International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Protein Structure Prediction and Examination of Nucleocapsid Protein from SARS - COV2 using Computational and Simulation Tools

Aakash[1], Ryali Bhargavi Sri Vidya[2], Dr. Ajijur Rehman[3]

[1, 2]M. Sc (Biochemistry & Medical Biotechnology), National Institute of Medical Sciences & Research Jaipur, Rajasthan, India

[3]Assistant Professor, Department of Allied Sciences, National Institute of Medical Sciences & Research Jaipur, Rajasthan, India

[2]Email: *vidyaryali3[at]gmail.com*
Contact Numbers: 7673915636, 8851962721

Abstract: ***Recently emergency of COVID - 19 Corona virus has resulted in WHO declared public health issues and international emergency concern. The purpose of research globally around they working towards establishing a great understanding of this particular viruses and developing treatments and vaccinations all over the world to prevent spread. Structural properties of a protein it provides an important resource for understanding how it functions. This paper provides structural details of Nucleocapsid Protein (s) of novel corona virus. The Nucleocapsid protein of ncov2 is prominent for confining with a cell receptor which is act as referee for the synthesis of virus and host membranes, and these activities being crucial for virus ingress in to the host cell. This paper studies primary, secondary and tertiary structures of Nucleocapsid protein of SARS - COV2 predicted using Expasy Program, PSIPRED, and Homology Modelling Methods respectively. The predicted structure was validated using PROCHECK by Ramachandran plot and also validated through ProSAWeb tool. Finally, this predicted structure will helping and to discovering efficient medicine against corona epidemic. In future, result of the structure homology modelling would be energy minimized and can - do MD (Molecular dynamics) simulations to examine the how they expected model behave structurally, dynamically, using several computational and simulation tools.***

Keywords: Nucleocapsid Protein, Secondary Structure Prediction, Tertiary Structure Prediction, PSIPRED, Homology Modelling, PROCHECK, Protparam, Compute pI Molecular Dynamics

1. Introduction

Protein have a vast influence on the molecular machinery of life Protein are essential to life in bodies protein fold into certain three - dimensional (3D) structure called the native structure. The function of protein rely on their native structure determine the 3D structure has become a major task of model biological research. The Primary structure of proteins is the basis for higher level structures and proteins functions; thus, the primary structure always provides the basis for studying and modelling of artificial compositions and mutations. The most popular analysis of the similarity within a protein family is archived via multiple sequence comparisons and alignments using standard software, for example, BLASTP. They are several approaches for the similarity analysis [1 - 3]. The amino - acid sequences of proteins can be obtained from public databank, for example, the Swiss - Protein [4]. Recent advances in large scale sequencing technology we have seen an exponential growth in the protein sequence informatic protein structure are primarily determined using X - ray crystallographic or nuclear magnetic resonance (NMR) Spectroscopy but this method are time consume expensive and not feasible for all protein the experimental approach to the determine protein function. (E. g.) Gene knockout targeted mutation and inhibition of gene expression study [5]. In a cell protein carryout various type of biological function by folding into particular 3D structure thus elucidating a protein a protein structure is the key to understanding. It's function which in turn is essential for any further related biological medical or pharmaceutical application currently experiment determination of the protein structure is still expensive. Currently, about 20, 000 experimental protein structures are deposited in the Protein Data Bank (PDB) [5]. Template base modelling (also called comparative modelling or protein threading) Template free modelling (also called AB - initio modelling or free folding). Among all current theoretical approaches, comparative modelling is the only method that can reliably generate a 3D model of a protein (target) from its amino - acid sequence [6, 7]. Various structural genomics initiatives were started in the last few years, aiming to speed up the elucidation of new protein structures [8]. Bioinformatic has a topic research in the field of computer science and information technology. Bioinformatics consider a biological data of database to store biological information such as Nucleotide, amino - acids sequence and proteins [9].

1.1 Protein Structure

In this section we introduce the basic definitions and facts about protein structure, the four different levels of protein structure, as well as provide details about protein structure databases. The Alpha Fold Protein Structure Database (Alpha FoldDB, https: //alphafold. ebi. ac. uk) is an openly accessible, extensive database of high - accuracy protein - structure predictions [10].

1.2 Primary Structure

Amino acids form the basic building blocks of proteins. Amino acids consist of a central carbon atom (Cα) attached by an amino (NH2), a carboxyl (COOH) group, and a sidechain (R) group. The side chain group differentiates the

various amino acids. In the case of proteins there are primarily 20 different amino acids that form the building blocks. A protein is a chain of amino acids linked with the peptide bonds. The primary structure of a protein is largely responsible for its function. A vast majority of genetic disease a due to abnormality in the amino acid sequence of the protein. The amino acid composition of the protein determines physical and chemical properties [11, 12].

1.2.1 Primary Protein Structure

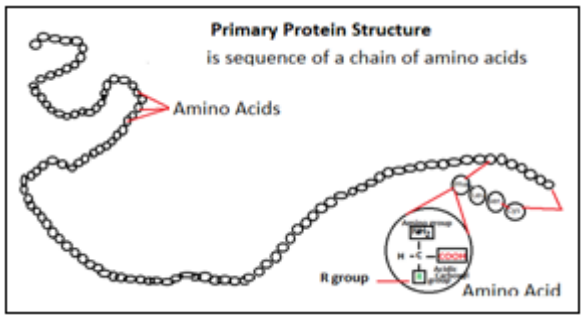


Figure 1

1.2.2 Secondary Protein Structure:

A Sequence of character we presenting the secondary structure of the protein describes the general 3D form of local regions. These regions oraganize themselves independntly from the rest of the protein into patterns of repeatdly occuring structural fragments. The most dominant local conformations of polypeptide chain are α - helices and β - sheets [11, 13].

α helix

- α - helix is the most spiral structure of protein. It has a rigid arrangement of polypeptide chain. α - Helical structure was proposed by Pauling and Corey (1951) which is regarded as one of the milestones in the biochemistry research.
- α helix is a tightly packed coiled structure with amino acid side chain.
- The α helix is stablized by extensive hydrogen bonding.
- The hydrogen bonds are individually weak but collectively they are strong enough to stabilize the helix.
- Each turn of α - helix contains 3.6 amino acid travels a distance of 0.54 nm.
- The splicing of each aminoacid is 0.15 nm [11, 12].

β Sheet

- In a β pleated sheet two or more segment of a polypeptide chain or 2 different polypeptide chains line up next to each other, forming a sheet - like structure held together by hydrogen bonds.
- The hydrogen bonds form between carbonyl and amino groups of backbone, while the Rgroups extend above and below the plane of the sheet.
- Represented as a series of flattened arrows.
- It is found in fibroin protein of silk fibres and featers of birds [11].

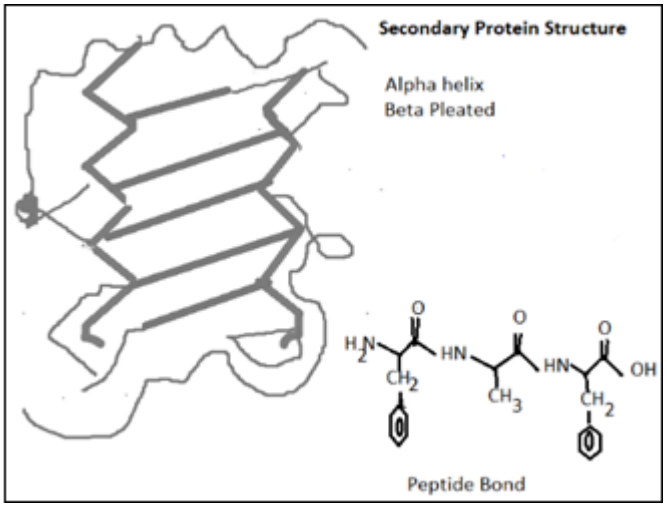


Figure 2

1.3 Tertiary Structure

The 3D arrangement of protein structure it is a compact structure with hydrophobic side chains held interior while the hydrophilic group are on the surface of the protein molecule 3D structure is a function of the interacting side chains between the different amino acids. Hence, the linear ordering of amino acids forms secondary structure; arranging secondary structures yields tertiary structure [11, 13].

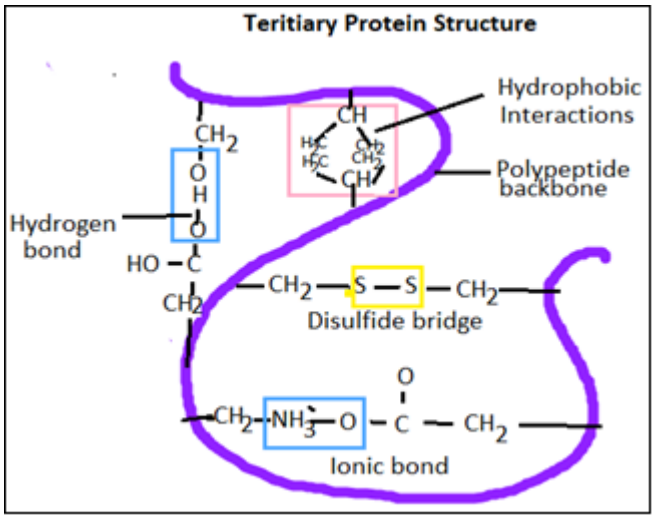


Figure 3

1.4 Quaternary Structure

- Quaternary structures represent the interaction between multiple polypeptide chains. The interaction between the various chains is due to the non - covalent interactions between the atoms of the different chains. Examples of these interactions include hydrogen bonding. Vander walls interactions, ionic bonding, and disulphide bonding.
- Research in computational structure prediction concerns itself mainly with predicting secondary and tertiary structures from known experimentally determined primary structure sequence. This is due to the relative ease of determining primary structure and the complexity involved in quaternary structure [11, 12, 13].

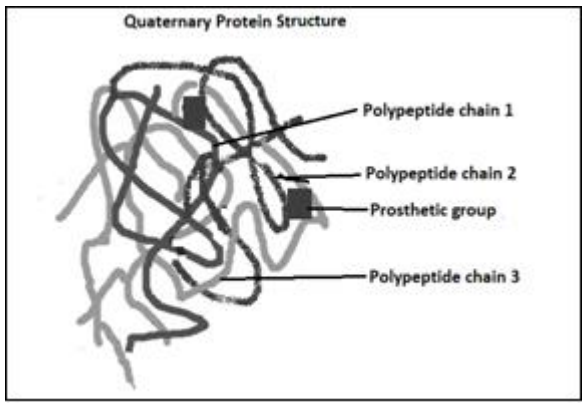


Figure 4

1.5 Protein sequence and structure databases

The large amount of protein sequence information experimentally determined structure information and structure classification information is stored in publicly available database in this section we review some of the databases that are used in field and provide their availability information in table; 1

Among all current theoretical approaches, comparative modelling is the only method that can reliably generate a 3D model of a protein (target) from its amino acid sequence [14, 15].

1.5.1Sequence databases

The universal protein resource (Uniport) is the most comprehensive warehouse containing information about protein.

Table: 1 [16]

Database link	Information	Availability
Uniport	Sequence	http: //www.pir. uniprot. org/
UniRef	Cluster sequences	http: //www.pir. uniprot. org/
NCBI	Nonredundant sequences	ftp: //ftp. ncbi. nlm. nih. gov/blast/db/
PDB	Structure	http: //www.rcsb. org/
SCOP	Structure classification	http: scop. mrc - lmb. cam. ac. uk/scop/
CATH	Structure classification	http: //www.cathdb. info/
FSSP	Structure classification	http: //www.ebi. ac. uk/dali/fssp/
ASTRAL	Compendium	http: //astral. berkeley. edu/

1.6 Protein databank

The protein databank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971 first open access, molecular data resource in biology [17]. Majority of journals required of PDB accession code and at least one funding agency (National Institute of General Medical Sciences). adopted by the guidelines published by the International Union of Crystallography (IUCr) requiring data deposition for all structures. The PDB is a key in the areas of structural biology, since 2003, the PDB has been managed jointly by the worldwide Protein Data Bank (wwPDB) consortium [18], including the US Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB; rcsb. org) [19], the Protein Databank Bank in Europe (PDB; pdbe. org) [20], Protein Data Bank Japan (PDBj; pdbj. org) [21]and BioMagResBank (BMRB; www.bmrb. wisc. edu) [22]. The PDB Core Archive houses 3D atomic coordinates of more than 144 000 structural models of proteins, DNA/RNA, and their complexes with metals and small molecules and related experimental data and metadata [23]. Many other databases use protein structures deposited in the PDB. Fore. g., SCOP and CATH classify protein structures, while PDB sum provides a graphic overview of PDB entries using information from other sources.

1.7 SCOP (structural classification of protein)

A manual classification of the protein structural domains based on similarity on their structure and amino acid sequence. It's helps to determine the evolutionary relationship between the protein. Nearly all proteins have structural similarities with other proteins, and in some of these cases, share a common evolutionary origin. First released to the public 27 years ago, the Structure Classification of Proteins (SCOP) database [24 - 27]. Work on the classic SCOP database concluded in 2009 with the release of SCOP. Since that time maintained the successor knowledgebase SCOP [28 - 30]. In order to provide ongoing updates to the hierarchy and classification of new protein structures from the protein databank (PDB) [31, 32]. Protein with same shape and some similarity of sequence and function are placed in families and are assumed to have a close common ancestor. SCOP2 [33], CATH [34], and ECOD [35], as well as to the sequence - based Pfam classification [36] are provided elsewhere [37]. Similar super families without compelling evidence of a common evolutionary origin are grouped into Folds, which are arranged into Classes based mainly on secondary structure content and organization [38]. Classification at the Super family level in particular depends on expert curation to integrate many types of information [39]. SCOP was created 1994 in the centre for protein engineering (CPE) and the laboratory of molecular biology (LMB).

Hierachical Organisation

a) Class: Type of folds, e. g., beta sheet
b) Fold: The different shape of domains within a class (topology).
c) Superfamily: The domains in a fold are grouped into superfamilies'which have at least a distant common ancestor (structural homology)
d) Family: The domains in a superfamily are grouped into families which have a more recent common ancestor (sequence homology)
e) Proteindomain: The domains in families are grouped into protein domains which are essentially the same protein (functionally identical)
f) Species: The domains in "protein domains" are grouped according to species (unique sequence)
g) Domain: Part of a protein. For simple proteins, it can be the entire protein

CATH:

CATH database is a semi - automated protein structure classification database like SCOP database CATH uses a consensus of three automated classification techniques to

break a chain into domain and classify them in the various structural categories. A Subset of proteins with different super families/families in the final cluster exhibit higher structural similarity than some proteins of the target CATH hierarchy [40].

CATH - GENE - 3D:
Is a classification of protein structures download from PDB we grouped protein domains into superfamilies' when there is sufficient evidence they have diverged from a common ancestor. Gene 3D uses the information in CATH to predict the location of structural domains on millions of protein sequence available in public database.

FSSP Database:
The FSSP is a structures classification database FSSP used on automatic classification exhaustive structure to structure alignment of protein using the DALI alignment FSSP does not provide a hierarchical classification like the SCOP and CATH database. There have been several studies analysing the relationship between the SCOP, CATH and FSSP database for representing the fold space for protein the major disagreement between the three database lies in the domain identification step rather than the domain classification step a high percentage of agreement exists between the SCOP, CATH and FSSP database especially at the fold level with sequence identity greater.

Corona virus Disease:
Corona Virus Disease (COVID - 19) is an epidemic disease discovered recently known as Corona virus (SARSCOV2). COVID - 19 belongs to Coronaviride family, mainly comprise harmful bacterium with zoonotic attribute, Severe Respiratory Syndrome (SARS - COV), Middle East Respiratory Syndrome (MERS - COV) of this group already started in 2003 and at present this COVID - 19 has transpired in china. These are single stranded RNA germs which could be secluded in various zoological genus. This lengthy strand of ribonucleic acid (RNA), which serves as the viruses' genetic material. When this virus infects a cell, it hijacks the molecular machinery to create long chains of proteins required by the virus to generate even more copies itself. The spike protein of SARS - COV2 can easily interact with host cell receptor ACE2.

The complete viral particle of a nucleocapsid (N) core encircled by an envelope contains 3 membrane proteins, Spike (S), membrane (M), envelope (E) which are equivalent to whole members of genre. The Spike (S) Protein, it biologically appears like a projection on the exterior of the viral particle, intermediates confining to host cells and membrane fusion. The Spike glycoproteins comprises 2 sub units (S1 and S2) Newest study says that, a spike variation, possibly appeared in November 2019, activated and leaping to human beings.

The 3D structure of spike protein is essential to discovery and development of antiviral drugs. Plenty of methods are available to visualize 3D structure of a protein like homology modelling, threading, and abinitio methods. One of the most robust, strong and widely used methods for interpreting 3D Structure is homology modelling. It is Comparative modelling, provides atomic resolution model, which models a structure deploy on the sameness of query sequence with given target protein. The modelled Structures which are generated by above techniques are static but by nature all proteins are dynamic in nature.

Structure Prediction and Examination of Nucleocapsid and Spike:

Protein structure prediction method:
Protein structure prediction is the inference of the 3D structure of the protein from the gene or amino acid sequence that is the prediction primary structure fromgene OC43. Protein structure prediction by using many tools of bioinformatics like expasy swiss model can involve sequence similarity searches multiple sequence alignment, identification characterization of domains.

a) Importance of Protein Prediction:
To understand the sequence of the unknown protein. To understand the shape and structure of the protein a protein structure allows it to perform it job for instance antibody are shaped like a Y. This help immune system protein bind to foreign molecule such as bacteria and virus. To understand the protein, function a protein function as we know protein act as an enzyme that catalyse chemical reaction provide a structure support protect against disease and cell signalling. Protein play the important of the body with the help of prediction know the function of the protein. Protein prediction also help to the drug design. One of the biggest goals in the structure bioinformatics is the prediction 3D structure. The goal is able to determine the shape of the structure and function of the protein.

b) Different strain of the Corona virus:
Corona virus disease is a respiratory infection disease caused by SARS - CoV - 2 virus in the human. That virus has spike (s) glycoprotein which bind to the host receptor and receptor present mostly in respiratory track and fuse the viral and cellular membrane and causing the respiratory infection disease.

Corona virus have different type of strain
229E (α Coronavirus)
NL63 (α Coronavirus)
OC43 (β Coronavirus)
HKU1 (β Coronavirus)

Human corona virus OC43 (HCoV - OC43):
Is a member of the species Betacoronavirus 1, which infects human and cattle. The infecting coronavirus is an enveloped positive - sense, single - stranded RNA virus that enters its host cell by binding to the N - acetyl - 9 - O - acetylneuraminic acid receptor. OC43 is one of seven coronaviruses known to infect humans. It is one virus responsible for the common cold. It has, like other coronaviruses from genus Betacoronavirus, subgenus Embecovirus, an additional shorter spike protein called hemagglutinin - esterase (HE).

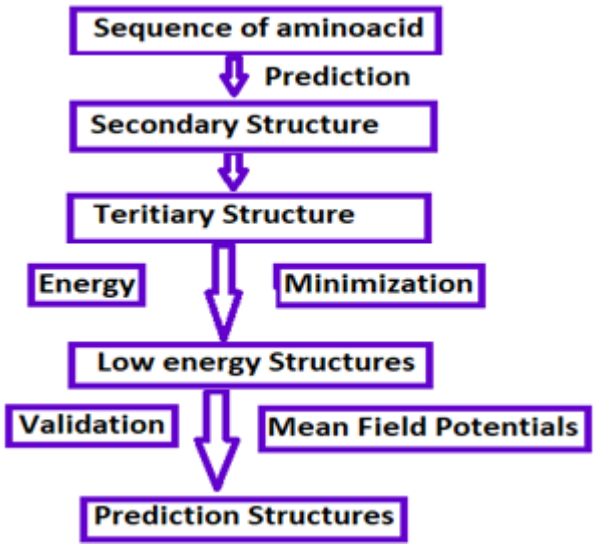
Virus classification:

Kingdom	Orthornavirae
Phylum	Pisuviricota
Class	Pisoniviricetes

Order	Nidovirales
Family	Coronaviridae
Genus	beta coronavirus
Species	betacoronavirus 1
Strain	human coronavirus OC43

Four HCoV - OC43 genotypes have been identified, with genotype the most likely arising from genetic recombination. The complete genome sequencing of genotypes C and D boot scan analysis shows recombination events between genotype D and C in the generation of genotype D genotype A molecular clock analysis using spike and nucleo capsid gene along with HCoV - 229E, A species in the genus Alpha coronavirus HCoV - OC43 is among the viruses that cause the common cold both viruses can cause severe lower severe respiratory tract infection.

Structure Prediction:



Aim and Objective

Study and analysis the protein structure Prediction, Sequence similarity. Secondary structure prediction solvent accessibility Prediction, Constructing & 3D models to atomic detail, and model validation.

- The contest Organizers compile a secret collection of experimentally verified protein structures.
- Report the structural and biophysical characterization of the SARS - CoV - 2 Nucleocapsid C - terminal domain.
- Validation of the sequence by the PROCHECK - Ramachandran plot analysis and Prosa sequence positions, Back bone conformations.
- Quality of sequence position and Lowest energy.
- Homology Modelling and Interprets the RMSD and RMSZ.

2. Review of Literature

Jumper, J., Evans, R et al., (2021), Studied that highly accurate protein structure prediction with Alpha Fold developed into the CASP14. In this study they develop first, protein predicting structures to near experimental the neural network, backbone analysis and RMSD interpret phylogenetic and covariation relationships without hard coding a particular correlation challenged proteins such as ORF8 and of SARS - CoV - 2 (T1064) the network searches and rearranges the secondary structures Alpha fold predictions in (GPU) minutes to hours depending on length of protein sequences were collected from Uniport (Swiss - prot&TrEMBL), soil and marine reference catalogue. The data analysis shows they remove clusters in the sequence, Cα atoms to measure the backbone accuracy (IDDT - Cα).

Oliverira Sc, de Magalhaeset al., (2020), In this study Immunoinformatic analysis of SARS - CoV - 2 nucleocapsid protein identification of COVID - 19 Vaccine Targets. In this study involving computer simulation for the identification of the epitopes recognized by antibodies and T cells are central to immunological applications such a drug design. The data collected for the sequences analysed by using Basic Local Alignment search tool specific for protein sequences (BLASTp) and Multiple sequence alignment are manually edited in PyBox shade. RMSD value for Pymol in this case analysis was able to identified epitopes conserved in several human Coronavirus N protein. Data analysed CoV - 2 epitopes associated with bioinformatics prediction points to specific regions of viral nucleocapsid that are target to human immune response vaccines that target human immune response toward these conserved epitopes its targeted MHC binding patterns in human coronavirus nucleocapsid protein.

Frank. Qi sheng Li, Han Xiao et al., (2005) In this study SUMOylating of the nucleocapsid protein of severe acute respiratory syndrome coronavirus. The SARS - CoV N protein was cloned and expressed in bacterial and mammalian cells. In Escherichia coli BL 21 cells, the protein was expressed as a single protein species. Biochemical characterization and mutagenesis studies demonstrated that the protein was post translationally modified by covalent attachment to the small ubiquitin - like modifier methods used in indirect immunofluorescence, western blot analysis, post translational modification of SARS - CoV N protein by SUMOylating. Detection of multiple protein species with a wide range of molecular masses when the N protein was expressed in mammalian cells indicates. Molecular weighted detected of dimer N protein detected it does not contain Sumo conjugate failure to detected the SUMOylates dimer in this study. In this data analysed that N - protein was dimerized compared to the monomers.

Rajendra Kumar Azad et al., (2021) The molecular assessment of SARS - CoV - 2 nucleocapsid phosphoprotein variants among Indian isolates. In this study analysed the N - Protein sequences of SARS - CoV 2 from Indian COVID - 19 patients and compared with the wuhan virus sequence. This study identified twenty mutations in N - Protein among Indian isolates and discussed their possible consequences and their protein identification they used clustal omega online tool to perform multiple sequence alignment to identify mutations. In this study analysed the mutations are spreading all over the N - Protein; however, a cluster is observed in the intrinsically disordered region 2 (or) Linker region present between NTD and CTD. There are seven mutations present at this location suggesting that this might be a mutational hotspot area of this protein. In this data analysis for using the secondary structure of prediction Chou

and Fasman this study identifies mutations at some residues which are in the close vicinity of these critical residues. And the predict that few of mutants might have differential interaction with the drugs, and that can possibly contribute to the drug resistance.

Kang S, Yang M et al., (2020) In this study the crystal structure of SARS - CoV 2 nucleocapsid N - terminal domain (termed as SARS - CoV 2 N - NTD) as a model for understanding the molecules interactions that govern SARS - CoV 2 N - NTD binding to ribonucleotides. Compared with other solved CoVs N - NTD, we characterized the specificity surface electrostatic potential features of SARS - CoV 2 N - NTD. Additionally, further demonstrated the potential unique nucleotide - binding pocket characteristics and development of new drugs that interfere with viral N Protein in SARS - CoV 2. Methods used cloning, expression and purification secondary structure predictions. It has been analysed that complete genome of SARS - CoV 2 (GenBank: MN908947, Wuhan - Hu - 1 Coronavirus) And it collects the data resource in National Centre for Biotechnology Information databank SARS - CoV, N protein - encoding regions are conserved among. And it found sequence identity SARS - CoV, MERS - CoV, and HCoV - OC43 respectively, as well as Coronavirus nucleocapsid protein. Until Now, Seven Coronavirus have been identified as human - susceptible viruses. With low pathogenicity cause mild respiratory symptoms.

Cortes - Sarabia K., Luna - Pineda, V. M et al., (2022) Utility of in silico - identified peptides in spike - S1 domain and nucleocapsid of SARS - CoV 2 for antibody detection in COVID - 19 patients and antibody production. This study search for characterize epitopes in the S and N proteins of SARS - CoV2 that could be used for antibody detection in serum samples derived from COVID - 19 patients using. Indirect ELISA and immune response induction in immunized animals for antibody production that could be used for diagnostic purposes. Immunogenic (or) antigenic sites in SARES - CoV 2 using immunoinformatic tools as IEDB, NetCTL serves, BC Pred, Vaxi Jen server, Protparam. New Zealand rabbits immunized with MAP8 peptides identification of immunogenic and antigenic peptides using in silico analysis. Bepi Pred - 20 to predict antigenic peptides. Twelve linear epitopes and sequential epitopes were identified, ten with surface accessibility, fifteen with antigenicity and seven B cell epitopes with not charges. The predicted molecules weight (MW) > 10 kDa. Methods which as used for standardization of Novel ELISA methods with high sensitivity and specificity. Another reported ten predicted peptides derived from the S, M and N proteins of SARA - CoV 2. The ELISA method used in this study has several similarities. ELISA methods using reported peptides for antibody detection in vaccinated (or) Variants of concern and peptide synthesis.

David Baker and Andrej Saliet al., (2001) The study it has structural genomics aims to study an efficient combination of experiment and prediction. Achieved by careful selection of target protein and their structure determination by X - ray Crystallography (or) NMR Spectroscopy. They are variety of target selection schemes, ranging from focusing on only novel folds to selecting all proteins in a model genome. Improvement in the accuracy of modelling approaches will require methods that finely sample protein conformational space using a free energy scoring functions that has sufficient accuracy to distinguish the native structure. Improvements in the sampling strategies may also be given the relatively long - time scale of protein folding. Mostly high accuracy comparative models are based on more than 50% sequence identity to their templates De novo Prediction.

Ali A. Dawood, Mohamod AA et al., (2021) In this study in order to compare the chemical structure with their biological behaviour at the levels of atoms and the Ligand - binding affinity. The analysis and the associated characteristics are largely responsible for nucleocapsid protein and ACE2 receptor that can be further changed for improved binding and selectivity. Protein modelling the SARS - CoV 2 NP Homology structural simulation was build based on the Swiss - Model server models. For molecular docking collected the sequence from Protein Data Bank. Then using PRODIGY software techniques, the resulting docking data was processed and analysed. There were a high correlation and identification between SARS - CoV 2 NPs, Bat - CoV - HKU3 - 2. The multiple sequence alignment (MSA) of the five series from various sources to the coronavirus family yield as given different results. The RMSD is the most common quantity measure of the correction between two atomic coordinates superimposed. Investigate new data processing to compare chemical structures with their biological behaviour, as well as affinity, at the level of atoms between NP and ACE 2 receptor.

Varsha Bhat and Jhinuk Chatterjee et al., (2021) In this study, a range of in silico toxicity prediction tools were used to evaluate the toxicity of some experimental compounds that have been recently reported in the literatures as potentially targeting SARS CoV - 2. They to their predicted toxicity profiles. In this Data collected from the three different applications (Web - based server). Were used for toxicity prediction for each of the listed parameters, the majority of the compounds were associated with low probabilities of hepatotoxicity and Carcinogenicity, and were predicted to be non - mutagenic and non - cardiotoxic. As efforts are ramped upto find SARS - CoV - 2 specific therapies, these results shown they lead optimisation by providing an indication of how the compounds may fare during invitro and in vivo toxicity testing, and their potential to be safe and effective drugs against Covid - 19.

3. Methods and Tools

a) Data Collection:

The amino acid sequence of Nucleocapsid protein and spike protein of COVID - 19 (Uniprot ID: P0DTC9) extracted from Uniprot Data Source (Nucleocapsid Fasta sequence ID: 39105221, Spike protein ID: 39105218 extracted from NCBI Data Source, which is available in www.ncbi. nlm. nih. gov PubMed, www.Uniport. org.

b) Primary Structure Prediction:

Primary structure prediction various tool is available this tool use four prediction the physical properties and using the sequence information. Several physico - chemical attributes

like composition of amino acid, atomic composition, Extinction Coefficient, Grand Average Hydrophobicity (GRAVY), etc relating to primary structure of P0DTC9 were anticipated by examining the amino acid sequence arrangement of the Nucleocapsid by server called of ExpasyProtParam server. The is available at web. expasy. org.

c) Secondary structure prediction:

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the secondary structure of protein sequence based only on knowledge of their primary structure for protein this means prediction the formation of protein structure such as alpha helices and beta strands. The secondary structure of spike protein was using PSIPRED server. The secondary structure was analysed with 3Dmodel structure to anticipate the structural characteristics of amino acid residues in various structural regions of the model obtained by modeller. There are several protein secondary structure prediction methods and the most important of these methods are:

- Gor method
- PsiPred method

d) Tertiary structure prediction:

There is possible method for prediction the tertiary structure of any protein sequence they are as follows. The template for the modelling structure can be take out by applying BLAST - PROTEIN (BLASTp) sequence similarity/search tool, which is available in NCBI website. BLAST - PROTEIN, performing the sequence similarity/search by taking Uniport: P0DTC9 as target sequence in the form of FASTA format. From the results of BLASTp, we extracted top three resultant proteins based on highest sequence similarity and their PDBID's are, 8fd5 (SAR - COV - 2) and they shared identity with queried protein as 99.50% respectively. The three - dimensional structure of the input protein sequence was obtained by modeller Homology and Comparative tool.

Homology modelling

Homology modelling also known as comparative modelling is to date the most reliable and well - established computational approach for predicting protein structure. Homology modelling predicts protein structure based on sequence homology with known structure if two protein shear a high enough sequence similarity then they have similar 3D structure. This tool uses the structure of homologous protein or protein fragment as an estimate for your protein structure and then models conformation differences that are likely to exits due to sequence divergence In general HM is very accurate the target sequence shares 40% or greater sequence identity to structure in the PDB and varies in accuracy with sequence between 20 - 30% sequence identity. The resultant structure model is all quality validate and can be used for functional annotation of genes molecular docking further experimental work such as protein engineering and drug design.

e) Structure Analysis and Visualization:

The Predicted structures obtained by Homology modelling can be visualized by SWISS MODEL via EXPASY web server based molecular visualization tool. The RMSD analysis of this structure analysed with template and typically energy minimized structure helpful to determine the accuracy of the model. The resultant structure of stereochemical stability attribute can be analysed with a tool called PROCHECK analysis.

4. Result and Discussion

a) Primary Structure Prediction

To analyse primary structure of a protein, Prot Param, Compute pI tool from a Expasy Server was used. Prot Param computes different kinds of physico - chemical attributes which can be extrapolate from a given protein input sequence. The input protein sequence can be taken up as Swiss rot/TrEMBL accession number/ID or it may be in form of raw sequence. Here, we used raw sequence and uploaded as FASTA format. And these calculations performed by ProtParam, which are based on either N - terminal amino acid (or) structure data.

The result generated by Expasy'sProtParam for the Nucleocapsid protein, contains 1347 amino acid residues with approximate molecular weight of 110414.81. The theoretical pI (pH at which protein remains stable) was anticipated to be 5.01 which is <pI=7 which says that protein is acidic by nature. Table 2 shows that several physico - chemical properties of Nucleocapsid proteins from SARS - COV2 obtained from Expasy ProtParam Tool.

Table 2: Physio - chemical properties of Nucleocapsid protein from SARS - COV2 obtained from Expasy ProtParam Tool

Parameters	Predicted Value
Molecular Weight	110414.81
Theoretical pI	5.01
Number of Positive Residues	0
Number of Negative Residues	0
Half - life mammalian reticulocytes (In Vitro)	4.4 hrs
Half - life yeast	>20 hours
Half - life E. Coli	>10 hours
Extinction coefficient	18750 - 0
Instability Index	45.03
Aliphatic Index	29.77
Gravy Index	0.830

The approximation is helpful to develop the storage system of a buffer for the amino acid of regard. The total number of positive residues (Arg+Lys) are 0 and the total number negative residues (Asp+Glu) are also 0. The half life cycle can be articulate as the time required to decay a protein to its half concentrated after the process of synthesis. The half - life of P0DTC9 for 3 morphons (human, yeast, E - coli) was approximately to be 4.4 hours for mammalian reticulocytes in vitro and 20 hours for yeast in vivo and more than 10 hours for E. coli in vivo. One of the prominent parameters given by Protparam is Extinction Co - efficient. There two types of Extinction Co - efficient, one that assumes that all cystine residues form cystine bonds and other assumes that all cystine residues as being reduced or unbounded. Cystine is a important amino acid that forms bonds with other amino acids mainly between cystine and it gives a instability to the protein. Formally, Extinction Co - efficient can be defined as

the absorbance of a Protein sample at 280nm measured in water with a spectrophotometer. And Extinction Co - efficient of P0DTC9 was approximately to be scaling 18750 based on composition of cystine.

Some of major one is discussion below.

N.1 Protein; nucleocapsid Protein (Human Coronavirus OC43).

Human coronavirus OC43 strain ATCC VR-759, complete genome

NCBI Reference Sequence: NC_006213.1

GenBank Graphics

>NC_006213.1:29079-30425 Human coronavirus OC43 strain ATCC VR-759, complete genome
ATGTCTTTTACTCCTGGTAAGCAATCCAGTAGTAGAGCGTCCTCTGGAAATCGTTCTGGTAATGGCATCC
TCAAGTGGGCCGATCAGTCCGACCAGTTTAGAAATGTTCAAACCAGGGGTAGAAGAGCTCAACCCAAGCA
AACTGCTACCTCTCAGCAACCATCAGGAGGGAATGTTGTACCCTACTATTCTTGGTTCTCTGGAATTACT
CAGTTTCAAAAGGGAAAGGAGTTTGAGTTTGTAGAAGGACAAGGTGTGCCTATTGCACCAGGAGTCCCAG
CTACTGAAGCTAAGGGGTACTGGTACAGACACAACAGACGTTCTTTTAAAACAGCCGATGGCAACCAGCG
TCAACTGCTGCCACGATGGTATTTTTACTATCTGGGAACAGGACCGCATGCTAAAGACCAGTACGGCACC
GATATTGACGGAGTCTACTGGGTCGCTAGCAACCAGGCTGATGTCAATACCCCGGCTGACATTGTCGATC
GGGACCCAAGTAGCGATGAGGCTATTCCGACTAGGTTTCCGCCTGGCACGGTACTCCCTCAGGGTTACTA
TATTGAAGGCTCAGGAAGGTCTGCTCCTAATTCCAGATCTACTTCGCGCACATCCAGCAGAGCCTCTAGT
GCAGGATCGCGTAGTAGAGCCAATTCTGGCAATAGAACCCCTACCTCTGGTGTAACACCTGACATGGCTG
ATCAAATTGCTAGTCTTGTTCTGGCAAAACTTGGCAAGGATGCCACTAAACCTCAGCAAGTAACTAAGCA
TACTGCCAAAGAAGTCAGACAGAAAATTTTGAATAAGCCCCGCCAGAAGAGGAGCCCCAATAAACAATGC
ACTGTTCAGCAGTGTTTTGGTAAGAGAGGCCCTAATCAGAATTTTGGTGGTGGAGAAATGTTAAAACTTG
GAACTAGTGACCCACAGTTCCCCATTCTTGCAGAACTCGCACCCACAGCTGGTGCGTTTTTCTTTGGATC
AAGATTAGAGTTGGCCAAAGTGCAGAATTTATCTGGGAATCCTGACGAGCCCCAGAAGGATGTTTATGAA
TTGCGCTATAACGGCGCAATTAGGTTTGACAGTACACTTTCAGGTTTTGAGACCATAATGAAGGTGCTGA
ATGAGAATTTGAATGCCTATCAACAACAAGATGGTATGATGAATATGAGTCCAAAACCACAGCGTCAGCG
TGGTCATAAGAATGGACAAGGAGAAAATGATAATATAAGTGTTGCAGTGCCCAAAAGCCGCGTGCAGCAA
AATAAGAGTAGAGAGTTGACTGCAGAGGACATCAGCCTTCTTAAGAAGATGGATGAGCCCTATACTGAAG
ACACCTCAGAAATATAA

1) ProtParam:

ProtParam is a tool available (http: //ca. expasy. org/tools/protparam. html) which allows the compute - action of various physical and chemical parameter for a given protein stored in SWISSPROT or TrEMBL or for a user entered sequence.

User - provided sequence:

10	20	30	40	50	60
ATGTCTTTTA	CTCCTGGTAA	GCAATCCAGT	AGTAGAGCGT	CCTCTGGAAA	TCGTTCTGGT
70	80	90	100	110	120
AATGGCATCC	TCAAGTGGGC	CGATCAGTCC	GACCAGTTTA	GAAATGTTCA	AACCAGGGGT
130	140	150	160	170	180
AGAAGAGCTC	AACCCAAGCA	AACTGCTACC	TCTCAGCAAC	CATCAGGAGG	GAATGTTGTA
190	200	210	220	230	240
CCCTACTATT	CTTGGTTCTC	TGGAATTACT	CAGTTTCAAA	AGGGAAAGGA	GTTTGAGTTT
250	260	270	280	290	300
GTAGAAGGAC	AAGGTGTGCC	TATTGCACCA	GGAGTCCCAG	CTACTGAAGC	TAAGGGGTAC
310	320	330	340	350	360
TGGTACAGAC	ACAACAGACG	TTCTTTTAAA	ACAGCCGATG	GCAACCAGCG	TCAACTGCTG
370	380	390	400	410	420
CCACGATGGT	ATTTTTACTA	TCTGGGAACA	GGACCGCATG	CTAAAGACCA	GTACGGCACC
430	440	450	460	470	480
GATATTGACG	GAGTCTACTG	GGTCGCTAGC	AACCAGGCTG	ATGTCAATAC	CCCGGCTGAC
490	500	510	520	530	540
ATTGTCGATC	GGGACCCAAG	TAGCGATGAG	GCTATTCCGA	CTAGGTTTCC	GCCTGGCACG
550	560	570	580	590	600
GTACTCCCTC	AGGGTTACTA	TATTGAAGGC	TCAGGAAGGT	CTGCTCCTAA	TTCCAGATCT
610	620	630	640	650	660
ACTTCGCGCA	CATCCAGCAG	AGCCTCTAGT	GCAGGATCGC	GTAGTAGAGC	CAATTCTGGC
670	680	690	700	710	720
AATAGAACCC	CTACCTCTGG	TGTAACACCT	GACATGGCTG	ATCAAATTGC	TAGTCTTGTT
730	740	750	760	770	780
CTGGCAAAAC	TTGGCAAGGA	TGCCACTAAA	CCTCAGCAAG	TAACTAAGCA	TACTGCCAAA
790	800	810	820	830	840
GAAGTCAGAC	AGAAAATTTT	GAATAAGCCC	CGCCAGAAGA	GGAGCCCCAA	TAAACAATGC
850	860	870	880	890	900

```
ACTGTTCAGC  AGTGTTTTGG  TAAGAGAGGC  CCTAATCAGA  ATTTTGGTGG  TGGAGAAATG

    910         920         930         940         950         960
TTAAAACTTG  GAACTAGTGA  CCCACAGTTC  CCCATTCTTG  CAGAACTCGC  ACCCACAGCT

    970         980         990        1000        1010        1020
GGTGCGTTTT  TCTTTGGATC  AAGATTAGAG  TTGGCCAAAG  TGCAGAATTT  ATCTGGGAAT

   1030        1040        1050        1060        1070        1080
CCTGACGAGC  CCCAGAAGGA  TGTTTATGAA  TTGCGCTATA  ACGGCGCAAT  TAGGTTTGAC

   1090        1100        1110        1120        1130        1140
AGTACACTTT  CAGGTTTTGA  GACCATAATG  AAGGTGCTGA  ATGAGAATTT  GAATGCCTAT

   1150        1160        1170        1180        1190        1200
CAACAACAAG  ATGGTATGAT  GAATATGAGT  CCAAAACCAC  AGCGTCAGCG  TGGTCATAAG

   1210        1220        1230        1240        1250        1260
AATGGACAAG  GAGAAAATGA  TAATATAAGT  GTTGCAGTGC  CCAAAAGCCG  CGTGCAGCAA

   1270        1280        1290        1300        1310        1320
AATAAGAGTA  GAGAGTTGAC  TGCAGAGGAC  ATCAGCCTTC  TTAAGAAGAT  GGATGAGCCC

   1330                                                    1340
TATACTGAAG ACACCTCAGA AATATAA
```

References and documentation are available.

Number of amino acids: 1347
Molecular weight: 110414.81
Theoretical pI: 5.01

CSV format

Amino acid composition:		
Ala (A)	401	29.8%
Arg (R)	0	0.0%
Asn (N)	0	0.0%
Asp (D)	0	0.0%
Cys (C)	300	22.3%
Gln (Q)	0	0.0%
Glu (E)	0	0.0%
Gly (G)	326	24.2%
His (H)	0	0.0%
Ile (I)	0	0.0%
Leu (L)	0	0.0%
Lys (K)	0	0.0%
Met (M)	0	0.0%
Phe (F)	0	0.0%
Pro (P)	0	0.0%
Ser (S)	0	0.0%
Thr (T)	320	23.8%
Trp (W)	0	0.0%
Tyr (Y)	0	0.0%
Val (V)	0	0.0%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 0
Total number of positively charged residues (Arg + Lys): 0

Atomic composition:

Carbon	C	4035
Hydrogen	H	6725
Nitrogen	N	1347
Oxygen	O	1668
Sulfur	S	300

Formula: $C_{4035}H_{6725}N_{1347}O_{1668}S_{300}$
Total number of atoms: 14075

Extinction coefficients:
This protein does not contain any Trp residues. Experience shows that this could result in more than 10% error in the computed extinction coefficient.

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 18750
Abs 0.1% (=1 g/l) 0.170, assuming all pairs of Cys residues form cystines

Ext. coefficient 0
Abs 0.1% (=1 g/l) 0.000, assuming all Cys residues are reduced

Estimated half - life:
The N - terminal of the sequence considered is A (Ala).

The estimated half - life is: 4.4 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:
The instability index (II) is computed to be 45.03
This classifies the protein as unstable.

Aliphatic index: 29.77

Grand average of hydropathicity (GRAVY): 0.830

FASTA Sequence of Nucleocapsid protein of SARS - COV2 (P0DTC9) in FASTA format take out from NCBI Website.

Statistical review of 12 unstable and 32 stable proteins that exhibit significant difference in appearance of certain dipeptides (a peptide composed of 2amino acid residues) in unstable proteins as analysed with stable proteins. An experiment is produced to correlate catabolic stability of proteins with attributes of their primary sequence here weight values of instability for a protein of template could accordingly be applied as an indicator for anticipating its stability characteristics.

Protparam also exhibits that the proteins with instability index of < 40 were stable and value > 40 were anticipated to be unstable. The instability index (II) of P0DTC9 was calculated as 45.03, indicating that protein will be unstable in vacutainer. Another characteristic is aliphatic index, defined as relative volume occupied side chains (alanine, valine, isoleucine, and leucine). Higher the value designated that higher stability of a protein. The aliphatic index of **P0DTC9** is anticipated to be 29.77. The GRAVY (Grand Average of Hydropathy) indicates that likelihood of the interplay of protein with water. The lower value GRAVY index yields higher likelihood that protein will not interplay with water. The GRAVY index for **P0DTC9**is anticipated to be 0.830. This score exhibits that the protein will not interplaying with water, which says that the nature of the protein is hydrophobic it is non immunogen.

The Protparam result of Nucleocapsid Protein (**P0DTC9**) comparison with Spike (**7sb3**) Protein Parameters and Prediction values are not similar

Number of amino acids: 1273
Molecular weight: 141178.47

Theoretical pI: 6.24

CSV format

Amino acid composition:		
Ala (A)	79	6.2%
Arg (R)	42	3.3%
Asn (N)	88	6.9%
Asp (D)	62	4.9%
Cys (C)	40	3.1%
Gln (Q)	62	4.9%
Glu (E)	48	3.8%
Gly (G)	82	6.4%
His (H)	17	1.3%
Ile (I)	76	6.0%
Leu (L)	108	8.5%
Lys (K)	61	4.8%
Met (M)	14	1.1%
Phe (F)	77	6.0%
Pro (P)	58	4.6%
Ser (S)	99	7.8%
Thr (T)	97	7.6%
Trp (W)	12	0.9%
Tyr (Y)	54	4.2%
Val (V)	97	7.6%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 110
Total number of positively charged residues (Arg + Lys): 103

Atomic composition:

Carbon	C	6336
Hydrogen	H	9770
Nitrogen	N	1656
Oxygen	O	1894
Sulfur	S	54

Formula: $C_{6336}H_{9770}N_{1656}O_{1894}S_{54}$

Total number of atoms: 19710

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 148960
Abs 0.1% (=1 g/l) 1.055, assuming all pairs of Cys residues form cystines

Ext. coefficient 146460
Abs 0.1% (=1 g/l) 1.037, assuming all Cys residues are reduced

Estimated half - life:
The N - terminal of the sequence considered is M (Met).

The estimated half - life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:
The instability index (II) is computed to be 33.01
This classifies the protein as stable.

Aliphatic index: 84.67
Grand average of hydropathicity (GRAVY): - 0.079

Table 3: SpikeProtein SARS - COV2

Parameters	Predicted Value
Molecular Weight	141178.47
Theoretical pI	6.24
Number of Positive Residues	103
Number of Negative Residues	110
Half - life mammalian reticulocytes (In Vitro)	30 hrs
Half - life yeast	>20 hours
Half - life E. coli	>10 hours
Extinction coefficient	148960 - 146460
Instability Index	33.01
Aliphatic Index	84.67
Gravy Index	- 0.079

The result generated by Expasy' Protparam for the spike protein contains 1273 amino acid residues with approximate molecular weight of 141178.47. The Theoretical pI stable anticipated to the 6.24 and positive residues (Arg+Lys) are 103 and negative residues (Asp+Glu) are 110. Half - life mammalian reticulocytes (In Vitro) 30 hours and 20 hours yeast in vivo and more than 10 hours for E - coli. Extinction Co - efficient approximately to be scaling between 148960 - 146460 based on composition of cystine. Proteins with instability index calculated as 33.01, aliphatic side chains are higher value designated high stability of protein it is anticipated 84.67. The GRAVY index is interplay with water it is anticipated to be - 0.079 the nature of protein is immunogen.

2) Expasy (translate tool):
Translate is a tool which allows the translation of a nucleotide DNA/RNA sequence to protein sequence. ExPASy (http: //www.expasy. org) has worldwide reputation as one of the main bioinformatics resources for proteomics. It has now evolved, becoming an extensible and integrative portal accessing many scientific resources, databases and software tools.

The Fasta sequence of DNA or RNA sequences are the translate into a both forward and reverse.

Primary protein structure predicts

Amino acid residues

5'3' Frame 1

MSFTPGKQSSSRASSGNRSGNGILKWADQSDQFRNV
QTRGRRAQPKQTATSQQPSGGNVVPYYSWFSGITQFQ
KGKEFEFVEGQGVPIAPGVPATEAKGYWYRHNRRSF
KTADGNQRQLLPRWYFYYLGTGGPHAKDQYGTDID
GVYWVASNQADVNTPADIVDRDPSSDEAIPTRFPPGT
VLPQGYYIEGSGRSAPNSRSTSRTSSRASSAGSRSRAN
SGNRTPTSGVTPDMADQIASLVLAKLGKDATKPQQV
TKHTAKEVRQKILNKPRQKRSPNKQCTVQQCFGKRG
PNQNFGGGEMLKLGTSDPQFPILAELAPTAGAFFFGSR
LELAKVQNLSGNPDEPQKDVYELRYNGAIRFDSTLSG
FETIMKVLNENLNAYQQQDGMMNMSPKPQRQRGHK
NGQGENDNISVAVPKSRVQQNKSRELTAEDISLLKKM
DEPYTEDTSET -

b) Secondary structure prediction:
Structure prediction methods and the most important of these methods are: Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the secondary structure of protein sequence based only on knowledge of their primary structure for protein this means prediction the formation of protein structure such as alpha helices and beta strands.

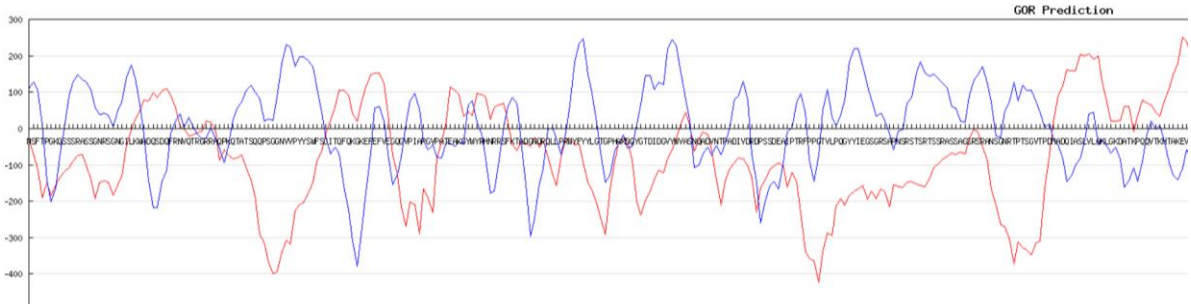
There are several protein secondary structure predictions.

1) GOR method (Garnier, osguthorpe and robson)
GOR is a tool available is GOR 4 (http: //npsa - pbil. ibcp. fr/cgi - bin/npsa - automat. pipage=npsa - gor4html) **39105218**

GOR is a method that assumes that amino acid up to 8 residues on each side influence the secondary structure of the central residues this tool is now in the fourth the accuracy of GOR when checked against a set of 267 protein of know structure is 64% this implies that 64% of amino acid were correctly predicted. The method works better for helix then for sheet because of sheet is depended on the longer - range interaction between non - adjacent sequence fragment.

ATGTTTTTGATACTTTTAATTTCCTTACCAACGGCTT
TTGCTGTTATAGGAGATTTAAAGTGTACTTCAG
ATAATATTAATGATAAAGACACCGGTCCTCCTCCTA
TAAGTACTGATACTGTTGATGTTACTAATGGTTT
GGGTACTTATTATGTTTTAGATCGTGTGTATTTAAA
TACTACGTTGTTTCTTAATGGTTATTACCCTACT
TCAGGTTCCACATATCGTAATATGGCACTGAAGGG
AAGTGTACTATTGAGCAGACTATGGTTTAAACCAC
CATTTCTTTCTGATTTTATTAATGGTATTTTTGCTAA
GGTCAAAAATACCAAGGTTATTAAAGATCGTGT
AATGTATAGTGAGTTCCCTGCTATAACTATAGGTAG
TACTTTTGTAAATACATCCTATAGTGTGGTAGTA
CAACCACGTACAATCAATTCAACACAGGATGGTGA
TAATAAATTACAAGGTCTTTTAGAGGTCTCTGTTT
GCCAGTATAATATGTGCGAGTACCCACAAACGATT
TGTCATCCTAACCTGGGTAATCATCGCAAAGAACT
ATGGCATTTGGATACAGGTGTTGTTTCCTGTTTATA
TAAGCGTAATTTCACATATGATGTGAATGCTGAT
TATTTGTATTTTCATTTTTATCAAGAAGGTGGTACTT
TTTATGCATATTTTACAGACACTGGTGTTGTTA
CTAAGTTTTTGTTTAATGTTTATTTAGGCATGGCGC
TTTCACACTATTATGTCATGCCTCTGACTTGTAA
TAGTAAGCTTACTTTAGAATATTGGGTTACACCTCT
CACTTCTAGACAATATTTACTCGCTTTCAATCAA
GATGGTATTATTTTTAATGCTGTTGATTGTATGAGT
GATTTTATGAGTGAGATTAAGTGTAAAACACAAT
CTATAGCACCACCTACTGGTGTTTATGAATTAAACG
GTTACACTGTTCAGCCAATCGCAGATGTTTACCG
ACGTAAACCTAATCTTCCCAATTGCAATATAGAAG
CTTGGCTTAATGATAAGTCGGTGCCCTCTCCATTA
AATTGGGAACGTAAGACATTTTCAAATTGTAATTTT
AATATGAGCAGCCTGATGTCTTTTATTCAGGCAG
ACTCATTTACTTGTAATAATATTGATGCTGCTAAGA
TATATGGTATGTGTTTTTCCAGCATAACTATAGA
TAAGTTTGCTATACCCAATGGCAGGAAGGTTGACC
TACAATTGGGTAATTTGGGCTATTTGCAGTCATTT

AACTATAGAATTGATACTACTGCAACAAGTTGTCA
GTTGTATTATAATTTACCTGCTGCTAATGTTTCTG
TTAGCAGGTTTAATCCTTCTACTTGGAATAAGAGAT
TTGGTTTTATAGAAGATTCTGTTTTTAAGCCTCG
ACCTGCAGGTGTTCTTACTAATCATGATGTAGTTTA
TGCACAACACTGTTTCAAAGCTCCTAAAAATTTC
TGTCCGTGTAAATTGAATGGTTCGTGTGTAGGTAGT
GGTCCTGGTAAAAATAATGGTATAGGCACTTGTC
CTGCAGGTACTAATTATTTAACTTGTGATAATTTGT
GCACTCCTGATCCTATTACATTTACAGGTACTTA
TAAGTGCCCCCAAACTAAATCTTTAGTTGGCATAGG
TGAGCACTGTTCGGGTCTTGCTGTTAAAAGTGAT
TATTGTGGAGGCAATTCTTGTACTTGCCGACCACAA
GCATTTTTGGGTTGGTCTGCAGACTCTTGTTTAC
AAGGAGACAAGTGTAATATTTTTGCTAATTTTATTT
TGCATGATGTTAATAGTGGTCTTACTTGTTCTAC
TGATTTACAAAAAGCTAACACAGACATAATTCTTG
GTGTTTGTGTTAATTATGACCTCTATGGTATTTTA
GGCCAAGGCATTTTTGTTGAGGTTAATGCGACTTAT
TATAATAGTTGGCAGAACCTTTTATATGATTCTA
ATGGTAATCTCTACGGTTTTAGAGACTACATAACAA
ACAGAACTTTTATGATTCGTAGTTGCTATAGCGG
TCGTGTTTCTGCGGCCTTTCACGCTAACTCTTCCGA
ACCAGCATTGCTATTTCGGAATATTAAATGCAAC
TACGTTTTTAATAATAGTCTTACACGACAGCTGCAA
CCCATTAACTATTTTGATAGTTATCTTGGTTGTG
TTGTCAATGCTTATAATAGTACTGCTATTTCTGTTC
AAACATGTGATCTCACAGTAGGTAGTGGTTACTG
TGTGGATTACTCTAAAAACAGACGAAGTCGTGGAG
CGATTACCACTGGTTATCGGTTTACTAATTTTGAG
CCATTTACTGTTAATTCAGTAAACGATAGTTTAGAA
CCTGTAGGTGGTTTGTATGAAATTCAAATACCTT
CAGAGTTTACTATAGGTAATATGGTGGAGTTTATTC
AAACAAGCTCTCCTAAAGTTACTATTGATTGTGC
TGCATTTGTCTGTGGTGATTATGCAGCATGTAAATC
ACAGTTGGTTGAATATGGTAGTTTCTGTGATAAC
ATTAATGCCATACTCACAGAAGTAAATGAACTACT
TGACACTACACAGTTGCAAGTAGCTAATAGTTTAA
TGAATGGTGTTACTCTTAGCACTAAGCTTAAAGATG
GCGTTAATTTCAATGTAGACGACATCAATTTTTC
CCCTGTATTAGGTTGTCTAGGCAGCGAATGTAGTAA
AGCTTCCAGTAGATCTGCTATAGAGGATTTACTT
TTTGATAAAGTAAAGTTATCTGATGTCGGTTTTGTT
GAGGCTTATAATAATTGTACAGGAGGTGCCGAAA
TTAGGGACCTCATTTGTGTGCAAAGTTATAAAGGC
ATCAAAGTGTTGCCTCCACTGCTCTCAGAAAATCA
GATCAGTGGATACACTTTGGCTGCCACCTCTGCTAG
TCTATTTCCTCCTTGGACAGCAGCAGCAGGTGTA
CCATTTTATTTAAATGTTCAGTATCGCATTAATGGG
CTTGGTGTCACCATGGATGTGCTAAGTCAAAATC
AAAAGCTTATTGCTAATGCATTTAACAATGCCCTTT
ATGCTATTCAGGAAGGGTTCGATGCAACTAATTC
TGCTTTAGTTAAAATTCAAGCTGTTGTTAATGCAAA
TGCTGAAGCTCTTAATAACTTATTGCAACAACTC
TCTAATAGATTTGGTGCTATAAGTGCTTCTTTACAA
GAAATTCTATCTAGACTTGATGCTCTTGAAGCGG
AAGCTCAGATAGATAGACTTATTAATGGTCGTCTTA
CCGCTCTTAATGCTTATGTTTCTCAACAGCTTAG
TGATTCTACACTGGTAAAATTTAGTGCAGCACAAG
CTATGGAGAAGGTTAATGAATGTGTCAAAAGCCAA
TCATCTAGGATAAATTTCTGTGGTAATGGTAATCAT
ATTATATCATTAGTGCAGAATGCTCCATATGGTT
TGTATTTTATCCACTTTAGTTATGTCCCTACTAAGTA
TGTCACAGCGAGGGTTAGTCCTGGTCTGTGCAT
TGCTGGTGATAGAGGTATAGCTCCTAAGAGTGGTT
ATTTTGTTAATGTAAATAATACTTGGATGTACACT
GGTAGTGGTTACTACTACCCTGAACCTATAACTGAA
AATAATGTTGTTGTTATGAGTACCTGCGCTGTTA
ATTATACTAAAGCGCCGTATGTAATGCTGAACACTT
CAATACCCAACCTTCCTGATTTTAAGGAAGAGTT
GGATCAATGGTTTAAAAATCAAACATCAGTGGCAC
CAGATTTGTCACTTGATTATATAAATGTTACATTC
TTGGACCTACAAGTTGAAATGAATAGGTTACAGGA
GGCAATAAAAGTCTTAAATCAGAGCTACATCAATC
TCAAGGACATTGGTACATATGAATATTATGTAAAA
TGGCCTTGGTATGTATGGCTTTTAATCTGCCTTGC
TGGTGTAGCTATGCTTGTTTTACTATTCTTCATATGC
TGTTGTACAGGATGTGGGACTAGTTGTTTTAAG
AAATGTGGTGGTTGTTGTGATGATTATACTGGATAC
CAGGAGTTAGTAATCAAAACTTCACATGACGACT
AA



2) PSIPRED METHOD:

PSIPRED protein structure prediction server allows users to submit a protein sequence performs a prediction of their choice and receive the results of the prediction both textually via e - mail and graphically via the web the user may select one of three prediction method to apply to their sequence PSIPRED a highly accurate secondary structure prediction method Figure: 1&2 Show psipred/show aatypes.

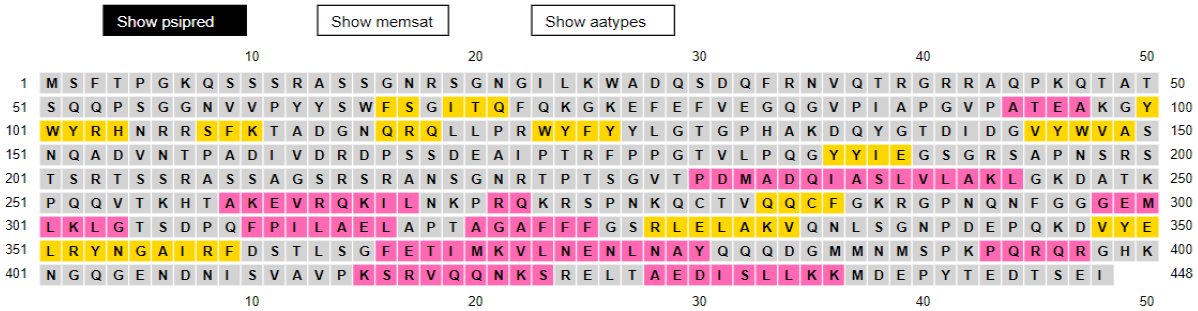


Figure 1: Sequence plot (Secondary structure of Nucleocapsid SARS CoV)

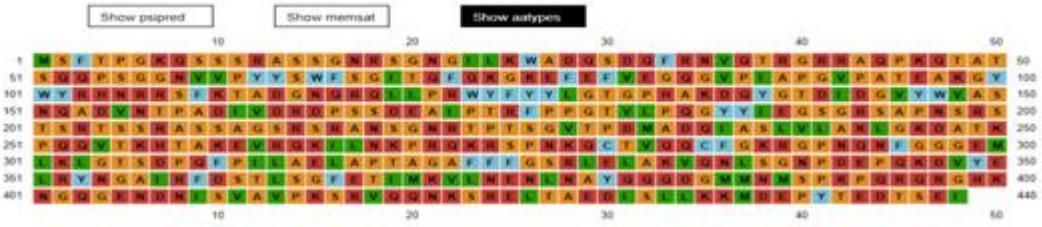


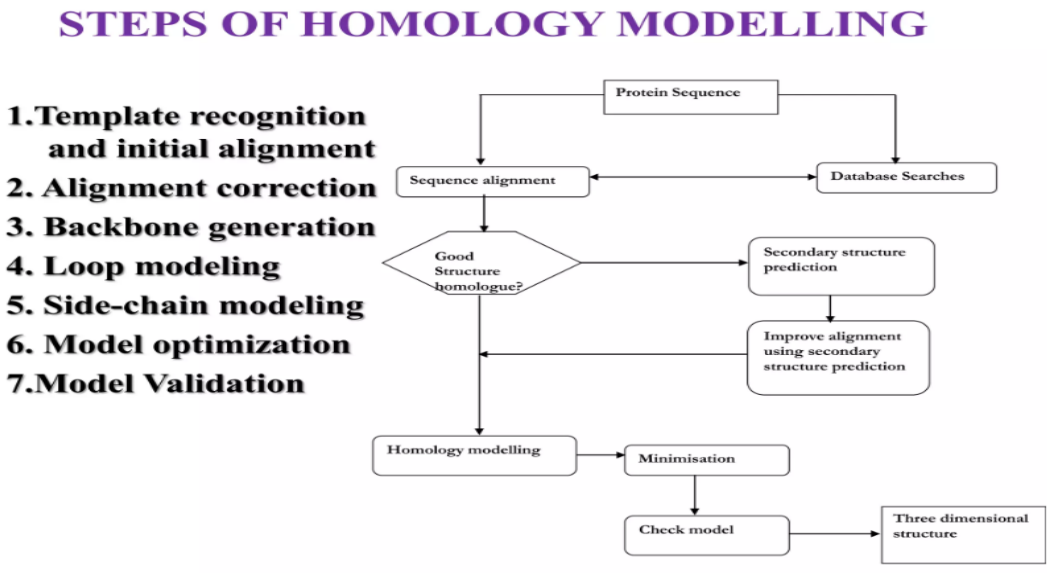
Figure 2: Sequence plot (Secondary structure of Nucleocapsid SARS CoV)

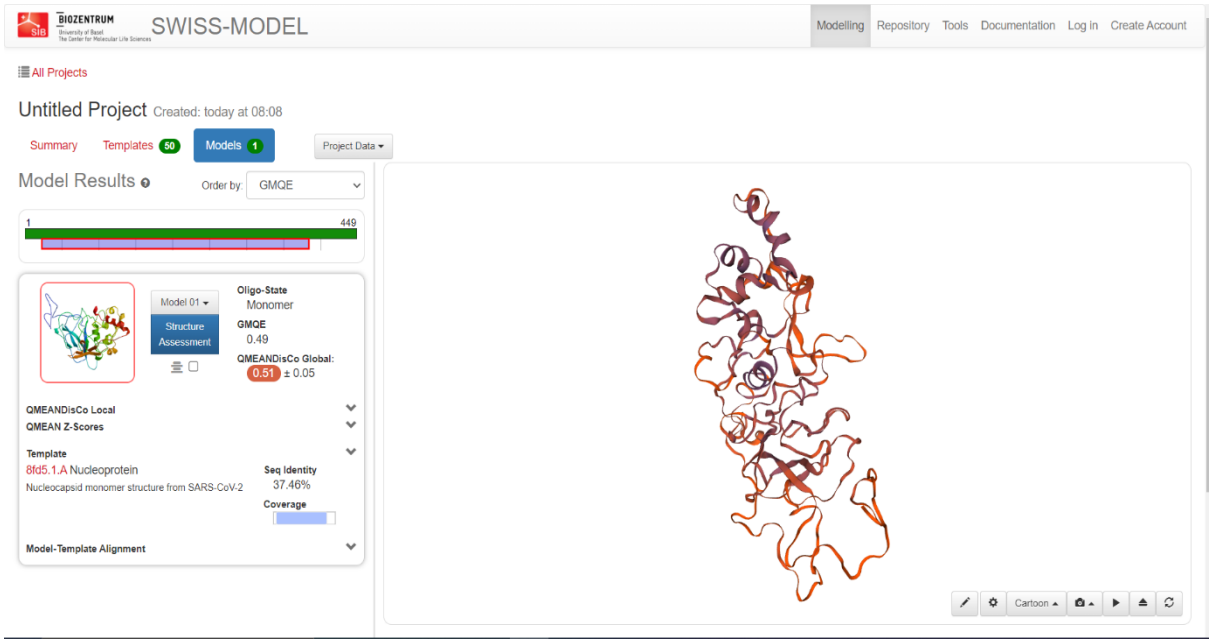
C) Tertiary structure prediction

There is possible method for prediction the tertiary structure of any protein sequence they are as follows: Homology modelling The 3D structure of P0DTC9 was obtained or modelled with comparative and Homology Modelling tool, Modeller 9.23. The target protein was given as query sequence to Blastp to determine the homologous sequence. The top three results extracted and which are shared 99% identity, such as PDB ID's: 8fd5. Therefore, these are selected as the structure Homology Modelling. The obtained 3D structure of the target Nucleocapsid from SARS - COV2 (Uniprot ID: P0DTC9) and superimposed with ID: 6coz RMSD (Root Mean Square Deviation) value (3.24) which shows Prominently Quality of the modelling structure. Super pose molecular viewer tool calculated RMSD.

Homology modelling

In general HM is very accurate the target sequence shares 40% or greater sequence identity to structure in the PDB and varies in accuracy with sequence between 20 - 30% sequence identity. The resultant structure model is all quality validate and can be used for functional annotation of genes molecular docking further experimental work such as protein engineering and drug design.





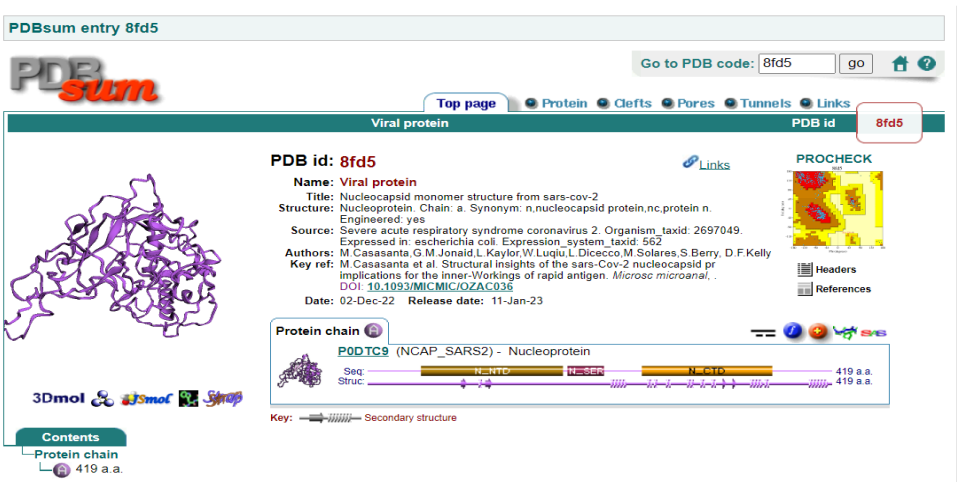
Protein structure validation report

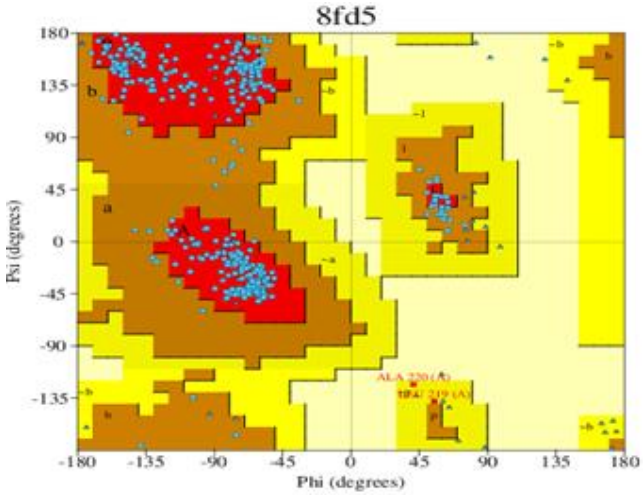
- Hence the prediction of protein structure which respect through biological functions and statistical significance.
- Structure we prediction complicated if the is no information of Template and Accurate.
- It should be identifying the (Active site amino acids and Hydrophobic patches, Hydrophobic fold.
- We should remove the errors of protein before prediction.
- The validation report it shows the Quality Indicators of 3D validation.

It recognizes the errors of protein and structure model errors.

- Side chain packing errors.
- Backbone conformations
- Phi/Psi angles of Ramachandran plot.

Validation report which as tells energy plot NMR result from PROSA.





PROCHECK statistics

1. Ramachandran Plot statistics

		No. of residues	%-tage
Most favoured regions	[A,B,L]	2954	88.8%*
Additional allowed regions	[a,b,l,p]	356	10.7%
Generously allowed regions	[~a,~b,~l,~p]	7	0.2%
Disallowed regions	[XX]	9	0.3%*
Non-glycine and non-proline residues		3326	100.0%
End-residues (excl. Gly and Pro)		41	
Glycine residues		234	
Proline residues		155	
Total number of residues		3756	

Based on an analysis of **118** structures of resolution of at least **2.0** Angstroms and *R*-factor no greater than **20.0** a good quality model would be expected to have over **90%** in the most favoured regions [A,B,L]

Figure 5: Ramachandran plot of P0DTC2 produced by PROCHECK rendering the occurrence of 89.9% amino acids in core part

The plot of Ramachandran Figure 5 of the model obtained by tool called PROCHECK renders number of residues in favoured region are 419 (89.9%). Number of residues in allowed region are 33 (9.5%). Number of residues in generously allowed region are 2 (0.6%). Number of residues in disallowed region are 0 (0%). These results tell that the model is stereo chemically stable. We can also verify 3D structure of using PROCHECK. The Figure 5 shows that 3D protein structure of predicted modelled structure generated by PROCHECK.

The validation is more important step in Homology Modelling. To do this, ProSA - Web tool is used. It is a Web - based Protein Structure Analysis tool, provides easily available online and user interface for the protein structure validation. ProSA computes an overall virtue score for a given input structure by means of Z - Score value. This Score also figured out in a graph as plot, which contains the Z - scores of the all experimentally set up a protein chains in a given Protein Data Bank (PDB) input file. The Z - Score returned by this web server for modelled structure is Nucleocapsid (- 5.28). Comparison of the Spike Glycoprotein is (- 12.28). The Z - Score plot shown in Figure 6, a class of structures from various sources (X - Ray, NMR) are differentiate by various colours. This will helps to verify whether Z - score modelled structure with in span of scores customarily retrieve for local proteins of similar size. But here the results of this Fig: 6 (8fd5) lower than the Fig: 7 (7sb3) chain.

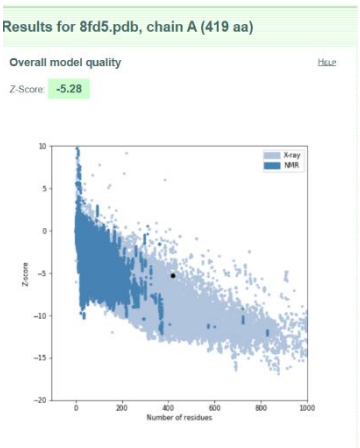


Figure 6: P0DTC9 (8fd5) Z - score plot generated by ProSA - Web tool

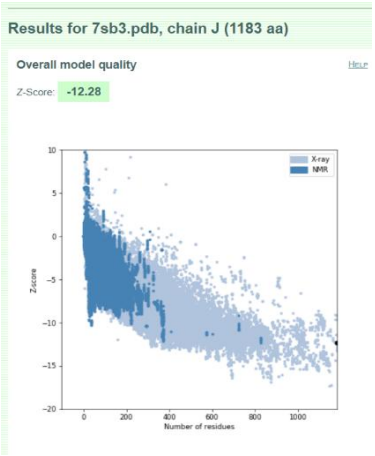


Figure 7: (7sb3) Z - score plot generated by ProSA - Web tool.

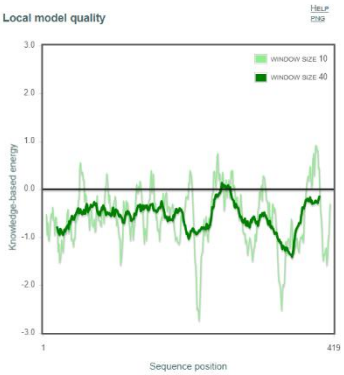


Figure 8: P0DTC9 (8fd5) residue scores plot generated by ProSA - Web

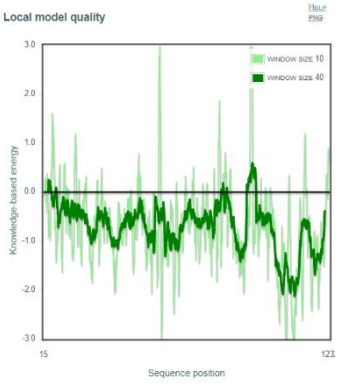


Figure 9: (7sb3) residue scores plot generated by ProSA - Web

Another graphical sequence i. From this p of given input structure. And this plot is fragment of 40 residues, since single residue evaluation. Window size of 40 - residue fragment denoted by the thick line; residue fragment is rendered in the plot as shown in the Figure 8 slightly touches. Window size 40 - residue fragment denoted by the thick line; residue fragment is rendered in the plot as shown in the Figure 9.

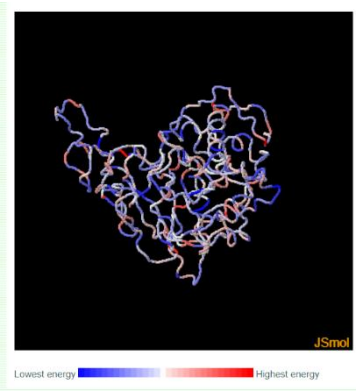


Figure 10: P0DTC9 (8fd5) Structure generated by ProSA - Web tool

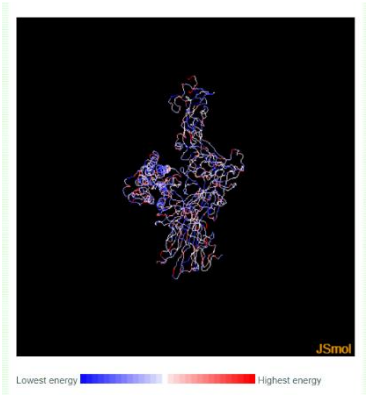


Figure 11: Structure (7sb3) score plot generated by ProSA - Web tool

In this Figure 10 and 11 shows the structure higher and lower energy the structure score plot generated by ProSA - Web. Comparison of the both structures are nucleocapsid (8fd5) is lowest energy than the Spike (7sb3) it has shown highest energy.

Model quality

Standard geometry

The Z score for a bond length (or angle) is the number of standard deviations the observed value is removed from the expected value. A bond length (or angle) with $|Z| > 5$ is considered an outlier worth inspection. RMSZ is the root - mean - square of all Z scores of the bond lengths (or angles) Nucleocapsid ID: P0DTC9 (8fd5). Table: 4

Mol	Chain	Bond RMSZ	Lengths #|Z| >5	Bond RMSZ	Angles #|Z| >5
1	A	0.74	0/3279	1.13	10/4422 (0.2%)

Torsion angles

Protein backbone

In this following table, the Percentiles column shows the percent Ramachandran outliers of the chain as a percentile score with respect to all PDB entties followed by that with respect to all EM entries. The Analysed column shows the number of residues for which the backbone conformations was analysed, and the total number of residues ID: P0DTC9. Table: 5

Mol	Chain	Analysed	Favoured	Allowed	Out liers	Percentiles
1	A	417/419 (100%)	401 (96%)	16 (4%)	0	100 100

Protein sidechains
In the following table, the Percentiles column shows the percent side chain outliers of the chain as a percentile score with respect to all PDB entries followed by that with respect to all EM entries.

The Analysed column shows the number of residues for which the side chain conformation was analysed, and the total number of residues. ID: P0DTC9. Table: 6

Mol	Chain	Analysed	Rotameric	Outliers	Percentiles
1	A	339/339 (100%)	338 (100%)	1 (0%)	92 95

All (1) residues with a non - rotameric side chain are listed below; Table: 7

Mol	Chain	Res	Type
1	A	197	SER

Sometimes side chains can be flipped to improve hydrogen bonding and reduce clashes. All (1) such sidechains are listed below; Table: 8

Mol	Chain	Res	Type
1	A	58	GLN

Map - model fit summary
The table lists the average atom inclusion at the recommended contour level (0.2) and Q - score for the entire model and for each chain. Table: 9

Chain	Atom inclusion	Q - score
All	0.5701	0.0060
A	0.5701	0.0060

- Study and analysis the protein structure prediction.
- To the similarity and dissimilarities of the protein conformations.
- Easily time consuming of data and manual setup of temperature and pH.
- The structure prediction leads to a vogue idea about the functioning of protein and its relation with other protein of its kind.
- It helps to design site - directed alterations with the goals of modifying its function.
- Better knowledge of formulations and upcoming challenging of computing methods under analysis of accuracy and compare the positions of various structure possible conformations.

5. Conclusion

The obtain ability of robust 3D structure of molecular target is very crucial for drug discovery. This analysis, Primary, secondary and 3D structures are of the Nucleocapsid protein of SARS - COV2 predicted using Expasy Program, PSIPRED, and further analysing the Homology Modelling Methods correspondingly. The structure examined of the predicted model was performed using Ramachandran plot and also validated through ProSAWeb tool. The structure examination of the nucleocapsid protein yields a valid platform for the designing specific antiviral medicine or drugs against SARS - COV2. In future, we can study molecular dynamics of above target protein to examine how the predicted model behave structurally, dynamically, thermodynamically by molecular dynamics and simulation tool in an appropriate specific protein structure.

Acknowledgement
The authors are thankful to NIMS university Jaipur Rajasthan India for providing various facilities during the research period

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J.1990. Basic local alignment search tool. J. Mol. Biol.215: 403 - 410. [Google scholar]
[2] Bairoch, A., and Apweiler, R.1999. The SWISS - PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res.27: 49 - 54. [Google scholar]
[3] Barker, W. C., Garavelli, J. S., Hou, Z., Huang, H., Ledley, R. S., McGarvey, P. B., Mewes, H. W., Orcutt, B. C., Pfeiffer, F., Tsugita, A., Vinayaka, C. R., Xiao, C., Yeh, L. S., and Wu, C.2001. Protein Information Resource: a community resource for expert annotation of protein data. NucleicAcids Res.29: 29 - 32. [Google scholar]
[4] Hayakawa, H., Koike, G., and Sekiguchi, M.1990. Expression and cloning of complementary DNA for a human enzyme that repairs O6 - methylguanine in DNA. J. Mol. Biol.213: 739 - 747. [Google scholar]
[5] Westbrook, J., Feng, Z., Chen, L., Yang, H. and Berman, H. M. (2003) The Protein Data Bank and structural genomics. Nucleic Acids Res., 31, 489 - 491.
[6] Tramontano, A., Leplae, R. and Morea, V. (2001) Analysis and assessment of comparative modelling predictions in CASP4. Proteins, 45 (Suppl.5).22 - 38.
[7] Marti - Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo F. and Sali, A. (2000) Comparative protein structure modelling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct., 29, 291 - 325.
[8] Brenner, S. E. (2001) A tour of structural genomics. Nature Rev. Genet., 2, 801 - 809.
[9] Shendure, J., &Ji, H. (2008). Next - generation DNA sequencing. Nature biotechnology, 26 (10), 1135 - 1145.
[10] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, GalabinaYordanova, David Yuan, OanaStroe, Gemma wood, AgataLaydon, Augustin Zidek, Tim Green, KathrynTunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, PushmeetKohil, Gerard Kleywegt, Ewan Birney, Demise Hassabis, Sameer Velankar, Alpha Fold Protein Structure Database: massively expanding the structural coverage of protein - sequence space with high - accuracy models, Nucleic Acids Research, Volume 50, issue D1, 7 January (2022).
[11] Dr. U. Satyanarayana. U. Chakrapani. Biochemistry with clinical concepts and case studies, 4th Edition (2013).
[12] Donald Voet, Judith G. Voet. Biochemistry, J Wiley & Sons, 2nd Edition, (1995).
[13] Carl Branden, & John Tooze. Introduction to protein structure, 2nd Edition,

[14] (1991 - 1999).
[15] Tramontano, A., Lepae, R. and Morea, V. (2001) Analysis and assessment of comparative modelling predictions in CASP4. Proteins, 45 (suppl.5).22 - 38.
[16] Marti - Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000 Comparative protein structure modelling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct., 29, 291 - 325.
[17] Rastogi, S. C., Rastogi, Parag, Mendiratta, Namita., (2022) Bioinformatics Methods and Applications.
[18] Protein Data Bank Protein Data Bank. Nat. New Bio.1971; 233. [PubMed].
[19] Berman H., Henrick K., Nakamura H. Announcing the worldwide Protein Data Bank. Nat. Struct. Biol.2003; 10: 980 - 980 [PubMed] [Google Scholar].
[20] Berman H. M., Westbook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Nucleic Acids Res.2000; 28: 235 - 242 [PubMed] [Google Scholar].
[21] Mir S., Althroub Y., Anyango S., Armstrong D. R., Berrisford J. M., Clark A. R., Conroy M. J., Dana J. M., Deshpande M., Gupta D. et al. PDBe: towards reusable data delivery infrastructure at protein data bank in Europe. Nucleic Acids Res.2017; 46: D486 - D492 [PubMed] [Google Scholar].
[22] Kinjo A. R., Bekker G. - J., SuzukiH., Tsuchiya Y., Kawabata T., ikegawa Y., Nakamura H. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. Nucleic Acids Res.2017; 45: D282 - D288 [PubMed] [Google Scholar].
[23] Ulrich E. L., Akutsu H., Doreleijers J. F., Harano Y., loannidis Y. E., Lin J., Livny M., Mading S., Maziuk D., Miller Z. et al. BioMagResBank. Nucleic Acids Res.2008; 36: D402 - D408 [PubMed] [Google Scholar].
[24] wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019 [Google Scholar].
[25] Murzin, A. G., Brenner, S. E., Hubbard, T. andChothia, C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247, 536 - 540.
[26] Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res., 30, 264 267.
[27] Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res., 32, D226 - D229.
[28] Andreeva, A., Howarth D., Chandonia, J. - M., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res., 36, D419 - D44425.
[29] Fox, N. K., Brenner, S. E. and Chandonia, J. - M. (2014) Scope: Structural Classification of Proteins - extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res., 42, D304 - D309.
[30] Chandonia, J. - M., Fox, N. K. and Brenner, S. E. (2017) SCOPe: manual curation and artefact removal in the structural classification of proteins - extended database. J. Mol. Biol., 429, 348 - 355.
[31] Chandonia, J. - M., Fox, N. K. and Brenner, S. E. (2019) Scope: classification of large macromolecular structures in the structural classification of proteins - extended database. Nucleic Acids Res., 47, D475 - D481.
[32] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. andBourne, P. E. (2000) The Protein Data Bank. Nucleic Acids Res., 235 - 242.
[33] Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M. et al. (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res., 49, D437 - D451.
[34] Andreeva, A., Kulesha, E., Gough, J. and Murzin, A. G. (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res., 48, D376 - D382.
[35] Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge., Rauer, C., Sen, N. et al. (2021) CATH: increased structural coverage of functional space. Nucleic Acids Res., 49, D266 - D273.
[36] Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B. - H. an Grishin, N. V. (2014) ECOD: an evolutionary classification of protein domains. PLoSComputBiol, 10, e1003926.
[37] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer,
[38] E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J. et al. (2021) Pfam: The protein family's database in 2021. Nucleic Acids Res., 49, D412 - D419.
[39] Fox, N. K., Brenner, S. E. and Chandonia, J. - M. (2015) The value of protein Structure classification information - surveying the scientific literature. Proteins Struct. Funct. Bioinforma., 83, 2025 - 2038.
[40] Levitt, M. and Chothia, C. (1976) Structural patterns in globular proteins. Nature, 261, 552 - 558.
[41] Chandonia, J. - M., Hon, G., Walker, N., Lo Conte, L., Koehl, P., Levitt, M. and Brenner's. (2004) The ASTRAL compendium in 2004. Nucleic Acids Res., 32, D189 - D192.
[42] Deng L, Zhong G, Liu C, Luo J, Liu H. MADOKA: an ultra - fast approach for large - scale protein structure similarity searching. BMC Bioinformatics 20, 662 (2019).
[43] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with Alpha Fold. Nature 596, 583 - 589 (2021).
[44] Oliveira SC, de Magalhaes MTQ and Homan EJ (2020) Immunoinformatic Analysis of SARS - CoV - 2 Nucleocapsid Protein and identification of COVID - 19 Vaccine Targets.

[45] Frank Qi sheng Li et al. /FEBS Letters 579 (2005) 2387 - 2396.
[46] Rajendra Kumar Azad, The molecular assessment of SARS - CoV - 2 Nucleocapsid Phosphoprotein variants among Indian isolates, Heliyon, Volume 7, issue 2, 2021, e06167, ISSN2405 - 8440.
[47] Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, He S, Zhou Z, Zhou Z, Chen Q, Yan Y. Crystal structure of SARS - CoV - 2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharmaceutical Sinica B.2020 Jul 1; 10 (7): 1228 - 38.
[48] Cortes - Sarabia, K., Luna - Pineda, V. M., Rodriguez - Ruiz, H. A. et al. Utility of in silico - identified - peptides in spike - S1 domain and nucleocapsid of SARS - CoV - 2 for antibody detection in COVID - 19 patients and antibody production in COVID - 19 patients and antibody production. Sci Rep 12, 15057 (2022).
[49] David Baker and Andrej Sali et al. Science 294, 93 (2001) Protein structure prediction and Structure Genomics.
[50] A Dawood, M Altobje, Z AlrassamMikrobio Zhu 83 (2), 82 - 92, (2021).
[51] Varsha Bhat and Jhinuk Chatterjee Alternatives to Laboratory Animals, Vol.49 (1 - 2) 22 - 32 (2021).