Structural dynamics of SARS-CoV-2 nucleocapsid protein induced by RNA binding

Helder Veras Ribeiro Filho[13], Gabriel Ernesto Jara[13], Fernanda Aparecida Heleno Batista[1], Gabriel Ravanhani Schleder[2], Celisa Caldana Tonoli[1], Adriana Santos Soprano[1], Samuel Leite Guimarães[1], Antonio Carlos Borges[2], Alexandre Cassago[2], Marcio Chaim Bajgelman[1], Rafael Elias Marques[1], Daniela Barreto Barbosa Trivella[1], Kleber Gomes Franchini[1], Ana Carolina Migliorini Figueira[1], Celso Eduardo Benedetti[1*] and Paulo Sergio Lopes de Oliveira[1*]

*Corresponding authors: paulo.oliveira@lnbio.cnpem.br, celso.benedetti@lnbio.cnpem.br

[1]Brazilian Biosciences National Laboratory, Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, SP, Brazil

[2]Brazilian Nanotechnology National Laboratory, Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, SP, Brazil

[3]Contributed equally

Keywords: SARS-CoV-2; nucleocapsid protein; N protein; COVID-19

Abstract

The nucleocapsid (N) protein of the SARS-CoV-2 virus, the causal agent of COVID-19, is a multifunction phosphoprotein that plays critical roles in the virus life cycle, including transcription and packaging of the viral RNA. To play such diverse roles, the N protein has two structured RNA-binding modules, the N- (NTD) and C-terminal (CTD) domains, which are connected by an intrinsically disordered region. Despite the wealth of structural data available for the isolated NTD and CTD, how these domains are arranged in the full-length protein and how the oligomerization of N influences its RNA-binding activity remains largely unclear. Herein, using experimental data from electron microscopy and biochemical/biophysical techniques combined with molecular modeling and molecular dynamics simulations, we show that, in the absence of RNA, the N protein forms structurally dynamic dimers with the NTD and CTD arranged in extended conformations. In the presence of RNA, however, the N protein assumes a more compact conformation where the NTD and CTD are packed together. We also provide an octameric model for the full-length N bound to RNA that is consistent with electron microscopy images of the N protein in the presence of RNA. Together, our results shed new light on the dynamics and higher-order oligomeric structure of this versatile protein.

Introduction

All coronaviruses, including the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the causative agent of the Coronavirus disease 2019 (COVID-19) pandemics, possess an organized nucleocapsid formed by a ribonucleoprotein (RNP) complex surrounded by a lipid envelope[1-3]. The major component of the RNP complex is the nucleocapsid (N) protein, one of the four structural proteins of coronaviruses and also the most abundantly expressed viral protein in infected host cells[2].

N proteins are conserved among coronaviruses and are known to play multiple roles in the virus life cycle[4]. In addition to packaging the viral genomic RNA, N proteins are required for genome replication, transcription and translation, and for the assembly of the RNPs into newly formed viral particles[5-13]. This functional diversity is intimately linked to the dynamic structure of the N protein and its ability to bind and alter the RNA structure[9,14].

Coronaviruses N proteins are composed of two structured and globular domains represented by the N- (NTD) and C-terminal (CTD) domains, both of which are capable of binding single-stranded RNA and DNA molecules[15-24]. The NTD has an extensive basic U-shaped RNA-binding cleft implicated in the binding and melting of transcription regulatory sequences (TRS) needed for transcription of sub-genomic RNAs [9,14]. The CTD is responsible for the protein dimerization and it also forms a positively charged groove thought to contribute to the recognition of the packaging signal (PS) and to the assembly of the RNP into the virion particle[13,17,23].

The NTD and CTD are connected by a central disordered serine and arginine-rich region, denoted as SR linker. This linker region is also proposed to play fundamental roles in protein oligomerization and function. Of note, the SR linker was shown to be modified by phosphorylation, which not only reduces the affinity of the protein for the RNA, but also drives a liquid phase separation of the N protein with the RNA and other virus proteins and

host cell components[25-27]. In addition to the CTD, the flexible C-terminal tail also seems to influence protein oligomerization by promoting protein tetramerization[28-30].

Due to its dynamic structure and multiple oligomeric organization depending on environmental conditions, and to the fact that the N protein is also modified by phosphorylation[25,26,31], no three-dimensional (3D) structures are available for coronaviruses full-length N proteins. Here, using electron microscopy and biophysical analysis to guide molecular modeling and molecular dynamics simulations, we propose structural models for the full-length SARS-CoV-2 N protein in the absence and presence of RNA that not only support current experimental data, but also provide a framework for understanding the multifunctional role of this protein.

Results

Dimers of full-length N adopts an extended conformation in the absence of RNA

To expand our knowledge on the structure of full-length N and to understand how the NTD and CTD are oriented to each other, the recombinant N protein produced in *E. coli* was purified by affinity and size-exclusion chromatography (SEC) and analyzed by negative stain electron microscopy (NSEM). We noticed that the N protein, which appeared as a single peak in the SEC and as a single band in denaturing gel electrophoresis (Figure 1A), contained traces of bacterial RNA in non-denaturing gels. Accordingly, treatments with RNase A, but not DNase I, removed most of the nucleic acid associated with the protein (Figure 1B). Notably, removal of the contaminating RNA with RNase A led to a change in the oligomeric state of the protein (Figure 1B).

To investigate how the N protein behaves in the absence of RNA, the recombinant protein was treated with urea and high salt concentration to remove bound RNA[25,32,33], and purified by affinity and SEC. The urea-treated N eluted as a major peak with a higher elution volume

compared with the untreated protein (Figure 1A) and showed a molecular mass (87,2 ± 3.9 kDa) and hydrodynamic radius (9.9 ± 1.3 nm) determined by SEC-MALS and DLS, respectively, consistent with a dimer in solution (Figure 1, C and D). These results are in line with literature data showing that the N protein readily oligomerizes into dimers[34,35]. The urea-treated N showed a circular dichroism profile comparable to that of the protein treated with RNase A (Figure 1E), indicating that the urea treatment did not substantially alter the protein secondary structure.

Negative stain images of the urea-treated N samples showed various amorphous particles with dimensions smaller than 5 nm (Figure 2A). The dimensions of these particles are thus not compatible with the hydrodynamic radius determined experimentally for the full-length N protein (Figure 1D); nevertheless, these particles could correspond to the NTD or CTD globular regions.

To investigate this hypothesis, we picked 1500 particles and performed a two-dimensional (2D) classification to calculate the area of each class average. We compared this distribution of areas with the distribution of areas of simulated reprojections derived from the monomeric NTD and dimeric CTD atomic structures (Figure 2B). Notably, we found that the area distributions of the NTD monomer or CTD dimer fitted well into the area distributions of the class averages (Figure 2B). This suggested that the particles observed in the NSEM images (Figure 2A) likely correspond to the NTD and CTD regions of the N protein dimer in extended conformations. This idea is supported by coarse-grained (CG) dynamics simulations of the N protein dimer, which show that the NTD regions move freely in relation to the CTD dimer, producing a variety of conformers (Figure 2C) with an estimated mean radius of gyration of 6.6 nm (Figure 2D), a value that is close to the radius of gyration of full-length N determined by SAXS and NMR[19,34]. Moreover, the CG simulations do not point to NTD-CTD contacts (Figure S1), supporting the idea that the large density of positive charges on the surface of the NTD and CTD causes repulsion of these domains, which is favored by the

flexibility of the SR linker. The simulated reprojections derived from representative 3D models of the CG simulation also illustrate the diversity of N protein conformations and show density regions of similar size of the particles observed in the NSEM images (Figure 2, A and C). The area distributions of the simulated reprojections of the full-length N dimers overlap with the area distributions of the class averages (Figure 2B). Together, our data suggest that dimers of full-length N show extended and highly flexible conformations in the absence of RNA.

Distinct forms of the N protein bound to RNA were revealed by NSEM

Because no 3D structures of full-length N in complex with RNA are available, we inspected the N protein samples not treated with urea or RNase A by NSEM. The NSEM micrographs showed that these protein preparations contained a myriad of particles and aggregates that are consistent with the DLS measurements (Figure 1D and Figure 3). The larger aggregates with a width of 20 – 30 nm resemble the helical structures of the linearly arranged N protein isolated from the transmissible gastroenteritis (TGEV) coronavirus[36] (Figure 3A). Smaller particles with less than 20 nm in diameter were also observed (Figure 3B, colored boxes). Because these particles were quite abundant and structurally diverse, and showed dimensions consistent with N protein particles isolated from SARS-CoV-2 and related virus[37-39], we classified them according to their size and shape (Figure 3C).

The 2D class averages analysis revealed seven particle classes (Figure 3B). Class-1 (toroid-like) particles comprise rounded particles with a weak central density. These particles, which were the smallest and most compact (between 7 and 10 nm wide), resemble those corresponding to the mouse hepatitis virus (MHV) N protein dimer[37,38].

Class-2 particles comprise square-shaped particles approximately 14 nm wide. Notably, these particles have not yet been associated with the SARS-CoV-2 N protein; however, they resemble the RNPs isolated from the human coronavirus 229E (HCV229E) and MHV3[40]. The class-2 particles also resemble those of the M1 influenza A protein and Schmallenberg

orthobunyavirus RNP, where each globular unit located at the square vertices corresponds to a single structured protein domain[41,42].

Class averages 3 to 5 comprise elliptical and rounded particles ranging from 11 to 19 nm in length, whereas class averages 6 and 7 comprise rod-like structures of 16-20 nm in length (Figure 3B).

Although the diversity of these particles hinders the use of classical single-particle approaches needed for 3D structure resolution, we used the 2D class average analysis to estimate particle dimensions and to guide molecular modeling and dynamics simulations to gain insights into the conformational states of full-length N bound to RNA. We focused our analysis on the most abundant particles represented by classes 1 and 2, which are also the particles that showed well-defined features and dimensions (Figure 3B).

N protein dimers adopt a packed conformation in the presence of RNA

As reported recently[34,35] and shown in Figure 1C, the N protein devoid of RNA is a dimer in solution, which is consistent with the notion that dimers are the fundamental oligomeric unit of full-length N. Because the toroid-like particles observed in Figure 3B have dimensions ranging from 7 to 10 nm, that is, almost half of the hydrodynamic radius determined for the urea-treated N protein (Figure 1D), we reasoned that these particles could represent N protein dimers in a more compact conformation due to the presence of RNA. This idea is supported by the observation that the RNase A-treated N, which still have RNA bound to it, as judged by the 260/280 nm absorption ratio (~1.8), exhibits a smaller hydrodynamic radius (5,9 ± 0,7 nm) compared with the urea-treated protein, which typically shows a 260/280 nm ratio of ~ 0.5 (Figure 1D).

To investigate this hypothesis, the urea-treated N was incubated with a polyA15 or polyA50 RNA and the hydrodynamic radius was determined by DLS. We found that the N protein incubated with polyA15, but not polyA50, showed a significant reduction in the

hydrodynamic radius (6.3 ± 0.6 nm), compared with the RNA-free N protein (Figure 1D). Notably, the hydrodynamic radius observed for the protein in the presence of polyA15 is comparable to that of the RNase A-treated protein, which still retains RNA (Figure 1D). Although the polyA50 RNA did not significantly altered the hydrodynamic radius of the protein (Figure 1D), protein oligomerization, in addition to protein compaction, might have occurred in the presence of such RNA. In line with this idea, SEC-MALS analysis of the urea-treated N protein incubated with the polyA50 showed, in addition to the N dimer peak, a small but detectable peak of approximately 340 kDa, which would be consistent with an N protein octamer (Figure 1F). Such peak was not observed with polyA15 (not shown). In addition, when inspected by NSEM, N protein samples incubated with polyA15 showed several round particles of less than 10 nm in diameter resembling class 1 particles, whereas samples incubated with polyA50 showed much larger particles of irregular shape (Figure S2). These results thus suggest that the N protein undergoes protein compaction and possibly oligomerization in the presence of RNA.

To understand how the RNA could drive protein compaction, we performed CG dynamics simulations of dimeric full-length N, in which protein-protein and protein-RNA contacts are mostly driven by electrostatic forces. The simulations were carried out in the presence of single-strand RNA molecules ranging from 10 to 70 nucleotides (nt) in length (Figure 4). We observed that while the CTD dimer, which has the highest density of positive charges, readily associates with the shortest RNA molecules (10 - 15 nt), the NTD units move freely, as observed in the simulations performed in the absence of RNA (Figures 2C and 4A). With 30 nt, we also observed the possibility of RNA simultaneous interaction with one NTD and the CTD dimer. As the length of the RNA molecules increases (40 to 70-nt), a single RNA molecule can interact simultaneously with the NTDs and CTD bringing these two domains in close proximity. Full compaction of the N protein dimer was observed with the 40 to 60-nt RNA models with a minimum radius of gyration of 3.7 nm (Figure 4B). However, the

presence of multiple RNA binding sites along the N protein could account for the compaction observed experimentally with the polyA15 RNA (Figure 1D). In fact, CG simulations of the N protein dimer with multiple RNA molecules of 15 nt also show a significant reduction in the protein radius of gyration relative to the RNA-free protein (Figure S3), which is thus in line with the DLS results (Figure 1D).

To better correlate the CG simulations with the negative stain images, we built a low-resolution 3D density map from the toroid-like particles (Figure 5). The comparison between the class averages used to build the 3D map and reprojections generated from the map is presented in Figure S4. By fitting the atomic model derived from the CG simulations performed with the 60 nt-long RNA into the 27.5 Å resolution density map (Figure S5A), we found that the NTDs are oriented side-by-side facing the CTD dimer (Figure 5B). All domains, including the flexible regions, contribute to shield the RNA segment. Interestingly, according to this model, there are no significant interface contacts between the CTD dimer and the NTDs (Figure 5, B and C). As already mentioned, most class-1 particles display a weak density at the center, which indicates the existence of empty space between the globular domains (Figure 5A). This feature was also noticed in some particles derived from the N protein sample incubated with polyA15 (Figure S2). These results support the idea that although the NTDs approach the CTD dimer in the presence of RNA, these domains do not fully interact with each other, corroborating experimental data[19,43]. It is noticeable that, in the proposed model, the interaction of the RNA with the NTD resembles the binding mode of the NTD to a 10-mer RNA reported recently[23] (Figure 5C, upper inset). Likewise, amino acid residues implicated in RNA binding in the CTD[17,22] are also involved in RNA interaction in our model (Figure 5C, lower inset). In addition, the intrinsically disordered regions represented by the N-terminal end and SR linker also make contacts with the RNA (Figure 5B). Interestingly, according to the model, the arginine and lysine residues of the SR linker

interact with RNA while the serine residues remain solvent-exposed and thus accessible to protein kinases.

Together, our data indicate that full-length N undergoes domain compaction in the presence of RNA and offer an interpretation of how the NTD and CTD pack together in the protein dimer.

The octameric model of full-length N bound to RNA

In addition to the toroid-like particles described above, we inspected the class-2 square-like particles, because these particles were also commonly found in the RNA-containing N preparations and presented a clear organization pattern with four rounded units connected by the edges (Figure 3B and Figure 6A).

To understand the structural organization of such particles in detail, we collected ~25.000 class-2 particles and classified them into 150 subclasses (Figure 6A). These particles are ~13 nm wide and their rounded units located at the square vertices have dimensions varying from 5 to 7 nm, features that are clear in the most populated class averages (Figure 6A). The dynamic nature of the full-length N protein is highlighted by structural variations even within the same class (Figure 6A).

The class average analysis also revealed particles with specific features; for instance, showing a gap between two adjacent rounded units, like a U-shaped particle (Figure 6A, right inset). Of note, several square-like particles showed a blurred appendix of similar dimensions as the toroid-like particles (Figure 6A, right inset). Other well-defined patterns comprise particles that seem to be composed of only three rounded units (Figure 6A, left inset). These findings suggest that the square-like particles are formed by independent N protein units, probably dimers. Moreover, the negative stain images illustrate the pleomorphic nature of the N protein oligomers. As non-identical particles are an obstacle to finely uncover the N protein structure through single particle averaging protocols, we used the CG analysis to

rationalize a possible N protein structural organization that could explain the square-like particles.

Considering that each of the globular corners of the square-like particles has dimensions of 6 nm across and are formed by four independent N protein units, we modeled four copies (octamer) of the most packaged N-protein dimer from the 60-nt-long RNA simulation by placing each copy at the vertices of a virtual square. Then, in CG simulations, we connected each N protein dimeric unit through their C-terminal tails (see next session and methods for details) and allowed approximation and accommodation of the oligomeric system (Figure 6B).

From the CG octameric structure, we built a simulated density map at 20 Å resolution, close to the resolution of the negative stain images, and generated reprojections from the 3D map at different orientations (Figure 6B). Interestingly, some reprojections are remarkably similar to the top view of the square-like particles and their dimensions. In addition, reprojections corresponding to the side view of the octameric structure (Figure 6B) are consistent with the NSEM images of class-6 and class-7 particles (rod-like particles, Figure 3B). Thus, by merging the original square-like particles with the set of rod-like particles, we built a low-resolution (26 Å) 3D density map for the square-like particles (Figure 5C and Figure S5). The 3D reprojections and original particle projections of both square and rod-like particles match appropriately, validating the reconstruction (Figures S6).

This map reveals four quasi-globular units connected through the edges in a planar configuration (Figure 6C). The volume of each globular unit resembles the volume of the toroid-like particles (Figure 6C), thus reinforcing the idea that each globular unit contains an N dimer. This structural organization is remarkably similar to the RNP particles purified from the MHV3 and HCV229E strains observed by negative staining[40]. Taken together, our data suggest that the square-like particles observed in NSEM images represent the top views of octamers of the N protein.

van der Waals forces guiding the C-terminal tail self-interaction

The ability of the N protein to form dimers, tetramers, and higher-order oligomers in solution has been reported previously[17,22,30,34], nevertheless, how exactly N dimers interact with each other to form such higher-order oligomers is presently unknown.

In vitro studies suggest that the C-terminal tail residues 343-402 in SARS-Cov-1 and 365-419 in SARS-CoV-2 are required for N protein tetramerization[28-30]. Considering the high degree of identity shared by these regions, a more restricted dimer-dimer interaction zone, comprising residues 365 to 402, can be reasoned. This region is predicted to be unstructured in all coronavirus N proteins (Figure S7). Its adjacent C-terminal end residues, predicted to form an alpha-helix, are also conserved in most coronaviruses (Figure S7). To identify potential protein-protein contacts involved in oligomerization, we performed all-atom molecular dynamics simulations of the C-terminal tail.

One μs simulations starting with separated C-terminal tail monomers, *a* and *b*, revealed that the two monomers adopt a similar folding, retaining two alpha helices named α1 (375-382) and α2 (400-419) (Figure S8), consistent with the secondary structure prediction (Figure S7). The simulations also revealed intrachain contacts in each C-terminal tail monomers (Figure 7, A and B). *Monomer a* shows persistent van der Waals interactions between residues 407-415 (helix α2) and residues 377-385 (helix α1) and its neighbor coil residues L382 and R385 (Figure 7, A and D). The contacts involving helix α2 (N406, L407, S410, M411, S413, A414) with the coiled region of *Monomer b* (mainly residues P383, V392, L395) are also observed (Figure 7, B and E). C-terminal tail dimer formation was observed in all the simulations and the interface was asymmetrically formed by helix α2 from one of the monomers (Figure 7F). Overall, the results indicate that van der Waals interactions involving mainly the hydrophobic sequence LLPAA (394 to 398) (Figure S9) are the major forces driving the C-terminal tail dimerization.

Discussion

The underlying mechanism by which the N protein associates with the RNA to form the RNP particles remains unknown. Likewise, the molecular basis of the N protein self-association is also poorly understood. The understanding of these processes, which are fundamental for the virus life cycle, has been hampered by the dynamic structural nature of the N protein. Thus, despite the wealth of structural data for the NTD and CTD regions, no atomic models are available for the full-length N protein. Here, by combining NSEM with biochemical/biophysical and *in silico* approaches, we propose structural models and their dynamics to explain how the full-length N protein self-associates in a dimeric form both in the absence and presence of RNA. We also propose a model of full-length N bound to RNA in an octameric organization that is reminiscent of RNP particles isolated from other coronaviruses[36,38].

The negative stain images of full-length N bound to bacterial RNA showed a wide range of particles of different sizes and shapes, including large helical-like structures, which reflect the pleomorphic nature of this protein. Although particles of N protein incubated with polyA15 are consistent with class 1 particles, thought to contain *E. coli* RNA, those derived from samples incubated with polyA50 showed irregular shape and were distinct from the well-organized class 2 particles, which are also believed to carry bacterial RNA. This reflects the dynamics of the N protein upon RNA binding and suggests that sequence-specific RNAs from *E. coli* can drive the packing of the N protein more effectively than the polyA molecules tested.

Previous electron microscopy studies have suggested that the RNPs isolated from MHV and SARS-Cov-1 virions display a helical structure[37,38]. More recently, though, two studies using cryo-electron tomography have proposed that native SARS-CoV-2 RNPs are highly heterogeneous and densely packed but locally ordered in the virus, with neighboring RNP units with dimensions around 14 nm organized in a ‘‘beads on a string’’ fashion[39,44]. Thus,

although further studies are still required to confirm a helical organization of the SARS-CoV-2 RNP, the coiled structures of the MHV RNPs of approximately 11 nm in diameter with a 4 nm empty space[37] are quite compatible with the dimensions of the square-like (class 2) particles reported here. To our knowledge, these square-like particles have not been described for SARS-CoV-2 or any other coronaviruses N protein, although they are remarkably similar to electron microscopy images of RNPs isolated from MHV_3 and HCV-229E[40]. We thus propose that the N protein octamers with the shape of the square-like particles described here could represent building blocks of the SARS-CoV-2 RNP structure, as suggested for the MHV RNP[38].

To form such higher-order structures with the RNA, the N protein is thought to also depend on protein-protein interfaces. Accordingly, the C-terminal tail of the coronavirus N proteins has been implicated in protein tetramerization[28-30]. Here, using molecular dynamics, we investigated how the C-terminal tail could play a role in the N protein oligomerization. We found that the C-terminal tail adopts a folded structure maintained by van der Waals intrachain contacts involving two helices. Moreover, the hydrophobic segment LLPAA appears to play an important role in tail-tail interaction, which agrees with HDX values for this segment reported previously (Figure 7F)[30]. The hydrophobic character of the C-terminal tail is thus thought to drive the formation of higher order oligomers (tetramer and octamers) of the N protein. This idea is supported by the fact that the hydrophobic IILLF segment located at the C-terminal tail of the Tula hantavirus N protein is essential for the protein oligomerization[45]. Peptidomimetic molecules could thus be used to disrupt this tail-tail interface to prevent N protein oligomerization.

The dynamic nature of N as an RNA-binding protein was also revealed by CG simulation models, which corroborated the NSEM images and radius of gyration of the protein in the absence and presence of RNA[19,34]. According to these models, in the absence of RNA, the N protein oligomerizes into dimers where the NTDs move freely relative to the CTD dimer, as

previously suggested[34]. Conversely, in the presence of RNA, the protein undergoes a compaction that brings the NTDs closer to the CTD dimer. This compaction was confirmed by DLS measurements of the protein with polyA15. Despite the proximity of the NTD and CTD upon RNA binding, we did not observe a direct contact between these domains, as suggested by the weak density at the center of the 3D density map of the toroid-like particles, which resemble MHV dimeric N protein particles[38]. The RNA-binding mode suggested by our CG simulations is also consistent with the RNA-binding mode determined experimentally for the NTD[23], and with the predicted RNA-binding surface of the CTD [17,22]. The CG models also suggested a role of the SR-linker in RNA binding, which is in line with previous findings[9,14,19]. Considering that phosphorylation of the SR-linker plays a key role in the coronavirus life cycle[33], it is noteworthy that, in our CG models, the positively charged residues of the SR-linker point towards the RNA, while the serine residues are solvent-exposed and thus prone to be phosphorylated by protein kinases.

In addition, the N protein dimer models provided here offer further insights into how a single protein dimer interacts with a single RNA molecule of varying lengths without considering higher-order oligomerization. These models aimed to simulate how the N protein binds to the viral genome, where several protein dimers are thought to compete for a short genome segment. Our CG models suggested that a genome segment ranging from 40-60 nt would be sufficient to occupy all the RNA-binding sites and induce protein compaction. However, such RNA-binding sites could be simultaneously occupied by multiple short RNA segments, as suggested by the GC models and confirmed experimentally with polyA15 and polyT10[35]. Likewise, longer RNA segments could drive protein oligomerization, as observed with polyA50 and polyT20[35].

In conclusion, our results shed new light on the dynamics and higher-order oligomeric structure of the SARS-CoV-2 N protein and provide a framework for understanding the multifunctional and versatile role of this protein.

Methods

Cloning procedures

The SARS-CoV-2 RNA was isolated from virus particles with the QIAmp viral RNA mini kit (Qiagen - USA) and reversely transcribed to cDNA with the High-Capacity Reverse Transcription Kit (Thermo - USA). The N protein sequence (GenBank QIG56001.1) was amplified from cDNA samples using primers SC2-protN28182-F (5'-AGTCTTGTAGTGCGTTGTTCG-3') and SC2-protN29566-R (5'-ATAGCCCATCTGCCTTGTGT-3') and cloned into pGEM-T Easy (PROMEGA - USA), generating plasmid pGEM-SC2-N. The N sequence was reamplified from pGEM-SC2N with forward 5'-AACAAGCTAGCATGTCTGATAATGGACCCCAAAATCAG-3' and reverse 5'-GGTCTGCGGCCGCTTAGGCCTGAGTTGAGTCAGCACTGCT-3' primers and subcloned into the *Nhe*I/*Not*I sites of a pET28a-TEV vector carrying a 6xHis-tag and TEV protease cleavage site at the N-terminus.

Protein Expression and Purification

The N protein was expressed in *E. coli* BL21 (DE3) cells (Novagen -USA) and purified by metal-affinity and SEC. Cells were grown at 37°C under agitation (200 rpm) in LB medium containing kanamycin (50 mg/L) to an optical density (OD600nm) of 0.8. Recombinant protein expression was induced by the addition of 0.1 mM isopropyl-thio-β-d-galactopyranoside (IPTG) for 16 h at 25 °C. After centrifugation, cells were resuspended in 50 mM sodium phosphate, pH 7.6, 300 mM NaCl, 10% v/v glycerol, 1 mM phenylmethylsulfonyl fluoride and incubated on ice with lysozyme (0.1 mg/mL) for 30 min. Bacterial cells were disrupted by sonication and the soluble fraction was loaded on a 5 mL HiTrap Chelating HP column (GE Healthcare - USA) previously equilibrated with same buffer. Proteins were eluted using a linear gradient (20 to 500 mM) of imidazole at a flow rate of 1 mL/min. Eluted fractions containing the N protein were pooled, concentrated and

loaded on a HiLoad 16/60 Superdex 200 column (GE Healthcare), previously equilibrated with 10 mM Tris, pH 8.0, 100 mM NaCl, at a flow rate of 0.5 mL/min.

To produce the N protein without nucleic acid contaminants, the *E. coli* cells were lysed in buffer A (50 mM sodium phosphate, pH 7.6, 500 mM NaCl, 20 mM Imidazole, 6 M Urea, 10% Glycerol). The suspension was sonicated and centrifuged at 18,000 x g for 45 min at 4 °C. The supernatant was applied on a HiTrap Chelating HP column equilibrated with the same buffer. Proteins were eluted with a linear imidazole gradient using buffer B (50 mM Sodium Phosphate, pH 7.6, 500 mM NaCl, 300 mM Imidazole, 3 M Urea, Glycerol 10%). Protein fractions were mixed and dialyzed against buffer C (50 mM sodium phosphate, pH 7.6, 500 mM NaCl, 10% Glycerol). Protein samples were centrifuged at 14000 x g for 10 min at 4 °C and subjected to molecular exclusion chromatography on a Superdex 200 16/60 column, equilibrated in buffer C, under a flow rate of 0.7 mL/min. Protein purity was analyzed by SDS-PAGE and protein concentration was determined by absorbance at 280 nm using the molar extinction coefficient calculated from the amino acid composition. Protein samples were concentrated and stored at –80 ºC.

Dinamic Light Scatering

N protein samples, ranging from 2 to 100 µM, were submitted to dynamic light scattering analysis in the ZetaSizer NanoS (Malvern) equipment. All the measurements were acquired following the equipment automatic setup, at 10 °C. Obtained data are shown as the average of at least three measurements.

SEC-MALS analysis

SEC-MALS analyses were performed on a Viscotek (Malvern, UK) OmniSEC equipment equipped with a SEC module coupled to a two-angle laser light scattering detectors, with a refractometer and viscometer. Samples of N protein at 32 µM were loaded onto a Superdex 200 10/30 column, equilibrated with 50 mM sodium phosphate, pH 7.6, 500 mM NaCl, under a flow rate of 0.3 mL/min. Samples of N protein (20 µM) in 5-fold excess polyA50 were

analyzed following the same protocol. The OmniSEC software was used to acquire and evaluate the data.

Circular dichroism analysis

Protein samples at 6 to 30 µM final concentration were diluted in 50 mM sodium phosphate buffer, pH 7.6, and analyzed by FAR-UV circular dichroism. All measurements were recorded on a Jasco J-810 Spectropolarimeter at 10°C, in the range of 197-260 nm. The CD signals were normalized to residual molar ellipticity using the equation $\theta = (mdeg.100.MW)/(mg/mL.l.NR)$, where mdeg = CD signal, MW = protein molecular weight, mg/mL = protein concentration in mg/mL, l = optical path in centimeters and NR = protein residues number.

Negative staining and imaging

Three µL of purified N (8 µM) in 50 mM sodium phosphate buffer, pH 7.6 were applied onto glow-discharged (15 mA, negative charge for 25 s) 400-mesh copper grids covered with a thin layer of continuous carbon film (01824 - Ted Pella, USA). After 1 min, the excess liquid was drained using a filter paper. The grids were stained twice with 3 µL uranyl acetate solution (2%) for 30 s. The excess solution was drained and the grids were allowed to dry at room temperature. Data collection was performed using a 200 kV Talos Arctica G2 transmission electron microscope (Thermo Fisher Scientific). Data set of 18148 micrographs were automatically acquired with EPU software and recorded on a Ceta 16M detector. The pixel size and defocus were 1.96 Å and -1.5 µm, respectively. The exposure time was 1 s in an accumulated dose of ~23e-/Å^2.

The same grid preparation protocol was applied for the RNA-free and polyA-bound N protein samples, except that the proteins were at a lower concentration (from 0.5 to 3.5 µM). Screening data was performed using a 120kV JEM-1400Plus transmission electron microscope (JEOL, Japan), equipped with OneView 16-Megapixel Camera (Gatan, USA), magnification of 80k and pixel size of 1.39-1.89Å. To prepare the polyA-bound samples, the

urea-treated N was mixed with the RNAs at 1:10 protein-RNA molar ratio and incubated for 2 h at room temperature, before inspection.

Image processing

A total of 17370 collected negative stained micrographs (4096 x 4096) from purified N protein preparations were preprocessed using Imagic-4d software system[46]. Raw micrographs were firstly submitted to *a posteriori* camera correction and then to a contrast transfer function (CTF) correction. For CTF correction, the amplitude spectrum of each image was determined and submitted to eigenvector analysis and automatic unsupervised classification into 2000 classes. The resulting class averages were used to determine the CTF parameters, which were passed to the individual images. Then, the CTF correction was applied by phase-flipping each image. A total of 15474 CTF-corrected micrographs were selected based on the quality of CTF estimation and defocus for further processing. All image processing programs, except Imagic, were run in the Scipion framework[47].

Micrographs were resized in Fourier space to 1024 x 1024 dimensions and submitted to particle picking using the Xmipp3 package[48]. The picking was performed in original 4096 x 4096 size, by an initial manual picking of ~550 particles from 40 micrographs, and proceeded by automatic picking, resulting in 178,021 particles in a 200 x 200 box size. The particles were subjected to a round of 2D classification in Relion to separate toroid, square and rod-like particles groups. A total of 24137 toroid-like particles (class 1), 33550 square-like particles (class 2) and 12552 rod-like particles (classes 6 and 7) were used for further processing.

Given the structural pleomorphism of purified N protein, we cautiously evaluate the use of classical single-particle protocols. Therefore, image processing was informed by 3D atomic models obtained from CG molecular dynamics simulations. Toroid-like particles were classified into 300 class averages using Relion[49] and an initial 3D model (C1 symmetry) was built using 30 class averages in Eman2[50]. Class averages placed at a 112 x 112 box with a

circular soft mask (0.52 radius and 0.05 drop-off) were band-pass-filtered (0.001 low pass and 0.1 high pass). Then, the initial 3D map was refined in Imagic using an iterative process of angular reconstitution and class average rotation and translation alignment to 3D reprojections, achieving 3D resolution convergence.

For square and rod-like particles, an *ab initio* reference-free initial volume was generated in Relion, followed by two independent rounds of 3D classification. We calculated the consensus[51] of both 3D classifications (resulting in a total of 9 3D classes) and selected a single stable 3D class corresponding to 17,803 particles (38.6% of the particle set). This 3D class was further auto-refined following the 0.143 FSC target in Relion, obtaining self-consistent convergence in 13 iterations. The 3D density unmasked map resolution was accessed by the Fourier Shell Correlation (FSC), using ½-bit threshold (Figure S4).

CG molecular dynamics simulations

CG molecular dynamics simulations were performed using CafeMol 3.1 software[52]. An initial N protein dimer all-atom model was built in YASARA software[53] using crystallographic structures of NTD monomer (PDB ID: 6VYO) and CTD dimer (PDB ID: 6WJI). For this, NTD monomers were placed at a distance that allowed a fully extended NTD-CTD linker conformation. Intrinsically disorders domains (N-terminal tail, NTD-CTD linker and C-terminal tail) were modeled using YASARA. For simulations in the presence RNA, an initial single strand RNA all-atom model was modeled without tertiary structure or base pairing in YASARA. Nucleotide sequence of all RNA of the same size used in simulations were the same and consist of a scrambled sequence, since the CG model used does not compute base specific interactions.

Five independent simulations with different random seeds were conducted for each dimeric condition (without RNA or in the presence of 15 to 70 nucleotides) by Langevin simulations. The temperature was set to 300 K. Local protein-protein and RNA-RNA interactions were modeled using AICG2+ and GO potentials[52], respectively. For non-specific interactions

between N protein monomers, and between N protein monomers and RNA, excluded volume and electrostatic interactions were considered. Electrostatic forces were computed using Debye–Hückel equation with a cutoff of 20 Å and ion strength of 0.1 M. To positively charged residues +1e was assign whereas for negatively charged –1e. Local GO potential was changed to Flexible Local Potential in intrinsically disorder regions. Each simulation was run for $1.5x10^7$ steps, with a time length of 0.2 for each step in CafeMol scale. For dimers simulations, only the last $5x10^6$ were used in MD analysis.

For N protein octamer simulation, we built an initial model using four N protein dimer copies from the last all-atom fitting model. Each dimer was placed 200 Å from the center of mass of the neighbor dimer. In CafeMol simulations, a harmonic spring (force coefficient of 0.5 and distance constraint of 5) was used to bring together C-terminal tails from neighbor dimers. The contacts involving each dimer was modeled using GO potential to maintain native structure of both protein and RNA. In addition to the harmonic spring, we used excluded volume and electrostatic interactions to model the contacts among the four dimeric units. The octamer simulation was run for $5x10^6$ steps, which was sufficient to stabilize the RMSD calculated from CA atoms. The last frame from the simulation was further analyzed.

Tools for comparing CG simulations with experimental data

For comparing CG simulation radius of gyration with experimental data, we estimated the radius of gyration of the simulated systems with Bio3D package using an in-house R script[54]. For comparing CG simulations structures with negative stained images, we compared reprojections of 3D atomic model with the particle projections from negative stained images. The reprojections from the CG models were built firstly by creating a 20 Å resolution simulated density map from the atomic models in Chimera[55] using *molmap* function. Then, the 3D density map was reprojected in different orientations (rotational angle 0 to 360 degress with step of 10 degrees and tilt angle 0 to 180 degrees with step of 10 degrees) with

Xmipp3 *create gallery* function in Scipion workflow. The area of 2D reprojections was determined using ImageJ software[56].

3D Atomic model fitting

From the previous independent simulation of the CG model of N protein dimer in complex with a 60-nt-long RNA were selected 100 structures with the lowest radius of gyration using only the structures domains (NTDs and CTDs). The correlated structures were removed by clustering. The structures of reduced set (~40 structures) were converted to all-atom structure model. The protein was reconstructed using PULCHRA[57] and the RNA was converted into all-atom model using DNAbackmap tool from CafeMol.

The structures were rigid-body docked on the density map by employing *colores* program from Situs package[58]. Three different docking calculations were performed by selecting three different structural segments of the system: 1) NTDs, CTDs and RNA heavy atoms; 2) NTDs and CTDs heavy atoms, and 3) all heavy. The docked structures at each group were ranked by its cross-correlation coefficient (CCC). The selected structure for the model fitting was that with the lowest CCC and with close spatial coordinates in all the docking calculation groups.

The VMD AutoPSF plugin was used to generate a topology for MD simulations and MDFF simulations. The complete all-atom model (protein-RNA) was solvated in a TIP3P water box of 134x191x163 $Å^3$ using Solvate plugin from VMD. Na^+ and Cl^- ions were added to neutralized and to adjust the ionic strength to 150 mM using the VMD plugin autoionize[59]. The complete system involved 396,013 atoms. The molecular interactions were described by CHARMM36 force field with CMAP corrections[60].

All simulations were performed with NAMD 2.13[61]. The system was equilibrated using the following protocol: 1) 10,000 minimization steps of water molecules and ions by restraining the protein and RNA; 2) 10,000 minimization steps of the complete system; 3)

200 ps of NVT MD fixing positions of the protein and the RNA and 4) 200 ps of NPT MD fixing positions of the protein and the RNA. The temperature was maintained in 300K using a Langevin thermostat with damping coefficient of 1 ps^{-1} and coupling only heavy atoms. In the NPT simulations, the pressure was maintained constant in 1 atm using Nose-Hoover barostat with a piston decay of 200 fs and a piston period of 400 fs. Particle Mesh Ewald (PME) method were applied for describing long-range electrostatic interactions[62]. A non-bonded cut-off of 10 Å was applied to calculate electrostatic and van der Waals potential (vdW). A shifting function was applied to electrostatic interactions for avoid truncation of the potential function. vdW interactions were smoothed by applying a switch function at distance 9 Å. The time step was 2 fs using SHAKE for all covalent hydrogen bonds of the protein and nucleic molecules[63] and SETTLE algorithm for rigid water molecules[64].

MD Flexible Fitting of the protein-RNA into the density map were performed following a multi-step protocol[65]. The protocol involved three steps: 1) 5 ns MDFF simulation, applying a scaling factor used in the energy potential of the map (ξ = 0.3 kcal mol^{-1}) on RNA and NTDs and CTDs. Also, the dihedral angles of α-helices and β-sheet of the NTDs and CTDs were constraint using harmonic restrains ($k_{\mu,protein}$= 400 kcal mol^{-1} rad^{-2}). 2) 5 ns MDFF using similar conditions of the step 1, but including the intrinsically disordered regions (N- and C-terminal tails and SR-linker) in the potential energy function of the map. 3) a minimization of 10,000 steepest descent steps, just restraining backbone position of the protein and the RNA with a force constant of 10 kcal mol^{-1} $Å^{-2}$. 4) a minimization of 10,000 steepest descent steps without any restraint. The time step was 1 fs. SETTLE algorithm was used for rigid water molecules[64].

All-atom molecular dynamics simulations

For C-terminal tail simulations, a folded linear protein was built with YASARA[52], from PSIPRED secondary structure prediction[66]. This structure shows one helix from residues 403 to 419 and was used as starting point for MD simulation of one C-terminal tail monomer. The

MD simulation of the monomer was to check the folding along the trajectory without the influence of another monomer. The initial secondary structure did not change along the trajectory and no folding was observed. When the trajectory is aligned using the helix as reference, the coil region moves randomly. After a clustering, the average structure of the simulation was used as starting point for building a dimer system with the monomers separated by ~70 Å using Packmol program[67]. The large distance was set in order to decrease any possible bias on the starting configuration.

The dimer was immersed into an octahedral box with 56,098 TIP3P water box, 150 Na^+ ions and 152 Cl^- ions. The dimer formation was study by running five replicates. However, this system was described by ~170,000 atoms, needing an important computational effort to perform a 1 μs MD simulation. Thus, five uncorrelated starting points of the dimer were picked from the dimer MD simulation. The criterion was to select those with a minimum distance of 30 Å between any atom of the monomers. All the selected dimers were re-solvated in octahedral box with 45,556 TIP3P waters, 121 Na+ ions and 123 Cl- ions in order to keep the system composition and the ionic strength (150 mM). All the systems were described using Amber ff14SB force field[68] and the topologies were generated using *tLeap* program from AmberTool20[69].

All the MD simulations were equilibrated by the following steps: 1) minimization of the solvent by 2500 steepest descent steps followed by 2,500 conjugate gradient steps; 2) minimization of the whole system by 2,500 steepest descent steps followed by 2500 conjugate gradient steps; 3) the system was heated from o to 300 K in 200 ps under NVT conditions, restraining the protein atom positions; 4) the density was equilibrated under 500 ps under NPT conditions, restraining the protein atom positions. The production step of all the simulation was run for 1 μs under NPT conditions. For all the simulations, the temperature was set to 300 K using a Langevin thermostat with a collision frequency of 5 ps^{-1}. In the case of the NPT simulations, the pressure was set to 1 atm using a Monte Carlo

barostat with pressure relaxation time of 1 ps^{-1}. The long-electrostatic interactions were calculated using Particle Mesh Ewald method[62] 9 Å. SHAKE algorithm was applied to all bonds, allowing a time-step of 2 fs. The simulations were run using GPU accelerated PMEMD program that is part of the AMBER18 package[70].

The contact map analysis was done using *nativecontacts* command using two masks, one of each monomer, including all atoms and a cut-off distance of 7 Å. The contact maps figures were done using Rstudio[71]. The secondary structure along trajectory was performed using *secstruct* command. The inter-monomers vdW and electrostatic interaction energy was calculated using *lie* command with a dielectric constant of 78. All the commands are part of Cpptraj program from AmberTools20[69]. The images of the C-terminal tail monomers and dimers were generated using pymol 2.3.0 (open-source build).

Acknowledgments

We thank LNNano/CNPEM for the use of electron microscopy facility (TEM-26919, TEM-27882). This work is part of the Rede Virus MCTI taskforce on COVID-19 funded by FINEP (grant number 01.20.0003.00), Brazilian Ministry of Science, Technology and Innovation. The authors acknowledge financial support from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), project number 17/18139-6. We thank David A. Case from Rutgers University for the AMBER18's license fee waiver.

Author contributions

HVRF and GEJ performed all-atom and coarse-grained molecular dynamics simulations; GEJ performed molecular dynamics flexible fitting simulations; HVRF, GRS and SLG processed negative stain electron microscopy images; FAHB and CCT conducted biophysical experiments; ACB, AC, ASS, CCT and SLG prepared electron microscopy grids; ACB and AC collected electron microscopy images; FAHB, CCT and ASS expressed and purified

proteins; MCB, REM, DBBT, KGF, ACMF, CEB and PSLO designed the experiments, analyzed the data and wrote the paper.

References

1. Lai, M. M. C. Corona Virus: Organization, Replication and Expression of Genome. *Annual Review of Microbiology* **44**, 303–303 (1990).

2. de Haan, C. A. M. & Rottier, P. J. M. Molecular Interactions in the Assembly of Coronaviruses. *Advances in Virus Research* **64**, 165–230 (2005).

3. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology* **19**, 155–170 (2021).

4. McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991–3018 (2014).

5. Stohlman, S. A. *et al.* Specific interaction between coronavirus leader RNA and nucleocapsid protein. *Journal of Virology* **62**, 4288–4295 (1988).

6. Almazán, F., Galán, C. & Enjuanes, L. The Nucleoprotein Is Required for Efficient Coronavirus Genome Replication. *Journal of Virology* **78**, 12683–12688 (2004).

7. Hsieh, P.-K. *et al.* Assembly of Severe Acute Respiratory Syndrome Coronavirus RNA Packaging Signal into Virus-Like Particles Is Nucleocapsid Dependent. *Journal of Virology* **79**, 13848–13855 (2005).

8. Schelle, B., Karl, N., Ludewig, B., Siddell, S. G. & Thiel, V. Selective Replication of Coronavirus Genomes That Express Nucleocapsid Protein. *Journal of Virology* **79**, 6620–6630 (2005).

9. Grossoehme, N. E. *et al.* Coronavirus N Protein N-Terminal Domain (NTD) Specifically Binds the Transcriptional Regulatory Sequence (TRS) and Melts TRS-cTRS RNA Duplexes. *Journal of Molecular Biology* **394**, 544–557 (2009).

10. Zúñiga, S. *et al.* Coronavirus Nucleocapsid Protein Facilitates Template Switching and Is Required for Efficient Transcription. *Journal of Virology* **84**, 2169–2175 (2010).

11. Hurst, K. R., Ye, R., Goebel, S. J., Jayaraman, P. & Masters, P. S. An Interaction between the Nucleocapsid Protein and a Component of the Replicase-Transcriptase Complex Is Crucial for the Infectivity of Coronavirus Genomic RNA. *Journal of Virology* **84**, 10276–10288 (2010).

12. Verheije, M. H. *et al.* The Coronavirus Nucleocapsid Protein Is Dynamically Associated with the Replication-Transcription Complexes. *Journal of Virology* **84**, 11575–11579 (2010).

13. Kuo, L., Koetzner, C. A., Hurst, K. R. & Masters, P. S. Recognition of the Murine Coronavirus Genomic RNA Packaging Signal Depends on the Second RNA-Binding Domain of the Nucleocapsid Protein. *Journal of Virology* **88**, 4451–4465 (2014).

14. Keane, S. C., Liu, P., Leibowitz, J. L. & Giedroc, D. P. Functional Transcriptional Regulatory Sequence (TRS) RNA Binding and Helix Destabilizing Determinants of Murine Hepatitis Virus (MHV) Nucleocapsid (N) Protein. *Journal of Biological Chemistry* **287**, 7063–7073 (2012).

15. Tang, T. K. *et al.* Biochemical and immunological studies of nucleocapsid proteins of severe acute respiratory syndrome and 229E human coronaviruses. in *Proteomics* vol. 5 925–937 (Proteomics, 2005).

16. Yu, I.-M., Oldham, M. L., Zhang, J. & Chen, J. Crystal Structure of the Severe Acute Respiratory Syndrome (SARS) Coronavirus Nucleocapsid Protein Dimerization Domain Reveals Evolutionary Linkage between Corona- and Arteriviridae. *Journal of Biological Chemistry* **281**, 17134–17139 (2006).

17. Chen, C.-Y. *et al.* Structure of the SARS Coronavirus Nucleocapsid Protein RNA-binding Dimerization Domain Suggests a Mechanism for Helical Packaging of Viral RNA. *Journal of Molecular Biology* **368**, 1075–1086 (2007).

18. Takeda, M. *et al.* Solution Structure of the C-terminal Dimerization Domain of SARS Coronavirus Nucleocapsid Protein Solved by the SAIL-NMR Method. *Journal of Molecular Biology* **380**, 608–622 (2008).

19. Chang, C.-K. *et al.* Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging. *Journal of Virology* **83**, 2255–2264 (2009).

20. Chang, C., Chen, C.-M. M., Chiang, M., Hsu, Y. & Huang, T. Transient Oligomerization of the SARS-CoV N Protein – Implication for Virus Ribonucleoprotein Packaging. *PLoS ONE* **8**, e65045 (2013).

21. Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D. & Huang, T. H. The SARS coronavirus nucleocapsid protein - Forms and functions. *Antiviral Research* **103**, 39–50 (2014).

22. Peng, Y. *et al.* Structures of the <scp>SARS</scp> -CoV-2 nucleocapsid and their perspectives for drug design. *The EMBO Journal* **39**, e105938 (2020).

23. Dinesh, D. C. *et al.* Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLOS Pathogens* **16**, e1009100 (2020).

24. Kang, S. *et al.* Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharmaceutica Sinica B* **10**, 1228–1238 (2020).

25. Carlson, C. R. *et al.* Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. *Molecular Cell* **80**, 1092-1103.e4 (2020).

26. Savastano, A., Ibáñez de Opakua, A., Rankovic, M. & Zweckstetter, M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nature Communications* **11**, 1–10 (2020).

27. Perdikari, T. M. *et al.* SARS-CoV-2 nucleocapsid protein phase-separates with RNA and with human hnRNPs. *The EMBO Journal* **39**, e106478 (2020).

28. Luo, H., Chen, J., Chen, K., Shen, X. & Jiang, H. Carboxyl Terminus of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Self-Association Analysis and Nucleic Acid Binding Characterization †. *Biochemistry* **45**, 11827–11835 (2006).

29. Lo, Y.-S. *et al.* Oligomerization of the carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. *FEBS Letters* **587**, 120–127 (2013).

30. Ye, Q., West, A. M. V., Silletti, S. & Corbett, K. D. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Science* **29**, 1890–1901 (2020).

31. Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nature Communications* **12**, 502 (2021).

32. Wang, Y. *et al.* Low stability of nucleocapsid protein in SARS virus. *Biochemistry* **43**, 11103–11108 (2004).

33. Peng, T. Y., Lee, K. R. & Tarn, W. Y. Phosphorylation of the arginine/serine dipeptide-rich motif of the severe acute respiratory syndrome coronavirus nucleocapsid protein modulates its

multimerization, translation inhibitory activity and cellular localization. *FEBS Journal* **275**, 4152–4163 (2008).

34. Zeng, W. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochemical and Biophysical Research Communications* **527**, 618–623 (2020).

35 Zhao, H. *et al.* Energetic and structural features of SARS-CoV-2 N-protein co-assemblies with nucleic acids. *iScience* **24**, 102523 (2021).

36. Risco, C., Antón, I. M., Enjuanes, L. & Carrascosa, J. L. The transmissible gastroenteritis coronavirus contains a spherical core shell consisting of M and N proteins. *Journal of Virology* **70**, 4773–4777 (1996).

37. Bárcena, M. *et al.* Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 582–587 (2009).

38. Gui, M. *et al.* Electron microscopy studies of the coronavirus ribonucleoprotein complex. *Protein and Cell* **8**, 219–224 (2017).

39. Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730-738.e13 (2020).

40. Macnaughton, M. R., Davies, H. A. & Nermut, M. v. Ribonucleoprotein-like structures from coronavirus particles. *Journal of General Virology* **39**, 545–549 (1978).

41. Dong, H., Li, P., Böttcher, B., Elliott, R. M. & Dong, C. Crystal structure of Schmallenberg orthobunyavirus nucleoprotein-RNA complex reveals a novel RNA sequestration mechanism. *RNA* **19**, 1129–1136 (2013).

42. Zhang, K. *et al.* Two polar residues within C-terminal domain of M1 are critical for the formation of influenza A Virions. *Cellular Microbiology* **17**, 1583–1593 (2015).

43. Chang, C. *et al.* Modular organization of SARS coronavirus nucleocapsid protein. *Journal of Biomedical Science* **13**, 59–72 (2006).

44. Klein, S. *et al.* SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. doi:10.1038/s41467-020-19619-7.

45 Kaukinen, P. *et al.* Oligomerization of Hantavirus N Protein: C-Terminal α-Helices Interact To Form a Shared Hydrophobic Space. *J. Virol.* **78**, 13669–13677 (2004).

46. van Heel, M. et al. Four-dimensional cryo-electron microscopy at quasi-atomic resolution: IMAGIC 4D . in 624–628 (2012). doi:10.1107/97809553602060000875.

47. de la Rosa-Trevín, J. M. *et al.* Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *Journal of Structural Biology* **195**, 93–99 (2016).

48. de la Rosa-Trevín, J. M. *et al.* Xmipp 3.0: An improved software suite for image processing in electron microscopy. *Journal of Structural Biology* **184**, 321–328 (2013).

49. Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180**, 519–530 (2012).

50. Tang, G. *et al.* EMAN2: An extensible image processing suite for electron microscopy. *Journal of Structural Biology* **157**, 38–46 (2007).

51. Sorzano, C. O. S. *et al.* Image Processing in Cryo-Electron Microscopy of Single Particles: The Power of Combining Methods. in *Methods in Molecular Biology* vol. 2305 257–289 (Humana Press Inc., 2021).

52. Kenzaki, H. *et al.* CafeMol: A coarse-grained biomolecular simulator for simulating proteins at work. *Journal of Chemical Theory and Computation* **7**, 1979–1989 (2011).

53. Krieger, E. & Vriend, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics* **30**, 2981–2982 (2014).

54. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* **22**, 2695–2696 (2006).

55. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).

56. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* vol. 9 671–675 (2012).

57. Rotkiewicz, P. & Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **29**, 1460–1465 (2008).

58. Wriggers, W. Conventions and workflows for using Situs. *Acta Crystallographica Section D: Biological Crystallography* **68**, 344–351 (2012).

59. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

60. Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 Dihedral Angles. *Journal of Chemical Theory and Computation* **8**, 3257–3273 (2012).

61. Phillips, J. C. *et al.* Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **153**, 044130 (2020).

62. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

63. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

64. Miyamoto, S. & Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).

65. Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: A practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**, 174–180 (2009).

66 Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

67. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164 (2009).

68. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

69. D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C., and P. A. K. AmberTools20. (2020).

70. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).

71. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Figures

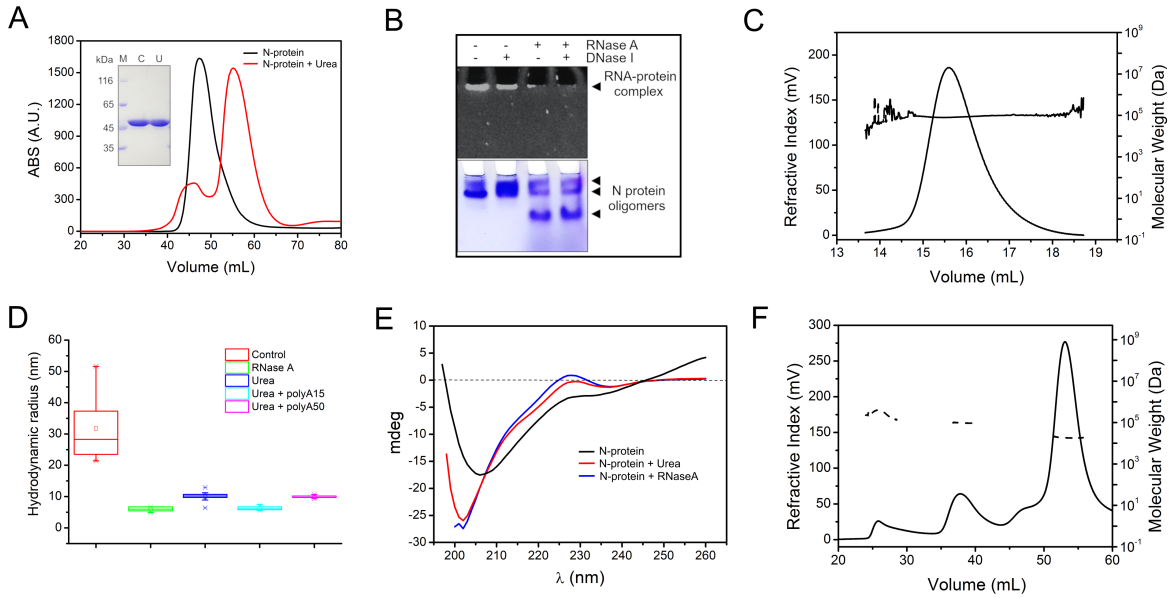

Figure 1. The N protein forms high molecular-weight complexes with RNA but is a dimer in solution in the absence of RNA. A- SEC showing that the untreated N protein sample (black line) elutes as a single peak in the void volume of the gel filtration column whereas the urea-treated sample (red line) elutes as a major peak of a higher elution volume (upper panel). The purity of these proteins samples was evaluated by SDS-PAGE, which shows single bands of the expected size for the recombinant N protein (inset). B- Native gel electrophoresis stained with ethidium bromide (upper panel) or Coomassie blue (lower panel) showing that the N protein forms high molecular-weight complexes with bacterial RNA, as treatments with RNase A, but not DNase I, dissolve these complexes (arrowhead). The RNase A treatment also changed the oligomeric forms of N protein (arrowheads) in solution (lower panel). C- SEC-MALS analysis showing that the N protein treated with urea is a dimer in solution, with a mean molecular mass of 87,2 ± 3.9 kDa. D- DLS measurements showing the hydrodynamic radius of recombinant N protein without any treatment (control), treated with RNase A or urea, in the absence or presence of polyA15 or polyA50. E- CD plot showing that N protein samples treated with RNase A or urea show similar CD curves (blue and red lines) compared to the N protein sample containing RNA (black line). F - SEC-MALS analysis showing that the N protein in the presence of polyA50 presents three main peaks (from the right to the left): non-bounded polyA50, N protein dimer and N protein octamer.

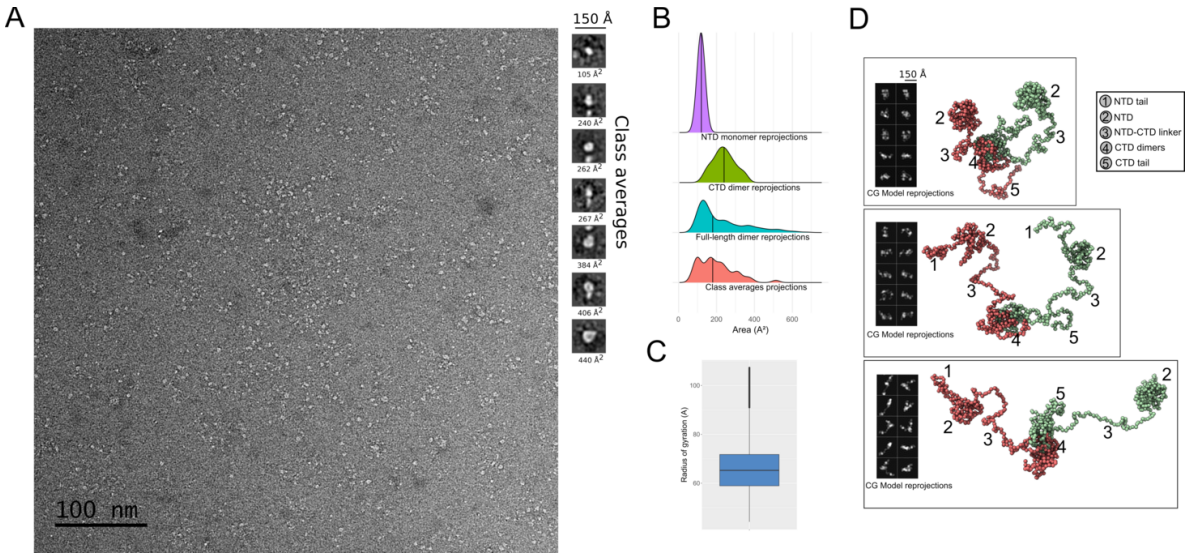

Figure 2. Dimers of full-length N show an extended conformation in the absence of RNA. A- Representative negative stain image of full-length N in the absence of RNA. On the right, representative class averages from a total of 50 classes produced by the 2D classification of 1500 particles. The area corresponding to each particle is presented. B- Density plots of area distribution estimated for NTD monomer, CTD dimer or full-length dimer reprojections or class averages projections. C- Representative frames of the minimum (upper), mean (center) or maximum (lower) radius of gyration obtained from the CG simulations. Two-dimensional simulated reprojections from each frame are presented at the left and the N protein structural domains are numbered. D- Boxplot of the radius of gyration estimated from five independent CG simulations of N protein dimer in absence of RNA.

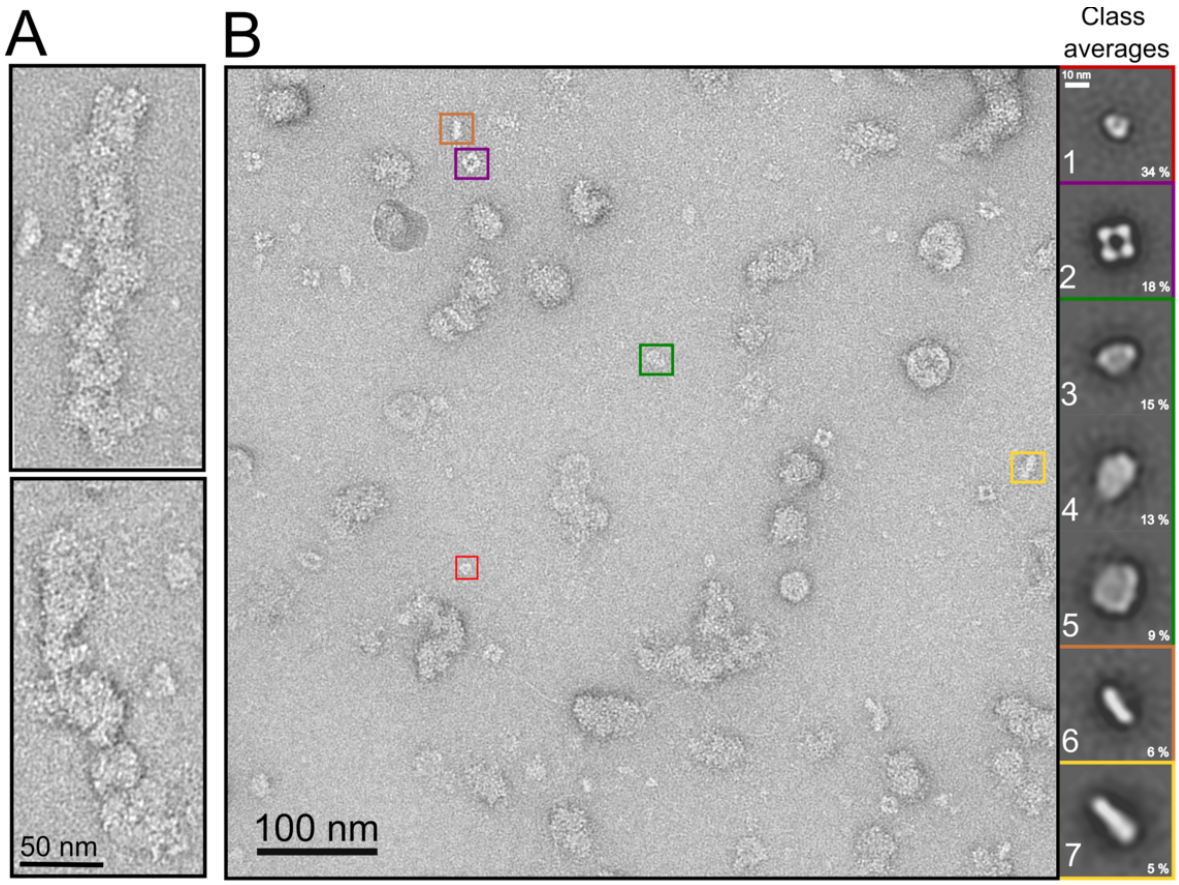

Figure 3. Representative negative stain micrographs of recombinant N protein containing RNA. A- Detail of large helical-like structures with 20 – 30 nm width. B- Representative micrograph showing a wide range of particles and aggregates. Smaller particles, with less than 20 nm in diameter are shown (colored boxes). In insets, class averages (1 to 7) of picked particles with sizes up to 20 nm in width from one hundred electron micrographs: class-1 (toroid-like particles), class-2 (square-shaped particles), classes 3 to 5 (elliptical and round-like particles), and classes 6 and 7 (rod-like particles). The percentage of each class average in relation to all picked particles are indicated.

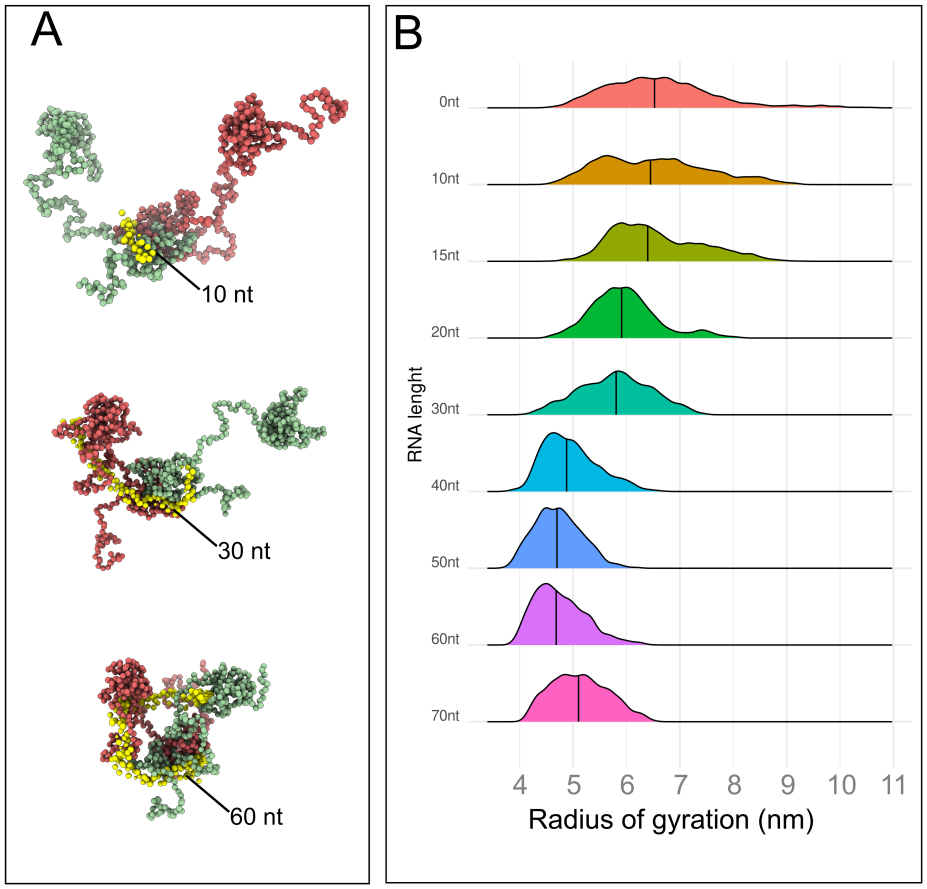

Figure 4. CG dynamics simulations of full-length N dimers in the presence of RNA of varying lengths. A- Representative frames of mean radius of gyration for protein-RNA complexes. N protein monomers are colored in red and green, whereas the RNA is colored in yellow. B- Density plot of the gyration radius distribution (in Å) calculated from five independent CG simulations of the N protein dimers in the presence of single-strand RNA of 15 to 70 nucleotides (nt) long.

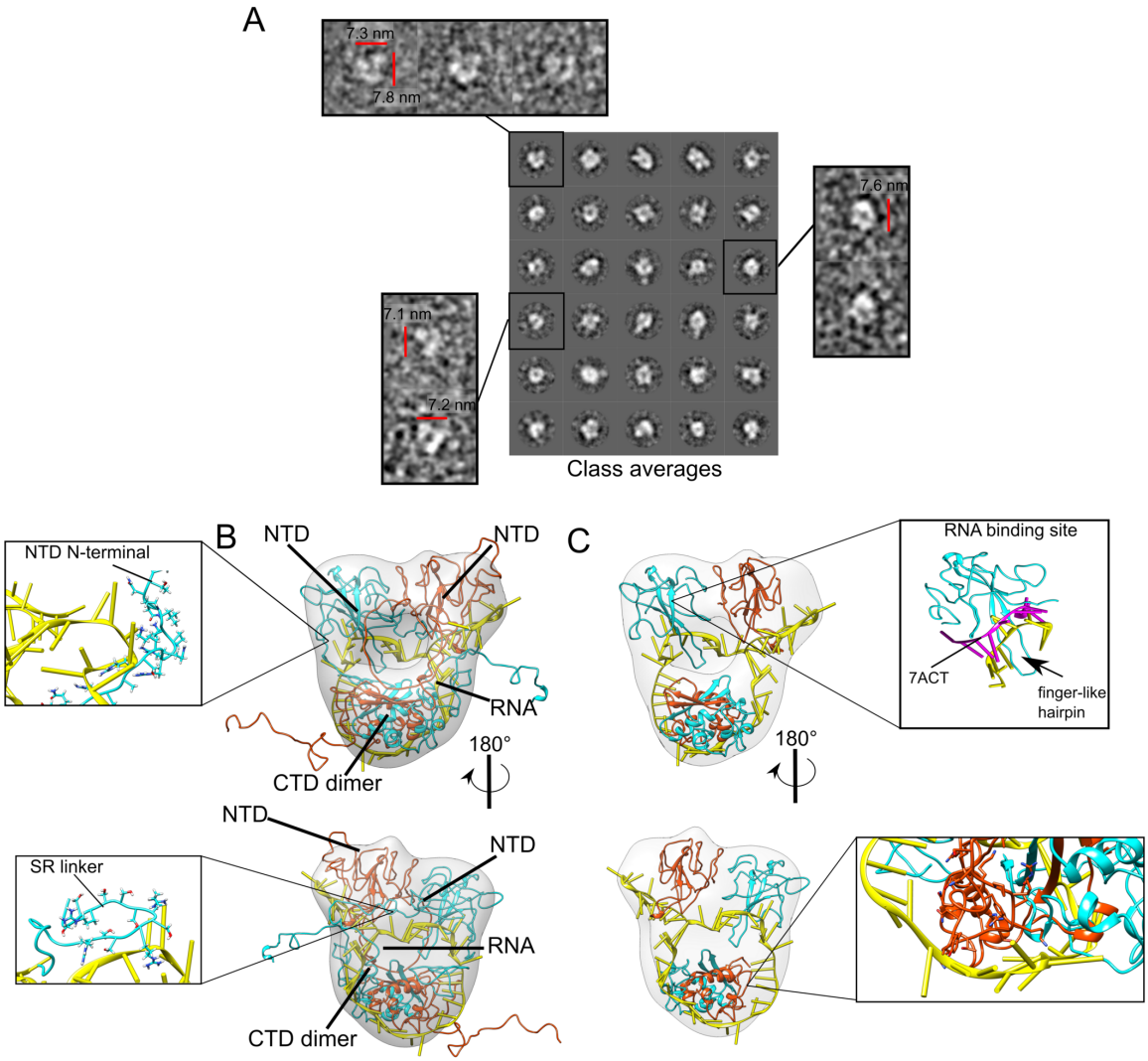

Figure 5. The full-length N dimer undergoes domain compaction in the presence of RNA. A- 2D classification of 24137 toroid-like particles picked with xmipp3. Thirty class averages used for the 3D model are shown on the right, whereas raw projections of particles representing the classes are shown on the left. B- 3D density map reconstruction of the toroid-like particles with the flexible fitting of the N dimer atomic model derived from the CG simulations performed with the 60 nt-long RNA. Upper inset shows interaction between NTD N-terminal (residues 26 to 42 in sticks) and lower inset shows SR linker (residues 183 to 195 in sticks) contact with RNA. C- Same as B, but only presenting structured regions of N dimer without its flexible regions. Upper inset shows the superposition of the NTD from the structural model with the NMR structure of NTD complexed with a 10-mer RNA (PDB code 7ACT). Lower inset shows residues implicated in RNA binding in the CTD dimer (248 to 280)[25,34] as sticks.

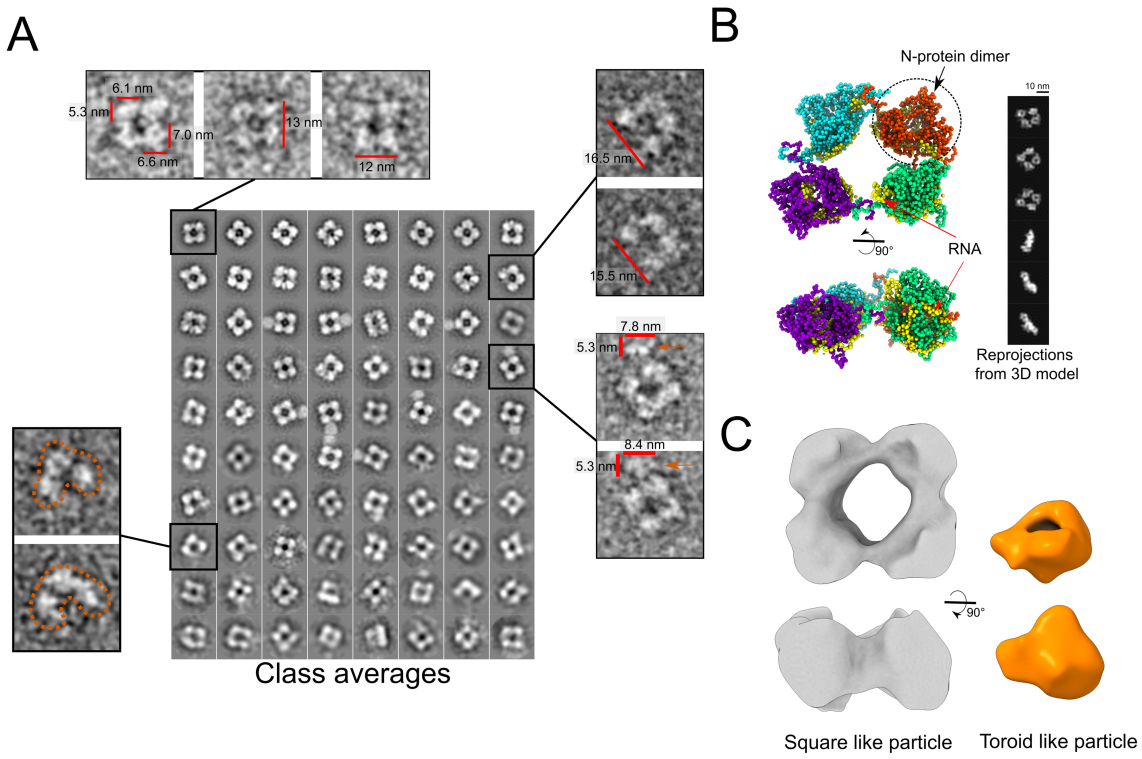

Figure 6. 2D Classification and 3D density map reconstruction of the square-like particles. A- 2D classification of square-like particles picked with xmipp3. Class averages are sorted by the number of class members (number increase up and left). Raw projections of particles that compose the classes are shown on insets. B- Representative frame of N protein octamer CG simulation in the presence of 60-nt RNA for each dimeric unit. Reprojections from the 3D structure are shown on the right. C- 3D-density map of the square-like particles in comparison to the map of toroid-like particles, both rendered at 3-sigma contour level.

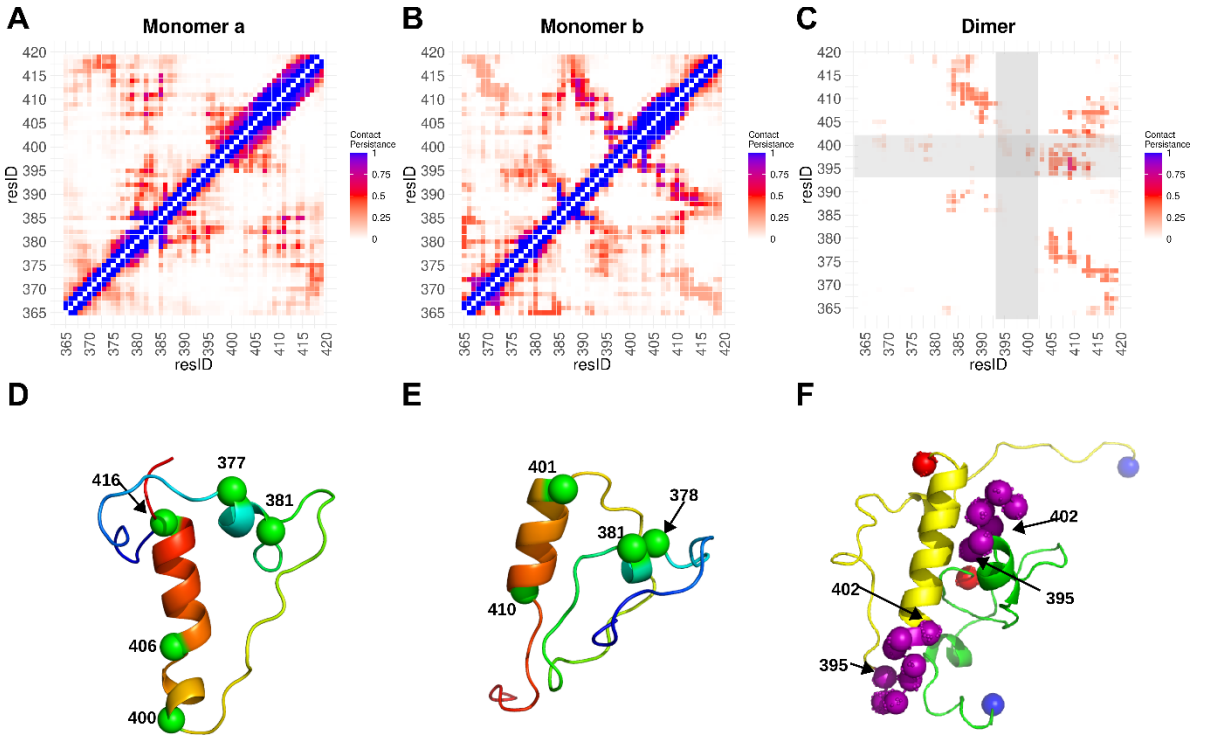

Figure 7. All-atom molecular dynamics simulation of the N protein C-terminal tail. A- and B- Contact maps for each individual monomer (a or b) or C- for C-terminal tail dimer. D-, E- and F- Representative structures of contact maps. The monomer contact maps (a or b) is an average of the whole trajectory of five MDs, considering onlythe last 50 ns of each trajectory for making the dimer contact map. The scale is defined as the contact persistence from 0 (none) to 1 (along all the simulation). D- and E- Structure of each monomer colored from N-terminal in red to C-terminal in blue, in greenVDW some Cα to reference residues number. F- A representative structure of the dimer. Each monomer is in new cartoon representation, one in yellow and the another in green. In purple, the hydrogen atoms from the residues 395 to 402 suggested to participate in the dimer interface (grey transparent regions in C, see Ye et al., 2020). Spheres in blue and in red indicate the n- and c-terminal of the C-terminal tail used in the simulation.