Journal of Medicine and Biology

**Research Article** | Vol 2 Iss 2

# In-silico Prediction and Docking of Tertiary Structure of Protein X, Multifunctional Proteins of Rabies Virus

**Tanmay Verma*, Dhara Patel, Sharav Desai and Dhananjay Meshram**

Department of Quality Assurance, Pioneer Pharmacy Degree College, Vadodara, India

***Corresponding author:** Tanmay Verma, Department of Quality Assurance, Pioneer Pharmacy Degree College, Vadodara, Gujarat-390019, India, Tel: +919099674397; E-mail: patel.dhara.j@gmail.com

## Abstract

**Rabies is a virus belonging to the genus of Lyssavirus and family of Rhabdoviridae. These virus can infect wide range of host including vertebrates, invertebrates, and plants. Infection with these virus causes mental confusion, anxiety, preserve intelligence, autonomic dysfunction hydrophobia and aerophobia periods of agitation, hyper excitation or drowsiness. Rabies virus mainly contain five protein namely nucleoprotein phosphoprotein, glycol protein large structural protein and matrix protein. Nucleoprotein encapsides the genome in ratio of 1 protein N per nine ribonucleotides protecting it from nucleus so it is important for organism. Sequence of nucleoprotein was obtain from uniprot database. By submitting FASTA sequence of nucleoprotein at protparam, physiochemical property are found homology modeling was carried out using phyre and server and it is refined by galaxy web generated structures were evaluated by Errat, Qmean and Procheck from the result model two was selected for nucleoprotein having best quality ligands were obtain from protein data bank using model two and ligands docking is done using PyRx autodock vina, most suitable ligands were found which can be bind to model with highest energy these data of ligands interaction can be used to design new drug for rabies.**

**Keywords**: ***Lyssavirus; FASTA; Errate; Qmean; Docking***

**Received Date:** May 11, 2020; **Accepted Date:** May 22, 2020; **Published Date:** May 29, 2020

## Introduction

Rabies virus belongs to the genus of Lyssavirus and the family of Rhabdoviridae. The family Rhabdoviridae consists of more than 100 single-stranded, negative-sense, non-segmented viruses that infect a wide variety of hosts, including vertebrates, invertebrates, and plants. Viruses like lagos bat, mokola, duvenhage and two subtypes of European bat lyssaviruses are antigenically and genetically similar with the rabies virus and also belongs to the genus Lyssavirus [1]. Dogs and Bats are the main reservoir of infection and is responsible for over 90% of deaths. All age groups are exposed to Rabies but the predominant group (50%) is 6 years - 15 years. This virus can be found in all countries, but Human mortality is highest in Asia and Africa [2]. Males are four times more common victims than females [3]. The state of India like Meghalaya, Manipur, Sikkim, Arunachal Pradesh, Nagaland, Dadra and Nagar Haveli have reported only occasional

**Citation:** Tanmay Verma, In-silico Prediction and Docking of Tertiary Structure of Protein X, Multifunctional Proteins of Rabies Virus. J Med Biol 2(2): 65-71.

hydrophobia deaths, while a considerable number is reported from Uttar Pradesh, West Bengal, Tamil Nadu, Andhra Pradesh and Maharashtra [3]. The clinical course of human rabies virus infection can be separated into five stages: Incubation, prodrome, neurological stage, coma, and death. Majority of cases are of Encephalitic rabies characterized by Hyperactivity appearing as nervousness, mental confusion and anxiety, preserved intelligence. Typical encephalitic rabies is identified by autonomic dysfunction, hydrophobia and aerophobia, and periods of agitation, hyper excitation, confusion, hyperactivity, and drowsiness. During paralytic rabies in humans, seizures are common and fever is usually high and constant. The pathology of rabies virus infection in the central nervous system has been reviewed and the only gross pathological lesion is congestion of the meningeal vessels; a mild cerebral oedema is also observed. Some of the histological features of rabies virus infection are 1) A perivascular accumulation of leukocytes, primarily in the spinal cord and brainstem. 2) Neuronal degeneration and neuronophagia, and 3) Glial proliferation [4,5]. The rabies virus is measuring just about 600Å × 1800Å. It is composed of an internal protein core or nucleocapsid, containing the nucleic acid, and an outer envelope, a lipid-containing bilayer covered with trans-membrane glycoprotein spikes. The virus genome encodes five proteins associated with either the ribonucleoprotein (RNP) complex or the viral envelope. The L (transcriptase), N (nucleoprotein), and NS (transcriptase-associated) proteins comprise the RNP complex, together with the viral RNA. These aggregate in the cytoplasm of virus-infected neurons and compose negri bodies, the characteristic histopathologic finding of rabies virus infection. The M (matrix) and G (glycoprotein) proteins are associated with the lipid envelope. The G protein forms the protrusions that cover the outer surface of the virion envelope and is the only rabies virus protein known to induce virus-neutralizing antibody [2]. Encapsidates the genome in a ratio of one protein N per nine ribonucleotides, protecting it from nucleases. If expressed without protein P it binds non-specifically RNA and therefore can bind its own mRNA. Interaction with protein P abolishes any non-specific RNA binding, and prevents phosphorylation. The soluble N-P complex encapsidates specifically the genomic RNA, with protein N protecting the genome like a pearl necklace. The encapsidated genomic RNA is termed the nucleocapsid (NC) and serves as template for viral transcription and replication. Protein N binds protein P in the NC through a different interaction, and can be phosphorylated. Subsequent viral replication is dependent on intracellular concentration of newly synthesized protein N. During replication, encapsidation by protein N is coupled to RNA synthesis and all replicative products are resistant to nucleases [6].

## Material and Methods

### Sequence retrieval, physiochemical properties, and secondary structure

UniProt make available the scientific community with a complete, superior and freely accessible source of protein sequence and functional information. UniProt database was helpful in obtaining the sequence of amino acids of Nucleoprotein.

Initially, web side was opened and the name of organism (Rabies Virus) was added to search box. Search was restricted to only the proteomes by selecting the proteomes before the search. Reference strand UP000008649 was used from which the sequence was retrieved. The sequence of Nucleoprotein with P06025 id was downloaded in FASTA form for further use. Then the sequence was copy to the ExPasy portal. ExPasy is the portal controlled by the Swiss Institute of Bioinformatics (SIB). It provides the wide range of data about the life science. It was mainly used as the protein analysis tool. By copied FASTA sequence it can provide the various parameters includes the number of amino acids and total number of atoms, molecular weight, amino acids and atomic composition, Number of negatively and positively charged residues. Also shows extinction coefficients, half-life, instability index, aliphatic index and finally grand average of hydrophobicity [7-9].

Scratch server (SSpro) was used for prediction of secondary structure of nucleoprotein. This server managed by the Donald Bren School of Informatics and Computer Science, California. Secondary structure prediction is based on protein information (sequence homology) and homologous protein's secondary structure (Structure homology). It also includes relative solvent accessibility, disordered regions, domain, disulfide bridges, single mutation stability, residue contacts versus average, individual residue contact and tertiary structure and many more useful tools. Amino acid sequence of nucleoprotein is submitted in FASTA format to server SSpro8. SSpro8 is advance version on SSpro. Instead of using three classes (Helix, strand and the rest) to assign the secondary structure of a protein, SSpro8 adopts the full DSSP (Dictionary of protein secondary structure prediction) 8-class output classification [10-12].

**Homology modelling, refinement and evaluation of the 3D structure**

Phyre2 was used for forecasting and building of protein structure. It is the network operated by the Imperial College of London [13]. On this network, sequence was submitted along with email address and job description. There are two modes available for the modelling; Intensive and normal. Intensive mode is selected for the construction of structure.

The prototypes produced by Phyre are based on finding a sequence alignment to a known structure, copying the coordinates and relabeling the residues according to submitted sequence based on the alignment. It uses profiles or PSSMs (Position-specific scoring matrix) generated by PSI-Blast (Position-specific iterative basic local alignment search tool) for both submitted sequence and the sequences of the recognized structures. Phyre performs a profile-profile matching algorithm together with predicted secondary structure matching. The only deviations Phyre creates to the mainstay of the known structure (template) is when modelling insertions or deletions which is done by searching loop library for compatible loops [14]. It is always suitable to achieve better quality of predicted structure because the quality of a structure produced by prediction program depends on the similarity between target and available structure. Galaxy Refine web server is based on a refinement method that has been successfully tested in critical assessment of techniques for protein structure prediction (CASP10), At Initial start it rebuilds side chains and performs side chain repacking and subsequent overall structure relaxation by molecular dynamics simulation improving the local structure quality [15-19].Total five structure are generated from this server. They all are evaluated based on several validation parameters like Z score and Errat quality. ERRAT is algorithm provide and maintain by University of California, USA. It was used to analyze improvements in the model building and refinements studies. Similar Tool PROSA from centre for applied molecular engineering, division, of bioinformatics and university of Salzburg, Austria. The models were submitted to above web tool in form of PDB file and all the calculation were done using $C^{\alpha}$ Potential. The Z score was considered for whole model quality and deviation of total energy according to an energy distribution derived from random conformation. QMEAN6 and ANOLEA score are calculated by using SWISS-MODEL server. ANOLEA score is the atomic empirical mean force, it shows packing quality of the model. Energy calculation were performed on protein chain evaluating non-local environment of each heavy atom in the molecule. Qualitative model energy analysis score for both global and local quality was calculated using the QMEAN6 tool. All prediction calculations were based on propensity scales for each of 20 amino acids. Each scale comprises of 20 values allotted to each of amino acid residues on the basis of their relative propensity to possess the property described by the scale [15-19].

**Protein data bank was used to search ligands**

For the docking, ligands were searched on www.rcsb.org. It is web site on internet which is operates the US data centre. Ligands were downloaded in sdf files. PyRx was used for the docking. Minimized ligands were easy to handle than after they all were converted in to PDBQT and ligands. Autodock Vina was run. Model 2 was selected as target for ligands.

## Results and Discussion

Nucleoprotein and its amino acid sequence was obtained from the Uniprot records. Sequence was submitted to ProtParam Tool (ExPasy server) for computing its various physiochemical properties (Figure 1A & 1B). The Nucleoprotein is made up of total 450 amino acids having the high amount of the serine (8.0%) and leucine (7.3%). Cysteine and Glutamine are present is less amount. From the sequence, estimated formula of Nucleoprotein is $C_{2268}H_{3496}N_{602}O_{675}S_{19}$ with 7060 total number of atoms. Extinction coefficient of protein is 48290 it means protein can absorb high amount of light at the 280 nm. Extinction coefficient or molar coefficient is proportional to the absorbance of light. Half-life of nucleoprotein in mammalian reticulocytes is around 30 hours and the instability index is 35.78. Instability index indicates stability of protein in the test tube. 43.10 is comes under the stable protein. To predict the secondary structure of protein, full DSSP classification was adopted by SSPro8 server. Predicted structure comprised of 47.55% of alpha helix, 8.66% turn having less than 1% of 3-10 helix and bend. It had 10.66% of extended strand as well, other 31.77 % were rest component. 3D structure was produced using phyre2 web and refined by the galaxy web with appropriate database to rebuild the side chains (Figure 2A & 2B). Galaxy Web server gave five structure; several validation parameters were compared to obtain most suitable model. Data of galaxy web is given in Table 1. From that Model 2 was selected for further use. Qualitative model energy analysis (QMEAN) is a collective scoring function linking the major geometrical aspects of protein structures. Model 2 had highest QMEAN score is given in Table 2. In the Z-sore plot of Nucleoprotein, the particular value that contains the Z-scores of all experimentally determined protein chains in current PDB is showed in (Figure 3A & 3B).


**Figure 1:** A) Amino acid Sequence of Nucleoprotein; B) The secondary structure predicted by SsPro8.
**Note:** H: alpha-helix; G: 3-10-helix; E: ex-tended strand; B: beta-bridge; T: turn; S: bend; C: the rest.

Plot of phi/psi dihedral angels amongst N-Cα and Cα-C peptide bond in the protein's mainstay is displayed by Ramachandran. From the drug bank and PDB, ligands were carefully chosen. All the ligands were minimized for their energy and were lastly altered into PDBQT format. According to their binding affinity to protein top ten were extracted. ERRAT plot shows error value for residues (Figure 4). The Y-axis represent the error value and the X-axis represent the amino acid residues of the protein model. An error value exceeding 99% confidence level indicates a poorly-modelled region. A) The initial model with the quality of 55.83%; B) The refined model 2 with the quality of 70.82%. The Z-score plot of Nucleoprotein determined by ProSA. The Z-score is -5.06, within the range of experimental native (Figure 5).

Ramachandran plot (Figure 6) of the fourth model is determined by Procheck. The most favored regions are marked as A, B, and L. The additional allowed regions are marked as a, b, l, and p. All non-glycine and proline residues are shown as filled black squares, whereas glycine (non-end) are shown as filled black triangles. Disallowed residues are colored red.

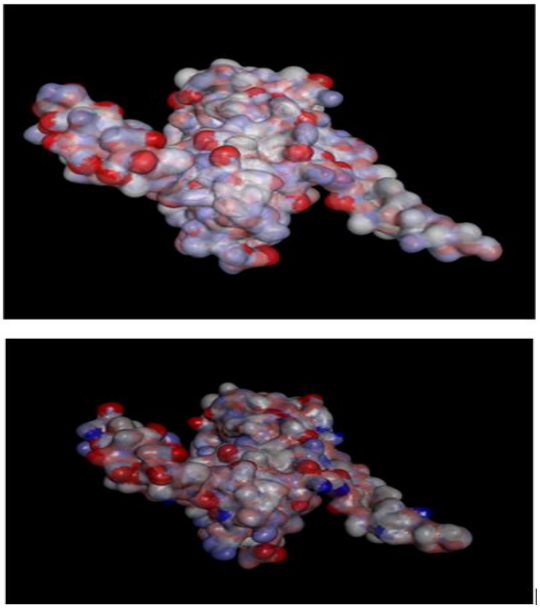

**Figure 2:** A) 3D model for nucleoprotein from Phyre2 unrefined; B) Refined structure by galaxy web.

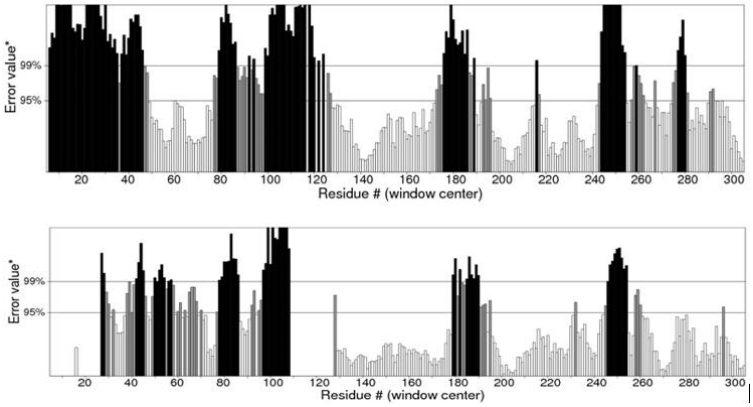


**Figure 3:** A) Errat plot of unrefined model; B) Errat plot of refined model.

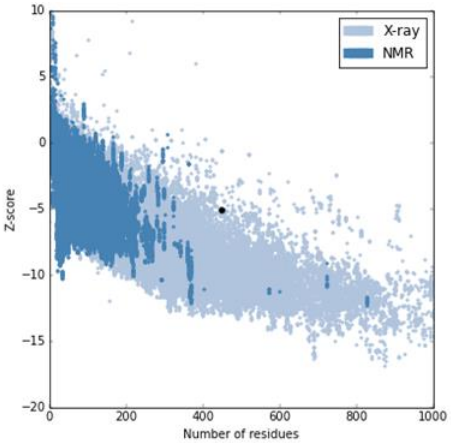


**Figure 4:** Z-score plot.

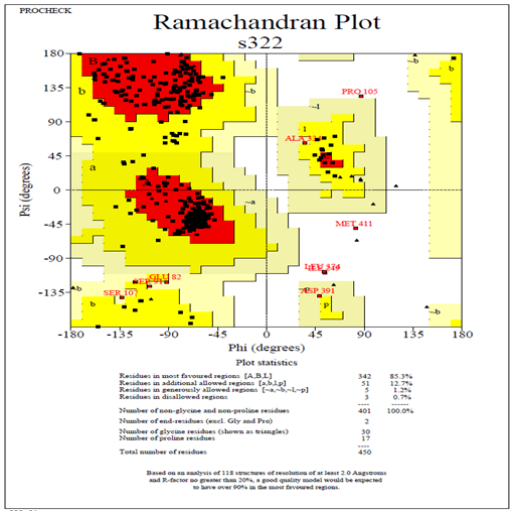


**Figure 5:** Ramachandran plot.

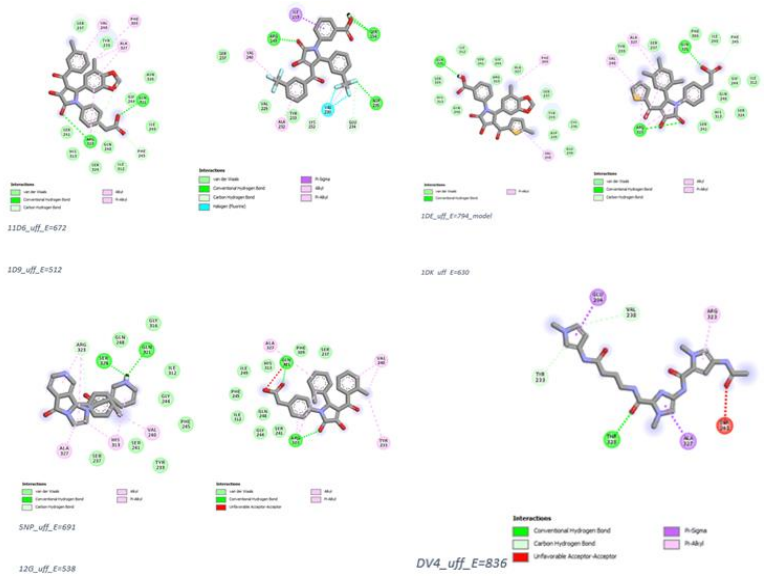


**Figure 6:** Docking interaction structure of lead molecules.

| Model | GDT-HA | RMSD | MolProbity | Clash Score | Poor Rotamers | Rama Favored |
|---|---|---|---|---|---|---|
| **Initial** | 1.0000 | 0.000 | 3.540 | 83.9 | 3.9 | 82.4 |
| **Model 1** | 0.9422 | 0.433 | 2.422 | 18.6 | 1.6 | 91.3 |
| **Model 2** | 0.9433 | 0.428 | 2.392 | 18.6 | 1.6 | 92.2 |
| **Model 3** | 0.9422 | 0.416 | 2.355 | 16.7 | 1.6 | 92.0 |
| **Model 4** | 0.9400 | 0.440 | 2.526 | 19.1 | 2.3 | 92.4 |
| **Model 5** | 0.9506 | 0.425 | 2.442 | 15.4 | 2.1 | 91.3 |

**Table 1:** Refined model produced by galaxy web.

| Scoring Function Term | Initial Model | | Refined Model 2 | |
|---|---|---|---|---|
| | Raw Score | Z-score | Raw Score | Z-Score |
| C_beta interaction energy | 0.003 | -5.508 | -0.005 | -2.484 |
| All-atom pairwise energy | -0.005 | -4.441 | -0.013 | -2.293 |
| Solvation energy | -0.617 | -1.727 | -0.714 | -0.172 |
| Torsion angle energy | 0.313 | -6.916 | 0.106 | -4.650 |
| Secondary structure agreement | 0.364 | -2.575 | 0.362 | -2.600 |
| Solvent accessibility agreement | 0.608 | -0.931 | 0.571 | -1.287 |
| QMEAN6 score | 0.492 | -6.921 | 0.570 | -5.020 |

**Table 2:** The QMEAN6 and component scores with respect to the experimental structure of similar size.

## Conclusion

From above work, it further recommended that 3D structure of Protein X could be utilized to model inhibitors with homologous matching in different microorganisms. This work also likewise demonstrated that the transactivation domain site of activity might be abused in ligand configuration to hinder the viral development, as it is watched that the protein X found to exceptionally fundamental for an infection for a few of the capacities.

## References


1. Berman HM, Westbrook J, Feng Z, et al. (2000) The protein data bank. Nucleic Acids Research 28(1): 235-242.
2. Boutet E, Lieberherr D, Tognolli M, et al. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the uniprot knowledge base: How to use the entry view. Methods in Molecular Biology 1374: 23-54.
3. Chou PY (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Advances in Enzymology and Related Areas of Molecular Biology 47: 45-148.
4. Colovos C, Yeates TO (1993) Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Science 2(9): 1511-1519.
5. Ikai A (1980) Thermostability and aliphatic index of globular proteins. The Journal of Biochemistry 88(6): 1895-1898.
6. Kelley LA, Mezulis S, Yates CM, et al. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10(6): 845-858.
7. Knox C, Law V, Jewison T, et al. (2010) DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. Nucleic Acids Research 39(suppl_1): D1035-D1041.
8. Ko J, Park H, Heo L, et al. (2012) GalaxyWEB server for protein structure prediction and refinement. Nucleic Acids Research 40(W1): W294-W297.
9. Kushwaha SK, Shakya M (2010) Protein interaction network analysis-approach for potential drug target identification in Mycobacterium tuberculosis. Journal of Theoretical Biology 262(2): 284-294.
10. Laskowski RA, MacArthur MW, Moss DS, et al. (1993) PROCHECK: A program to check the stereochemical quality of protein structures. Journal of Applied Crystallography 26(2): 283-291.
11. Liu J, Li X, Dong GL, et al. (2008) In silico analysis and verification of S100 gene expression in gastric cancer. BMC Cancer 8(1): 261.
12. Murray D, Doran P, MacMathuna P, et al. (2007) In silico gene expression analysis-an overview. Molecular Cancer 6(1): 50.
13. Recanatini M, Bottegoni G, Cavalli A (2004) In silico antitarget screening. Drug Discovery Today: Technologies 1(3): 209-215.
14. Romero P, Wagg J, Green ML, et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biology 6(1): R2.
15. Sahoo M, Lingaraja Jena SD, Kumar S (2016) Virtual screening for potential inhibitors of NS3 protein of Zika virus. Genomics & Informatics 14(3): 104-111.
16. Shin JM, Cho DH (2005) PDB-Ligand: A ligand database based on PDB for the automated and customized classification of ligand-binding structures. Nucleic Acids Research, 33(suppl_1): D238-D241.
17. Trott O, Olson AJ (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry 31(2): 455-461.
18. UniProt Consortium (2015) UniProt: A hub for protein information. Nucleic Acids Research 43(D1): D204-D212.
19. Wiederstein M, Sippl MJ (2007) ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35(suppl_2): W407-W410.