Bioinformatic Application in COVID-19

6

Gurjot Kaur, Soham Mukherjee, and Shreya Jaiswal

Abstract

COVID-19 pandemic has seen massive application of in silico approaches to decipher the virus infectivity in humans as well as for the purpose of drug discovery. Data sharing has been a key to worldwide scientific collaboration and thus accelerating measures to counter the pandemic. Our chapter describes the main in silico approaches, their progression especially database generation and molecular modelling during the first year of the pandemic and emphasizes how these applications will contribute to pandemic-associated scientific discoveries, giving bioinformatics an important role for future tragedies.

Keywords

Bioinformatics · COVID-19

6.1 Introduction

COVID-19 pandemic started in December 2019 and was caused due to SARS-CoV-2 virus. Till the end of 2020, no specific treatment could be obtained. Drug and vaccine developments are progressing with the help of bioinformatic approaches. But first, let us delineate the historical use of in silico approaches in viral outbreaks.

G. Kaur (✉) · S. Jaiswal
School of Pharmaceutical Sciences, Shoolini University, Solan, India
e-mail: gurjotkaur@shooliniuniversity.com

S. Mukherjee
Faculty of Applied Sciences and Biotechnology, Shoolini University, Solan, India

S. Hameed, Z. Fatima (eds.), *Integrated Omics Approaches to Infectious Diseases*,
https://doi.org/10.1007/978-981-16-0691-5_6

6.1.1 Historical Facts in Previous Viral Outbreaks

Interestingly, COVID-19 is not the first viral disease where bioinformatic approaches were applied although the application of these approaches has magnified during the pandemic. The initial application was seen in Zika and Ebola virus as discussed below. Valuable lessons were learnt and amplified during the COVID-19 pandemic.

6.1.1.1 Zika Virus (2015–2016) and Ebola Virus (2014–2016) Outbreaks

Zika virus is caused by a flavivirus transmitted primarily by *Aedes* mosquitoes with mild symptoms lasting 2 to 7 days such as fever, rash, conjunctivitis, muscle and joint pain, and malaise or headache. Interestingly, Zika virus outbreak has been reported multiple times in the last century with the most recent one in 2015 in Brazil. As a countermeasure, Brazilian and American scientists came together for an open drug discovery collaborative effort. The scientific community recorded application of various computational strategies for drug repurposing and heavy usage of molecular docking methods on the viral proteins. These reports included protein homology modelling, X-ray crystallization structures, novel ligand, and protein discovery under the OpenZika project. However, due to a lack of corroborating in vitro and animal studies, many projects were shut down. On the other hand, the most recent Ebola virus outbreak took place in West Africa although it was first discovered in 1976 with fruit bats of the Pteropodidae family as natural hosts. It was previously called Ebola haemorrhagic fever, a severe, often fatal illness with a mortality rate of 50%. Drug discovery process for Ebola consisted of computational pharmacophore analysis of Ebola-active compounds, machine learning, in vitro testing, and generation of FDA-approved drugs (see review [1]). Table 6.1 summarizes the landmark computational studies for the two virus outbreaks.

6.2 COVID-19 Pandemic

By 2020, the COVID-19 patients became a major calamity as shown by COVID-19-positive cases and mortality on the WHO dashboard (https://covid19.who.int/). The WHO declared a state of global health emergency to coordinate scientific and medical efforts to rapidly develop a cure for patients [13]. The governments implemented social distancing and lockdowns to curb the spread. Many existing antiviral medications have been tried in hopes to slow or even cure the severely affected patients and thus decrease the morality rate. One of the very promising strategies, therefore, was to use bioinformatic approaches as shown in Ebola and Zika outbreaks.

COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a new strain of coronaviruses that has been isolated from the Huanan Seafood Market, Wuhan, China, in December 2019. The identified primary reservoir is horseshoe bat, and transfer to human takes place through unknown intermediate hosts [14]. In general, the family of coronaviruses can cause respiratory,

Table 6.1 Computational approaches in Ebola and Zika drug discovery

Reference	Computational approach used in the study	Finding of the study
Ebola virus		
Brown, Lee [2]	• Structure-based screening (in silico) of millions of drug-like small molecules (approximately 5.4 million) targeting binding pocket of EBOV viral protein 35 (VP35), followed by in silico molecular docking of hits and NMR-based binding studies	• Several compounds, especially pyrrolidinones, capable of binding to VP35 with high affinity and inhibiting polymerase co-factor activity identified
Ekins, Freundlich [3]	• Common features of pharmacophore analysis of four FDA-approved compounds exhibiting in vitro and in vivo activity against EBOV • Receptor-ligand pharmacophore analysis consisting of hydrophobic and hydrogen bonding features • In silico molecular docking of the FDA-approved drugs onto VP35 structure	• Pharmacophore model for EBOV actives suggested common chemical features in the compounds and potentially similar target/mechanism • Receptor pharmacophore and molecular docking-supported VP35 could be a likely target for other FDA-approved drugs and analogues
Litterman, Lipinski [4]	• Computational validation of 55 small molecules with activity against EBOV and drug likeness property evaluation • Pan assay interference compounds (PAINS) computational filter to identify potentially problematic structures	• Usage of medicinal chemistry, in silico and in vitro, and in vivo data could provide better measures for Ebola drug development
Veljkovic, Loiseau [5]	• Virtual screening was conducted of drugs against Ebola using EIIP/AQVN-based criterion	• The selected (EIIP/AQVN) criterion was used as an efficient filter in screening for the inhibitors of Ebola virus infection. Drugs (approved and experimental) were selected by the criterion and represented the valuable source of therapeutics for the treatment of Ebola virus disease
Ekins, Freundlich [6]	• Bayesian machine learning computational models generated with molecules from EBOV replication assay and viral pseudotype entry assay. Models used to score drug libraries to identify potential inhibitors • Pharmacophore analysis conducted for best hits and further in vitro testing performed to check efficacy	• Based on scoring by Bayesian models, three distinct molecules were identified which were previously not covered in literature pertaining to EBOV actives • Ligand pharmacophore analysis and further in vitro tests supported efficacy of the hits against EBOV
Zika virus		
Ekins, Perryman [7]	• Dengue virus crystal structure was used as template in homology model of ZIKV envelope protein. In silico screening performed through molecular docking of selected	• Top 10 compounds from Prestwick Chemical Library including three antivirals (ritonavir, indinavir, and saquinavir) and a few antimalarials

(continued)

Table 6.1 (continued)

Reference	Computational approach used in the study	Finding of the study
	compounds from Prestwick Chemical Library to identify hits for in vitro testing	were identified as potential hits based on conformation scores
Ekins, Liebler [8]	• Homology models of ZIKV proteins generated through evolutionary relationships among flaviviruses. Best homology models screened according to GQME, QMEAN4 scoring, and Ramachandran plot analysis • Zika virion surface illustration generated by combining homology model with dengue virus envelope symmetry data	• Homology models provided to serve as starting points for docking studies. Qualitative analysis of ZIKV virion compared to dengue virus cryo-EM virion outlined
Ekins, Liebler [8]	• Distributed computing on millions of devices to run docking of drugs (compounds) against structures (crystal) and homology models of Zika proteins	• Homology modelling provided a way to accelerate new drug development in the absence of crystal structure of Zika proteins
Sahoo, Jena [9]	• Molecular docking analysis of four FDA-approved drugs (for dengue virus) and structure-based virtual screening of novel drug-like compounds against NS3 protein of ZIKV, which is essential for viral replication	• Berberine was identified as the best FDA-approved antiviral with potential to be repurposed. Top 10 novel drug-like compounds sharing properties with berberine were identified, among which best hits were found to exhibit high binding affinity and interactions with key residues
Mottin, Braga [10]	• The study of molecular behaviour of helicase domain of NS3 (NS3h) in ZIKV in concert with ssRNA to elucidate NS3h activity and inhibition by molecular dynamics simulations	• Molecular dynamic trajectories demonstrated presence of ssRNA and stabilize the RNA binding loop in NS3h so it maintains closed conformation. RNA binding loop conformation is also suggested to affect optimal ligand binding
Sharma, Murali [11]	• Induced fit docking of EGCG onto binding site of envelope protein to elucidate key interactions, molecular dynamic simulations, and drug-like property calculations	• Interaction of EGCG with various residues was found, and it also provides protein conformation stabilization as analysed from molecular dynamics trajectories
Ramharack and Soliman [12]	• 'Per-residue energy decomposition pharmacophore' models for NS5 polymerase and NS5 methyltransferase using the molecules ribavirin and BG323. Virtual screening conducted using ZINC database and molecular docking performed for screened candidates	• Virtual screening from ZINC database lead to identification of 23 candidates for NS5 polymerase and 18 candidates for methyltransferase. Compounds with the best docking scores were identified as potential actives (ZINC39563464 for NS5 polymerase and ZINC64717952 for NS5 methyltransferase)

VP35 viral protein 35, *NMR* nuclear magnetic resonance, *FDA* US Food and Drug Administration, *EBOV* Ebola virus, *EIIP* electron ion interaction potential, *AQVN* average quasi valence number, *EVD* Ebola virus disease, *ZIKV* Zika virus, *GQME* global model quality estimation, *QMEAN4* qualitative model energy analysis with linear combination of four statistical potential terms, *cryo-EM* cryogenic electron microscopy, *EGCG* epigallocatechin-3-gallate

gastrointestinal, hepatic, and central nervous system diseases. SARS-CoV-2 and its nearest neighbours in the phylogenetic tree, i.e. SARS-CoV and MERS-CoV, cause severe respiratory diseases. Its widespread transmission is due to travel and human to human contact [15]. SARS-CoV-2 is currently known to be sensitive to heat and UV rays and effectively destroyed with 75% ethanol, acetic acid, and chlorine-containing disinfectants [16].

6.2.1 Current Bioinformatic Efforts in COVID-19

We have highlighted the bioinformatic efforts that have been used in drug discovery research for COVID-19 till end of 2020.

6.2.1.1 Genomic Efforts: Sequencing Efforts

The whole genome sequences for the virus, SARS-CoV-2, have been isolated from patients from several countries including Brazil, China, Germany, and the USA. They were made publicly available as soon as they were sequenced to accelerate scientific research. There are currently more than 300 samples online. All the sequence samples were found to be closely related with few mutations pointing towards a common ancestor. For example, the Brazilian genome differed by three mutations to the Wuhan reference strain, and two of these three mutations were shared with a German sample. Efforts have been made to provide complete genome sequencing and thus aid the analysis of how the gene is translated to protein and what could be the protein functions especially in SARS-CoV-2 infectivity pathway. Genome sequencing is the starting point for all analysis regarding structure and function of resulting proteins. In addition, it provides knowledge about the origins of the SARS-CoV-2 virus and thus transmission profile. For the scientific research to rapidly move forward, it was crucial that the whole genome sequencing of SARS-CoV-2 is performed at a fast pace.

SARS-CoV-2 belongs to genus *Betacoronavirus* and subgenus *Sarbecovirus*. The virion or the infecting particle consists of an envelope containing a single positive-stranded RNA. The first genome, accession number NC_045512.2, was isolated from a patient in Wuhan, Hubei province, China, and named SARS-CoV-2 Wuhan-Hu-1. Current GenBank sequences and next-generation sequences stand at 39751 and 4266, respectively (data taken from NCBI-NLM SARS-CoV-2 Resources on November 12, 2020). Current estimates suggest a genome size of 29.9 kb and 11 open reading frames or ORFs. The organization of genes encoding the various proteins is shown below.

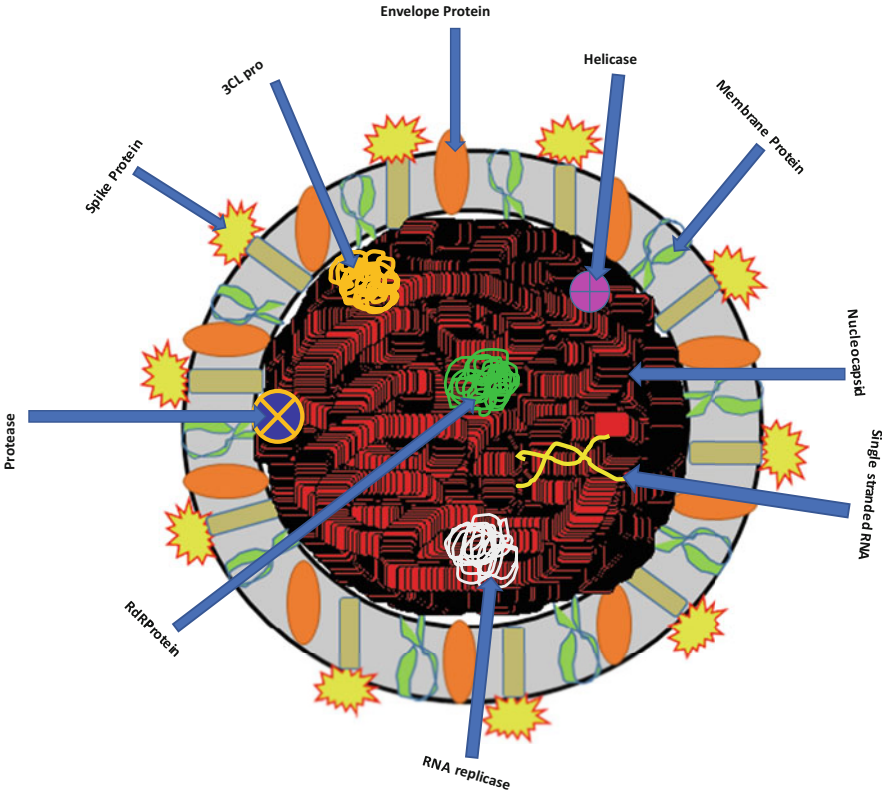

Fig. 6.1 Proteins on SARS-CoV-2 virion

5′-leader-UTR-replicase-ORFab-Spike (S)-Envelope (E)-Membrane (M)-Nucleocapsid (N)-3′UTR-poly (A) tail-3′-UTR end.

Figure 6.1 illustrates the cellular location of different proteins on SARS-CoV-2 virion. Unsurprisingly, the genome of SARS-CoV-2 is very similar to SARS-CoV (82%), bat- CoV-RaTG13 (96%), and bat-SL-CoVZC45 (86.9%). Main difference lies in the longer branch length to the bat viruses. Although mutations are being observed between various SARS-CoV-2 strains isolated from patients, the reported similarity is 99.98%. Phylogenetic analysis of 160 genomes has shown three main variants (classified as A, B, and C ancestral types) with certain mutations in specific variants, i.e. synonymous mutations T29095C and T8782C are identified in type A and type B, respectively, and non-synonymous mutations C28144T (Leu to Ser) and G26144T (Gly to Val) are detected in type B and type C, respectively [17]. This knowledge is being employed to identify genome-based community hotspots. For some mutations, radical changes in functionality of the protein, host specificity, or virus infectivity have been seen (Table 6.2). Mutational hotspots are located at positions 1397, 2891, 14408, 17746, 17857, 18060, 23403, and 28881. Mutational

Table 6.2 Common SARS-CoV-2 proteins and current structure models

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
Mpro	Cysteine protease participates in the viral replication and cleaves the viral ORF1ab polyprotein at 11 sites	Mutation: R60C obtained from Mpro of Vietnam isolate revealed that the point mutation affects protease stability and binding of inhibitor [18]	6 LU7 6 W63 3M3V 6Y84 7BRO 6WQF 6 M03 6Y2E 6Y2F 6Y2G 6Y7M 6YB7 6LZE 6M0K 6YNQ 6YVF 6WNP 7BUY 7BRR 6WTT 7BRP 6YZ6 7BQY 6YT8
Spike glycoprotein (S)	The host-cell membrane is fused with the viral membrane. During the maturation of virus, Cleaving of spike protein is done to its subunits: virus is attached to the cell membrane through S1 subunit by interacting with host receptor. Fusion of the virus with human cell membranes is mediated by Ace2 and S2	Mutation: D614G, G476S, and V483A (most frequent) in RBD. They slow down the development of therapeutics Mutation: S19P and E329G may determine the host specificity of SARS-CoV-2. Important for molecular recognition, PCR testing kits, antiviral specificity, and vaccine development	6VYB 6VSB 6LXT 6LVN 2AJF 6VW1 6LZG 6M0J 6 W41 6YLA 6YM0 6WAQ 7BZ5 Complex of spike (6ACD) and ACE2 (2AJF) Complex of spike (6VXX) and ACE2 (1R42)

(continued)

Table 6.2 (continued)

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
Polyprotein 1ab (ORF1ab-266–21,555 nucleic acids): NSP1–16 (ORF1ab)	Replicase with multiple functions. Polyprotein consisting of 15 non-structural proteins, auto-proteolytically cleaved into multiple enzymes that form replicase-transcriptase machinery consisting of RNA-dependent RNA polymerase (RdRp), helicase, 3′–5′ exonuclease, endoRNAse, and 2′-O-ribose methyltransferase. NSP15 encodes a nidoviral uridylate-specific endoribonuclease that interacts with the NSP7/NSP8 complex. The antiviral activity of the STAT1 transcription factor is antagonized by ORF6 by sequestering IMPα/β1 on the rough ER/Golgi membrane. Human ubiquitin system with multiple members of the Cullin-2 E3 ligase complex interacts with the ORF10	Mutation: S24L in ORF8 protein strengthens the folding stability of the spike protein and ORF8 protein and shows female dominated pattern Mutation: A stabilizing mutation at position 321 in the endosome-associated-protein-like domain in NSP2 protein Mutation: A destabilizing mutation at position 192 in the NSP3 protein—useful for differentiating SARS-CoV-2 from SARS-related coronaviruses. Mutation: S723G and P1010I in the transmembrane helical segments of NSP2 and NSP3 determine the host specificity of SARS-CoV-2	NSP10–NSP16 complex: 6 W75 6YZ1 NSP15: 6VWW 6 W01 6WXC 6WLC NSP16–NSP10 complex: 6W4H 6 W61 6WKS 6WQ3 6WRZ 6WVN 6WJT 6WKQ NSP3: 6VXS 6WEN 6 W02 6W6Y 6WCF 6WEY 6WOJ 6YWL 6YWK 6YWM NSP7 and NSP8 complex: 6WIQ 6WQD 6WTC 6YHU NSP9: 6W4B 6WXD 6W9Q
RdRp or RNA-dependent RNA polymerase	It is required for the replication of viral genome and nucleic acid metabolism	Mutation: C14408T mutation is adjacent to the drugs targeting RdRp hydrophobic cleft	6 M71

(continued)

Table 6.2 (continued)

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
Helicase	Required for viral genome replication and nucleic acid metabolism	Mutations: P504L and Y541C affected functional domain of NSP13 and even change the shape of ATP binding site in the helicase [19]	6XEZ
3′–5′ exonuclease	Required for viral genome replication and nucleic acid metabolism	None yet reported	None yet available
EndoRNAse (NSP15)	Required for viral genome replication and nucleic acid metabolism	None yet reported	7K1O 7K1L 6WXC 6 W01 6WLC 6X1B 6VWW
2′-O-ribose methyltransferase (encoded by non-structural NSP16)	Required for viral genome replication and nucleic acid metabolism. It binds N7-methyl guanosine cap and methylates the ribose 2′-O position of the first and second nucleotide of viral mRNA, which is essential to evade the immune system	Mutation: L111F, A116S, P236L, and two mutations localized towards termini, namely, P7S and N285D, have been identified in Indian isolates. These mutations are likely to alter protein dynamics, stability, and secondary structure and render pharmacological agents infective [20]	6YZ1 6W4H 6 W75 6 W61
Envelope protein (E)	This envelope protein self-assembles itself in the host membranes by forming ion channels and also acts as viroporin. It has a critical role in viral assembly and releases exerting viral pathogenicity	Mutation: Non-synonymous mutations in envelope protein observed in ~0.4% of SARS-CoV-2 whole genomes which might potentially affect propagation [21]	7K3G
Nucleocapsid protein (N)	This protein is responsible for the interaction of spike, envelope, and membrane proteins, which forms the nucleocapsid of the virus	Mutation: Several SARS-CoV-2 strains/substrains have exhibited mutations on serines 186, 197, and 202 which are phosphorylation sites involved in cell cycle control [22]	6M3M 6WZO 6WZQ 6WJI 6YUN 6ZCO 6WKP 7ACT 7ACS 2GIB 1SSK 7C22

(continued)

Table 6.2 (continued)

Protein target	Function of the protein	Mutations	PDB ID of 3D structure
			2CJR 6YI3 2JW8 2OFZ 2OG3 7 CE0 7CDZ 6VYO

Original article references are given. Other references used for the table can be found in reviews: [23, 24]

variants are crucial to assess any possible drug resistance and COVID-19 clinical presentation. This information is also crucial for designing COVID-19 vaccines as well as rapid diagnostic assays. Structural genomic analysis has shown that the viral genome is composed of four structural proteins (spike-envelope-membrane-nucleocapsid) and two non-structural proteins (main protease and RdRp) [23].

6.2.1.2 Protein Structure-Based Methods

Homology Modelling of SARS-CoV-2 Proteins

Considerable work has been accomplished in the development of homology models of SARS-CoV-2 proteins. Leading the way is the main viral protease or Mpro. Other proteins that have been heavily modelled through homology modelling are spike (S) protein, ACE2 human target, and RdRp. Table 6.2 provides the current list of protein targets and highlights whether a specific target is over- or underutilized for the computational studies.

The knowledge of three-dimensional structure of proteins is crucial to develop drugs that can modulate the protein's action. The structural information, obtained through X-ray crystallography or cryogenic electron microscope or nuclear magnetic resonance, provides information on binding sites and the mechanism of action/inhibition. The three-dimensional structures are also crucial to molecular docking studies as they serve as starting point. These models have been used for docking to understand mechanisms of viral infection and possible treatments. Currently, there are over 400 (total) structures available for various SARS-CoV-2 targets on RCSB PDB.

Molecular Docking Studies

Docking simulations have been performed for treatment through both small molecules and antibodies. Small molecule strategies require a target protein structure and screening of molecules that bind this protein using molecular docking and validation using molecular dynamic simulations. Many studies have been performed

in a very short span of time with the rapidly made available targets (Table 6.2). In the second method, antigen-antibody docking simulation has predicted high-affinity binding of human antibodies to SARS-CoV-2 proteins. Human antibody CR3022 is shown to possess a high affinity for spike protein [24].

Unfortunately, many early studies lacked proper validation of molecular docking data (redocking or molecular dynamics), and thus reproducibility of the results is highly doubtful. This trend was seen for many drug or herbal chemical candidates in the early days of COVID-19. Studies that came a little later validated the data by molecular dynamics and provided important parameters regarding drug/herbal chemical stability inside the binding pocket of protein target and possible hydrogen bonding and hydrophobic interactions during the simulations [25–28]. Very few direct validations of molecular docking protocol by redocking with the known ligand, i.e. native ligand on the X-ray crystal structure, were performed in the early days [1, 24, 29].

Drug Repurposing

As drug discovery is a long process and initiating it from a completely new drug candidate may result in a long wait while the COVID-19 pandemic was gaining momentum, drug repurposing seemed to be a much rapid alternative, previously employed for many related (MERS and SARS-CoV) and unrelated diseases. In this process, many drugs have been screened [1, 24]. These candidate molecules may be used in SARS-CoV-2 enzyme assays, antagonism of protein mechanism, or decelerating viral infectivity. It is hypothesized that drugs previously used in other viral diseases could also be used. An important aspect would be to make sure that the drugs are, in general, available for further experimentation and development in case they show promise. It will be futile to develop drugs that are only available as structural moieties or are at an investigational stage only. A simple strategy to drug repurposing that shows promise is screening a class of compounds rather than random drug screening. FDA-approved HIV-1 protease inhibitors, e.g. atazanavir and ritonavir, and hepatitis C NS3/4A protease inhibitors, e.g. lopinavir, have been successfully docked into the Mpro active site of SARS-CoV-2. Currently, China and South Korea have treated COVID-19 patients with Kaletra, the combination of lopinavir 200 mg/ritonavir 50 mg with some benefits [30].

6.2.1.3 Database Generation

It is needless to say that databases are very important to accelerate research in the current times especially because they provide a reference point to scientists worldwide. Virtual databases provide the advantage of large storage capacity while being continuously updated. In SARS-CoV-2, this has provided a better understanding of the virus's origin after extensive comparisons between genomic data online. Genomic comparisons also helped in producing specific primers for RT-PCR detection kits as early as possible. Many databases have proven extremely useful in the pursuit of treatment strategies for COVID-19. While primary databases such as nucleotide sequence databases and protein sequence databases are extensively used to submit as well as obtain sequencing information for COVID-19 targets, secondary databases

Table 6.3 COVID-19-specific online databases

Database	Information provided by the database	References
2019nCoVR	For studies on classification of virus (viral taxonomy), evolution (changes) of genome, molecular diagnosis, and development of drugs	[31]
The Genome Detective Coronavirus Typing Tool	Used for the tracing of new viral mutations in SARS-CoV-2 genome sequences accurately	[32]
CoronaVR	It is used for the identification of potential epitope-based (B cells and T cells) vaccine candidates, siRNA-based (subsection of therapeutic section) therapeutic regimens, and diagnostic primers, that can be used in testing kits to identify COVID-19 virus	[33]
CoV-AbDab	It contains a relevant set of data which includes cross-neutralization antibody/nanobody origin evidences, full variable domain sequence, assignments of germline, regions of epitope, links to relevant PDB entries, comparative modelling (homology models), and literature. It is also used for the aid in preceding exposure diagnosis or to assist in predicting the efficacy of vaccine. It is used for the understanding of antibody at molecular level, which is able to attach with a *Betacoronavirus* antigen that could be relevant in treating COVID-19 (SARS-CoV-2) infection	[34]

such as Prosit, PRINTS, Pfam, InterPro, PhenomicDB, or a genotype-phenotype database provide support to understand protein structure and functionality, while molecular structure databases such as Protein Data Bank, SCOP, CATH, and PubChem are important to obtain structural information of molecular targets. A few COVID-19-specific databases are listed in Table 6.3.

Due to the widespread effect of the COVID-19 pandemic across countries and to allow easy sharing of scientific information between scientists, these databases have come up. While the WHO COVID-19 database provides general information on sociodemographics of the mortality and infected individuals, ViPR or Virus Pathogen Database and Analysis Resource has been updated to include information on SARS-CoV-2 and contains SARS-CoV-2-related data, tools, and analysis. CoV-AbDab provides data on COVID-19 antibodies. The International Nucleotide Sequence Database Collaboration (INSDC) has released a public statement entitled 'INSDC Statement on SARS-CoV-2 sequence data sharing during COVID-19', which highlights the importance of sharing SARS-CoV-2 sequence data within the international scientific community. The INSDC recommends that all researchers working with SARS-CoV-2 sequence data submit both their raw and consensus—or assembled—SARS-CoV-2 data to the INSDC databases, which are freely available to the scientific community. COVID-19 data portal EMBL-EBI (https://www.covid19dataportal.org/) enables data sharing throughout the globe. The initiative facilitates international collaboration to accelerate scientific discovery, monitor the

pandemic, and help develop treatments and a vaccine for the new coronavirus. Other SARS-CoV-2 resources can be accessed by NCBI at https://www.ncbi.nlm.nih.gov/sars-cov-2/.

6.2.1.4 Other Approaches

It should be noted that previous virus research did not employ many of the above discussed tools. Ebola and Zika viral research were the first application of bioinformatic research in the viral disease field, and COVID-19 marks the introduction of drug discovery by using in silico approaches as constructive options. All the current bioinformatic tools have helped scientist in accelerating drug/vaccine development for SARS-CoV-2. Since the advance of bioinformatics in 1990, many tools exist that are used in addition to genome sequencing and molecular modelling. Softwares (both online and offline) are being extensively utilized for sequence analysis, complete genome sequencing, expressed sequence tags, identification of unknown genes, discovery of splice variants, causes of differences between viruses, pharmacogenetics, next-generation sequencing, etc. Multiple alignment tools to perform alignment of experimentally obtained genetic sequences are important for sequence comparisons and sequence-based database searches and additionally help in phylogenetic analysis. This is evident from the use of RNA sequencing from broncho-alveolar fluid samples of SARS-CoV-2 patients to identify its origins. The phylogenetic analysis revealed almost 90% similarity of virus sequences to betacoronaviruses from bat [35]. BLAST or Basic Local Alignment Search Tool has also been widely used to compare sequences whether nucleic acid or amino acid. BLAST uses the already available information on biological sequences from organisms to understand the genetic relationship with other species. For example, SARS-CoV-2 genome sequence similarity with viral metagenomes in pangolin has been seen. In particular, the availability of these bioinformatic approaches was very important in the discovery of newer drugs for COVID-19 as an understanding of genetic sequence needed to follow up by the protein analysis, i.e. function of the gene. This was made possible by comparing related sequences and thus similar functionality. Once the protein function could be determined (coupled to experimental evidence), the already known drugs against related targets could be taken for drug repurposing. In addition, drugs could be developed using computational approaches once this basic understanding is obtained [29].

In addition, due to the rapid research in the field of COVID-19, a comprehensive repository of knowledge about SARS-CoV-2, its proteins, mechanism of infection, and more has been created online. Vaccine design is also utilizing computational methods to design multiepitopes against SARS-CoV-2. The vaccine design includes prediction of potential epitopes from antigenic protein sequences and construction of vaccine, followed by molecular docking simulation to assess the binding affinity to the protein. For example, antigenic epitopes from spike glycoprotein, nucleocapsid, ORF3a, and non-structural proteins have been attempted (reviewed in [24]). Vaccine design has benefitted immensely by the constructed structural genomic and interactomic road maps that describe viral infection molecular mechanisms. An example is the X-ray crystal structure of RBD in complex with human antibody

CR3022 (6 W41) for epitope engineering. T- and B-cell epitopes have been identified too through informatics (details in [23]). Development of SARS-CoV models using ECFP6 descriptors and the Bayesian algorithm has been attempted to develop an assay control software (Ekins 2020). The method is fast, does not require crystal structures, and enables scoring of small-molecule structures against many models simultaneously. System pharmacology approaches have also provided important insights into promising antiviral drugs against SARS-CoV-2 based on pathogenesis mechanism and host specificity. Novel algorithms are being used in COVID-19 research. These are mainly network-based algorithms and expression-based algorithms. Network-based algorithms generate networks, for example, drug target network or human protein interactome, and using these networks, putative drug candidates are identified. One such study generated HCoV-host interactome and integrated various drugs specific for targets [36]. Some expression-based algorithms have also aided drug repurposing by linking ACE2 to SARS-CoV-2 in the early days and thus providing two potential repurposed drugs that change ACE2 expression. Functional analysis of genomes is done in parallel to identify the cellular functions of gene products coupled to transcriptomics, proteomics, metabolomics, phenomics, and even systems biology. See reviews [23, 24] for details. Table 6.4 provides a few examples of applications of all the above-mentioned approaches in COVID-19.

6.3 Conclusion

There are many lessons that can be learnt from the application of in silico approaches in the viral pandemic of COVID-19. On the positive side, molecular modelling provides the advantage of reduced cost for faster development of a drug candidate. This is propelled by databases and collaborative efforts of scientists. Scientists can thus directly compare the various drug candidates in these databases and assess for further development.

This was the third instance of application of these techniques, and therefore, important lesson would be to rely more on high-resolution crystal structures rather than homology models. Otherwise, it will be very difficult to process the huge repository of docking results generated every time. Also, database generation should become a priority to increase the collaborative capacity and, therefore, data validation by scientists throughout the world. In addition, if the molecular docking is performed on commercially available drugs and herbal chemicals, it would help us limit the number of docked molecules yet keeping the feasibility of taking the drug candidate to the next stage of drug development. Molecular docking of chemicals that are not readily available leads to extra steps of extraction or synthesis that could be avoided. It is thus understandable that a collaborative approach where scientists team up to work on different aspects such as bioinformatics, chemical synthesis, in vitro testing, and in vivo testing will be fruitful. Interestingly, the comprehensive information about virus pathogenesis in the human body is still not available and thus hinders progress in drug development.

Table 6.4 Examples of bioinformatic approaches used in COVID-19

Bioinformatic approach used in COVID-19 research	Reference examples and details
Whole genome sequencing	[37]—RT-PCR, next-generation sequencing on specimens coupled to characterization at molecular level, phylogenetic analysis, and B- and T-cell epitope prediction for the sequences of Indian SARS-CoV-2. Complete (29,851 nucleotides) genomes with 99.98% identity with Wuhan seafood market pneumonia virus (accession number: NC 045512) were obtained. In the receptor binding domain, prediction of linear B-cell epitopes of spike protein of S1 domain and a conformational epitope was identified. Results confirmed two different introductions into the country
Sequence comparisons	[18]—The SARS-CoV-2 genome reported from 13 different countries was investigated to identify mutations. High identity (>99%) was seen amongst the 13 genomes. By sequence comparison, country-specific distinctive mutations in the vital proteins of SARS-CoV-2 were identified (vital proteins, i.e. replicase polyprotein, spike glycoprotein, envelope protein, and nucleocapsid protein). Mutational effects on function were investigated using various in silico approaches
Algorithms	[38]—Development of simple algorithm to help the early detection in the patients who are infected by SARS-CoV-2. Identification of individual predictors of COVID-19 development and establishment of prediction model with a learning set of 322 COVID-19 patients (training datasheet/set) and a set of 317 COVID-19 patients were validated. The analysis identified age, lactate dehydrogenase, and CD4 count as good predictors of COVID-19 progression giving an algorithm of (age × LDH)/CD4 count
Machine learning	[39]—These models were built for the prediction of protein-protein interactions between the viral proteins and proteins present in humans. A total of 1326 potential human target proteins of SARS-CoV-2 were predicted by the proposed ensemble model and validated using gene ontology and KEGG pathway enrichment analysis
Systems biology	[40]—166 modified herbal Chinese formulae and 1212 phytoconstituents that were containing b-sitosterol, stigmasterol, and quercetin were chosen. Using complex system entropy and unsupervised hierarchical clustering, 18 herbal formulae showed promise as COVID-19 candidates. 12 clusters of molecules yielded 8 pharmacophore families of structures following scaffold analysis, self-organizing mapping, and cluster analysis
Drug repurposing	[41]—The amalgamation of drugs, viz. pirfenidone and melatonin, was identified for COVID-19 by using system biology and artificial intelligence-based approaches. GUILDify v2.0 web server was also used to confirm the results

(continued)

Table 6.4 (continued)

Bioinformatic approach used in COVID-19 research	Reference examples and details
Vaccine development	[42]—Characterization of spike glycoprotein to obtain immunogenic epitopes using immunoinformatic approach. 13 major histocompatibility complex-(MHC) I and 3 MHC-II epitopes having antigenic properties were chosen as they were linked to specific linkers to build vaccine components and molecularly docked on Toll-like receptor-5 to get binding affinity

In conclusion, due to the heavy inundation with incomplete or unvalidated molecular modelling studies, it is warranted that proper steps are taken to manage pseudoscience propagation.

6.4 Future Outlook

Application of bioinformatic approaches in Ebola, Zika, and COVID-19 pandemic has shown that these approaches can be used for planning and developing antiviral drugs and support pandemic situations when properly streamlined. Target identification, interaction mapping, understanding structure activity relationships, and molecular docking and dynamics could provide important tools for elucidating underlying pathogenesis and its targeting by novel drug candidates. When coupled to proper experimental testing, the drug development will have far-reaching results. We are able to increase the safety of the drug molecules, i.e. by understanding the physical-chemical properties as well as probability of success. Time must be spent on elucidating underlying mechanisms so as to provide relief to clinical symptoms of COVID-19.

Acknowledgements GK conceptualized and wrote the chapter. SM and SJ prepared the Tables and Fig. 6.1. This work is part of the project MCB200120 funded through COVID-19 HPC Consortium.

References

1. Ekins S, Mottin M, Ramos P, Sousa BKP, Neves BJ, Foil DH et al (2020) Deja vu: stimulating open drug discovery for SARS-CoV-2. Drug Discov Today 25(5):928–941
2. Brown CS, Lee MS, Leung DW, Wang T, Xu W, Luthra P et al (2014) In silico derived small molecules bind the filovirus VP35 protein and inhibit its polymerase cofactor activity. J Mol Biol 426(10):2045–2058
3. Ekins S, Freundlich JS, Coffee M (2014) A common feature pharmacophore for FDA-approved drugs inhibiting the Ebola virus. F1000Res 3:277
4. Litterman N, Lipinski C, Ekins S (2015) Small molecules with antiviral activity against the Ebola virus. F1000Res 4:38

5. Veljkovic V, Loiseau PM, Figadere B, Glisic S, Veljkovic N, Perovic VR et al (2015) Virtual screen for repurposing approved and experimental drugs for candidate inhibitors of EBOLA virus infection. F1000Res 4:34
6. Ekins S, Freundlich JS, Clark AM, Anantpadma M, Davey RA, Madrid P (2015) Machine learning models identify molecules active against the Ebola virus in vitro. F1000Res 4:1091
7. Ekins S, Perryman AL, Horta AC (2016) OpenZika: an IBM World Community Grid project to accelerate Zika virus drug discovery. PLoS Negl Trop Dis 10(10):e0005023
8. Ekins S, Liebler J, Neves BJ, Lewis WG, Coffee M, Bienstock R et al (2016) Illustrating and homology modeling the proteins of the Zika virus. F1000Res 5:275
9. Sahoo M, Jena L, Daf S, Kumar S (2016) Virtual screening for potential inhibitors of NS3 protein of Zika virus. Genomics Inform 14(3):104–111
10. Mottin M, Braga RC, da Silva RA, Silva J, Perryman AL, Ekins S et al (2017) Molecular dynamics simulations of Zika virus NS3 helicase: insights into RNA binding site activity. Biochem Biophys Res Commun 492(4):643–651
11. Sharma N, Murali A, Singh SK, Giri R (2017) Epigallocatechin gallate, an active green tea compound inhibits the Zika virus entry into host cells via binding the envelope protein. Int J Biol Macromol 104(Pt A):1046–1054
12. Ramharack P, Soliman MES (2018) Zika virus NS5 protein potential inhibitors: an enhanced in silico approach in drug discovery. J Biomol Struct Dyn 36(5):1118–1133
13. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A et al (2020) World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). Int J Surg 76:71–76
14. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ et al (2020) The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. Mil Med Res 7(1):11
15. Ralph R, Lew J, Zeng T, Francis M, Xue B, Roux M et al (2020) 2019-nCoV (Wuhan virus), a novel coronavirus: human-to-human transmission, travel-related cases, and vaccine readiness. J Infect Dev Ctries 14(1):3–17
16. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579(7798):270–273
17. Forster P, Forster L, Renfrew C, Forster M (2020) Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A 117(17):9241–9243
18. Khan MI, Khan ZA, Baig MH, Ahmad I, Farouk AE, Song YG et al (2020) Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: an in silico insight. PLoS One 15(9): e0238344
19. Ugurel OM, Mutlu O, Sariyer E, Kocer S, Ugurel E, Inci TG et al (2020) Evaluation of the potency of FDA-approved drugs on wild type and mutant SARS-CoV-2 helicase (Nsp13). Int J Biol Macromol 163:1687–1696
20. Azad GK (2020) Identification of novel mutations in the methyltransferase complex (Nsp10-Nsp16) of SARS-CoV-2. Biochem Biophys Rep 24:100833
21. Hassan SS, Choudhury PP, Roy B (2020) SARS-CoV2 envelope protein: non-synonymous mutations and its consequences. Genomics
22. Tung HYL, Limtung P (2020) Mutations in the phosphorylation sites of SARS-CoV-2 encoded nucleocapsid protein and structure model of sequestration by protein 14-3-3. Biochem Biophys Res Commun 532(1):134–138
23. Chellapandi P, Saranya S (2020) Genomics insights of SARS-CoV-2 (COVID-19) into target-based drug discovery. Med Chem Res:1–15
24. Wang X, Guan Y (2020) COVID-19 drug repurposing: a review of computational screening methods, clinical trials, and protein interaction assays. Med Res Rev
25. Rolta R, Yadav R, Salaria D, Trivedi S, Imran M, Sourirajan A et al (2020) In silico screening of hundred phytocompounds of ten medicinal plants as potential inhibitors of nucleocapsid

phosphoprotein of COVID-19: an approach to prevent virus assembly. J Biomol Struct Dyn:1–18
26. Costa AN, de Sa ERA, Bezerra RDS, Souza JL, Lima F (2020) Constituents of buriti oil (*Mauritia flexuosa* L.) like inhibitors of the SARS-Coronavirus main peptidase: an investigation by docking and molecular dynamics. J Biomol Struct Dyn, 1–8
27. Krupanidhi S, Abraham Peele K, Venkateswarulu TC, Ayyagari VS, Nazneen Bobby M, John Babu D et al (2020) Screening of phytochemical compounds of Tinospora cordifolia for their inhibitory activity on SARS-CoV-2: an in silico study. J Biomol Struct Dyn:1–5
28. Ahmad S, Abbasi HW, Shahid S, Gul S, Abbasi SW (2020) Molecular docking, simulation and MM-PBSA studies of *nigella sativa* compounds: a computational quest to identify potential natural antiviral for COVID-19 treatment. J Biomol Struct Dyn:1–9
29. Villas-Boas GR, Rescia VC, Paes MM, Lavorato SN, Magalhaes-Filho MF, Cunha MS et al (2020) The NEW Coronavirus (SARS-CoV-2): a comprehensive review on immunity and the application of bioinformatics and molecular modeling to the discovery of potential anti-SARS-CoV-2 agents. Molecules 25(18):4086
30. Lim J, Jeon S, Shin HY, Kim MJ, Seong YM, Lee WJ et al (2020) Case of the index patient who caused tertiary transmission of COVID-19 infection in Korea: the application of Lopinavir/ritonavir for the treatment of COVID-19 infected pneumonia monitored by quantitative RT-PCR. J Korean Med Sci 35(6):e79
31. Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK et al (2020) The 2019 novel coronavirus resource. Yi Chuan 42(2):212–221
32. Cleemput S, Dumon W, Fonseca V, Abdool Karim W, Giovanetti M, Alcantara LC et al (2020) Genome detective coronavirus typing tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics 36(11):3552–3555
33. Gupta AK, Khan MS, Choudhury S, Mukhopadhyay A, Sakshi, Rastogi A et al (2020) CoronaVR: a computational resource and analysis of epitopes and therapeutics for severe acute respiratory syndrome coronavirus-2. Front Microbiol 11:1858
34. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM (2020) CoV-AbDab: the coronavirus antibody database. Bioinformatics
35. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y et al (2020) Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B 10(5):766–788
36. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 6:14
37. Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM et al (2020) Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res 151(2 & 3):200–209
38. Li Q, Zhang J, Ling Y, Li W, Zhang X, Lu H et al (2020) A simple algorithm helps early identification of SARS-CoV-2 infection patients with severe progression tendency. Infection 48 (4):577–584
39. Dey L, Chakraborty S, Mukhopadhyay A (2020) Machine learning techniques for sequence-based prediction of viral-host interactions between SARS-CoV-2 and human proteins. Biom J
40. Luo L, Jiang J, Wang C, Fitzgerald M, Hu W, Zhou Y et al (2020) Analysis on herbal medicines utilized for treatment of COVID-19. Acta Pharm Sin B 10(7):1192–1204
41. Artigas L, Coma M, Matos-Filipe P, Aguirre-Plans J, Farres J, Valls R et al (2020) In-silico drug repurposing study predicts the combination of pirfenidone and melatonin as a promising candidate therapy to reduce SARS-CoV-2 infection progression and respiratory distress caused by cytokine storm. PLoS One 15(10):e0240149
42. Bhattacharya M, Sharma AR, Patra P, Ghosh P, Sharma G, Patra BC et al (2020) Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): Immunoinformatics approach. J Med Virol 92(6):618–631