Identification and molecular characterization of mutations in Nucleocapsid Phosphoprotein of SARS-CoV-2

Gajendra Kumar Azad Corresp. 1

1 Department of Zoology, Patna University, Patna, Bihar, India

Corresponding Author: Gajendra Kumar Azad
Email address: gkazad@patnauniversity.ac.in

SARS-CoV-2 genome encodes four structural protein that include, Spike glycoprotein, Membrane protein, Envelope protein and Nucleocapsid Phosphoprotein (N protein). The N protein interacts with viral genomic RNA and helps in packaging. As the SARS-CoV-2 spread to almost all countries worldwide within 2-3 months; it also acquired mutations in its RNA genome. Therefore, this study was conducted with an aim to identify the variations present in N protein of SARS-CoV-2. Here, we analysed 4163 reported sequence of N protein from United States of America (USA) and compared with first reported sequence from Wuhan, China. Our study identified 107 mutations that reside all over the N protein. Further, we show the high rate of mutations in intrinsically disordered regions (IDRs) of N protein. Our study show 45% residues of IDR2 harbour mutations. The RNA binding domain (RBD) and dimerization domain of N protein also have mutations at key residues. We further measured the effect of these mutations on N protein stability and dynamicity and our data reveals that multiple mutations can cause considerable alterations. Altogether, our data strongly suggests that N protein is one of the mutational hotspot proteins of SARS-CoV-2 that is changing rapidly and these mutations can potentially interferes with various aspects of N protein functions including its interaction with RNA, oligomerization and signalling events.

TITLE

Identification and molecular characterization of mutations in Nucleocapsid Phosphoprotein of SARS-CoV-2

AUTHORS

Gajendra Kumar Azad[1#]

[1]Assistant Professor, Department of Zoology, Patna University, Patna-800005, Bihar (India)

[#]Corresponding Author:

Gajendra Kumar Azad

Email address: gkazad@patnauniversity.ac.in

Keywords: COVID-19; SARS-CoV-2; Mutations; Nucleocapsid Phosphoprotein (N protein); Infectious diseases; USA

ABSTRACT

SARS-CoV-2 genome encodes four structural protein that include, Spike glycoprotein, Membrane protein, Envelope protein and Nucleocapsid Phosphoprotein (N protein). The N protein interacts with viral genomic RNA and helps in packaging. As the SARS-CoV-2 spread to almost all countries worldwide within 2-3 months; it also acquired mutations in its RNA genome. Therefore, this study was conducted with an aim to identify the variations present in N protein of SARS-CoV-2. Here, we analysed 4163 reported sequence of N protein from United States of America (USA) and compared with first reported sequence from Wuhan, China. Our study identified 107 mutations that reside all over the N protein. Further, we show the high rate of mutations in intrinsically disordered regions (IDRs) of N protein. Our study show 45% residues of IDR2 harbour mutations. The RNA binding domain (RBD) and dimerization domain of N protein also have mutations at key residues. We further measured the effect of these mutations on N protein stability and dynamicity and our data reveals that multiple mutations can cause considerable alterations. Altogether, our data strongly suggests that N protein is one of the mutational hotspot proteins of SARS-CoV-2 that is changing rapidly and these mutations can potentially interferes with various aspects of N protein functions including its interaction with RNA, oligomerization and signalling events.

INTRODUCTION

In the late December, 2019, Wuhan, the Hubei province of China, reported a surge in hospitalisation due to pneumonia like symptoms (Zhu et al., 2020). The causative agent was identified as a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that shares close similarity with earlier known SARS-CoV (Chen et al., 2020). The SARS-CoV-2 is highly contagious that lead to its rapid spread worldwide, and in March 2020, the World Health Organization (WHO) declared the outbreak a pandemic. The disease caused by SARS-CoV-2 has been named as coronavirus disease 19 (COVID-19) that exhibits mild to severe respiratory distress in the infected individuals. As of 28th June, 2020 the COVID-19 has affected all countries worldwide with close to 10 million reported cases and 0.5 million confirmed deaths. Further, the epidemiological studies revealed that the mortality rate from COVID-19 is significantly higher among individuals over 60 years of age with weak immunity (Liu et al., 2020).

The SARS-CoV-2 has positive sense, single stranded RNA genome of approximately 29.8 kb (Wu et al., 2020b). The majority of viral genome encodes non-structural proteins that are proteolytically processed from a single Orf1ab polypeptide. SARS-CoV-2 genome also encode four structural proteins, including the Spike glycoprotein (S), Membrane protein (M), Envelope protein (E) and Nucleocapsid Phosphoprotein (N) (Wu et al., 2020a). The S, M and E proteins are located in the lipid bilayer of the virus and contribute to the formation of viral envelope; however, the N protein contributes to the viral genomic RNA packaging and remains embedded in the central core of the virion. N protein binds with viral genomic RNA and forms helical structure to maintain the structural integrity of RNA genome (Chang et al., 2014). This is one of the most abundant structural proteins encoded by the SARS-CoV-2 genome. The SARS-CoV-2 N protein resembles N protein from other RNA viruses, known to modulate host intracellular machinery and also involved in the regulation of virus life cycle.

The crystal structure of N protein revealed two distinct domains at N and C terminus (Kang et al., 2020). The domain present towards the N- terminus also known and RNA binding domain (RBD). The C-terminal side harbours dimerization domain which interacts with other N protein to make dimer. Apart from these two domains there are three intrinsically disordered regions (IDRs) at N and C terminal ends as well as between the RBD and dimerization domain of N protein. Since, this protein plays critical role in packaging of SARS-CoV-2 RNA genome, the mutations in N protein or interfering its function can lead to diverse outcome on viral life cycle (Rabi Ann Musah, 2005; Chenavas et al., 2013). Here, we compared the N protein sequences obtained from USA with first reported sequence from China to identify the variations present between them. We identified 107 mutations and their impact on N protein structure and function are discussed.

MATERIALS AND METHODS

Sequence retrieval from NCBI-virus-database

The NCBI-virus- database stores the deposited sequences of SARS-CoV-2 which is updated regularly as the new sequences are reported. As of 23rd June 2020, 4163 SARS-CoV-2 sequences of N protein are being deposited from USA. We downloaded these sequences and used them for analysis in this study. The first reported N protein sequence from Wuhan was used as reference sequence or wild type sequence (Wu et al., 2020b). The protein accession identification number of reference sequence used in this study is YP_009724397 and rest of the 4163 IDs (reported from USA) are mentioned in supplementary table 1.

Multiple sequence alignment by Clustal-Omega programe

To identify the mutations present in the SARS-CoV-2 N protein reported from USA, we did multiple sequence alignments and compared them with the first reported N protein sequence (YP_009724397) from Wuhan, China. The multiple sequence alignment was performed using Clustal Omega tool (Madeira et al., 2019).

Calculation of free energy and vibrational entropy between wild type and mutant N proteins

In order to measure the impact of mutations identified in this study on the structural dynamicity and stability of N protein, we calculated the differences in free energy (ΔΔG) and vibrational entropy (ΔΔSvib) ENCoM between wild type and mutants. This analysis was performed by DynaMut programe (Rodrigues, Pires & Ascher, 2018). To perform DynaMut protein modelling we used RCSB protein ID: 6VYO(Kang et al., 2020) for RBD molecular modelling and RCSB protein ID: 6WJI for dimerization domain molecular modelling of N protein. The positive ΔΔG represents stability and negative ΔΔG represents destabilisation of protein structure upon mutation. Similarly, positive ΔΔSvibENCoM represents increase in flexibility and negative ΔΔSvibENCoM represents increase in rigidity of protein structure. DynaMut also provide the visual representation of fluctuation in protein structure. The blue colour represents gain in rigidity and red colour represents the gain in flexibility upon mutation.

RESULTS

Identification of mutations in IDR1, IDR2 and IDR3 of N-protein

The crystal structure of N protein of SARS-CoV-2 has been recently solved (Kang et al., 2020), the structural details show it is comprised of three distinct regions; the N-terminal domain (contains RNA binding domain), C-terminal domain (contains dimerization domain) and IDRs as shown in figure 1. There are three IDRs in N protein; IDR1 (at the N-terminal end), IDR2 (between RBD and CTD) and IDR3 (at the C-terminal end). IDR2 is also referred as linker region (LKR) because it connects RBD and dimerization domain of N protein. In order to identify the variations present in N protein of SARS-CoV-2 reported from the USA, we performed multiple sequence alignments. Here, we used Clustal Omega programe to align 4163 N protein polypeptide sequences from USA and compared them with the first reported sequence from Wuhan, China.

Our analysis identified eighteen mutations in IDR1 (Table1). The IDR1 is present from 1-43 residues towards the N- terminal end of N protein. These eighteen mutations correspond to approximately 40% (18 out of 43) of the residues of IDR1. Among these the most frequently

mutated residues are Gly and Arg (both are mutated at four positions) and Pro residue is mutated at three different positions in IDR1 (Table 1).

Similar analysis with IDR2 identified thirty six mutations which correspond to approximately 45% of residues of IDR2 (Table 2). The IDR2 is present from 181-256 residues of the N protein and connects RBD and dimerization domains. The most frequently mutated residue in IDR2 was found to be Ser, it is mutated at twelve positions. Further, the Ala, Gly and Arg residues are mutated at five positions, respectively.

Similarly, we identified fifteen mutations in IDR3 (Table 3). The IDR3 is present from 365-419 residues towards the C- terminal end of N protein. Most notable mutations are Thr and Ala residues are mutated at three positions and Pro, Asp, and Gln are mutated at two positions, respectively (Table 3). Altogether, we identified sixty nine mutations in intrinsically disordered regions IDR1, IDR2 and IDR3 of N protein.

Identification of mutations in RBD and dimerization domain of N-protein

The RBD of N-protein starts from 44th residue till 180th residue. We mapped the mutation in this region of N protein and our analysis revealed presence of twenty two mutations (Table 4). These twenty two mutations also correspond to approximately 16% of the residues of RBD. Our mutational analysis shows the most frequently mutated residues are Pro and Ala at five positions and Asp at three positions as shown in table 4.

Similar analysis with the dimerization domain of N protein revealed that it harbours sixteen mutations. The dimerization domain of N-protein starts from 257th residue till 364th residue. Our mutational analysis shows Thr is mutated at four positions and Asp at three positions. Further, only 14 % residues are mutated in this domain which is least among all other regions of the N-protein identified here. Altogether, we identified thirty eight mutations in RBD and dimerization domain of N protein.

Mutations causes alteration in dynamic stability of N protein

In order to understand the effect of mutations on the stability of the protein we calculated the differences in free energy (ΔΔG) between wild type and mutants. We performed this analysis using DynaMut programe. The positive ΔΔG corresponds to increase in stability while negative ΔΔG corresponds to decrease in stability. We performed this analysis with all of the mutations that reside in RBD and dimerization domain of N protein. The IDRs do not have proper 3D structure therefore; this analysis is not accurate for those regions. Our data revealed the noticeable increase or decrease in free energy in various mutations as shown in table 6. The top

five positive and negative ΔΔG values are highlighted in table 6. The maximum increase in ΔΔG was observed for T271I (1.184 kcal/mol) and the highest negative ΔΔG was obtained for I292T (-1.952 kcal/mol), both of these mutations reside in dimerization domain of N protein.

We also measured the change in vibrational entropy energy ($\Delta\Delta S_{Vib}$ENCoM) between the wild type and the mutants present in RBD and dimerization domain of N protein (Table 6).Vibration entropy contributes to the configurational-entropy of the proteins (Goethe, Fita & Rubi, 2015). The negative $\Delta\Delta S_{Vib}$ENCoM of mutant N protein corresponds to the increase in rigidification and positive $\Delta\Delta S_{Vib}$ENCoM corresponds to gain in flexibility of the protein structure. The maximum positive $\Delta\Delta S_{Vib}$ENCoM was obtained for P364L (0.256 kcal.mol-1.K-1) and negative $\Delta\Delta S_{Vib}$ENCoM was obtained for G284E (-0.844 kcal.mol-1.K-1). The variation in vibrational entropy between wild type and mutant can also be visualised as shown in figure 2. The blue colour corresponds to rigidification in protein structure and red colour corresponds to gain in flexibility upon mutation. The top three positive and negative $\Delta\Delta S_{Vib}$ENCoM are shown in figure 2 (A-F). Altogether, the data obtained from ΔΔG and $\Delta\Delta S_{Vib}$ENCoM strongly suggests that the mutations identified in this study can influence N protein stability and dynamicity.

Intramolecular interactions are altered due to mutations in N protein

Next, we sought to closely analyse the changes in the intramolecular interactions in some of the mutants that exhibited significant alterations in ΔΔG. We compared the intramolecular interaction for T271I (ΔΔG: 1.184 kcal/mol) and I292T (ΔΔG: -1.952 kcal/mol) as these two mutants showed maximum variations among thirty eight mutants present in RBD and dimerization domain of N protein (Table 4 and 5). Our data clearly showed the variations in the interactions mediated by wild type and mutant residues in the pocket, where these amino acids resides as shown in figure 3A-B (T271I) , and 3C-D (I292T). Altogether, our data strongly suggests that the mutants identified in our study are affecting the dynamic stability as well as intramolecular interactions in the N protein.

DISCUSSIONS

SARS-CoV-2 is an RNA virus, a causative agent of COVID-19. This virus spread worldwide within a span of few months and during its spread it also acquired mutations. One such example is D614G on Spike glycoprotein(Korber et al., 2020) of SARS-CoV-2. This mutation is the most frequent mutation observed in most of the countries including USA, European countries, India and others indicating that it has now become the dominant pandemic form in many countries (Korber et al., 2020). Even though the appearance of diversity among various protein encoded

by SARS-CoV-2 is low, owing to its quick global spread, enable this virus for accelerated natural selection events that might lead to generation of rare favourable variants. Hence, the studies to understand the mutations in SARS-CoV-2 warrant thorough investigations. This study was performed with an aim to identify mutations in N protein which is one of the main structural proteins of SARS-CoV-2. Here, we analysed 4163 sequences of N protein from USA and identified 107 mutations upon comparison from first reported sequences of the same protein from Wuhan, China. We also observed around 64% (69 out of 107) of these mutations reside in the IDRs of N protein. Among IDRs, the IDR2 harbours 36 mutations that correspond to the most number of mutations observed in a single distinct region of the N protein.

Earlier studies demonstrated that Ser and Arg-rich linker region (IDR2) plays indispensible role in intracellular signalling events primary by phosphorylation at Ser residues (Wootton, Rowland & Yoo, 2002; McBride, van Zyl & Fielding, 2014). The wild type LKR/ IDR2 contains sixteen Ser residues, and our study revealed that out of those, twelve serine residues are mutated (table 2). Therefore, we can safely assume that these mutations of Ser residues might contribute to alteration of phosphorylation dependent signalling.

We also measured ΔΔG and $\Delta\Delta S_{Vib}$ENCoM for the mutants that reside in the RBD and dimerization domain of N protein. The four mutants that exhibited highest values for ΔΔG and $\Delta\Delta S_{Vib}$ENCoM identified in our study are T271I, I292T, G284E and P364L. Since, all of them are in the dimerization domain; therefore, it is possible that these mutations might lead to alteration in the dimerization potential of N protein.

Evidences indicate that the N protein of coronaviruses functions as an RNA chaperones (Zúñiga et al., 2007, 2010) and also contributes to packaging and maintenance of the RNA genome. Hence, the drugs that can either inhibit the interactions of RNA with N protein or interfere with dimerization of N protein can be a potential antiviral candidates (Lo et al., 2013). One such drug is Nucleozin and its derivatives that targets ribonucleoprotein formation in influenza virus by interfering N protein oligomerization (Gerritz et al., 2011). Furthermore, a recent study was conducted to identify inhibitors of SARS-CoV-2 N protein, identified various promising candidate drugs including Conivaptan, Ergotamine, Venetoclax and Rifapentine (Onat Kadioglu, 2020). Most of these candidate drugs interact with the residues that are either mutated or are in the close vicinity of the mutations identified in our study. Altogether, the mutation revealed in this study can interfere with various aspects of N protein functions that include oligomerization, interaction with RNA and interference in N protein mediated signalling events.

CONCLUSION

In this study we identified 107 mutations in N protein of SARS-CoV-2 reported from USA. Further, we demonstrate these mutations can potentially alter dynamic stability of N protein. Altogether, the data presented here, warrants further investigations to understand its impact on SARS-CoV-2 phenotype and drugs that target N protein.

ACKNOWLDEGEMENTS

We would like to acknowledge the Department of Zoology, Patna University, Patna, Bihar (India) for providing infrastructural support for this study.

FIGURE AND TABLE LEGENDS

Figure 1: The schematic structure of Nucleocapsid Phosphoprotein (N protein) of SARS-CoV-2. N protein is a structural protein encoded by SARS-CoV-2 RNA genome. It is 419 residues in length and contains RNA binding domain and Dimerization domain towards N and C-terminus, respectively. N protein also has three intrinsically disordered regions (IDRs), IDR1, IDR2 and IDR3.

Figure 2: Visual representation of Δ Vibrational Entropy Energy between Wild-Type and Mutant N protein. The amino acids residues are colored according to the vibrational entropy change as a consequence of mutation of N protein. **BLUE** represents a rigidification of the structure and **RED** a gain in flexibility. (A-C) represents the top three mutants that show rigidification in structure upon mutation. (D-F) represents the top three mutants that show gain in flexibility upon mutation. Each panel also shows the mutation and the location of the residues.

Figure 3: Visual representation of interatomic interactions contributed by T271I and I292T of N protein. Both of these mutants showed maximum positive and negative ΔΔG among mutants present in RBD and dimerization domain of N protein. (A-B) represents threonine to isoleucine substitution at 271st position; (C-D) represents isoleucine to threonine substitution at 292nd position. Wild-type and mutant residues are represented in light-green color. The interactions made by wild type and mutant residues are highlighted in each panel.

Table 1: The table show the location and details of mutations identified in IDR1 of N protein. These mutations are present among the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this

analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

Table 2: The table show the location and details of mutations identified in IDR2. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

Table 3: The table show the location and details of mutations identified in IDR3. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

Table 4: The table show the location and details of mutations identified in RBD of N protein. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

Table 5: The table show the location and details of mutations identified in dimerization domain of N protein. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

Table 6: The table show the ΔΔG and ΔΔSvib ENCoM of the mutants present in RBD and dimerization domain of N-protein. DynaMut programe was used to calculate both parameters. The top five positive and negative ΔΔG values are highlighted in bold digits. The top three positive and negative ΔΔSvib ENCoM values are highlighted in bold digits.

REFERENCES

Chang CK, Hou MH, Chang CF, Hsiao CD, Huang TH. 2014. The SARS coronavirus

306 nucleocapsid protein - Forms and functions. *Antiviral Research*. DOI:
307 10.1016/j.antiviral.2013.12.009.
308 Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T,
309 Zhang X, Zhang L. 2020. Epidemiological and clinical characteristics of 99 cases of 2019
310 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*. DOI:
311 10.1016/S0140-6736(20)30211-7.
312 Chenavas S, Cre´pin T, Delmas B, Ruigrok RWH, Slama-Schwok A. 2013. Influenza virus
313 nucleoprotein: Structure, RNA binding, oligomerization and antiviral drug target. *Future*
314 *Microbiology*. DOI: 10.2217/fmb.13.128.
315 Gerritz SW, Cianci C, Kim S, Pearce BC, Deminie C, Discotto L, McAuliffe B, Minassian BF, Shi
316 S, Zhu S, Zhai W, Pendri A, Li G, Poss MA, Edavettal S, McDonnell PA, Lewis HA, Maskos
317 K, Mo¨rtl M, Kiefersauer R, Steinbacher S, Baldwin ET, Metzler W, Bryson J, Healy MD,
318 Philip T, Zoeckler M, Schartman R, Sinz M, Leyva-Grado VH, Hoffmann HH, Langley DR,
319 Meanwell NA, Krystal M. 2011. Inhibition of influenza virus replication via small molecules
320 that induce the formation of higher-order nucleoprotein oligomers. *Proceedings of the*
321 *National Academy of Sciences of the United States of America*. DOI:
322 10.1073/pnas.1107906108.
323 Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, He S, Zhou Z, Zhou Z, Chen Q, Yan Y,
324 Zhang C, Shan H, Chen S. 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein
325 RNA binding domain reveals potential unique drug targeting sites. *Acta Pharmaceutica*
326 *Sinica B*. DOI: 10.1016/j.apsb.2020.04.009.
327 Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, Foley B, Giorgi EE,
328 Bhattacharya T, Parker MD, Partridge DG, Evans CM, Silva T de, LaBranche CC,
329 Montefiori DC, Group SC-19 G. 2020. Spike mutation pipeline reveals the emergence of a
330 more transmissible form of SARS-CoV-2. *bioRxiv*. DOI: 10.1101/2020.04.29.069054.
331 Liu K, Chen Y, Lin R, Han K. 2020. Clinical features of COVID-19 in elderly patients: A
332 comparison with young and middle-aged patients. *Journal of Infection*. DOI:
333 10.1016/j.jinf.2020.03.005.
334 Lo YS, Lin SY, Wang SM, Wang CT, Chiu YL, Huang TH, Hou MH. 2013. Oligomerization of the
335 carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. *FEBS*
336 *Letters*. DOI: 10.1016/j.febslet.2012.11.016.
337 Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter
338 SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in
339 2019. *Nucleic acids research*. DOI: 10.1093/nar/gkz268.

McBride R, van Zyl M, Fielding BC. 2014. The coronavirus nucleocapsid is a multifunctional protein. *Viruses*. DOI: 10.3390/v6082991.

Onat Kadioglu MSHJGTE. 2020. Identification of novel compounds against three targets of SARS CoV2 coronavirus by combined virtual screening and supervised machine learning . *Bull World Health Organ*. DOI: 10.2471/BLT.20.251561.

Rabi Ann Musah. 2005. The HIV-1 Nucleocapsid Zinc Finger Protein as a Target of Antiretroviral Therapy. *Current Topics in Medicinal Chemistry*. DOI: 10.2174/1568026043387331.

Rodrigues CHM, Pires DEV, Ascher DB. 2018. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Research*. DOI: 10.1093/nar/gky300.

Wootton SK, Rowland RRR, Yoo D. 2002. Phosphorylation of the Porcine Reproductive and Respiratory Syndrome Virus Nucleocapsid Protein. *Journal of Virology*. DOI: 10.1128/jvi.76.20.10569-10576.2002.

Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, Sheng J, Quan L, Xia Z, Tan W, Cheng G, Jiang T. 2020a. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host and Microbe*. DOI: 10.1016/j.chom.2020.02.001.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020b. A new coronavirus associated with human respiratory disease in China. *Nature*. DOI: 10.1038/s41586-020-2008-3.

Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*. DOI: 10.1056/NEJMoa2001017.

Zúñiga S, Cruz JLG, Sola I, Mateos-Gómez PA, Palacio L, Enjuanes L. 2010. Coronavirus Nucleocapsid Protein Facilitates Template Switching and Is Required for Efficient Transcription. *Journal of Virology*. DOI: 10.1128/jvi.02011-09.

Zúñiga S, Sola I, Moreno JL, Sabella P, Plana-Durán J, Enjuanes L. 2007. Coronavirus nucleocapsid protein is an RNA chaperone. *Virology*. DOI: 10.1016/j.virol.2006.07.046.

Table 1(on next page)

IDR1 Mutations

The table show the location and details of mutations identified in IDR1 of N protein. These mutations are present among the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

1 Table 1:

S.No.	Wild type residue	Residue position	Mutated residue
1	Asp	3	Tyr
2	Asn	4	Asp
3	Pro	6	Thr
4	Gln	9	His
5	Pro	13	Leu
6	Arg	14	His
7	Gly	18	Cys
8	Gly	19	Arg
9	Pro	20	Leu
10	Asp	22	Tyr
11	Ser	23	Thr
12	Gly	30	Ala
13	Glu	31	Asp
14	Arg	32	Leu
15	Gly	34	Leu
16	Ala	35	Thr
17	Arg	36	Leu
18	Arg	40	Cys
19	Arg	40	Leu

2

Table 2(on next page)

IDR2 mutations

The table show the location and details of mutations identified in IDR2. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

1 Table 2:

S. No.	wild type residue	Position of mutation	Mutated residue
1	Ser	183	Tyr
2	Arg	185	Cys
3	Arg	185	Leu
4	Ser	187	Leu
5	Ser	188	Leu
6	Ser	190	Ile
7	Arg	191	Leu
8	Asn	192	Ser
9	Ser	193	Ile
10	Ser	194	Leu
11	Arg	195	Ile
12	Ser	197	Leu
13	Pro	199	Ser
14	Ser	202	Asn
15	Arg	203	Lys
16	Arg	203	Met
17	Gly	204	Arg
18	Thr	205	Ile
19	Ala	208	Gly
20	Arg	209	Lys
21	Arg	209	Thr
22	Ala	211	Ser
23	Gly	212	Cys
24	Asn	213	Tyr
25	Gly	215	Ser
26	Ala	218	Val
27	Ala	220	Thr
28	Gln	229	His
29	Ser	232	Arg
30	Ser	232	Thr
31	Met	234	Ile
32	Ser	235	Pro
33	Ser	235	Phe
34	Gly	236	Val
35	Gly	238	Cys
36	Gly	243	Cys
37	Thr	247	Ala
38	Lys	249	Arg
39	Ser	250	Phe

40	Ala	252	Ser
41	Ser	255	Ala

2

Table 3(on next page)

IDR3 mutations

The table show the location and details of mutations identified in IDR3. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

1 Table 3:

S. No.	wild type residue	Position of mutation	Mutated residue
1	Pro	365	Ser
2	Pro	365	Leu
3	Asp	377	Tyr
4	Asp	377	Gly
5	Thr	379	Ile
6	Gln	380	His
7	Ala	381	Val
8	Pro	383	Ser
9	Pro	383	Leu
10	Gln	386	Lys
11	Gln	386	His
12	Thr	391	Ile
13	Thr	393	Ile
14	Ala	397	Ser
15	Ala	398	Val
16	Asp	399	Glu
17	Ser	413	Ile
18	Ser	416	Leu

2

3

4

Table 4(on next page)

RBD mutations

The table show the location and details of mutations identified in RBD of N protein. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

1 Table 4:

S. No	wild type residue	Position Of mutation	Mutated residue
1	Pro	46	Ser
2	Glu	62	Val
3	Pro	67	Ser
4	Asp	81	Tyr
5	Ala	90	Ser
6	Ala	119	Ser
7	Pro	122	Leu
8	Ala	125	Thr
9	Asp	128	Tyr
10	Asn	140	Thr
11	Pro	142	Ser
12	Asp	144	Tyr
13	Asp	144	His
14	Ile	146	Phe
15	Pro	151	Leu
16	Ala	152	Ser
17	Asn	154	Tyr
18	Ala	156	Ser
19	Gln	163	Arg
20	Thr	166	Ile
21	Lys	169	Arg
22	Ser	180	Ile

2

Table 5(on next page)

Dimerization domain mutations

The table show the location and details of mutations identified in dimerization domain of N protein. These mutations are present in the N protein of SARS-CoV-2 reported from USA. The first reported sequence of N protein from Wuhan, China was used as wild type sequence for this analysis. The table shows only those residues that have variation, rest of the sequences are identical among all samples.

1 Table 5:

S.No.	wild type residue	Residue position	Mutated residue
1	Val	270	Leu
2	Thr	271	Ile
3	Gly	284	Glu
4	Gln	289	His
5	Ile	292	Thr
6	Gln	294	Leu
7	Asp	297	Val
8	Pro	309	Leu
9	Met	322	Ile
10	Ser	327	Leu
11	Thr	329	Met
12	Thr	334	Ile
13	Asp	340	Gly
14	Asp	340	Asn
15	Asp	348	Tyr
16	Thr	362	Ile
17	Pro	364	Leu

2

Table 6(on next page)

ΔΔG and ΔΔSvib ENCoM calculations

The table show the ΔΔG and ΔΔSvib ENCoM of the mutants present in RBD and dimerization domain of N-protein. DynaMut programe was used to calculate both parameters. The top five positive and negative ΔΔG values are highlighted in bold digits. The top three positive and negative ΔΔSvib ENCoM values are highlighted in bold digits.

1 Table 6:

S.No.	Mutant	PDB ID	ΔΔG (kcal/mol)	ΔΔSvibENCoM (kcal.mol-1.K-1)
1	E62V	6VYO	0.105	0.091
2	P67S	6VYO	**-0.486**	0.16
3	D81Y	6VYO	0.454	-0.425
4	A90S	6VYO	0.274	0.043
5	A119S	6VYO	0.073	-0.069
6	P122L	6VYO	**-0.166**	-0.049
7	A125T	6VYO	**-0.565**	-0.022
8	D128Y	6VYO	**0.846**	-0.236
9	N140T	6VYO	0.318	-0.177
10	P142S	6VYO	0.26	-0.17
11	D144Y	6VYO	0.291	-0.293
12	D144H	6VYO	-0.036	0.06
13	I146F	6VYO	0.708	**-0.837**
14	P151L	6VYO	**0.771**	-0.14
15	A152S	6VYO	0.298	-0.051
16	N154Y	6VYO	-0.096	-0.063
17	A156S	6VYO	0.428	-0.256
18	Q163R	6VYO	-0.092	-0.017
19	T166I	6VYO	0.194	-0.055
20	K169R	6VYO	0.231	0.077
21	V270L	6WJI	0.679	-0.194
22	T271I	6WJI	**1.184**	-0.472
23	G284E	6WJI	0.553	**-0.844**
24	Q289H	6WJI	0.18	**0.181**
25	I292T	6WJI	**-1.952**	**0.186**
26	Q294L	6WJI	0.447	-0.078
27	D297V	6WJI	-0.113	-0.072
28	P309L	6WJI	**0.887**	**-0.524**
29	M322I	6WJI	**-0.348**	0.045
30	S327L	6WJI	**0.894**	-0.259
31	T329M	6WJI	0.569	-0.189
32	T334I	6WJI	0.236	-0.115
33	D340G	6WJI	0.398	-0.114
34	D340N	6WJI	0.194	-0.088
35	D348Y	6WJI	0.136	-0.121
36	T362I	6WJI	0.396	0.047
37	P364L	6WJI	-0.061	**0.256**

2

Figure 1

The schematic structure of Nucleocapsid Phosphoprotein (N protein) of SARS-CoV-2.

N protein is a structural protein encoded by SARS-CoV-2 RNA genome. It is 419 residues in length and contains RNA binding domain and Dimerization domain towards N and C-terminus, respectively. N protein also has three intrinsically disordered regions (IDRs), IDR1, IDR2 and IDR3.

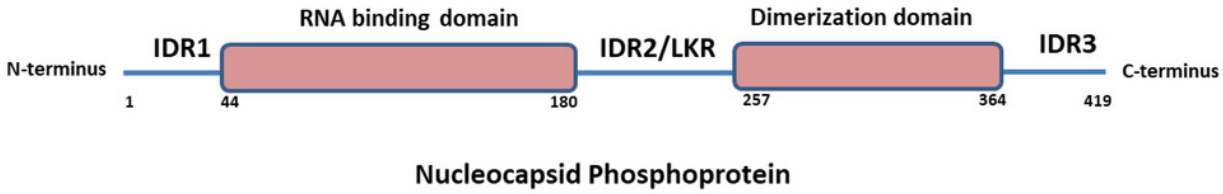


Figure 2

Visual representation of Δ Vibrational Entropy Energy between Wild-Type and Mutant N protein.

The amino acids residues are colored according to the vibrational entropy change as a consequence of mutation of N protein.**BLUE** represents a rigidification of the structure and **RED** a gain in flexibility. (A-C) represents the top three mutants that show rigidification in structure upon mutation. (D-F) represents the top three mutants that show gain in flexibility upon mutation. Each panel also shows the mutation and the location of the residues.

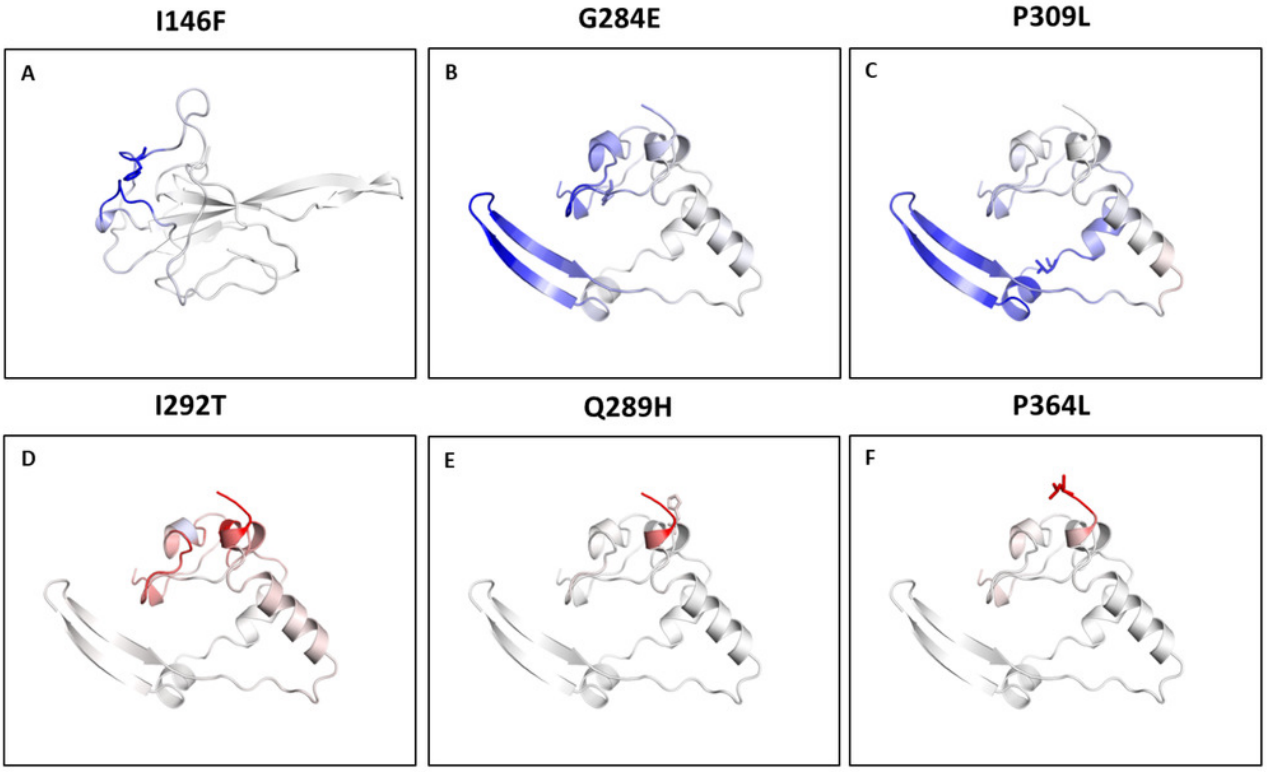
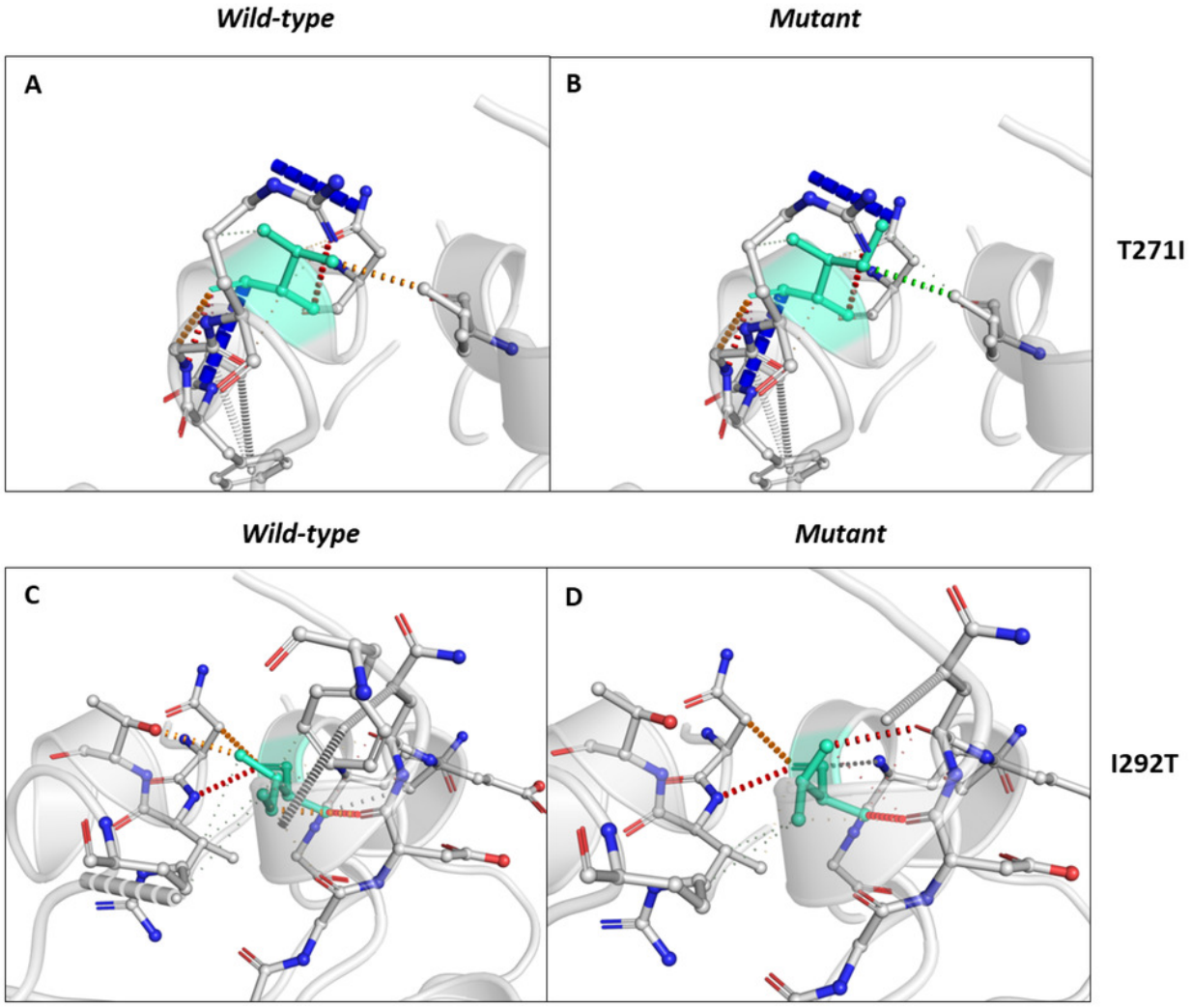


Figure 3

Analysis of interatomic interactions.

Visual representation of interatomic interactions contributed by T271I and I292T of N protein. Both of these mutants showed maximum positive and negative ΔΔG among mutants present in RBD and dimerization domain of N protein. (A-B) represents threonine to isoleucine substitution at 271st position; (C-D) represents isoleucine to threonine substitution at 292nd position. Wild-type and mutant residues are represented in light-greencolor. The interactions made by wild type and mutant residues are highlighted in each panel.


Wild-type
Mutant
A
B
T271I
Wild-type
Mutant
C
D
I292T