Computational and Structural Biotechnology Journal 18 (2020) 1884–1890

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Genus-specific pattern of intrinsically disordered central regions in the nucleocapsid protein of coronaviruses

Sailen Barik

3780 Pelham Drive, Mobile, AL 36619, USA

ARTICLE INFO

Article history:
Received 5 June 2020
Received in revised form 7 July 2020
Accepted 8 July 2020
Available online 17 July 2020

Keywords:
Intrinsic disorder
Coronavirus
RNA-binding
Phylogenetic
Infection
Host tropism
Ser/Arg-rich
Phosphorylation

ABSTRACT

The nucleocapsid (N) protein is conserved in all four genera of the coronaviruses, namely alpha, beta, gamma, and delta, and is essential for genome functionality. Bioinformatic analysis of coronaviral N sequences revealed two intrinsically disordered regions (IDRs) at the center of the polypeptide. While both IDR structures were found in alpha, beta, and gamma-coronaviruses, the second IDR was absent in deltacoronaviruses. Two novel coronaviruses, currently placed in the *Gammacoronavirus* genus, appeared intermediate in this regard, as the second IDR structure could be barely discerned with a low probability of disorder. Interestingly, these two are the only coronaviruses thus far isolated from marine mammals, namely beluga whale and bottlenose dolphin, two highly related species; the N proteins of the viruses were also virtually identical, differing by a single amino acid. These two unique viruses remain phylogenetic oddities, since gammacoronaviruses are generally avian (bird) in nature. Lastly, both IDRs, regardless of the coronavirus genus in which they occurred, were rich in Ser and Arg, in agreement with their disordered structure. It is postulated that the central IDRs make cardinal contributions in the multitasking role of the nucleocapsid protein, likely requiring structural plasticity, perhaps also impinging on coronavirus host tropism and cross-species transmission.

1. Introduction

The devastating global pandemic of coronavirus (CoV) disease that started in 2019 (hence named COVID-19) is caused by 'severe acute respiratory syndrome coronavirus 2′ (SARS-CoV-2), so designated because of its similarity to the previous SARS-CoV or SARS-CoV-1 that appeared in 2002 [1]. The extreme morbidity and lethality caused by SARS-CoV-2 has created a flurry of interest in coronavirus origin, phylogeny, virulence, and zoonosis. Genome sequencing of a large number of naturally occurring coronaviruses from a variety of host species and studies of their specificity and zoonosis, have led to the establishment of four major genera, namely alphacoronavirus, betacoronavirus, gammacoronavirus and deltacoronavirus [2,3]. Such studies revealed that the alpha and beta coronaviruses are of bat origin and infect only mammals, whereas gamma and delta are of avian origin and also infect birds, and rarely mammals. The human pathogen SARS-CoV-2, deadly agent of the 2019 epidemic, is a betacoronavirus, likely transmitted from bats in the Wuhan province of China [3–5].

Such studies also established the diversity and commonality of the viral genes in the coronavirus genera and their relative orders. All coronaviruses contain a single-stranded positive-sense (mRNA sense) genome, packaged in the virion along with four major structural proteins, including the nucleocapsid protein, N, that wraps the genome RNA [6]. Because of its apparently complex and multiple roles in the CoV life cycle, the N protein and its functional domains have been actively investigated and previously reviewed [7–13]. Collectively, these studies have suggested the existence of three broadly conserved domains [7,11,13]: the so-called N-terminal and C-terminal domains (NTD and CTD) (~129 aa each in length), and a central domain (~residue 150–250) with undefined boundary, although these domains display substantial divergence in length as well as sequence and may have overlapping functions. The NTD and CTD are both involved in RNA-binding, and the CTD is additionally important for dimerization of the N protein [13,14]. In contrast, the central region of the N protein, considered a linker between the two terminal domains, is a relatively uncharted territory. A serine/arginine-rich (S/R-rich) stretch

Abbreviations: CoV, Coronavirus; CoVID-19, Coronavirus disease-2019; IDR, Intrinsically disordered region; MERS, Middle East Respiratory Syndrome; N, Nucleocapsid; SARS, Severe acute respiratory syndrome; TGEV, Transmissible Gastroenteritis Virus.

E-mail address: barikfamily@gmail.com

https://doi.org/10.1016/j.csbj.2020.07.005
2001-0370/

has been recognized close to the NTD and specific amino acid residues have been shown to participate in binding viral RNA [15]. Phosphorylation is a common mechanism of post-translational regulation of proteins, and N protein has also been shown to be phosphorylated [10,16–19]; in fact, several earlier entries in the GenBank recognized it as 'nucleocapsid phosphoprotein'. The SR-rich sequence is phosphorylated by glycogen synthase kinase-3 [10], but the full gamut of phosphorylation on N and its significance and regulation deserves additional studies.

The higher order structure of recombinant NTD and CTD of the coronaviral N protein has been determined by X-ray crystallography, but that of the full-length N remains elusive. While the alpha-helix and the beta-sheet are traditionally recognized as major elements of a protein's secondary structure, intrinsically disordered regions (IDRs) are being increasingly recognized as important for protein function [20,21]. The IDRs by definition lack a committed structure and are there conformationally flexible and promiscuous, which also makes them accessible for interaction with other macromolecules (RNA, protein, DNA) [20,22,23]. Preliminary studies have noted three IDRs in the 422-residue long N protein of SARS-CoV-1 [7,8,13], one each at the amino- and carboxy-termini (aa 1–44 and aa 366–422), and another one (aa 182–247) within the central region. Collectively, these regions have been implicated in interactions with viral RNA and proteins (interaction with self, forming N-N homodimer, and with the integral membrane protein, forming N-M heterodimer) [8,9,18,24]. The central IDR and secondary structure elements (α-helix and extended β-strand) were found to be conserved in several coronaviruses, although their phylogenic distribution or distinction was not studied [7,8]. An interesting recent bioinformatic study analyzed the percentage IDR content of all proteins of several coronaviruses, and found the highest IDR content in the N protein, followed by M [25]. The authors then used the percent disorder values to classify the coronaviruses into three arbitrary, nonphylogenetic categories (i.e. different from the genera) and proposed a correlation between percent disorder, hardness of the virion shell, and fecal-oral transmission [25]. The molecular biological mechanism of this unorthodox hypothesis, however, remains unclear. Here, focus was placed on the central IDR region of the CoV N protein, dissecting it in greater detail, which revealed remarkable genus-specific arrangements as well as relationships with the S/R clusters.

2. Bioinformatic methods

2.1. Sequence retrieval, comparison and analysis

All coronaviral N protein sequences were retrieved from NCBI GenBank 'Protein' RefSeq source, using search terms 'nucleocapsid' and 'Coronaviridae'. Among ~134 sequences retrieved, redundant (duplicate) submissions were visually inspected and removed. When identical sequences of coronaviruses from the same species and the same geographical area were found, even though they were independent isolates by different research groups, only one of them was retained. Overall, the goal was to analyze nonredundant but representative sequences from diverse species and environments. In reverse search, sequence of a particular genus (such as Alphacoronavirus) and/or from a specific host isolate was employed in Protein BLAST. The search was limited to 'Coronaviridae (taxid:11118)' in the 'Organism' box, to eliminate any non-coronaviral sequences with significant homology, even though they might constitute a potentially interesting study in the future. Multiple sequence alignments were performed by Clustal Omega [26] at the EMBL-EBI web server [27], as described [28].

2.2. Disorder analysis

Multiple programs are currently available for the prediction of intrinsic disorder regions (IDRs) in protein sequences [29–36], essentially all of which are rooted on the principle that structurally disordered regions contain a preponderance of charged and hydrophilic residues, and exhibit lower sequence complexity. Most algorithms also focus on physicochemical properties and machine learning classifiers [30,32,34–36]. Some second-generation methods use a meta-approach, in which predictions from multiple predictors are weighed out, but such analyses may be slower. In extensive previous analyses [28], PrDOS was preferred for of its simple graphical user interface (GUI), significant server speed at all times of the day and night, and choice of false positive rate. It is also a hybrid of both template-based and machine-based predictions, and allows batch analysis of multiple sequences. Finally, it is relatively unbiased, without favoring and disfavoring specific motifs such as disulfide bonds [34]. Operationally, a scoring matrix is generated after two-rounds of PSI-BLAST search of sequence databases, and the profiles are used for a template-based search for a homolog with known disorder status in the PDB. For sequences lacking a known structural homolog (e.g. the central part of the coronaviral N protein), a support vector machine (SVM) algorithm is used to obtain the position-specific scoring matrix. Nevertheless, as in my previous experience [34], a single N sequence was subjected to disorder prediction by multiple programs, and the PrDOS results was found to agree with all of them, including several meta-predictors (data not shown), such as MetaDisorder [33].

2.3. Disorder probability plot

The PrDOS results were downloaded in the CSV format, converted into Excel files, and then the location of the IDR noted (Table 1).

The disorder probability graphs (Fig. 1) were also plotted using Excel. Since the N sequences of different coronaviruses are of diverse length (~342–468 residues), their IDRs are also positionally shifted. It was, therefore, necessary to manually align the IDR peaks in order to compare the different sequences, and as a result the amino acid numbers on the X-axis, spanning ~ 121 residues (Fig. 1), are not actual amino acid numbers on the polypeptide (that are displayed in Table 1), but rather offered as a length scale for the displayed sequence to facilitate referral of the IDR peaks. To reiterate, IDR analysis by PrDOS was performed on the full-length N protein sequence, but only the IDR region was displayed in the graph (Fig. 1).

3. Results

3.1. Intrinsically disordered regions (IDRs) of the coronaviral N protein

Before embarking on the analysis of N protein IDRs and their potential phylogenetic distribution, it was important to ascertain that the full-length N sequences obey the phylogeny of the viruses to which they belong. Thus, N protein sequences of all coronaviruses were collected in a genus-blind manner, and subjected to multiple alignment by Clustal Omega. When the sequences in the Clustal output were then labeled for each viral genus (as categorized in the GenBank annotation), they were in fact found to be clustered by genus (data not shown). In order words, all the alpha sequences were together, all beta were together, and so on. Thus, the sequence homology of the full-length N proteins independently matched viral phylogeny, which assured that if a relation-

Table 1
N protein sequences analyzed.

Coronavirus (CoV) genus	GenBank Accession#	Graph color legend (Fig. 1)	ID region analysed (aa residue#)
Alpha-CoV			
Lucheng Rn rat CoV	**YP_009336487.1**	Series 1	147–266
Feline CoV	**YP_004070199.1**	Series 2	138–257
Canine enteric CoV K378	**Q04700.1**	Series 3	141–260
TGEV	ANR94935.1	Series 4	144–263
Mink CoV strain WD1127	**YP_009019186.1**	Series 5	141–259
Ferret CoV	**YP_009256201.1**	Series 6	143–262
Human CoV 229E	AGW80953.1	Series 7	144–263
Rhinolophus bat CoV HKU2	**YP_001552240.1**	Series 8	130–249
BtRf-AlphaCoV/YN2012	**YP_009200739.1**	Series 9	131–250
Beta-CoV			
Pipistrellus bat CoV HKU5	**YP_001039969.1**	Series 1	150–270
Tylonycteris bat CoV HKU4	**YP_001039960.1**	Series 2	149–269
Bat (unclassified species) CoV	**YP_009361864.1**	Series 3	150–270
MERS CoV	**YP_009047211.1**	Series 4	150–270
Bat Hp-betaCoV/Zhejiang2013	**YP_009072446.1**	Series 5	155–275
Bat CoV BM48-31/BGR/2008	**YP_003858591.1**	Series 6	160–280
SARS-CoV-2	**QHD43423.2**	Series 7	161–281
SARS-CoV-1	P59595.1	Series 8	162–282
Horseshoe bat SARSr-CoV	Q3LZX4.1	Series 9	160–280
Gamma-CoV			
Canada goose CoV	**YP_009755908.1**	Series 1	152–272
Turkey CoV	**YP_001941174.1**	Series 2	146–266
AIBV	**Q8JMI6.1**	Series 3	146–266
Beluga whale CoV SW1	YP_001876448.1	Series 4	145–234
Bottlenose dolphin CoV	QII89031.1	Series 5	145–234
Delta-CoV			
Common moorhen CoV	**YP_005352885.1**	Series 1	132–252
Munia CoV HKU13-3514	YP_002308510.1	Series 2	127–247
Magpie-robin CoV HKU18	**YP_005352858.1**	Series 3	126–246
Porcine CoV HKU15	YP_009513025.1	Series 4	131–251
Thrush CoV HKU12-600	**YP_002308501.1**	Series 5	127–247
White-eye CoV HKU16	**YP_005352842.1**	Series 6	128–248
Night heron CoV HKU19	YP_005352867.1	Series 7	122–231
Wigeon CoV HKU20	YP_005352875.1	Series 8	130–239

Representative N proteins of the four genera of coronaviruses, the IDR of which, spanning the indicated amino acid numbers, are plotted in Fig. 1. Color codes for Fig. 1 are designated here as Excel-assigned Series#. A subset of these sequences, marked in bold, are aligned in Fig. 2, in the same order as shown (i.e., top to bottom). ID = Intrinsic Disorder; aa = amino acid; TGEV = Transmissible Gastroenteritis Virus; MERS = Middle East Respiratory Syndrome; AIBV = Avian Infectious Bronchitis Virus.

ship between IDR and genus is discovered in future studies, it would likely be meaningful.

Upon confirming N protein phylogeny, disorder analysis was conducted. Bioinformatic search for IDRs in full-length CoV N predicted their location in the following areas: (1) The N- and C-terminal regions (data not shown), confirming previous reports [7,13]. They were ignored in this study for two reasons: (a) They already received considerable attention in the past; (b) Disorder predictor algorithms use a sliding window of 9–12 amino acids to smooth the prediction values along the full length of the protein sequence; as a result, the predictions for ~12 residues at the two termini are not very reliable. This also makes it difficult to demarcate the exact boundary of specific IDR sequence predicted for the termini [37]; however, this is not an issue for internal IDRs as they are flanked by long non-IDR sequences on both sides.

Detailed examination of the internal IDRs in all coronaviruses revealed several patches of internal IDR, but most had two significant ones, approximately 30–40 amino acids long, rising above the baseline probability (Fig. 1). As the lengths of the N polypeptide differed in diverse coronaviruses, so did the location of the two IDRs; however, the noticed trend was that the first IDR was slightly longer (40 aa) than the second (30 aa). When some sliding of the sequences was allowed against one another to align the peaks, a general pattern emerged in the majority of the N proteins.

First and foremost, two peaks were found in alpha, beta, and gamma coronaviruses (Fig. 1A–C), with two apparent exceptions that are described later. The deltacoronaviruses, in contrast, showed little or no second peak (Fig. 1D). Only after this step, the sequences were analyzed for disorder, which also showed a genus-specific pattern, as presented in Fig. 1. Interestingly, a minority of sequences at the boundary between two genera in the Clustal alignment, often produced slightly mixed IDR peak shapes. For instance, some beta sequences near the gamma boundary had shorter IDR lengths, somewhat resembling the IDR shapes of gamma. Such overlaps are to be expected since phylogeny is a seamless evolutionary process, and RNA viruses in particular constantly generate various intermediate strains, which in fact forms the basis of the 'quasi-species' phenomenon [40]. Nevertheless, these boundary sequences were excluded from the graphic plot in Fig. 1, because they were not fully representative of any genus and would only blur the plot.

Two sequences that did not exactly fit the two-peak pattern of gammacoronaviruses, but were classified as gamma in the literature, had distorted second IDR peaks, although they were above the baseline probability (Fig. 1C). These two coronaviral species were isolated from beluga whale [38] and bottlenose dolphin [39]. Nonetheless, they are not delta viruses, in which the second peak does not exist at all. This is why they are placed under the gamma genus, but tentatively, labeling them as 'deviant' (Figs. 1C, 2). The significance of these two sequences are discussed later (Section 4).

3.2. Relationship between the central IDRs and Ser/Arg-rich sequences

As mentioned earlier, a stretch of N protein sequence just downstream of the NTD was noted to contain S/R-rich motif [13],

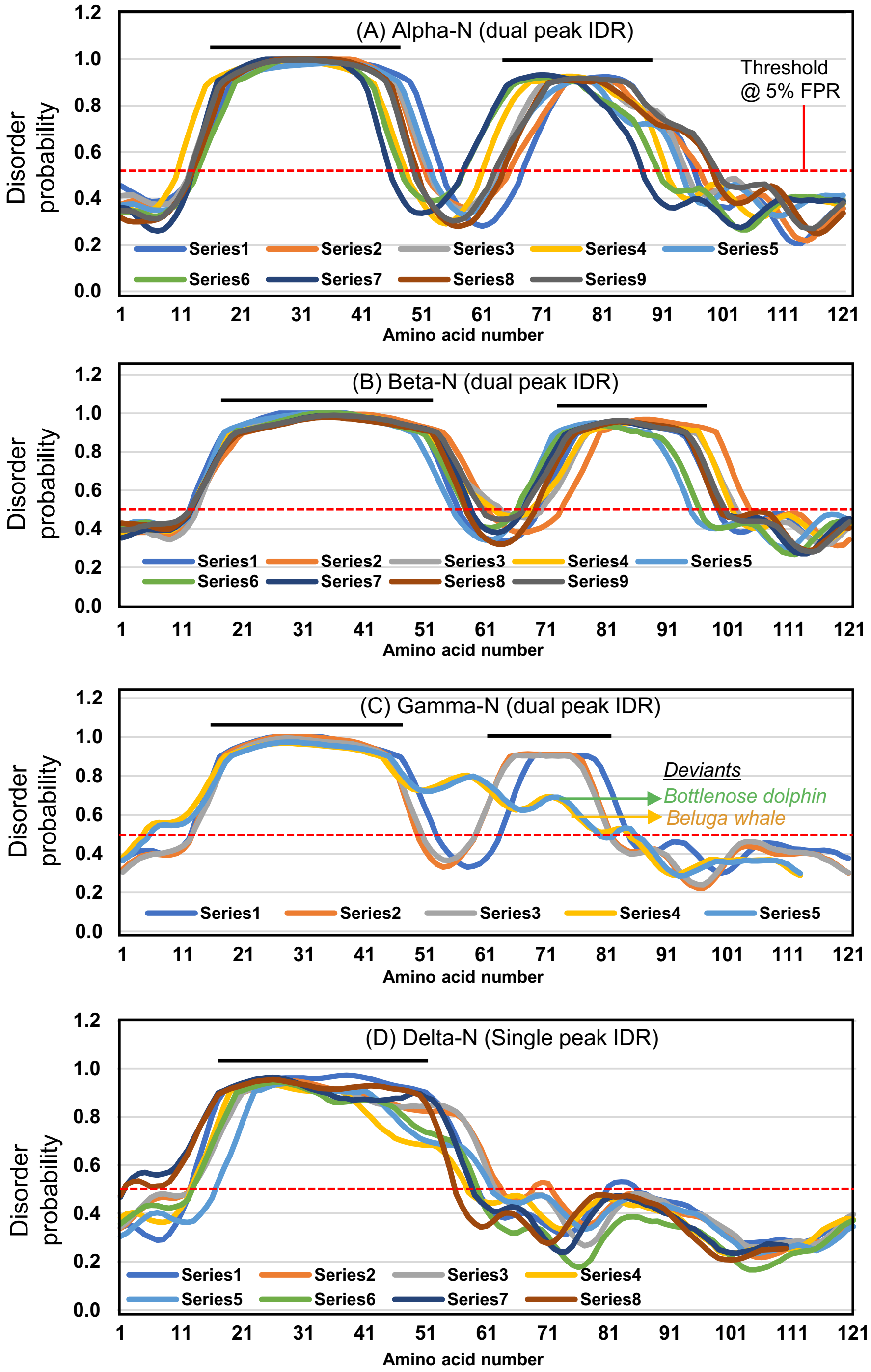
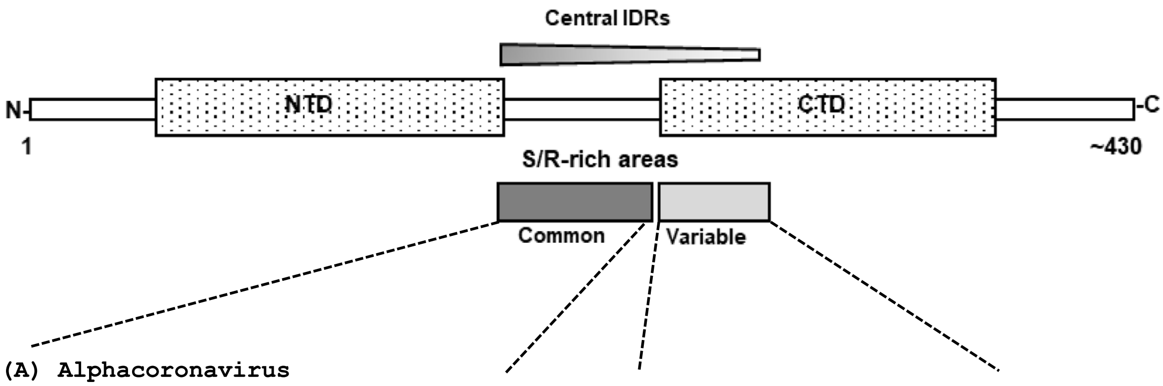


Fig. 1. Intrinsically disordered region (IDR) at the center of the coronaviral N protein. IDR was predicted by several methods, as described previously [28] and in Materials and Methods, which produced similar results, and those from PrDOS are presented (See Fig. 2 for the corresponding primary structures). The cut-off threshold was set at the relatively stringent FPR (false positive rate) of 5%, corresponding to a disorder probability of 0.5 (Y-axis). Thus, only the areas of disorder probability >0.5 (above the red dotted line) were considered as significant. The "Series" designations of the color-coded graphs are listed in Table 1. As explained in the Materials and Methods, the amino acid numbers (X-axis) do not refer to the full-length protein, but only to the displayed sequence portion, so as to provide a scale of the relative location and length of the two IDR peaks. However, the actual residue numbers are listed in Table 1. Note that alpha, beta, and gamma coronaviruses have two central IDRs, whereas deltacoronaviruses have only one, roughly corresponding to the first. Two highly similar coronaviruses, isolated from beluga whale and bottlenose dolphin, are currently considered in the gamma category [38,39], but their weak second IDR is marked as 'deviants' in this genus. They are presented in the same box as gamma to clearly demonstrate the difference in the peak, but also see Fig. 2, where their sequences are placed in a separate category to indicate this difference. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

although its phylogenetic distribution was not explored. The IDRs in general are poor in bulky hydrophobic and aromatic amino acids but rich in small and polar amino acids, and thus, may readily contain Ser and Arg [20,29,34]. Once the two IDR peaks in the central region were identified, it was worth determining whether or to what extent they matched the S/R-rich sequences. To find this



```
(A) Alphacoronavirus
159 VPNASRPQSRAQSQTRSQNTSRASSVSRNGSRAPSVDR--TKEDLKAVVSQLLAEMGISKNSKQTQNQKK---Q--KGSTPAATPHPNQDG-- 242
153 -VNQSRDNSRSPSQSRSQSRNRSQSRGRQ------QSNNKKDDSVEQAVLAALKKLGVDTE-KQQQRSRSKSKERSNSKTRDTTPKNENKH-- 235
156 -------NVNSRSNSRSRSRGRSRSNTRS------NDNQSSSQDIATAVAAALAQMGFAPKETQKNKSRSKSRDRAKSSER-PVPKNENKH-- 232
156 -------NVGSRNNSRSRSRGRSKSNNRS------NNNNSNSGDIATAVAAALAQMGFAPKDTQKNKSRSKSRDRSKSRER-PTPNNEN---- 230
154 --NRSRNNSRSGSQSRSVSRNRSQSRGRH------HSNNQN-NNVEDTIVAVLEKLGVTD----KQRSRSKPRERSDSKPRDTTPKNANKH-- 231
142 ------VEATSRSASRSQSRSRNQSQSRSGAQTPRVQQSSQAVDIVAAVKQALADLGIASSQSKSQSGKNTPKPRSRAVSPAPAPKPARKQMD 228
143 -------ESTSRSASRSQSRSRNQSQSRSGQQTPRFQQQSQSVDIVAAVKQALADLGISGTQNKPQSGRNTPKQKSRAVSPAPSPKQTRKQMD 228
              : . :** . .*  *  *        ..     .:  .:   * .:*.       :  :.       .      *:  .

(B) Betacoronavirus
162 IEGTGGNSQSSSRASS--RSSSRSSSRNGRSNNSSRNASPAPHGVGDVVGAGTLSVLLDLQKRLADLEAGKGN-KQPKVITKKDAQAAKQKMRH 251
162 IEGTGGNSQSSSRASSNSRNSSRSSSRGGRSTSNSRGTSPVSHGVGSAESLAALPLLLDLQKRLADLESGKSK--QPKVVTKKDAAAAKNKMRH 253
162 IEGTGGNSQSSSRASSASRNSSRSNSRGSRSGNSSRGTSPGPSGVGAVG---GEMLYLDLLNRLQALESGKTKQAQPKVITKKDAVAAKNKMRH 252
162 IEGTGGNSQSSSRASSVSRNSSRSSSQGSRSGNSTRGTSPGPSGIGAVG---GDLLYLDLLNRLQALESGKVKQSQPKVITKKDAAAAKNKMRH 252
167 VEGSRSTSSASSAASSRSNSRSRNSSL-SRSSSPGRGLPPA----TD-PNGVLAALLLTKLEALDAKVNGPK---QPPVVTKKTAAEIAAKPRQ 251
172 AEGSRGGSQASSRSNSRSRGNSRNSTP-S----SSRGSSPARMAA-G-GDTALALLLLDRLNQLESKVSGKTPQ-QSQVVTKKTAAEASKKPRQ 257
173 AEGSRGGSQASSRSSSRSRNSSRNSTP-G----SSRGTSPARMAGNG-GDAALALLLLDRLNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQ 260
     **: . *.:** :.*  .. **..:  .      *.  *               : *   : *     *     *  .:*** *     * *:

(C) Gammacoronavirus
164 IPTSRGRSGRNSVASSREQSRNGSRAATPERRGRTSSVNEDLIQKAADIILKNQGKSGRITKAKANEMAERRYCKRVLAPGKDVDS--- 249
158 IPLHRGRSGRSTAASSAASSRAPSRD---GSRGRRSGSEDDLIARAAKIIQDQQKKGSRITKAKADEMAHRRYCKRTVPPGYKVDQVFG 243
158 IPLNRGRSGRSTAASSAASSRPPSRE---GSRGRRSGSEDDLIARAAKIIQDQQKKGSRITKAKADEMAHRRYCKRTIPPGYKVDQVFG 243
    **  ******.:.***  .**  **      *** *. ::*** :**.** .:* *..*******:***.******.: ** .**.

(D) Deltacoronavirus
131 -DGPVQGFRIDPFQSRGRSVERGQQDRRAQS-AEPQRPKQRERSRSAPSRAA-PKQHAIPKRVLLKGKTISGTFGKRSPAKANVGSPDT------ 216
123 GDGVNTGLRVDPFQNRGRTMDRGPNQRSSSAPVDSGNRNQRARSKSAPSRTN-PKTHQIPKRTLTKGKTISQVFGKRTPGGANVGSA-------- 208
140 ---------VDPFNTRGRQLERAPGPRSMSMDTRPAANQPRKRDQSAPAAVRRKTQHQAPKRTLQKGKTISQVFGHRSRTGANVGSADTEKSGMA 225
140 ---------VDPFNSRGRTMDRSNGPRSMSADPRSQTNQPRRRDQSAPAAVRRKTQHQAPKRTLQKGKTISQVFGNRSRTGANVGSADTEKTGMA 225
             :***:.*** ::*.   *  .        : * *.:***: .   . *  ***.* ****** .**:*:   *****

(E) Unique coronavirus (Deviant Gamma or a new genus?)
149 DGPERGFFIERSRSRSRSNSQADNRSKSRERSQSSDKPRKEDRGRSKSPAQKNPGSGGKITKEKAALMVSRAYSKRTLPPNKPVDQ 234
149 DGPERGFFIERSRSRSRSNSQADNRSKSRERSQSSDKPRKEDRGRSKSPAQKNPGSGGKITKEKAALMVSRAYSKRTLPPNKPVDQ 234
    **************************************************************************************
```

Fig. 2. Representative primary structures of the centrally located intrinsically disordered regions (IDRs, in bold letters) in N proteins and the Ser/Arg residues (shown in red color) inside them. The identity and accession number of the sequences are listed in Table 1 in bold. The numbers are actual amino acid numbers in the full-length N protein. To draw attention to the abundant Lys (K) residues inside the IDRs, they have been colored blue. The two underlined Ser (in SRGGS) are GSK3 phosphorylation sites [18]. All four established coronavirus genera (A–D) are shown, accompanied by a schematic map of the protein, depicting only the relevant features (top). While the terminal IDRs are not shown, the central IDR area is indicated by the parallelogram, shaded in gradient to indicate a greater propensity of disorder on its N-terminal side, gradually tapering off towards the C-terminal side. For example, note that the alpha, beta, and gamma coronaviruses have two IDRs, whereas deltacoronaviruses have one, roughly corresponding to the first (also see Fig. 1). The concentration of Ser/Arg parallels this trend, being more concentrated in the common N-terminal half than in the C-terminal. The two coronaviruses, isolated from beluga whale and bottlenose dolphin are tentatively placed in a separate category (E) due to their apparently unique IDR arrangement (Also see Fig. 1, where they are placed as 'deviants' in the gammacoronavirus box for graphic illustration of the difference). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

out, the exact central IDR regions of several viruses were aligned and the Ser and Arg residues noted. For congruity, these sequences were common to those in Fig. 1 (Table 1).

As shown (Fig. 2), the Ser/Arg residues are most abundant in the first IDR peaks of all types of coronaviruses, without exception. The occurrence of these residues in the second IDR peak was more sporadic and showed some phylogenetic trend. Specifically, the second peak in alphacoronaviruses was also S/R-rich (Fig. 2A), but the abundance largely disappeared in the second peaks of beta and gamma (Fig. 2B, C). In the two 'deviant' gamma viruses, the stretched first peak was S/R-rich but the density tapered off towards the C-terminal end of the peak (Fig. 2E). A similar trend was noticed in the C-terminal end of the second IDR of both alpha and beta. In such S/R-poor IDR areas other polar amino acids seemed to predominate, Lys being the most noticeable (unmarked in Fig. 2).

One can conclude that the prevalence of Ser/Arg does not uniformly contribute to disorder, but likely serves other or additional roles, such as phosphorylation of the Ser residues [10,18].

4. Summary and discussion

In this short paper, the intrinsically disordered regions (IDRs) at the center of coronaviral nucleocapsid (N) proteins were explored. When the various features of the IDRs were dissected, they approximated a genus-specific pattern, and in a sense, appeared as diverse as the various genera. It is thus tempting to infer that the IDRs have implications in coronaviral phylogeny, pathology and transmission. Indeed, a very recent study demonstrated IgG, IgG and IgM antibodies against N in the sera of recovering CoVID-19 patients, suggesting that N is a potent antigen for host immunity and for diagnosis [41]. In the same study, use of light scattering and small angle X-ray scattering confirmed the high disorder content in the central region of purified SARS-CoV-2N protein, accompanied by flexibility and solvent accessibility, classic properties of IDR [41]. A notable finding was that the flexible linker did not adopt a fully extended conformation, suggesting the existence of residual structures within the linker, which is in agreement with the valley between the two IDR peaks shown here for SARS-CoV-2 and other betacoronaviruses (Fig. 1).

Since coronaviruses contain positive-strand RNA, recombinant viruses can be cloned relatively easily (in comparison to negative-strand ones), and subjected to reverse genetic analysis. Selected IDRs can be deleted, and Ser residues in them can be mutated to non-phosphorylatable amino acids [10,18], to test whether these changes affect virus growth and if so, how, perhaps by altering the protein's ability to interact with RNA and/or other proteins. It will be curious to find if loss of disorder in the second IDR brings the human tropism of SARS-CoV-2, a beta virus, closer to that of delta, which infects birds, in aspects of intracellular host-virus interaction. In a previous study, replacement of full-length N gene of the mouse hepatitis virus (MHV, a coronavirus) with that of bovine coronavirus (BCoV) generated a severely defective recombinant, which could be rescued by reverting mutations in the central S/R-rich region [12]. Note that the classical SR-family proteins, such as SF2/ASF, are dedicated RNA-binding proteins, many of which are splicing factors and have relatively long stretches of SR/RS repeats [42,43]. However, the S/R-rich domains of coronaviral N proteins do not have such long repeats, which may have lowered their exclusivity in RNA-binding but expanded their interaction repertoire into proteins. It is conceivable that this trade-off has fine-tuned their interacting sequences for multiple partners, specific for each coronavirus, making their study challenging and rewarding. Evidently, a more comprehensive analysis of mutations in this area, accompanied by studies of phosphorylation, may yield important results.

The two 'deviant' gammacoronaviruses deserve special mention. The beluga whale coronavirus (BwCoV) was discovered in 2008 in the liver of a single dead beluga whale (*Delphinapterus leucas*) in captivity, and characterized by genome sequencing. Historically, this was the first report of a complete coronavirus genome from a marine mammal [38]. The whale suffered a pulmonary disease and liver failure, but the full disease phenotype could not be studied because the virus was not cultured. Thus, it also remains unknown if beluga whale is the natural host of the virus. Six years later, a closely related virus was discovered in fecal samples from multiple Indo-Pacific bottlenose dolphins (*Tursiops aduncus*) and was named BdCoV [39]. For the record, the beluga whale, despite its name, is phylogenetically close to dolphins, both belonging to the cetaceans. Comparative genome analysis revealed that BwCoV and BdCoV were highly similar, the major difference being in their spike (S) protein sequences (only ~74% amino acid identities). By careful analysis of synonymous and nonsynonymous substitutions, Woo et al concluded that BdCoV may be evolving rapidly and that its transmission to bottlenose dolphins is a recent event [39]. Although both viruses showed higher homology with gammacoronaviruses than with any other genera, the authors were cognizant of the high similarity of the two viruses and their unique ability to infect mammals, whereas gamma viruses infect birds. These authors proposed that both viruses be classified as a distinct species, Cetacean coronavirus, in the *Gammacoronavirus* genus, whereas the canonical gammacoronaviruses will include the bird coronaviruses only. Based on this background, it was gratifying to find that the central IDR profile of these two viruses also reflect their uniqueness among the gamma viruses (Fig. 1C). From another perspective, the slouched second IDR in these viruses can be seen as a continuation of the first IDR gradually going downhill, but in either case, the deviation from the two-peak IDR profile is obvious.

To the best of my knowledge, these results constitute the first attempt to correlate the intrinsic disorder of an orthologous protein with the established coronaviral classification system. With its uncanny ability to recombine, the coronavirus will certainly continue to change and evolve into newer strains [44], which can be appended to these studies in the future to test the generality of the findings. Phylogenetic expansion or regression of disorder and its effect on evolutionary fitness is currently an uncharted field, and these studies may draw attention to this area. They can also form the basis of mutational analysis of the IDRs, such as swapping of IDRs between two different genera. In preventive medicine, a recombinant coronavirus with a deleted IDR may be attenuated for growth and serve as a vaccine candidate.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Liu YC, Kuo RL, Shih SR. COVID-19: the first documented coronavirus pandemic in history. Biomed J 2020;S2319–4170:30044–5. https://doi.org/10.1016/j.bj.2020.04.007.

[2] Woo PCY, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, et al. Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol 2012;86:3995–4008. https://doi.org/10.1128/JVI.06540-11.

[3] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019;17:181–92. https://doi.org/10.1038/s41579-018-0118-9.

[4] Helmy YA, Fawzy M, Elaswad A, Sobieh A, Kenney SP, Shehata AA. The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control. J Clin Med 2020;9:E1225. https://doi.org/10.3390/jcm9041225.

[5] Zhou P et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3. https://doi.org/10.1038/s41586-020-2012-7.

[6] Masters PS. The molecular biology of coronaviruses. Adv Virus Res 2006;66:193–292. https://doi.org/10.1016/S0065-3527(06)66005-3.

[7] Chang CK, Sue S-C, Yu T-H, Hsieh C-M, Tsai C-K, Chiang Y-C, et al. Modular organization of SARS coronavirus nucleocapsid protein. J Biomed Sci 2006;13:59–72. https://doi.org/10.1007/s11373-005-9035-9.

[8] Chang CK, Hsu YL, Chang YH, Chao FA, Wu MC, Huang YS, et al. Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: Implications for ribonucleocapsid protein packaging. J Virol 2009;83:2255–64. https://doi.org/10.1128/JVI.02001-08.

[9] Chang CK, Hou MH, Chang CF, Hsiao CD, Huang TH. The SARS coronavirus nucleocapsid protein – forms and functions. Antivir Res 2014;103:39–50. https://doi.org/10.1016/j.antiviral.2013.12.009.

[10] Peng TY, Lee KR, Tarn WY. 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 J 2008;275:4152–63. https://doi.org/10.1111/j.1742-4658.2008.06564.x.

[11] Hurst KR, Koetzner CA, Masters PS. Identification of in vivo-interacting domains of the murine coronavirus nucleocapsid protein. J Virol 2009;83:7221–34. https://doi.org/10.1128/JVI.00440-09.

[12] Hurst KR, Ye R, Goebel SJ, Jayaraman P, Masters PS. An interaction between the nucleocapsid protein and a component of the replicase-transcriptase complex is crucial for the infectivity of coronavirus genomic RNA. J Virol 2010;84:10276–88. https://doi.org/10.1128/JVI.01287-10.

[13] McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses 2014;6:2991–3018. https://doi.org/10.3390/v6082991.

[14] 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 2006;45:11827–35. https://doi.org/10.1021/bi0609319.

[15] Stohlman SA, Baric RS, Nelson GN, Soe LH, Welter LM, Deans RJ. Specific interaction between coronavirus leader RNA and nucleocapsid protein. J Virol 1988;62:4288–95. PMCID: PMC253863.

[16] Jayaram J, Youn S, Collisson EW. The virion N protein of infectious bronchitis virus is more phosphorylated than the N protein from infected cell lysates. Virology 2005;339:127–35. https://doi.org/10.1016/j.virol.2005.04.029.

[17] Surjit M, Kumar R, Mishra RN, Reddy MK, Chow VT, Lal SK. The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated and localizes in the cytoplasm by 14–3-3-mediated translocation. J Virol 2005;79:11476–86. https://doi.org/10.1128/JVI.79.17.11476-11486.2005.

[18] Wu CH, Yeh SH, Tsay YG, Shieh YH, Kao CL, Chen YS, Wang SH, Kuo TJ, Chen DS, Chen PJ. Glycogen synthase kinase-3 regulates the phosphorylation of severe acute respiratory syndrome coronavirus nucleocapsid protein and viral replication. J Biol Chem 2009;284:5229–39. https://doi.org/10.1074/jbc.M805747200.

[19] Fang S, Xu L, Huang M, Li FQ, Liu DX. Identification of two ATR-dependent phosphorylation sites on coronavirus nucleocapsid protein with nonessential functions in viral replication and infectivity in cultured cells. Virology 2013;444:225–32. https://doi.org/10.1016/j.virol.2013.06.014.

[20] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208. https://doi.org/10.1038/nrm3920.

[21] Uversky VN, Dunker AK. Understanding protein non-folding. Biochim Biophys Acta 1804;2010:1231–64. https://doi.org/10.1016/j.bbapap.2010.01.017.

[22] Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 2008;18:756–64. https://doi.org/10.1016/j.sbi.2008.10.002.

[23] Nishikawa K. Natively unfolded proteins: an overview. Biophysics 2009;5:53–8. https://doi.org/10.2142/biophysics.5.53.

[24] He R, Dobie F, Ballantine M, Leeson A, Li Y, Bastien N, et al. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem Biophys Res Commun 2004;316:476–83. https://doi.org/10.1016/j.bbrc.2004.02.074.

[25] Goh GK, Dunker AK, Foster JA, Uversky VN. Shell disorder analysis predicts greater resilience of the SARS-CoV-2 (COVID-19) outside the body and in body fluids. Microb Pathog 2020;144:. https://doi.org/10.1016/j.micpath.2020.104177104177.

[26] Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder-a breakthrough invention of evolution?. Curr Opin Struct Biol 2011;21:412–8. https://doi.org/10.1016/j.sbi.2011.03.014.

[27] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539. https://doi.org/10.1038/msb.2011.75.

[28] Barik S. Bioinformatic analysis reveals conservation of intrinsic disorder in the linker sequences of prokaryotic dual-family immunophilin chaperones. Comput Struct Biotechnol J 2017;16:6–14. https://doi.org/10.1016/j.csbj.2017.12.002.

[29] Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 2007, 35(Web Server issue), W460-464. DOI: 10.1093/nar/gkm363.

[30] Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. Bioinformatics 2008;24:1344–8. https://doi.org/10.1093/bioinformatics/btn195.

[31] Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 2010;1804:996–1010. https://doi.org/10.1016/j.bbapap.2010.01.011.

[32] Deng X, Eickholt J, Cheng JA. comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 2012;8:114–21. https://doi.org/10.1039/c1mb05207a.

[33] Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinf 2012;13:111. https://doi.org/10.1186/1471-2105-13-111.

[34] Atkins JD, Boateng SY, Sorensen T, McGuffin LJ. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. Int J Mol Sci 2015;16:19040–54. https://doi.org/10.3390/ijms160819040.

[35] Li J, Feng Y, Wang X, Li J, Liu W, Rong L, et al. An overview of predictors for intrinsically disordered proteins over 2010–2014. Int J Mol Sci 2015;16:23446–62. https://doi.org/10.3390/ijms161023446.

[36] Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 2017;74:3069–90. https://doi.org/10.1007/s00018-017-2555-4.

[37] Radivojac P, Obradović Z, Brown CJ, Dunker AK. Prediction of boundaries between intrinsically ordered and disordered protein regions. Pac Symp Biocomput 2003:216–27. PMID: 12603030.

[38] Mihindukulasuriya KA, Wu G, St Leger J, Nordhausen RW, Wang D. Identification of a novel coronavirus from a beluga whale by using a panviral microarray. J Virol 2008;82:5084–8. https://doi.org/10.1128/JVI.02722-07.

[39] Woo PCY, Lau SK, Lam CS, Tsang AK, Hui SW, Fan RY, et al. Discovery of a novel bottlenose dolphin coronavirus reveals a distinct species of marine mammal coronavirus in Gammacoronavirus. J Virol 2014;88:1318–31. https://doi.org/10.1128/JVI.02351-13.

[40] Domingo E, Holland JJ. RNA virus mutations and fitness for survival. Annu Rev Microbiol 1997;51:151–78. https://doi.org/10.1146/annurev.micro.51.1.151.

[41] Zeng W, Liu G, Ma H, Zhao D, Yang Y, Liu M, et al. Biochemical characterization of SARS-CoV-2 nucleocapsid protein. Biochem Biophys Res Commun 2020;527:618–23. https://doi.org/10.1016/j.bbrc.2020.04.136.

[42] Zahler AM, Lane WS, Stolk JA, Roth MB. SR proteins: a conserved family of pre-mRNA splicing factors. Genes Dev 1992;6:837–47. https://doi.org/10.1101/gad.6.5.837.

[43] Long JC, Caceres JF. The SR protein family of splicing factors: master regulators of gene expression. Biochem J 2009;417:15–27. https://doi.org/10.1042/BJ20081501.

[44] Cyranoski D. Profile of a killer: the complex biology powering the coronavirus pandemic. Nature 2020;581:22–6. https://doi.org/10.1038/d41586-020-01315-7.