structure of ssGP is submitted to this tool for evaluation.

(https://prosa.services.came.sbg.ac.at/prosa.php)

The Ramachandra plot is used in PROCHECK to assess the stereochemical quality of the protein structure. The torsional angles phi () and psi () of the individual residues of a peptide are plotted in the Ramachandran plot. The favored, allowed and disallowed regions for torsion angle values are pre-marked on the plot based on experimentally solved protein structures. Based on the percentage of residues which fall in these regions portray the reliability level of the predicted structure. The modeled structure of ssGP is submitted to this tool to check its stereo chemical quality. (https://prosa.services.came.sbg.ac.at/prosa.php)

Binding Site Prediction

Protein function annotation and drug discovery require the identification of protein binding sites. The COACH-D server is used to anticipate the ssGP binding location. It shows the top five protein-ligand complex structures that are comparable to the three-dimensional structure of ssGP, the query protein. Each complex's ligand as well as the target binding site residues are known. (https://yanglab.nankai.edu.cn/COACH-D/)

Virtual screening of similar ligands:

Compounds similar to Coach-D server ligands were screened and information taken from PubChem db. (https://pubchem.ncbi.nlm.nih.gov/).

Docking analysis:

iGemDock server is a docking program. It screens the given compounds against the given therapeutic target and visualizes and ranks them based on pharmaceutical interactions and a score system based on energy. It creates a profile of electrostatic, vander walls, and hydrogen bonding interactions between proteins and their compounds (14, 15).

The compounds collected from PubChem database and the known ligands were docked against the target protein ssGP using iGemDock server. (http://gemdock.life.nctu.edu.tw/). (16)

3. Results And Discussion

Primary structure analysis:

The properties of the ssGP protein are predicted from its primary sequence using Protparam tool. It was predicted to have 32 negatively charged residues and 40 positively charged residues and the extinction coefficients is around 59275 M^{-1} cm^{-1}, at 280 nm.The estimated half-life is 30 hours. The instability index is computed to be 31.94 and hence the protein is classified as stable protein. The aliphatic index is predicted as 74.85 and the grand average of hydropathicity is -0.448.

Secondary structure analysis:

From SOPMA tool it was predicted that the protein has an alpha helix of 18.67% and extended strands are 22.89%. This indicates that the protein falls under alpha & beta class of proteins.

Tertiary Structure prediction:

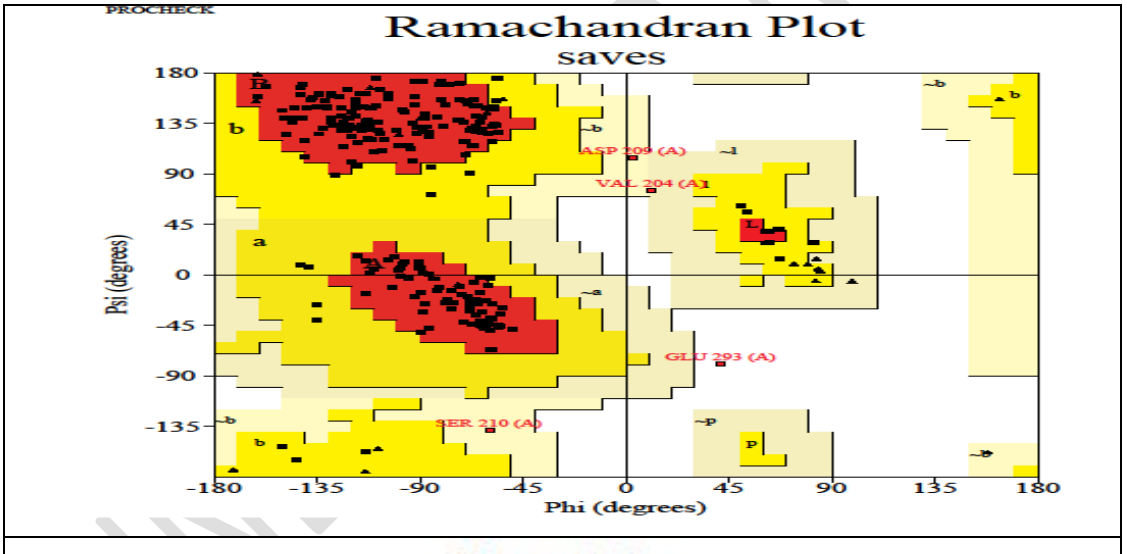
The tertiary structure of the ssGP was predicted using the SWISS Model and the resulting protein structure had the Q-Mean Value of 0.71(+/-0.05) indicating the high probability that the modeled structure is close to the experimental Structure of similar size. The template used

by the tool for modeling is Envelope glycoprotein from Ebola Virus (PDB ID: 7JPI) which is having sequence identity of 68.81% with the protein sequence.

Fig. 1 3D Model of ssGP protein

Evaluation of the modeled structure:

Structural Evaluation and Stereo Chemical Analysis of the ssGP tertiary structure model was carried out using Pro-Check and ProSA-web servers.



Plot statistics

Residues in most favoured regions [A,B,L]	200	88.9%
Residues in additional allowed regions [a,b,l,p]	21	9.3%
Residues in generously allowed regions [~a,~b,~l,~p]	3	1.3%
Residues in disallowed regions	1	0.4%
Number of non-glycine and non-proline residues	225	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	20	
Number of proline residues	17	
Total number of residues	264	

Fig. 2 Ramachandran plot of 3D Structural model of ssGP protein

In the ProCheck Ramachandran Plot (Fig. 2) as more than 98% of the residues lie in the favoured and allowed regions and only a single residue is seen in the disallowed region, the predicted structural model can be accepted.

From 2 plots displayed by ProSA-web server, one which indicates z-score of the model and the other which shows the energy distribution across the residues of ssGP structure, the reliability of the model can be known.

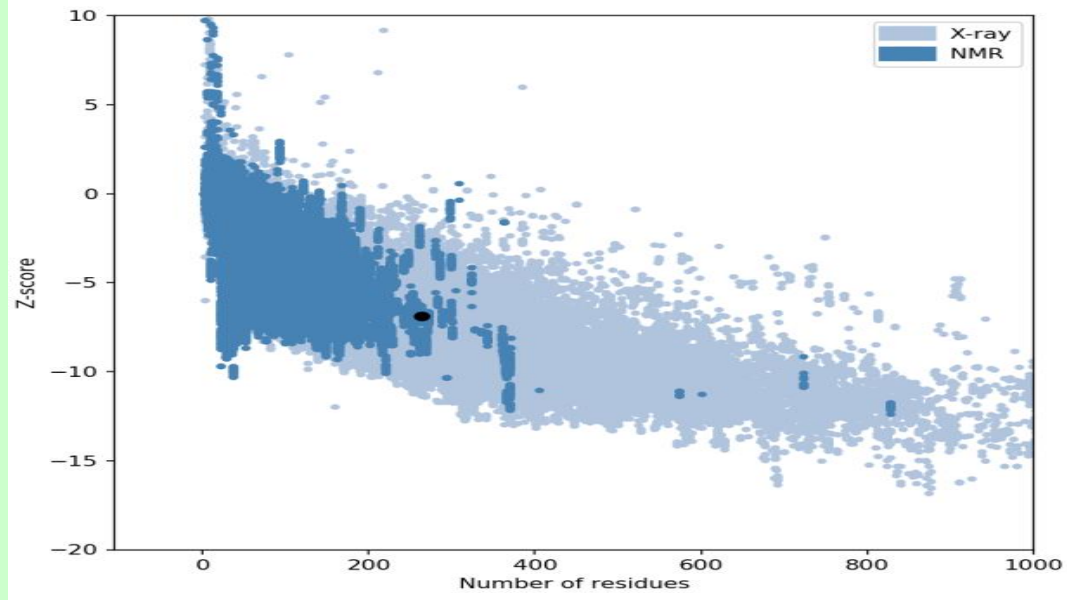


Fig. 3 Z-Score Plot, where the black dot indicates score of Structural model of ssGP protein.

The Z-score of the predicted model, depicted in the first plot of z score v/s number of residues of the experimentally determined structures (Fig. 3), is -6.88 which shows that the protein model's overall quality is high is close to experimentally determined structures and hence is reliable.

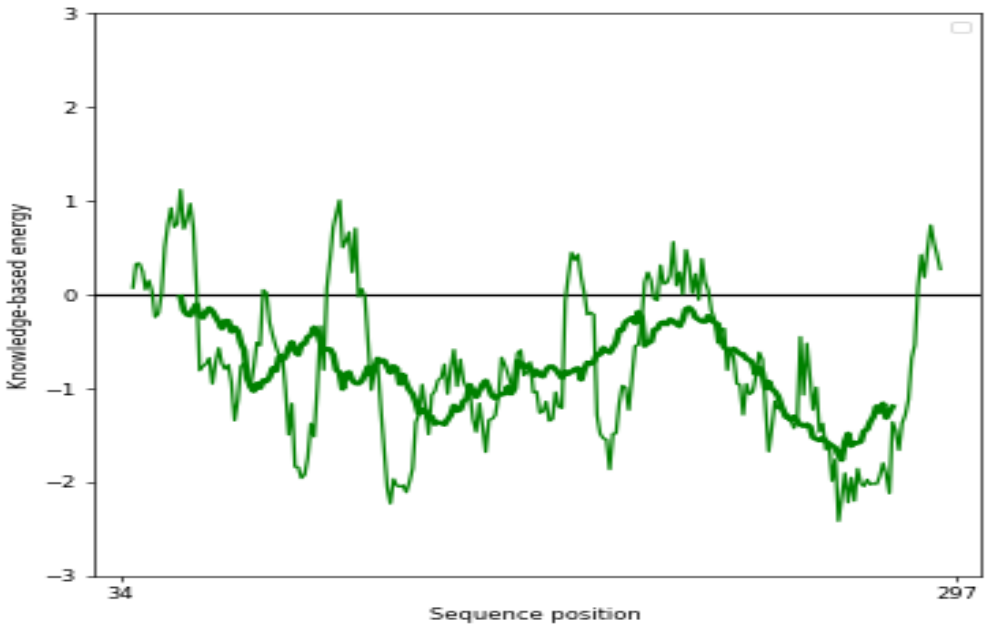


Fig 4 In the structural model of the ssGP protein, the plot shows the energies as a function of amino acid sequence position.

The second plot (Fig 4) plots energy as a function of amino acid sequence position to indicate local model quality. The thin line denotes a smaller window of 10 residues shows that both the ends of the protein as well as three smaller inner regions have positive energy but generally as this single residue energies are of limited value for model evaluation, average energy for stretch of residues are calculated. The thick line which specifies the average energy calculated over each 40-residue fragment moving window shows that all the residues have negative energy which denotes that the modeled structure is a stable one.

Binding site prediction:

Protein-ligand binding site is predicted through Coach-D. From the results (Table 1), five similar structure clusters with bound ligand and the binding site residues were known. As the 1st,2nd and 4th PDB templates have DNA and forms of DNA as ligands they are not considered and only the 3rd cluster is considered.

Table 1 Binding site prediction summary from COACH-D

Rank	C-score	Cluster size	PDB template	Ligand	Energy	Predicted binding residues
1	0.51	7	3csyl	k-mer	-1.8	38,124,125,126,148
2	0.09	2	3csyl	k-mer	-2.8	222,225,229
3	**0.07**	**2**	**5jq7A**	**T0R**	**-1.7**	**32,34,68,69,152,154**

4	0.05	2	1e3mB	QNA	-1	121,123,128

For the 3rd cluster, the template structure is the Crystal structure of Ebola glycoprotein in complex with toremifene ligand (Fig 5). The binding energy and the binding residues of the target to the ligand are known from Table 1. Toremifene is a nonsteroidal triphenylethylene antiestrogen. The PubChem CID of this compound is 3005573.

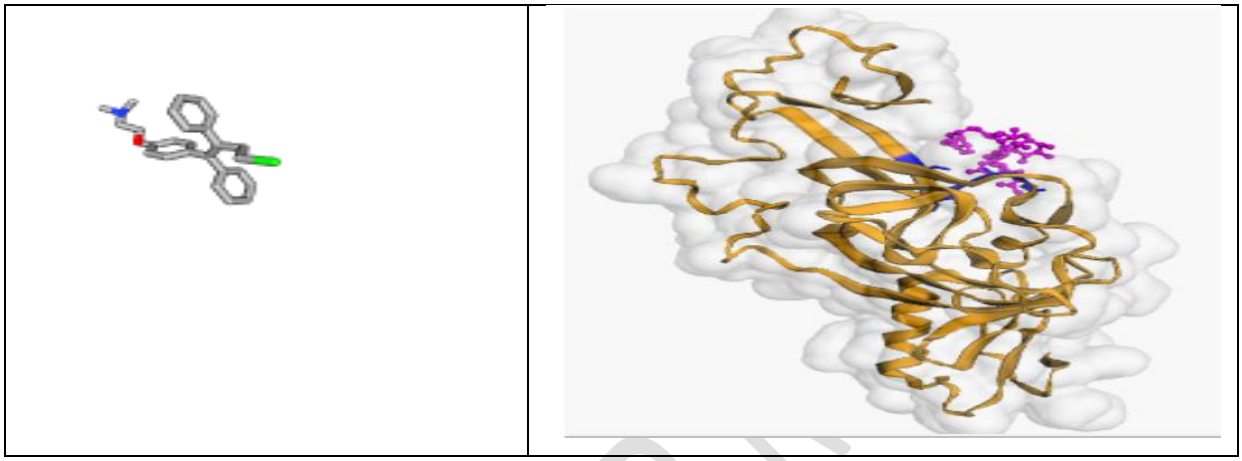


Fig 5. a) toremifene ligand b) Complex structure of toremifene with the target ssGP protein (binding site is the blue region in the ssGP protein)

Compounds which have similar structure to Toremifene were retrieved from PubChem database.

Molecular Docking

Using iGemDock, 21 compounds collected from PubChem along with Toremifene were docked against the target protein ssGP and the binding energy is noted (Table. 2). Of the submitted compounds, the ligand N-acetyllactosamine (PubChem CID 439271) has the least binding energy of -97.49 kcal/mol.

Table 2 Molecular docking results showing binding energy and energy distribution of ssGP-ligand complexes

		Compound	Energy	VDW	HBond	Elec
1	☑	model_01-899-1.pdb	-65.2	-27.77	-37.43	0
2	☑	model_01-2800-0.pdb	-82.48	-78.78	-3.7	0
3	☑	model_01-5376-1.pdb	-77.66	-74.16	-3.5	0
4	☑	model_01-5516-1.pdb	-88.62	-88.62	0	0
5	☑	model_01-82313-0.pdb	-69.21	-45.51	-23.7	0
6	☑	model_01-84265-2.pdb	-71.61	-53.08	-18.53	0
7	☑	model_01-108092-0.pdb	-71.03	-67.53	-3.5	0
8	☑	model_01-439174-2.pdb	-77.04	-57.49	-19.55	0
9	☑	model_01-439271-1.pdb	-97.49	-70.27	-27.22	0
10	☑	model_01-439281-0.pdb	-73.73	-51.92	-21.81	0
11	☑	model_01-439544-0.pdb	-97.24	-72.83	-24.41	0
12	☑	model_01-440552-0.pdb	-71.63	-40.83	-30.8	0
13	☑	model_01-644170-2.pdb	-75.92	-55.1	-20.83	0
14	☑	model_01-1458955-2.pdb	-82.46	-78.87	-3.6	0
15	☑	model_01-1548953-2.pdb	-85.86	-85.86	0	0
16	☑	model_01-2733526-0.pdb	-72.23	-69.75	-2.49	0
17	☑	model_01-3032583-0.pdb	-86.22	-81.62	-4.6	0
18	☑	model_01-3035198-1.pdb	-78.25	-72.66	-5.59	0
19	☑	model_01-6152516-2.pdb	-80.66	-77.63	-3.03	0
20	☑	model_01-11096158-1.pdb	-73.35	-45.47	-27.87	0
21	☑	model_01-NAG 24139-2.pdb	-77.04	-57.72	-19.32	0
22	☑	model_01-T0R 3005573-1.pdb	-81.32	-81.32	0	0

The target protein ssGP residues which bind to the the ligand N-acetyllactosamine are LYS 65, GLY 103, ARG 165, GLU 202, THR 207, GLU 101, TRP 195 (Fig 6).

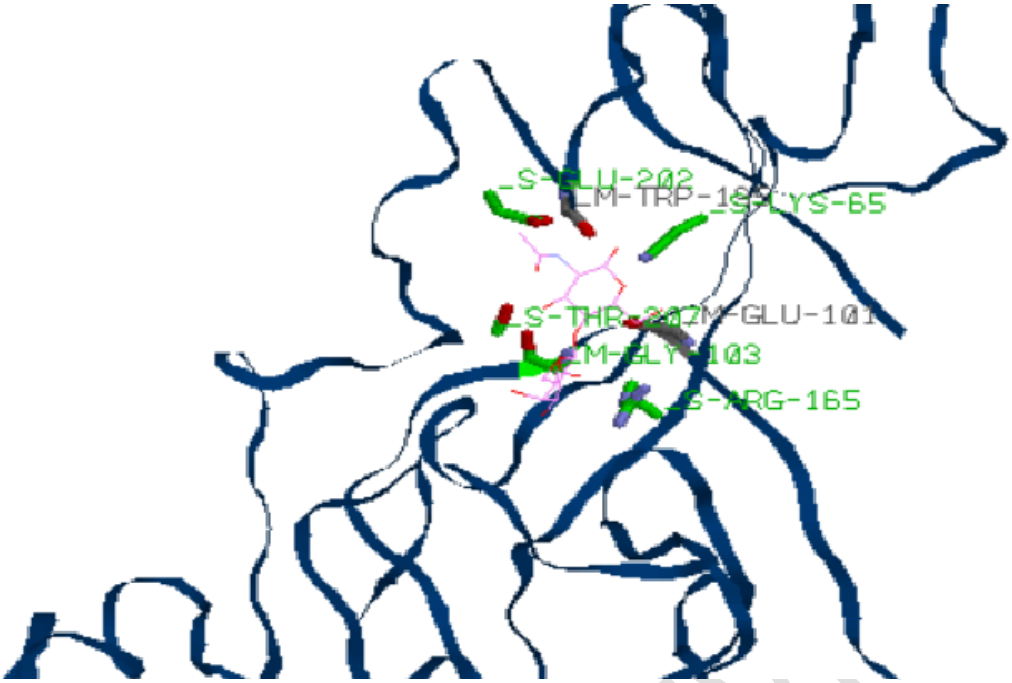


Fig 6. Complex structure of ssGP & N-acetyllactosamine ligand highlighting ssGP residues which interact with the ligand.

CONCLUSION:

Ebola virus disease is a rare zoonotic disease with great epidemiological risk factors. There is a growing concern that Reston ebolavirus though now shows only asymptomatic infections in humans may anytime turn pathogenic to humans. It causes severe disease in monkeys and pigs. Understanding its structure, mode of infection, adaptions to enter the host cell, structure and function of its proteins are all needed to prevent any subsequent widespread of this disease.

This paper includes studies related to structural characterization and identifying inhibitor of small secreted glycoprotein (ssGP) of Reston ebolavirus strain Philippines-96 using *in silico* analysis. The ssGP had 332 amino acids and the molecular weight is 37495.69 and its predicted instability index is 31.94 which classifies the molecule as stable protein. Secondary structure analysis depicts the number of alpha-helix and extended strands to be 18.67% and 22.89% respectively indicating that it belongs to alpha and beta protein class. The homology modeling is done through the Swiss model and the modeled structure is evaluated by ProCheck server and ProSA-web server. As nearly 98% residues lie in favoured and allowed regions of Ramachandran plot and the z score value is -6.88 for the local quality of the model it is known that the modeled structure is reliable and close to the experimentally determined structures. Binding site predictions are done through Coach-D and an inhibitor toremifene was identified. Analogs to toremifene were retrieved from PubChem database and including

toremifene and all 22 compounds were docked against ssGP in iGemDock sever. N-acetyllactosamine (PubChem CID 439271) compound had the least binding energy of -97.49 kcal/mol and hence it can be considered as a potent inhibitor to ssGP. The inhibiting capacity of this compound can be increased by lead optimization techniques and its pharmacological characteristics has to be investigated and this compound can be considered in designing anti-viral drugs against Reston Ebola virus.

References:

1 Formenty, Pierre. "Ebola Virus Disease". Emerging Infectious Diseases, 2014, pp. 121-134. Elsevier, doi:10.1016/b978-0-12-416975-3.00009-1.

2 Saxena, Aaruni, and Mauricio Ferri. "Clinical Management Of Ebola Virus Disease: Current And Future Approaches". *Topics In Medicinal Chemistry*, 2015, pp. 1-36. *Springer International Publishing*, doi:10.1007/7355_2015_5003.

3 Artika, I Made et al. "Pathogenic Viruses: Molecular Detection And Characterization". *Infection, Genetics And Evolution*, vol 81, 2020, p. 104215. *Elsevier BV*, doi:10.1016/j.meegid.2020.104215.

4 "What Is The Ebola Virus?". Council On Foreign Relations, 2021, https://www.cfr.org/backgrounder/what-ebola-virus.

5 *Special Pathogens Branch CDC (2008-01-14). "Known Cases and Outbreaks of Ebola Hemorrhagic Fever". Center for Disease Control and Prevention. Archived from the original on 2008-08-29. Retrieved 2008-08-02.*

6 McCormick & Fisher-Hoch 1999, p. 300

7 McCormick & Fisher-Hoch 1999, pp. 307–309

8 Editorial Team. (2009). Ebola Reston virus detected pigs in the Philippines. Euro Surveill. 14 pii=19105.

9 Outbreak news. (2009). Ebola Reston in pigs and humans, Philippines. Wkly. Epidemiol. Rec. 84, 49–50

10 Mehedi, M.; Falzarano, D.; Seebach, J.; Hu, X.; Carpenter, M.S.; Schnittler, H.J.; Feldmann, H. A new Ebola virus nonstructural glycoprotein expressed through RNA editing. J. Virol. 2011, 85, 5406–5414. [CrossRef]

11 Sanchez, A.; Trappier, S.G.; Mahy, B.W.; Peters, C.J.; Nichol, S.T. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc. Natl. Acad. Sci. USA 1996, 93, 3602–3607. [CrossRef] [PubMed]

12 Iwasa, A.; Shimojima, M.; Kawaoka, Y. sGP serves as a structural protein in Ebola virus infection. J. Infect. Dis. 2011, 204 (Suppl. 3), S897–S903. [CrossRef] [PubMed]

13 Feldmann, H.; Volchkov, V.E.; Volchkova, V.A.; Stroher, U.; Klenk, H.D. Biosynthesis and role of filoviral glycoproteins. J. Gen. Virol. 2001, 82 Pt 12, 2839–2848. [CrossRef]

14 Chandramohan, V., Nagaraju, N., Rathod, S. *et al.* Identification of Deleterious SNPs and Their Effects on Structural Level in CHRNA3 Gene. *Biochem Genet* **53,** 159–168 (2015). https://doi.org/10.1007/s10528-015-9676-y

15 Shanmuga Priya, V.G., Swaminathan, P., Muddapur, U.M. *et al.* Peptide Similarity Search Based and Virtual Screening Based Strategies to Identify Small Molecules to Inhibit CarD–RNAP Interaction in *M. tuberculosis*. *Int J Pept Res Ther* **25,** 697–709 (2019). https://doi.org/10.1007/s10989-018-9716-7

16 Priya VG, Muddapur U, Mehta M (2012) Computational analysis of *M. tuberculosis*–CarD protein. Adv Life Sci Technol 6:8–15