VirusDis. (April–June 2021) 32(2):286–297
https://doi.org/10.1007/s13337-021-00655-w

ORIGINAL ARTICLE

# In silico approach to understand the epigenetic mechanism of SARS-CoV-2 and its impact on the environment

**Sahar Qazi[1] · Kayenat Sheikh[1] · Khalid Raza[1]**

Received: 10 September 2020 / Accepted: 18 January 2021 / Published online: 5 April 2021


**Abstract** The novel coronavirus (2019-nCoV) has led to the apex pandemic in 2020, responsible for the recent sequential spread. The 2019-nCoV has been discerned to be a Beta-BAT-SARS-CoV-2 lineage. The gene ontology (GO) identifies the virus to be localized in the Golgi apparatus with a vital molecular function of *binding* and viral progression. The source organism is almost all bats, further suggesting that the host of this virus is bat rather than civets or snakes, and has motifs which are perfect matches to various human and mouse genomic motifs such as—*zinc fingers*, *DNA-binding domains,* and *basic helix-loop-helix* factors. It has basic clusters of orthologs (COGs)—*Superfamily I DNA and RNA helicases and helicase subunits* and *Predicted phosphatase homologous to the C-terminal domain of histone macroH2A1* respectively hinting at the *epigenetic alterations* which could be the reason behind the "novelty" the virus. Our study discerns that the SARS-CoV-2 endorses the epigenetic mechanism essential for its replication and reproduction in the host organism. Furthermore, we identified six non-toxic disinfectants with higher pharmacokinetics and pharmacodynamics properties, namely *Quaternary Ammonium*, *Octanoic acid*, *Citric acid*, *Phenolics*, *1,2-Hexanediol*, and *Thymol*, that bind to lyases, nuclear receptors, fatty acids binding family, enzymes, and family AG protein-coupled receptors indicating that they target the nucleocapsid (N) protein, envelope (E) protein, membranous proteins of the novel coronavirus, thus, killing it from the surfaces when sprayed and are not harmful to the biological environment.

**Keywords** 2019-nCoV · Coronaviruses · Epigenetics · Beta-BAT-SARS-CoV-2 · Disinfectants · Toxicology analysis

## Introduction

The word "Nido' originates from a Latin word referring to "*the nest*", hence, Coronaviruses belong to the *Nidovirales* order. They account for the apex virus groups encapsulating –*Coronaviridae, Arteriviridae, and Roniviridae* families, wherein the *Coronavirinae* consist of two subfamilies namely—the *Coronaviridae*and the *Torovirinae* families. The *Coronavirinae* family is further dichotomized into four groups—alpha, beta, gamma, and delta coronaviruses which infect humans and animals, respectively. The *Nidovirales* order viruses are non-segmented, enveloped and positive-sense RNA viruses composed of gargantuan-sized RNA genomes ($\sim$ 30 kb) [27]. Common features of the viruses which exist within the *Nidovirales*order are—(a) a highly conserved genomic composition with a systematic arrangement of the replicase gene first and then the structural and the accessory genes; (b) ribosomal frameshifting for the expression of non-structural genes; (c) enzymatic activities occurring within the replicase-transcriptase polyprotein; and (d) downstream gene expression by the formation of 3′ nested sub-genomic mRNAs [11].

The novel coronavirus (2019-nCoV) was detected in early December 2019 in the city of Wuhan, China, and has impacted since then the entire globe. It has been tagged to

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s13337-021-00655-w.

✉ Khalid Raza
kraza@jmi.ac.in

[1] Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India

be fatal for people above the age of 60 and people affected with co-morbidities such as—diabetes, immunological disorders, cancer, etc. [35]. This pandemic has caused researchers to shift to its impact on human health. Recently published studies have found that this virus belongs to the beta-BAT-SARS-CoV-2 and has some essential amino acids for viral replication namely—alanine (ALA), aspartic acid (ASP), phenylalanine (PHE), leucine (LEU), asparagine (ASN), threonine (THR) and valine (VAL) respectively [27].

COVID-19 has not just affected human health, but also the biological environment. With national lockdowns exerted in numerous countries, it has been observed that pollution has decreased and the air quality index (AQI) has also improved dramatically [25]. However, the Centers for Disease Control and Prevention (CDC) has suggested a list of disinfectant sprays and other utilities that have been reported to destroy animal and human coronaviruses [7]. It is not new to know that these disinfectants are mainly pesticides and are highly toxic. To kill bacteria and viruses, heavyweight compounds that have a higher lethality index (LD) are selected to be promoted as disinfectant sprays [26]. Nonetheless, to kill such harmful germs, it is evident that these disinfectant compounds may also be harmful to the human skin and the environment.

This paper aims to identify the genomic analogy of the *Wuhan-based*SARS-CoV-2 by rectifying the essential structural motifs, network biology of the virus with other bio-cellular entities, functional annotation of the entire genome. We also executed analysis to determine the epigenetic alterations which are essential for its replication and reproduction in the host organism and the phylogenetics of the novel coronavirus. Further, we executed toxicity analysis of the disinfectants used to sanitize the environment for the prevention of SARS-CoV-2 infections and suggested few disinfectants which are non-toxic and have zero environmental impact.

## Materials and methods

### Sequence retrieval

For analyzing the genome of the novel coronavirus we utilized the completely sequenced *Wuhan-based* SARS-CoV-2 genome (accession ID—NC_045512.2) available on NCBI, while, for proteomic analysis, we deployed the SARS-CoV protein sequence (UniProtID: P0C6X7/RNA-dependent RNA polymerase (RdRp) accession ID: YP009725307.1).

### Sequence similarity, alignment and phylogenetic analysis

For identifying sequences that are highly similar to the query sequence of RdRp, we employed the Basic Local Alignment Search Tool (BLAST) [1, 6], and essentially significant application found in NCBI. For similar protein sequences to RdRp, BLASTp was performed (when the query and the sequences searched against both are proteins). The database chosen was non-redundant (nr) protein sequences and additional algorithm parameters were chosen to be the default. BLOSUM62 substitution matrix was used [29]. ClustalW was brought into use for multiple alignments of sequences and phylogenetic analysis while the Jalview application version 2.11.0 for viewing was used [34].

### Motif identification, enrichment and comparison

For motif identification, we deployed CTCFBSDB software [5, 43], while for enrichment MEME suite's CentriMo software [4] and comparison TomTom software available in MEME suite [16] was used.

### Genomic annotation

A pipeline was developed using Rapid Annotation using Subsystem Technology (RAST) software [3] for genomic annotation purposes.

### Proteomic analysis

We subjected the SARS-CoV protein sequence (UniProt ID: P0C6X7) to WebMGA [37] software to identify the cluster of orthologs to annotate its function. For the identification of protein interactors of SARS-CoV, we employed ConsensusPathDB [21]. The disordered regions prediction was executed using Protein Disorder prediction System (PrDoS) software [19].

### Structure prediction, evaluation and gene ontology analysis

COACH [39, 40] and TM-align [31, 42] software were used for active site prediction, I-TASSER software [38, 41] was deployed for developing the tertiary RdRp structure (PDB ID: QHD43415_11). Gene ontology (GO) [42] was used to identify the molecular, biological, and cellular functionalities of the RdRp structure.

### Identification of disinfectants against COVID-19

United States Environmental Protection Agency (EPA) database was searched for various disinfectants that have been reported to be used against the COVID-19 virus. The simplified molecular-input line-entry system (SMILES) of these compounds was retrieved from PubChem [23].

### Toxicology analysis of disinfectants for COVID-19

Virtual Rat [32] was employed to check and predict absorption, distribution, metabolism, excretion, and toxicity disinfectant compounds that are safe, don't have any skin allergy or irritation are environment friendly as well. To identify the target class of these compounds, we deployed SwissTarget software [14] that predicted the enzymatic classes to which these compounds bind the most.

## Results and discussion

### Sequence analysis

After viewing the global health accounts regarding the patients enormously affected with COVID-19, we strived to intersperse computational approaches and biological dynamics to comprehend the novel Coronavirus. The genome of SARS-CoV-2 comprises 10 genes (29,903 nucleotides). The first gene orf1ab *(Gene ID: 43,740,578)* accounts for two-thirds of the entire genome synthesizes orf1ab polyprotein (leader protein, nsp2, nsp3, nsp4, 3C-like proteinase, nsp6, nsp7, nsp8, nsp9, nsp10, RNA-dependent RNA polymerase, helicase, 3′-to-5′ exonuclease, endoRNAse, 2′-O-ribose methyltransferase, nsp11) [9]. In this analysis report, *RNA-dependent RNA polymerase (RdRp) (Accession ID: YP009725307.1)* is the key investigative protein. According to Cheng et al. [10], Orf1a/b nsp12(932 a.a) RdRp replication and transcription occur to produce genome and sub-genome-sized RNAs of both polarities in SARS-CoV.

#### *Similarity assay*

In total, around 100 sequences were shown; most related 25 sequences of protein (manually selected) that belonged to a different strain are outlined in Table S1 provided as a supplementary file. The resultant data indicates an overall 95 percent identity. Approximately, 30 sequences had 100 percent identity against the query sequence. The rest of the identity lied in 99–96%. It is interesting to note here that orf1ab polyprotein (Bat coronavirus RaTG13) had 99.57% identity, and two non-structural polyprotein 1ab (Bat SARS-like coronavirus) had 100 percent identity, while the rest of others were in the range of 96–98 percent identity. The outcome suggests that SARS-CoV-2 is altogether not a new species, it is a novel virus evolving from the group of related viruses belonging to the Coronavirus family. The source organisms are almost all bats, further suggesting that the host of this virus is actually bat rather than civets or snakes as suggested by many recent studies.

The graphical summary of the blastp result for query RdRp is depicted in Fig. 1. The query is represented by a turquoise blue line and the red lines represent the rest of the matching 25 sequences. Here it is clear that the alignment is $> = 200$. The sequences belonging to the Corona_RPol_N superfamily because of the highly similar region between 1 and 375 A.A.

The lineage (Taxonomy) of the selected similar sequences discern that they all belong to the Betacoronavirus category of viruses, of which 28 are SARS related coronaviruses and 11 are SARS related coronavirus 2. Bat-SARS-like coronavirus were 4 and 5 were SARS coronavirus MA15 (Supplementary Fig. S1).

#### *Alignment assay*

The complete FASTA sequences of the 25 sequences were aligned (932 A.A of the query). We observations that the similarity fluctuates in different proportions in different sections of the sequence. However, the ORF1ab polyprotein (Bat coronavirus RaTG13) shows higher similarity than another group of viruses indicating more evolutionary similarity/belongingness. Here, the grouping was done based on phylogenetic similarity.

#### *Phylogenetic study*

The average distance tree showed a lineage of RdRp SARS-CoV-2 with SARS-CoV polyprotein and Bat coronavirus RaTG13 whereas the Bat SARS-like coronaviruses were distant from the query protein (Fig. 2). On the contrary, the neighbourhood joining tree (BLOSUM62) showed that the RdRp SARS-CoV-2 was more evolutionarily closer to the other coronaviruses and not the ORF1ab polyprotein of SARS-CoV (Fig. 3). Bat coronavirus RaTG13 remained in a similar kind of relationship as in the average distance tree. It continues to show the highest similarity besides the other SARS coronaviruses.

### Genomic analysis

#### *Motif identification*

Our main aim was to identify various motifs present in the coronavirus genome, thus, we deployed CTCFBSDB

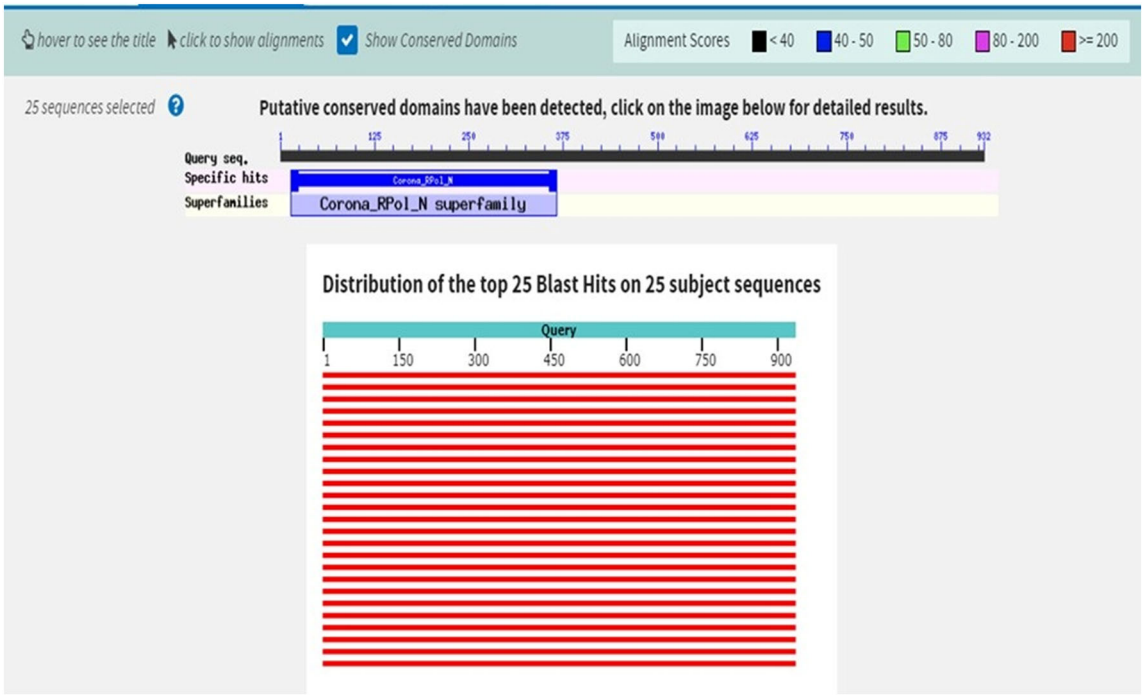


**Fig. 1** Graphical representation of blastp result for query RdRp (ID: YP009725307)

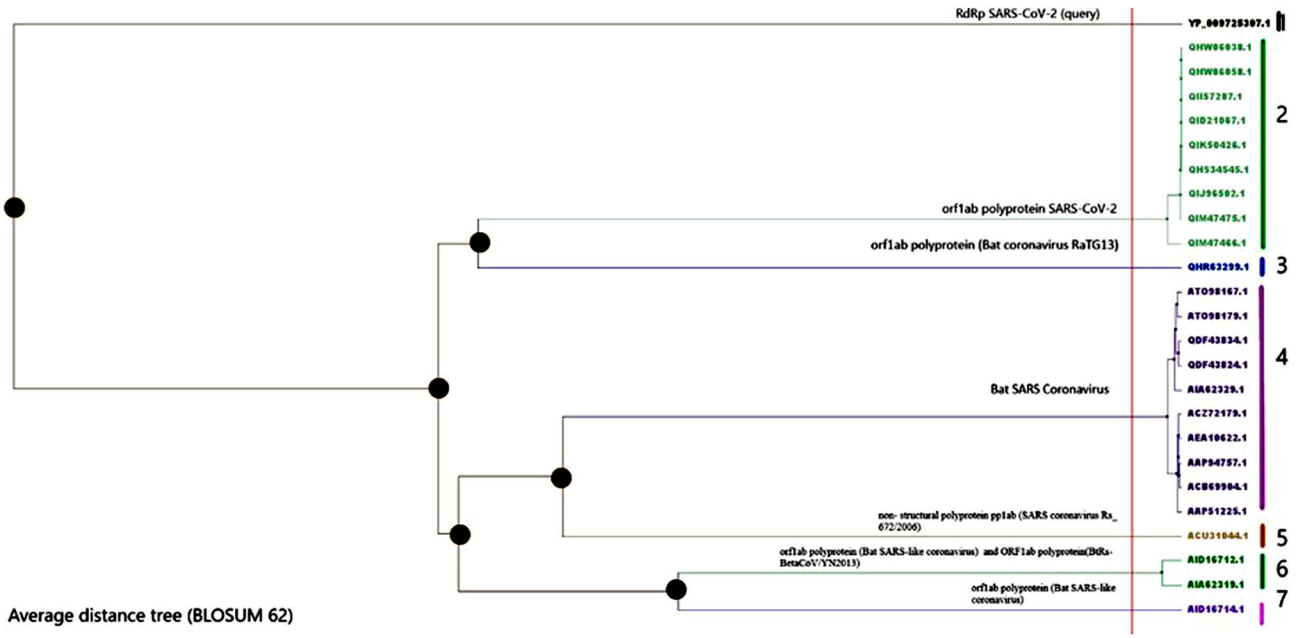


**Fig. 2** The average distance phylogenetic tree

software which is based on CCCTC-binding factor (CTCF), a highly evolutionarily conserved transcription regulator omnipresent from the fruit fly to human beings and binds to myriad DNA sequences via the combinatorial deployment of 11-zinc fingers depending on the biological context. These have been discerned to play an important role in gene expression control and represent a group of diverged DNA sequences. Studies also showcase them to be linked to epigenetic mechanisms namely—genomic imprinting and X-chromosome inactivation [5, 43]. Henceforth, we subjected the genomic sequences of the *Wuhan-based* SARS-CoV-2 genome (accession Id—NC_045512.2) to CTCFBSDB are figured out six essential motifs. Out of these six, only four motifs, including

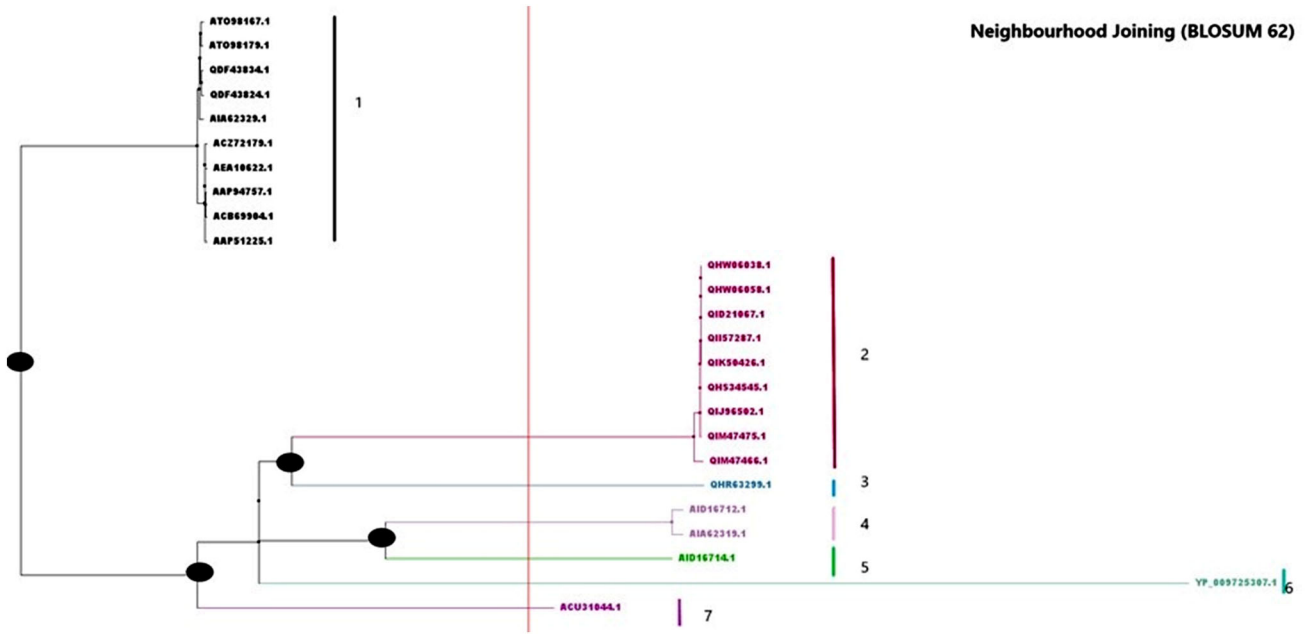


**Fig. 3** The Neighbourhood joining tree

CACCATCTGGTGTT, TACTGAGCAGGTGGTGCTGA, GAATTGCA, and TGGCAACCGGGTGTTGCTAT, were highly likely to be present in the Wuhan-based SARS-CoV-2 genome as these hits were highly matching to various other motifs present in the repositories such as JASPAR 2020 [12]. Table 1 represents the detailed description of the identified motifs along with their confidence score.

*Motif enrichment*

It was observed that the nucleotide pair of A-T was 0.3101 while the C-G pair had 0.1899. To identify the various other organisms which have similar motifs, we subjected the 4 best scoring motifs for motif comparison. It was observed that only three motifs were perfectly matching to various zinc fingers, DNA-binding domains, and basic helix-loop-helix factors present in the human and mouse genome (refer Supplementary Table S2 for details).

*Genomic annotation*

Much crucial information such as the characteristic features, closest neighbors, GC content which significantly affects the genome functioning and species ecology was identified. Also, GC content showcased a quadratic relationship with viral genomic size, the larger the genomic size, the GC content reduces significantly due to greater biochemical expenditure of GC base formation. We identified 38% GC content in the Wuhan-based 2019-nCoV genome which shares no closest neighbors. Howbeit, there were 12 characteristic features present in the coronavirus genome which were mainly coding sequence regions. Supplementary Table S3 represents the 12 characteristic coding sequences along with their length. It must be noted that the function is hypothetical; it could be due to the new strain of this coronavirus.

**Table 1** Six major identified motifs in the complete Wuhan-SARS-CoV-2 genome

| S.No | Motif source | Motif sequence | Motif symbol | Motif Start location | Motif length | Motif orientation | Score |
|---|---|---|---|---|---|---|---|
| 1 | EMBL_M1 | CACCATCTGGTGTT | M1 | 10,563 | 14 | + | 13.7721 |
| 2 | EMBL_M2 | GGAATTGCA | M2 | 19,466 | 9 | + | 12.404 |
| 3 | REN_20 | GTTAGCACCATAGGGAAGTC | M3 | 29,040 | 20 | − | 1.5473 |
| 4 | MIT_LM2 | TGCACTCAAGAGGGTAGCC | M4 | 859 | 19 | − | 8.54841 |
| 5 | MIT_LM7 | TACTGAGCAGGTGGTGCTGA | M5 | 5767 | 20 | − | 13.8196 |
| 6 | MIT_LM23 | TGGCAACCGGGTGTTGCTAT | M6 | 20,965 | 20 | + | 11.2214 |

## Proteomic analysis

*Cluster of orthologs identification and function annotation*

The Cluster of Orthologs (COG) protein was generated by comparing SARS-CoV-2predicted and known proteins in all completely sequenced genomes to infer sets of orthologs. Generally, every COG is composed of a group of proteins found to be orthologous across at least three lineages and is most probable to correspond to an ancient conserved domain. Two basic clusters linked to our coronavirus query sequence are *Superfamily I DNA, RNA helicases, helicase subunits,* and *predicted phosphatase homologous to the C-terminal domain of histone macroH2A1*. Superfamily 1 helicases are nucleic acid motor proteins which combine ATP hydrolysis to translocation, unwinding of DNA or RNA which are essential for cellular DNA and RNA metabolism, thus have apex range of nucleic acid events such as—DNA replication, recombination, and repair as well as many aspects of RNA metabolism [15]. MacroH2A1 consists of a domain 66% homolog to histone H2A has a unique structure, wherein a *C*-terminal adapter links the histone fold domain to a macrodomain bulging out from the condensed structure of the nucleosome affecting the functioning and organization of the surrounding chromatin, and is highly conserved in many unrelated proteins in the entire span of the animal kingdom, vertebrates and some invertebrates [28]. It must be noted that epigenetic changes in macroH2A variants are involved in human metabolic and oncologic diseases [20]. This hints at the fact that this 2019-nCoV might have undergone a few epigenetic modifications such as—histone modifications and DNA methylation during the evolutionary period, which makes it completely different from the existing human coronavirus. Table 2 describes two clusters of orthologs identified for the 2019-nCoV protein sequence. Previously reported studies have discerned that RNA type viruses, such as SARS-CoV have strong affiliations with RNA modifications such as –N6-methyladenosine (m6A) (most common type of epitransciptomic alteration)and N6,2′-O-dimethyladenosine (m6Am) modifications (m6A/m). These epigenetic changes are pivotal in viral replication process and affect the host innate immune system [2, 18, 22, 24].

*Protein interaction analysis*

Since protein annotation analyses provided us with a hint of the 2019-nCoV sequence to have histone modifications, which is an epigenetic alteration, could be the underlying mechanism for its "novel" nature. Apart from its ancestors, we subjected the 2019-nCoV protein sequence for a network analysis that could identify its nearest interactors and the type of interactions formed. Two unusual physical interactions between the nucleocapsid (N) protein of SARS-CoV-2 and SMT2 suppressor of mif two 3 homolog 1yeast, and Spike (S) protein of SARS interacts with Angiotensin I converting enzyme 2 (ACE2) RNA precursors were found (Fig. 4). SUMOylation, which is a post-translational modification (PTM) in proteins, is involved in controlling cellular processes namely—signal transduction, replication, chromosome segregation, and DNA repair. The interaction of N protein with SUMO1 (SMT2 suppressor of mif two 3 homolog 1 yeast) again provides enough evidence to our hypothesis that coronavirus sequence has epigenetic modifications as Small Ubiquitin-like MOdifier (SUMO) substrates often are inclusive of transcription factors and epigenetic regulators required for the epigenetic regulation of gene expression [13] and in the initiation and maintaining of heterochromatin silencing [8, 20, 36]. Furthermore, the spike protein (S) interaction with the ACE2 RNA precursor also indicates the epigenetic link. DNA hypo-methylation of ACE2 is crucially responsible for hemodynamic abnormality which causes abnormal cardiac output and a high peripheral resistance reflecting the identification of abnormal regulation of hereditable alteration in the gene expression occurring in the unavailability of underlying DNA sequence [17].

*Identification of disordered protein regions*

It is being reiterated here that the coronaviruses have various structural and accessory proteins arranged in a systematic manner encapsulating the ORFs, spike (S) protein, Envelope (E), and nucleocapsid (N). It was rectified that numerous disordered residues in the nucleoprotein (N) with the least disordered residues in the ORF, spike protein, and envelope. Figure 5 depicts the disordered residues present in the structural and accessory proteins of the 2019-nCoV.

**Table 2** The cluster of orthologs for the novel coronavirus (2019-nCoV)

| Name | No. of ORFs | Coverage | Abundance | Description |
|---|---|---|---|---|
| COG1112 | 1 | 0.385919 | 0.413452 | Superfamily I DNA and RNA helicases and helicase subunits |
| COG2110 | 1 | 0.547486 | 0.586546 | Predicted phosphatase homologous to the C-terminal domain of histone macroH2A1 |

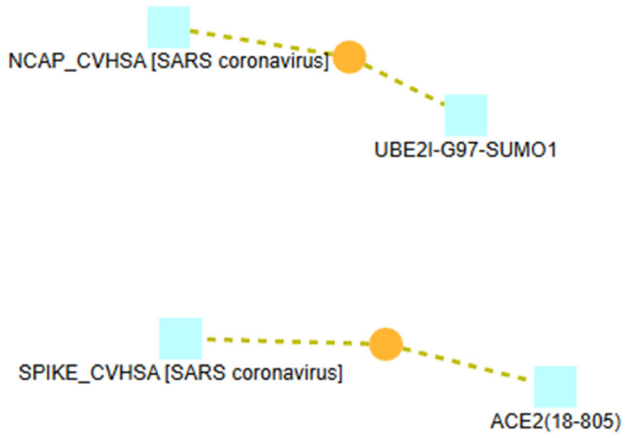


**Fig. 4** SARS-CoV-2 interactions

## Structural analysis

### *Active site prediction*

Here, the PDB hit 3H5SB has the highest C-score of 0.08 with the maximum cluster size (number of templates) of 27, the corresponding ligand is H5SB00 and the consensus binding residues are 11 in number. The positions of the binding residues in the chain are 593, 598, 601, 687, 688, 691, 759, 760, 813, 815, and 830. In almost all the predicted COACH PDB hits, the positions of the binding residues are almost the same suggesting that RdRp might show similar residue binding properties as them. The COACH results for the structure are shown in Supplementary Table S4, while the top 10 structural analogs as found in PDB (TM-Align) are listed in Supplementary Table S5.

The TM-score, IDEN, and coverage of the PDB Hit 6nurA are highest with the value of 0.85, 0.963, and 0.851, respectively. The RMSD value of 0.65 is the lowest of all. It indicates that out of all the hits, 6nurA structural and functional properties might be coinciding with that of RdRp structural and functional properties. There is a drop in TM-score significantly after this and ranges from 0.51 to 0.45, while the other parameters like RMSD, IDEN, and coverage have random values. However, the comparisons between the other hits and the query can be further stretched to find one or more similar properties.

### *Structure prediction*

The predicted structure had a TM-score of 0.80 as presented in Fig. 6. Such a high TM-score promises high confidence in the prediction of structure and it can be relied upon for further structural analysis. The structure of the

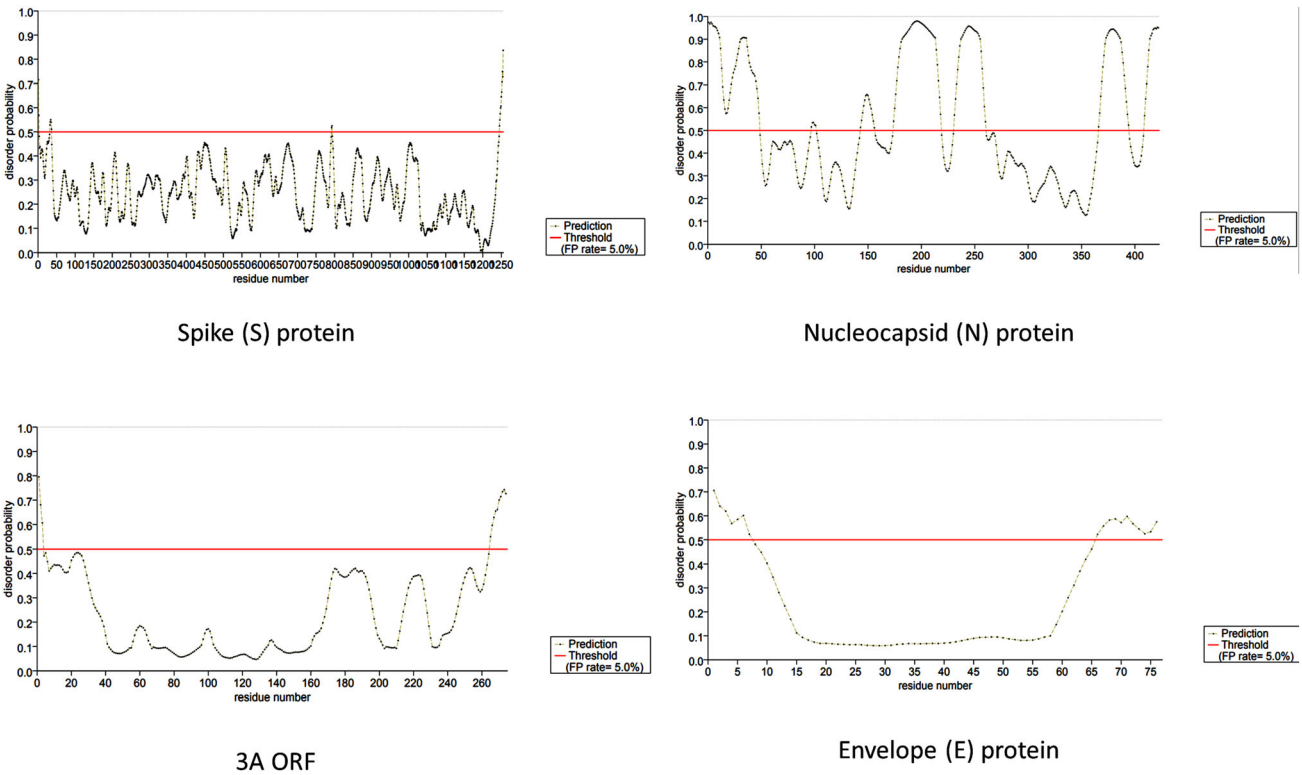


**Fig. 5** Graphical representation of disordered residues present in the structural and accessory proteins of 2019-nCoV sequences

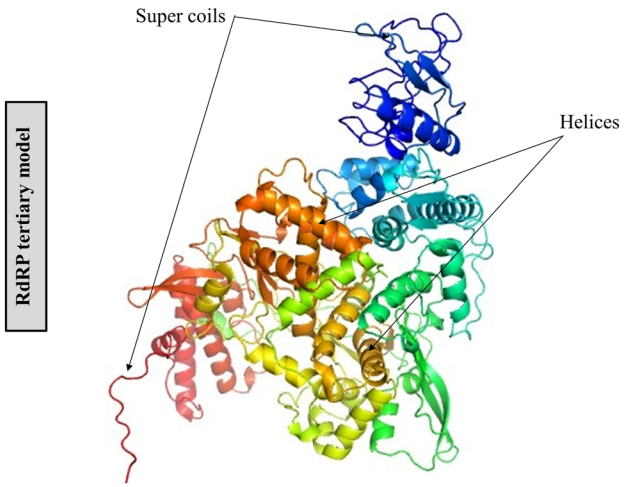


**Fig. 6** The RdRp structure predicted by I-TASSER

Wuhan-based 2019-nCoV depicted that it has a majority of helices and supercoiled regions with a few beta-strands.

*Gene ontology analysis*

Gene Ontology (GO) analysis allows predicting the biological aspects of a query, i.e., RdRp. The GO on I-TASSER predicted the terms like molecular functions (MF) [30], biological process (BP), and cellular component (CC). The list of high C-score MF terms is given in Table 3. The network of interconnection among the MF GO terms is presented in Supplementary Fig. S2.

Mostly, RdRp molecular function terms that are broadly associated are binding (identical protein binding, single-stranded RNA binding, double-stranded RNA binding). Hence, this molecular property of RdRp protein can be exploited to understand how and why it binds to proteins. It can be analyzed to know if some complementary sequence in the form of a drug binds to it, does the entire machinery of the virus to invade and multiply in a host collapses or not *catalytic activities* (Lys48-specific deubiquitinase activity, helicase activity, 3′-5′-exoribonuclease activity) *transferase activities*(mRNA (nucleoside-2′-O-)-methyltransferase activity, mRNA (guanine-N7-)-methyltransferase activity, RNA-directed 5′-3′ RNA polymerase activity). Table 4 enlists the GO terms related to BP, while Table 5 enlists the GO terms and the C-score. The RdRp protein can be located in almost all the parts of a virus and a host cell. However, the exomer complexes such as the Golgi apparatus and cytoplasmic vesicles have more traces of it (Supplementary Fig. S3). The host intracellular part has more abundant RdRp than the cytoplasmic viral factory. In addition to this, the top 5 enzyme homologies were obtained as shown in Supplementary Table S6. The C-score for all the hits was the same 0.060, however, the TM-score was highest, 0.433 for PDB ID 1s4fA suggesting that the enzymatic properties of the query might be its counterpart.

*Evaluation analysis of the predicted structure*

Out of the 10 were subjected to ProQ analysis for the goodness of model prediction. The results for the same are enlisted in Supplementary Table S7 [33]. The 5 structural analogs of PDB hit are mentioned in Supplementary Table S5. It is evident from the result that 6nurA (SARS-CoV-2 NSP12 bound to NSP7 and NSP8 co-factors) has the highest LG score and MaxSub score, which means that model 1 is extremely good.

The ProQ analysis results for enzyme homologs model goodness are enlisted in Table 6. The 2fhbA (Crystal Structure Analysis of Klebsiella pneumoniae pullulanase complexed with maltose) was a better model than 1LLwA. Structural studies on the synchronization of catalytic canters in glutamate synthase: complex with 2-oxoglutarate. However, the RMSD value was greater for the latter [33].

**Table 3** GO terms related to molecular function of QHD43415_11 from I-TASSER

| GO Term | C-score | Name |
|---|---|---|
| GO:1,990,380 | 0.92 | Lys48-specific deubiquitinase activity |
| GO:0,042,802 | 0.92 | Identical protein binding |
| GO:0,004,483 | 0.92 | mRNA (nucleoside-2′-O-)-methyltransferase activity |
| GO:0,004,482 | 0.92 | mRNA (guanine-N7-)-methyltransferase activity |
| GO:0,004,386 | 0.92 | Helicase activity |
| GO:0,003,968 | 0.92 | RNA-directed 5′-3′ RNA polymerase activity |
| GO:0,003,727 | 0.92 | Single-stranded RNA binding |
| GO:0,003,725 | 0.92 | Double-stranded RNA binding |
| GO:0,000,175 | 0.92 | 3′-5′-Exoribonuclease activity |

**Table 4** GO terms related to the biological process of QHD43415_11 from I-TASSER

| GO Term | C-score | Name |
|---|---|---|
| GO:0,080,090 | 1.00 | Regulation of the primary metabolic process |
| GO:0,051,171 | 1.00 | Regulation of nitrogen compound metabolic process |
| GO:0,039,657 | 1.00 | Suppression by virus of host gene expression |
| GO:0,039,507 | 1.00 | Suppression by virus of host molecular function |
| GO:0,031,326 | 1.00 | Regulation of the cellular biosynthetic process |
| GO:0,010,556 | 1.00 | Regulation of macromolecule biosynthetic process |
| GO:0,039,653 | 0.99 | Suppression by virus of host transcription |
| GO:0,039,503 | 0.99 | Suppression by virus of host innate immune response |
| GO:0,039,604 | 0.74 | Suppression by virus of host translation |
| GO:0,044,260 | 0.73 | Cellular macromolecule metabolic process |
| GO:0,044,238 | 0.73 | Primary metabolic process |
| GO:0,043,412 | 0.73 | Macromolecule modification |
| GO:0,039,579 | 0.73 | Suppression by virus of host ISG15 activity |
| GO:0,006,807 | 0.73 | Nitrogen compound metabolic process |
| GO:0,039,644 | 0.68 | Suppression by virus of host NF-kappaB transcription factor activity |
| GO:0,039,548 | 0.68 | Suppression by virus of host IRF3 activity |

**Table 5** GO terms related to Cellular Component for the QHD43415_11 from I-TASSER

| GO term | $C\text{-score}^{GO}$ | Name |
|---|---|---|
| GO:0,034,044 | 1.00 | Exomer complex |
| GO:0,033,646 | 1.00 | Host intracellular part |
| GO:0,016,020 | 1.00 | Membrane |
| GO:0,005,794 | 1.00 | Golgi apparatus |
| GO:0,039,714 | 0.73 | Cytoplasmic viral factory |

**Table 6** ProQ analysis result for the goodness of model for RdRp enzyme homologs

| S.No | PDB ID | Predicted LG score | Predicted MaxSub |
|---|---|---|---|
| 1 | 111wA | − 0.190 | − 0.447 |
| 2 | 2fhbA | 5.954 | 0.428 |
| 3 | 2vkzG | − 0.835 | − 0.113 |
| 4 | 1s4fA | − 0.835 | − 0.113 |
| 5 | 2q1fA | − 0.162 | − 0.379 |

## Identification and toxicity analysis of the disinfectants used against COVID-19

The United States Environmental Protection Agency database was used to retrieve 479 compounds that have been reported as disinfectants against all coronaviruses. Disinfectants that have been reported to kill human coronaviruses are mainly fifteen, namely *Quaternary ammonium*, *hydrogen peroxide*, *tetraacetylethylenediamine*, *phenolics*, *isopropanol*, *sodium carbonate peroxyhydrate*, *ethanol*, *sodium hypochlorite*, *octanoic acid*, *sodium chlorite*, *sodium dichloroisocyanurate*, *dodecylbenzenesufonic acid (L-lactic acid)*, *chlorine dioxide*, *hypochlorous acid*, *citric acid*, *potassium peroxymonosulfate*, *1,2-hexanediol*, *thymol*, *silver ion ([$Ag^{+}$]),* and *glycolic acid* (Supplementary Table S8). The simplified molecular-input line-entry system (SMILES) is the simplest single line form for specifying the molecular structures using ASCII strings. Sodium hypochlorite, Dodecylbenenesufonic acid, and silver ion SMILES structures were not available in the PubChem database. Thus, we selected the remaining compounds for toxicity analysis. It was observed that all these disinfectants are fair in the pharmacokinetics (PK) and physicochemical properties, howbeit, pharmacodynamics (PD) showcase that 30% of these disinfectant compounds showcase slightly toxic activity. The toxicity mainly refers to possible skin damages or irritation, and ocular and nasal irritation when these compounds are in direct contact with the human body. The four less toxic disinfectant compounds are—(a) *Tetraacetylethylenediamine*, (b) *Sodium carbonate peroxyhydrate*, (c) *Chlorine Dioxide,* and (d) *Glycolic acid*. These four disinfectants mainly bind to proteases, enzymes, and kinase families. The best suitable compounds that had no toxic activity and displayed a greater ADMET efficacy are—(a) *Quaternary Ammonium*, (b) *Octanoic acid*, (c) *Citric acid*, (d) *Phenolics*, (e) *1,2-Hexanediol*, and (f) *Thymol*. These six compounds mainly bind to lyases, nuclear receptors, fatty acids binding family, enzymes, and family AG protein-coupled receptors. Our analysis indicates that these six disinfectant compounds target the nucleocapsid (N) protein, envelope (E) protein, membranous proteins of the novel coronavirus,

thus, killing it from the surfaces when sprayed. Figure 7 presents a pie chart representations of the six best suitable disinfectants.

## Conclusion

The coronavirus disease (COVID-19) pandemic has caused researchers to shift to its impact on human health and the environment. In this paper, with the help of in silico analysis, we conclude that the Wuhan-based novel coronavirus is altogether not a new species, but a virus evolving from the group of related viruses belonging to the *Coronavirus* family. The source organisms are almost all bats, further suggesting that the host of this virus is bat rather than civets or snakes. The sequence similarity fluctuates in different proportions in different sections, howbeit, the ORF1ab polyprotein shows higher similarity than other groups of viruses indicating more evolutionary similarity/belongingness. The phylogenetic analysis of the 2019-nCoV discerns that there is an ancestral lineage of RdRp SARS-CoV-2 with SARS-CoV polyprotein and Bat coronavirus RaTG13. The genomic analysis started with motif identification in the 2019-nCoV and figured out six essential motifs, out of which 4 motifs (*CACCATCTGGTGTT*, *TACTGAGCAGGTGGTGCTGA*, *GGAATTGCA*, and *TGGCAACCGGGTGTTGCTAT*) were highly likely to be present in the Wuhan-based SARS-CoV-2 genome. Nucleotide base-pair composition of A-T was 0.3101 while C-G pair had 0.1899, and only three motifs—M1, M2, and M6 were perfectly matched to various *zinc fingers*, *DNA-binding domains,* and *basic helix-loop-helix* factors present in the human and mouse genome referring to the fact that these viral motifs are very similar to humans and mouse motifs, and these motifs have had some alteration which made it "novel" in this Wuhan-based 2019-nCoV. Furthermore, there were 12 characteristic features present in the 2019-nCoV genome mainly in the coding sequence regions, producing hypothetical proteins. The GC content was about 38% and showcased a quadratic relationship with viral genomic size.

The proteomic analysis retrieved two basic clusters of orthologs (COGs) for the Wuhan-based 2019-nCoV namely—*Superfamily I DNA and RNA helicases and helicase subunits* and *Predicted phosphatase homologous to the C-terminal domain of histone macroH2A1* hinting at the epigenetic alterations which could be the reason behind the "novelty" of the virus. To examine if any epigenetic mechanism was present, we identified two unusual physical interactions between the nucleocapsid (N) protein of SARS-CoV-2 and SMT2 suppressor of mif two 3 homolog 1yeast, while Spike (S) protein of SARS interacts with Angiotensin I converting enzyme 2 (ACE2) RNA precursors which confirmed our hypothesis of epigenetic modification in the Wuhan-based 2019-nCoV. Also, it was noted that numerous disordered residues in the nucleoprotein (N) of the 2019-nCoV protein have major disordered

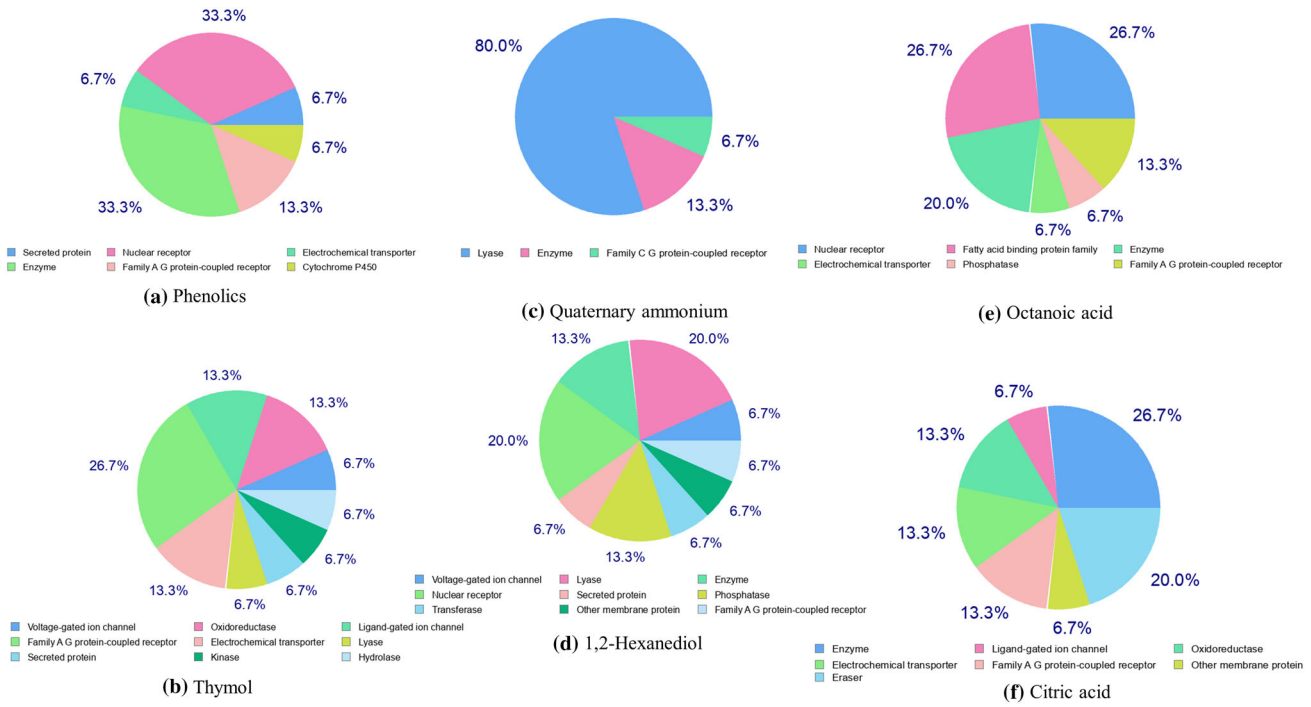


**Fig. 7** Pie chart representation of best six disinfectant compounds used against COVID-19

regions with the least disordered residues in the ORF, spike protein, and envelope.

Structural prediction of the 2019-nCoV displayed a TM-score of 0.80 and major subunits of the structure were helices and supercoiled regions with a few beta-strands. The gene ontology (GO) of the RdRp states that its molecular functions are associated with *binding,* the biological aspect is that it is vital for the virus progression and is localized in the membrane, exomer complex, Golgi apparatus, host intracellular region. Among the various disinfectants that have been listed as effective against COVID-19 virus, we identified only six non-toxic with higher PK and PD properties compounds namely *Quaternary Ammonium*, *Octanoic acid*, *Citric acid*, *Phenolics*, *1,2-Hexanediol*, and *Thymol*, that bind to lyases, nuclear receptors, fatty acids binding family, enzymes, and family AG protein-coupled receptors indicating that they target the nucleocapsid (N) protein, envelope (E) protein, membranous proteins of the novel coronavirus, thus, killing it from the surfaces when sprayed. Although disinfectant compounds *Tetraacetylethylenediamine*, *Sodium carbonate peroxyhydrate*, *Chlorine Dioxide,* and *Glycolic acid* have also been reported to kill the coronavirus, our study discerns that they are slightly toxic and mainly bind to proteases, enzymes, and kinase families. Even though these disinfectants have the potential to kill the virus, but they are slightly toxic and may cause a potential threat to the human body and the surrounding living bodies such as pets, plants, and birds.

**Acknowledgements** SQ is supported by DST-INSPIRE Fellowship funded by the Department of Science & Technology (DST), Govt. of India.

**Declarations**

**Conflict of interest** Authors declare that there is no any conflict of interest in the publication of this manuscript.

## References

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 1997;25(17):3389–402.
2. Atlante S, Mongelli A, Barbi V, et al. The epigenetic implication in coronavirus infection and therapy. Clin Epigenet. 2020. https://doi.org/10.1186/s13148-020-00946-x.
3. Aziz RK, Bartels D, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
4. Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. Nucl Acids Res. 2012;40:e128.
5. Bao L, Zhou M, Cui Y. CTCFBSDB: a CTCF binding site database for characterization of vertebrate genomic insulators. Nucl Acids Res. 2008;36:D83–7.
6. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform. 2009;10:421.
7. CDC. Cleaning and disinfecting your home: Everyday steps and extra steps when someone is sick. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/disinfecting-your-home.html. 2020.
8. Chang PC, Cheng YC, et al. The chromatin modification by SUMO-2/3 but not SUMO-1 prevents the epigenetic activation of key immune-related genes during Kaposi's sarcoma associated herpesvirus reactivation. BMC Genomics. 2013;14:824.
9. Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol. 2020;92(4):418–23.
10. Cheng VC, Lau SK, Woo PC, Yuen KY. Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. Clin Microbiol Rev. 2007;20(4):660–94.
11. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Coronaviruses. 2015:1-23.
12. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Santana-Garcia W. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucl Acids Res. 2020;48(D1):D87–92.
13. Garcia-Dominguez M, Reyes JC. SUMO association with repressor complexes, emerging routes for transcriptional control. BiochimBiophysActa. 2009;1789:451–9.
14. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, Zoete V. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucl Acids Res. 2014;42(W1):W32–8. https://doi.org/10.1093/nar/gku293.
15. Gilhooly NS, Gwynn EJ, Dillingham MS. Superfamily 1 helicases. Front Biosci Schol. 2013;5:206–16.
16. Gupta S, Stamatoyannopolous JA, et al. Quantifying similarity between motifs. Genome Biol. 2007;8(2):24.
17. Homes L Jr, Lim A, et al. DNA methylation of candidate genes (ACE II, IFN-γ, AGTR 1, CKG, ADD1, SCNN1B and TLR2) in essential hypertension: a systematic review and quantitative evidence synthesis. Int J Environ Res Publ Health. 2019;16(23):4829.
18. Imam H, Khan M, Gokhale NS, et al. N6-methyladenosine modification of hepatitis B virus RNA differentially regulates the viral life cycle. Proc Natl Acad Sci. 2018;115(35):8829–34. https://doi.org/10.1073/pnas.1808319115.
19. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. NAR. 2007;35(Web Server issue):W460–4.
20. Jueliger S, Lyons J, et al. Efficacy and epigenetic interactions of novel DNA hypomethylating agent guadecitabine (SGI-110) in preclinical models of hepatocellular carcinoma. Epigenetics. 2016. https://doi.org/10.1080/15592294.2016.1214781.
21. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. onsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Research 2011;39 (suppl_1):D712–D717
22. Kennedy EM, Bogerd HP, Kornepati AVR, et al. Posttranscriptional m 6 A editing of HIV-1 mRNAs enhances viral gene expression. Cell Host Microbe. 2016;19(5):675–85. https://doi.org/10.1016/j.chom.2016.04.002.
23. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucl Acids Res. 2019;47(D1):D1102–9.
24. Lichinchi G, Gao S, Saletore Y, et al. Dynamics of the human and viral m6A RNA methylomes during HIV-1 infection of T cells. Nat Microbiol. 2016;1(4):1. https://doi.org/10.1038/nmicrobiol.2016.11.

25. Mahato S, Pal S, Ghosh KG. Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi. India Sci Total Environ. 2020;730:139086. https://doi.org/10.1016/j.scitotenv.2020.139086.
26. National Pesticide Information Center (NPIC). Disinfectants and COVID-19. http://npic.orst.edu/ingred/ptype/amicrob/covid19.html. 2020.
27. Qazi S, Sheikh K et al. A coadunation of biological and mathematical perspectives on the pandemic COVID-19: a review. Preprints 2020, 2020040007. https://doi.org/10.20944/preprints202004.0007.v1.2020.
28. Re OL, Vinciguerra M. Histone MacroH2A1: a chromatin point of intersection between fasting. Senescence Cellul Regener Genes. 2017;8(12):367.
29. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9(2):173–5.
30. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725–38.
31. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucl Acids Res 2012;40(Web Server issue), W471–77.
32. Tseng YJ, Su B-H, Hsu M-T, Lin OA. Steps toward a virtual rat: predictive absorption, distribution, metabolism, and toxicity models. ACS Symp Ser. 2016. https://doi.org/10.1021/bk-2016-1222.ch014.
33. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci Publ Protein Soc. 2003;12(5):1073–86.
34. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics (Oxford, England). 2009;25(9):1189–91.
35. Whiting K. An expert explains: how to help older people through the COVID-19 pandemic. World Economic Forum. https://www.weforum.org/agenda/2020/03/coronavirus-covid-19-elderly-older-people-health-risk/. Accessed on 20th March, 2020. 2020.
36. Wilkinson KA, Henley JM. Mechanisms, regulation and consequences of protein SUMOylation. Biochem J. 2010;428:133–45.
37. Wu S, Zhu Z, et al. WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics. 2011;12:444.
38. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucl Acids Res. 2015;43(W1):W174–81.
39. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucl Acids Res 2013;41(Database issue), D1096–103.
40. Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics (Oxford, England). 2013;29(20):2588–95.
41. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2015;12(1):7–8.
42. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucl Acids Res. 2017;45(W1):W291–9.
43. Ziebarth JD, Bhattacharya A, Cui Y. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucl Acids Res 41(D1):D188–94. 2013.