Advances in Experimental Medicine and Biology 1368

Bairong Shen *Editor*

Translational Informatics

Prevention and Treatment of Viral Infections

Springer

Advances in Experimental Medicine and Biology

Volume 1368

Advances in Experimental Medicine and Biology provides a platform for scientific contributions in the main disciplines of the biomedicine and the life sciences. This series publishes thematic volumes on contemporary research in the areas of microbiology, immunology, neurosciences, biochemistry, biomedical engineering, genetics, physiology, and cancer research. Covering emerging topics and techniques in basic and clinical science, it brings together clinicians and researchers from various fields.

Advances in Experimental Medicine and Biology has been publishing exceptional works in the field for over 40 years, and is indexed in SCOPUS, Medline (PubMed), Journal Citation Reports/Science Edition, Science Citation Index Expanded (SciSearch, Web of Science), EMBASE, BIOSIS, Reaxys, EMBiology, the Chemical Abstracts Service (CAS), and Pathway Studio.

2020 Impact Factor: 2.622

Bairong Shen
Editor

Translational Informatics

Prevention and Treatment of Viral Infections

Springer

Editor
Bairong Shen
Institutes for Systems Genetics,
Frontiers Science Center for
Disease-Related Molecular Network,
West China Hospital, Sichuan University,
Chengdu, Sichuan, China

ISSN 0065-2598 ISSN 2214-8019 (electronic)
Advances in Experimental Medicine and Biology
ISBN 978-981-16-8968-0 ISBN 978-981-16-8969-7 (eBook)
https://doi.org/10.1007/978-981-16-8969-7

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

About the Editor

Bairong Shen is professor and executive director general of the Institutes for Systems Genetics, West China Hospital, Sichuan University. He received his PhD in chemistry from Fudan University in 1997. Dr. Shen was appointed Associate Professor of Physical Chemistry at Fudan University in 1999 for his accomplishments in theoretical and computational surface chemistry. In the early 2000s, Dr. Shen started his new exploration into biomedical informatics and related computational biology in his postdoctoral research at the University of Tampere, Finland. His success in the new paradigm of biological research won him a competitive faculty position in the European university as an Assistant/Associate Professor of Bioinformatics since 2004. He joined the Soochow University by founding the University's Center for Systems Biology in 2008. In Finland and China, Dr. Shen has taught more than 10 different courses in biomedical informatics and systems biology and published more than 100 peer-reviewed articles in competitive journals which covered the medical genetic areas including cancer biomarker discovery, biomedical informatics and the basic exploration in physics, chemistry, biology, and computational science. His recent researches focus on biomedical informatics and systems biology of complex diseases and healthcare.

Chapter 1
Databases, Knowledgebases, and Software Tools for Virus Informatics

Yuxin Lin, Yulan Qian, Xin Qi, and Bairong Shen

Abstract Virus infection is a common social health issue. In the past decades, serious virus infectious events have caused great loss in people's life and the economics. The nature of rapid widespread and frequent variation increases the difficulty for precision viral prevention and treatment. In the era of big data and artificial intelligence (AI), advances in bioinformatics techniques bring unprecedented opportunities for virus informatics study, which contribute to the systems-level modeling of virus biology. In this chapter, data resources including virus-related databases and knowledgebases are introduced. Bioinformatics models and software tools for multiple sequence alignment, evolutionary analysis, and genome-wide research of viruses are summarized and emphasized. Translational applications of recently developed data-driven and AI-assisted methods to viral cases such as SARS-CoV-2, HBV/HCV, and influenza virus are discussed. Finally, the concept and significance of virus informatics are highlighted for both virus surveillance and health promotion.

Yuxin Lin and Yulan Qian contributed equally with all other contributors.

Y. Lin
Department of Urology, The First Affiliated Hospital of Soochow University, Suzhou, China
e-mail: linyuxin@suda.edu.cn

Y. Qian
Department of Pharmacy, The First Affiliated Hospital of Soochow University, Suzhou, China

X. Qi
School of Chemistry and Life Sciences, Suzhou University of Science and Technology, Suzhou, China
e-mail: qixin@usts.edu.cn

B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_1

Keywords Virus informatics · Databases and knowledgebases · Bioinformatics tools · Precision medicine · Systems health

1.1 Introduction

Viruses are composed of a core of nucleic acid (either DNA or RNA, but never both) surrounded by a protein shell or coat. They are extremely small and may have a variety of structural shapes. Due to the lack of independent metabolic and enzyme systems, viruses are not capable to survive on their own, and they must live in the host cells to get enough materials and energy to help continue their biological activities, including replication, transcription, and translation. When a virus enters the host cell, it could produce a new generation based on the genetic information contained in its nucleic acid and affect the function of the host cell [1, 2].

In the past few decades, a spate of virus infection events such as SARS (severe acute respiratory syndromes), MERS (Middle East respiratory syndrome), and COVID-19 (coronavirus disease 2019) has occurred, which caused great loss in people's life and the social security. Currently, people still face the dilemma that the virus is rapidly evolving to be more adaptive to the changeable environments and will affect public health for a long time. To fight against it, much experimental and clinical effort has been put into the prediction and prevention of virus infection, meanwhile novel therapeutic regimens are proposed for the precision and personalized treatment of patients with virus-induced diseases [3, 4].

In the era of big biomedical data and artificial intelligence (AI), the innovation in informatics techniques provides unprecedented opportunities for holistic monitoring of virus–host interactions [5], and it gradually promotes the transition of virus research from experiment-oriented biological discovery to knowledge-guided systems modeling and validation [6]. With the accumulation of data resources and the development of bioinformatics algorithms, a number of computational tools and platforms are available in these years for translational virus studies including computer-aided drug and vaccine design, infection severity prediction, population-level infection prevention, and global health management [7, 8].

In this chapter, several well-established databases and knowledgebases related to viral sequence, structure, and interactions between virus and the host are first introduced for computational virus research. Then the state-of-the-art bioinformatics methods and software tools are listed based on their purposes for translational applications, e.g., multiple sequence alignment, phylogenetic and evolutionary understanding, and genome-wide exploration. According to the current social situation and hotspot, case studies associated with the control of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), HBV/HCV (hepatitis B/C virus), and influenza virus infection are emphasized and discussed. A new concept, i.e., virus informatics, is finally proposed and highlighted for systems-level virus surveillance and health promotion.

1.2 Databases and Knowledgebases for Computational Virus Research

1.2.1 Data Resources for Multiple Virus Groups

There have been many databases and online platforms developed for comprehensive analysis of multiple virus groups. As illustrated in Table 1.1, ViPR (The Virus Pathogen Database and Analysis Resource) is an advanced database for virology research [9]. It covers basic information including sequences, gene annotations, protein structures, and other clinical metadata for more than ten families of human pathogenic viruses. In addition to the browsing of virus records, the database provides powerful tools for statistical analysis, multiple sequence alignment, and phylogenetic tree construction [9]. Virus Variation Resource is another database belonging to NCBI (National Center for Biotechnology Information) for seven viral types, including virus of influenza, Dengue, West Nile, Ebola, MERS, Rotavirus A, and Zika. The functions such as sequence query, genetic annotation, and multiple sequence alignment are also available for downstream analysis [10].

Since the sequence information is essential for virus study, ViruSurf and RVDB (Reference Viral Database) are two well-constructed databases for viral sequences collected from heterogeneous sources [11, 12]. In particular, ViruSurf mainly focuses on recent outbreak viruses such as SARS-CoV, MERS-CoV, Ebola, and SARS-CoV-2 [11], whereas RVDB is a reference database for all viral, virus-related, and virus-like nucleotide sequences from eukaryotes [12]. ViMIC (Virus Mutations, Integration sites and Cis-effects) is a recently developed database of virus mutations, integration sites, and cis-effects for human diseases. It contains comprehensive information associated with virus mutation entries, viral integration sites, targeted genes, and sequence data of eight viruses in more than 70 diseases [13]. Apart from the sequence information, protein–protein interactions (PPIs) between virus and the host are of paramount significance for decoding the molecular mechanisms and helping develop precision strategies for virus intervention. Based on this goal, databases including HVIDB (Human-Virus Interaction Database) and Viruses.STRING are proposed to provide the opportunity for predicting and investigating complex PPIs for comprehensive virus–host pathogenetic understanding [14, 15].

In recent years, posttranslational modifications and circular RNA (circRNA) regulations are found to be important in virus activities, and the dysfunction may be a potential clue for human diseases. VPTMdb (Viral Posttranslational Modification Database) is the first database with posttranslational modification data in viruses and infected host cells [16]. VirusCircBase (Virus CircRNA Database) contains more than 11,000 circRNAs related to 23 viral species for the analysis of viral circRNAs in the context of human diseases and public health [17]. Besides, considering the integration of viruses into human genome is a key driven signature for disease development, VISDB (Viral Integration Site Database) is a manually curated

Table 1.1 Publicly available data resources for virus research

Title	Description	Citation
Data resources for multiple virus groups		
ViPR	• A virus pathogen database and analysis platform for virology research • URL: https://www.viprbrc.org/	[9]
Virus Variation Resource	• A data resource in NCBI for study of seven viral groups • URL: http://www.ncbi.nlm.nih.gov/genome/viruses/variation/	[10]
ViruSurf	• An integrated database for viral sequence study • URL: http://gmql.eu/virusurf/	[11]
RVDB	• A reference viral database with sequence data of different virus species from eukaryotes • URL: https://github.com/ArifaKhanLab/RVDB/	[12]
ViMIC	• A comprehensive database of virus mutations, integration sites, and cis-effects for human diseases • URL: http://bmtongji.cn/ViMIC/	[13]
HVIDB	• A comprehensive database for human–virus protein–protein interaction understanding • URL: http://zzdlab.com/hvidb/	[14]
Viruses.STRING	• A comprehensive database of virus–host protein–protein interactions • URL: http://apps.cytoscape.org/apps/stringapp/	[15]
VPTMdb	• A database for posttranslational modification of viruses • URL: http://vptmdb.com:8787/VPTMdb/	[16]
VirusCircBase	• A database of circular RNAs for virus • URL: http://www.computationalbiology.cn/ViruscircBase/	[17]
VISDB	• A database of reported integration sites of virus in the human genome • URL: https://bioinfo.uth.edu/VISDB/	[18]
Virus type-specific data platforms		
2019nCoVR	• The 2019 novel coronavirus resource for data sharing and analysis of COVID-19 • URL: https://bigd.big.ac.cn/ncov/	[19]
Virus-CKB	• An integrated knowledge base and online platform for drug discovery against COVID-19 • URL: https://www.cbligand.org/g/virus-ckb/	[20]
DockCoV2	• A database for potential drug discovery against SARS-CoV-2 • URL: https://covirus.cc/drugs/	[21]
CoV3D	• A database for coronavirus protein structures with high resolution for SARS-CoV-2 study • URL: https://cov3d.ibbr.umd.edu/	[22]
OxCOVID19	• A database of time–series data for COVID-19 pandemic • URL: https://covid19.eng.ox.ac.uk/	[23]
GISAID	• An initiative and program for all influenza data sharing • URL: N/A	[24]
IRD	• An influenza research database with genome sequence data and associated characteristic information of influenza virus • URL: http://www.fludb.org/	[25]

(continued)

Table 1.1 (continued)

Title	Description	Citation
FluReassort	• A database for genomic reassortment study of influenza viruses • URL: https://www.jianglab.tech/FluReassort/	[26]
BioHealthBase	• A database for study and analysis of interactions between influenza virus and host pathogen • URL: http://www.biohealthbase.org	[27]
Ebolabase	• A database with interactive data between Zaire Ebola virus and human proteins for drug repurposing • URL: http://ebola.bicpu.edu.in/	[28]
OvirusTdb	• A comprehensive database of oncolytic viruses for cancer therapeutics • URL: https://webs.iiitd.edu.in/raghava/ovirustdb/	[29]
HDVdb	• A comprehensive database for hepatitis D virus • URL: http://hdvdb.bio.wzw.tum.de/	[30]
HRRD	• A manually curated database for regulatory data between HPV and host RNA • URL: www.hmuhrrd.com/HRRD/	[31]

database integrating reported integration sites of viruses in the human genome, and it is a useful resource for virus–disease association study [18].

1.2.2 Virus Type-Specific Data Platforms

As described in Table 1.1, several databases and knowledgebases are designed specifically to given virus types. Due to the widespread infection of SARS-CoV-2 around the world, the newly developed 2019nCoVR (2019 Novel Coronavirus Resource), Virus-CKB (Viral-associated Disease-specific Chemogenomics Knowledgebase), DockCoV2 (Drug Database for SARS-CoV-2), CoV3D (Coronavirus Protein Structures), and OxCOVID19 (Oxford COVID-19 Database) are used for the sharing and analysis of SARS-CoV-2 data [19–23]. Among them, 2019nCoVR is a comprehensive resource for 2019 novel coronavirus. It provides a wide range of information related to SARS-CoV-2 from published literatures, news, and scientific papers and integrates functionalities for analysis and visualization of genome variation using the collected SARS-CoV-2 chains for genome research, drug development, and precision therapy of infection [19]. Virus-CKB is an integrated online platform and knowledgebase for drug discovery against SARS-CoV-2. It contains virus-related disease-specific chemogenomics data, which would help the computer-aided pharmacology target mapping to predict drugs for SARS-CoV-2 treatment [20]. Similarly, DockCoV2 is also a useful database, aiming at speeding up the potential drug discovery against SARS-CoV-2. Currently, a total of 3109 drugs are included in the database, and the data can be easily searched and downloaded for scientific use [21]. CoV3D is a data resource for coronavirus protein structures, and it will assist structure-based vaccine design for SARS-CoV-2 prevention [22]. Compared with these studies, OxCOVID19 is a database with dynamic and time–series data for the better understanding of the impact on COVID-19 pandemic [23].

Besides SARS-CoV-2, there are some publicly available databases for virus-level analysis of influenza, cancer, and other complex diseases. For example, GISAID (Global Initiative of Sharing All Influenza Data), IRD (Influenza Research Database), FluReassort (Influenza Virus Reassortment), and BioHealthBase are databases for sharing all influenza data, investigating the genome sequence, genomic reassortment, and influenza virus–host interactions, respectively [24–27]. Among them, GISAID is one of the famous databases initiated by many top scientists and Nobel Prize winners. It focuses on promoting the international sharing of all influenza virus sequences, relevant clinical and epidemiological data of human viruses, and geographic and species-specific data related to poultry and other animal viruses. The purpose of GISAID is to help researchers understand how viruses evolve, spread, and even become potentially major epidemic diseases [24]. Ebolabase (Ebola Virus Database) is an Ebola virus-specific database with interactive data between Zaire Ebola virus and human proteins [28]. For virus-mediated cancer studies, OvirusTdb (Repository of Oncolytic Viruses) is a manually curated database with comprehensive information of oncolytic viruses for cancer therapeutics [29]. HDVdb (Hepatitis D Virus Database) and HRRD (HPV-RNA Relationship Database), respectively, are developed for the prevention of HDV (hepatitis D virus) and HPV (human papilloma virus) infection. Among them, the collection of HDV genomic sequences and associated variability signatures in HDVdb would increase the discovery of effective drugs or vaccines against HDV infection [30] and hepatocellular carcinoma development. In HRRD database, the regulatory relationship between HPV and different types of RNAs (e.g., messenger RNAs, microRNAs, long non-coding RNAs) could improve the understanding of HPV-induced carcinogenesis and prognosis [31].

1.3 Bioinformatics Models and Tools for Translational Virus Analysis

1.3.1 Multiple Sequence Alignment

Multiple sequence alignment is an essential step for virus genome sequence analysis. Currently, a large number of computational tools have been proposed and improved for multiple sequence alignment based on different principles such as progressive, consistent, or evolutionary theories [7]. As shown in Table 1.2, MUSCLE (multiple sequence comparison by log-expectation) is a powerful software program for performing multiple alignments of protein sequences with high accuracy and throughput based on a newly defined log-expectation score. Compared with the existed approaches, it consumes a shorter calculation time and achieves the highest

Table 1.2 Bioinformatics tools and programs for virus analysis

Title	Description	Citation
Multiple sequence alignment		
MUSCLE	• A high accuracy tool for multiple sequence alignment • URL: http://drive5.com/muscle/	[32]
ProbCons	• Multiple sequence alignment based on probabilistic consistency analysis • URL: http://probcons.stanford.edu/	[33]
StatAlign	• An extendable software package for multiple sequence alignment based on Bayesian theory • URL: https://dl.acm.org/doi/10.1093/bioinformatics/btn457/	[34]
JABAWS:MSA	• A comprehensive tool integrating different algorithms for multiple sequence alignment • URL: http://www.compbio.dundee.ac.uk/jabaws/	[35]
MSAViewer	• A quick and easy visualization tool for multiple sequence alignment data • URL: http://msa.biojs.net/index.html/	[36]
Phylogenetic and evolutionary understanding		
BIONJ	• An improved method of the neighbor-joining algorithm for phylogenetic tree analysis • URL: http://www.atgc-montpellier.fr/bionj/	[37]
IQ-TREE	• A stochastic algorithm for the estimation of maximum-likelihood phylogenies • URL: http://www.cibiv.at/software/iqtree/	[38]
PhyloBayes3	• A program for phylogenetic reconstruction or molecular dating based on Bayesian analyses • URL: http://www.atgc-montpellier.fr/phylobayes/	[39]
BEAST2	• An improved platform for Bayesian evolutionary analysis • URL: http://www.beast2.org/	[40]
PhyloSuite	• An integrated platform for evolutionary phylogenetics studies • URL: http://phylosuite.jushengwu.com/	[41]
MEGA5	• An improved version for molecular evolutionary genetics analysis • URL: http://www.megasoftware.net/	[42]
Genome-wide exploration		
gff2ps	• A bioinformatics tool for visualizing annotations of genomic sequences • URL: http://genome.imim.es/software/gfftools/GFF2PS.html/	[43]
IBS	• An illustrator of biological sequences for organization representation • URL: http://ibs.biocuckoo.org/	[44]
VirION2	• An improved framework with sequencing and informatics workflow for the study of virus genomic diversity • URL: https://dx.doi.org/10.17504/protocols.io.6q9hdz6/	[45]
VVV	• A viral variant visualizer for the visualization of viral genetic diversity • URL: https://github.com/ALFLAG/Viral_Variant_Visualiser/	[46]

(continued)

Table 1.2 (continued)

Title	Description	Citation
RDP4	• A recombination detection program for the analysis of recombination patterns in virus genomes • URL: http://web.cbio.uct.ac.za/~darren/rdp.html/	[47]
ITN—VIROINF	• Linking virology and bioinformatics to understand interactions between virus and host • URL: https://viroinf.eu/	[5]
DAMIAN	• A computational tool for cohort-based study of microorganisms • URL: https://sourceforge.net/projects/damian-pd/	[48]
INSaFLU	• An automated online bioinformatics tool for surveillance of influenza virus • URL: https://insaflu.insa.pt/	[49]

result accuracy [32]. ProbCons (Probabilistic Consistency) is a famous probabilistic consistency-based method for multiple sequence alignment and comparison [33]. Based on Bayesian analysis, StatAlign is an extendable tool with user-friendly interfaces for joint Bayesian estimation of multiple sequence alignments and evolutionary trees [34]. In addition, integrated resources and visualization tools are popular for practical use. For example, the JABAWS framework integrates five multiple sequence alignment approaches and provides web services for bioinformatics analysis [35]. MSAViewer is a quick and easy JavaScript component for visualization of alignment data [36].

1.3.2 Phylogenetic and Evolutionary Understanding

The construction of phylogenetic tree is an effective way to understand the origin and evolution of viruses. At present, there have been a lot of algorithms developed for estimating and measuring phylogenetic trees, including neighbor joining, maximum parsimony, maximum likelihood, and Bayesian inference [7]. As shown in Table 1.2, BIONJ improves the traditional neighbor-joining algorithm by a simple model of sequence data and achieves overall better performance for phylogenetic analysis [37]. For maximum-likelihood phylogenies, fast tree inference methods are needed due to the large size of phylogenomics data. Hence, the IQ-TREE is proposed, and it gets high computation efficiency and accuracy for the estimation of maximum-likelihood phylogenies [38]. PhyloBayes3 and BEAST2 are two well-established programs both constructed from the theory of Bayesian inference [39, 40]. Among them, PhyloBayes3 is a powerful software package, using models of amino acid replacement and nucleotide substitution for phylogenetic reconstruction as well as molecular dating analysis [39]. BEAST2 is an improved platform of BEAST1. Compared with the old version, this new release integrates many recently published models and enhanced the abilities of data transmission for Bayesian evolutionary analysis [40].

There are also some integrated tools for phylogenetic studies. For example, PhyloSuite is a desktop program for sequence data management and evolutionary phylogenetic research [41]. It integrates a variety of phylogenetics-related bioinformatics tools and allows the streamline of analysis from basic data recognition to precise annotation of phylogenetic trees [41]. Based on the methods in evolutionary bioinformatics, MEGA5 (Molecular Evolutionary Genetics Analysis version 5) is proposed with newly added functions for evolutionary tree prediction, substitution model selection, sequence identification, and evolutionary rate estimation [42]. The software tool provides graphical interfaces for mining public datasets, conducting sequence alignment and building phylogenetic trees [42].

1.3.3 Genome-Wide Exploration

Genome-wide exploration of viruses, including genome annotation, detection of variation, and coronavirus recombination, are important for the analysis of virus function and the relationship between different viruses. As shown in Table 1.2, several computational programs can be used to annotate viral genetic sequences and visualize the genetic diversity of viruses. For example, gff2ps is an offline program that focuses on visualizing annotations of genomic sequences in files with General Feature Format (GFF). In a GFF file, each genomic sequence feature is shown in a single-line record, and the type of the feature on the genomic sequence is specified. Although many tools have been developed for reading GFF files, gff2ps tends to be popular due to its flexibility and high quality in data processing and result representation [43]. IBS (Illustrator of Biological Sequences) is another efficient tool for annotation and visualization of either nucleotide or protein sequences. Compared with gff2ps, IBS has both local and online versions for users to operate [44]. VirION2 and VVV (Viral Variant Visualizer) are powerful platforms for study and visualization of viral genetic diversity [45, 46]. Here, VirION2 is an improved tool integrating both short-/long-read sequencing and informatics techniques to investigate the genetic diversity of viruses [45], whereas VVV can be applied to analyze and visualize genetic variants of viruses from next generation sequencing data [46]. Since the recombination frequently occurs in viral genomes, the identification of significant recombination patterns is necessary be considered. Currently, the latest released RDP4 (recombination detection program version 4) holds the power for detecting and visualizing recombination events in genome sequence alignments of viruses, and it also equips with many novel functions for recombination analysis [47].

In addition, ITN-VIROINF implements important virological models to build a comprehensive computational platform linking virology and bioinformatics to help understand the interactions between viruses and their hosts [5]. DAMIAN (Detection & Analysis of viral and Microbial Infectious Agents by NGS) is a bioinformatics resource for the detection and cohort-based analysis of microorganisms including viruses. Besides the known sequence signatures, the tool allows screening novel pathogens [48]. Based on whole-genome sequencing data, INSaFLU is an automated online tool specific to influenza surveillance, e.g., genome sequence annotation, variant detection, phylogenetic tree analysis, and it will help the discovery of potential drugs and the decoding of key pathways associated with influenza evolution [49].

1.4 Data-Driven and AI-Assisted Studies for the Control of Viral Infection

Nowadays, data-driven and AI-assisted approaches make the procedure for system-level identification and analysis of changeable viral signatures more accurate. As summarized in Fig. 1.1, the approaches often start with a collection of omics and clinical data such as viral sequence, gene expression profiles, and interactive or regulatory associations among molecules at different levels. Then functional features will be selected for computational modeling using the methods of mathematics, biological networks, machine learning, and AI algorithms. Finally, key factors including candidate biomarkers, pathway signatures, targets for drug, and vaccine design could be identified for translational researches on viruses such as SARS-CoV-2, HBV/HCV, influenza virus, and other viruses for human diseases.

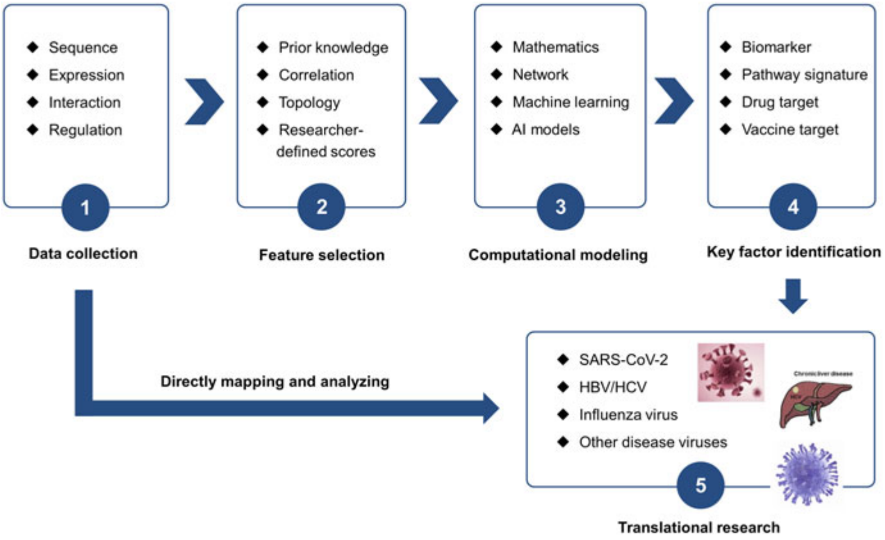

Fig. 1.1 The schematic pipeline for computer-aided virus research

1.4.1 SARS-CoV-2

Since the end of 2019, the infection of SARS-CoV-2 has seriously threatened the normal order of human life and has caused great damages in both social and economic fabrics. To fight against it, bioinformatics seems to be a powerful tool for virus studies in terms of genome sequence annotation, host recognition, candidate drug target identification, and computer-aided vaccine design.

Li et al. performed transcriptome profiling using lung and blood samples from patients infected with SARS-CoV-2 and investigated core gene expression signatures in the pathogenesis of pneumonia induced by SARS-CoV-2 infection [50]. Based on weighted gene correlation network analysis, two significant gene modules associated with clinical traits of COVID-19 patients were extracted, and the over-activation of cytokine release syndrome mediated by immune systems was identified as the potential mechanism in acute phase of SARS-CoV-2 infection [50]. Vastrad et al. downloaded high-throughput sequencing data from public database and identified key genes and significant pathways for COVID-19 diagnosis based on integrated network topology and functional enrichment analysis [51]. The result indicated that the abnormally expressed genes were mainly involved in viral transcription and immune-related signaling, and a panel of ten genes may be used as candidate diagnostic biomarkers and molecular targets for COVID-19 prediction and prevention [51]. Similar to this idea, Xie et al. identified differentially expressed genes from infected SARS-CoV-2 cell lines and constructed a PPI network for hub gene mining. The validation result showed that CXCL2, IL6, and CCL20 could serve as latent biomarkers in the prediction of SARS-CoV-2 infection [52].

Another outstanding advantage of bioinformatics is to help screen candidate targets for immune responses and vaccine design. For example, Grifoni et al. developed a novel approach integrating both sequence homology and bioinformatics methods for inferring potential immune targets of SARS-CoV-2. Using parallel bioinformatics predictions, a priori potential B- and T-cell epitopes were identified, and it would promote the vaccine design with high efficacy [53]. Since coronavirus nsp1 (non-structural protein 1) protein is an important player with versatile roles in virus–host interactions, Min et al. analyzed the characteristics of nsp1 in SARS-CoV-2 by bioinformatics and further explored its special function in manipulating translation of host mRNA [54]. In addition to nsp1, ACE2 (angiotensin-converting enzyme 2) is also a notable star in COVID-19 because it mediates the process of SARS-CoV-2 into human host cells. Barker et al. selected a series of bioinformatics methods to identify and compare the cell-specific expression of ACE2 among trachea, lung, and small intestine and found that the expression of ACE2 in different cell types was highly heterogeneous [55]. These results gave deep insights into the drug and vaccine design for future translational applications.

1.4.2 HBV/HCV

Accumulating evidences demonstrated that the infection of HBV and HCV are risk factors for the development of liver cancer. Bioinformatics techniques, especially network-based and AI-assisted models, are therefore developed and applied to discover key signatures for the prediction or prognosis of HBV/HCV-associated liver diseases [56].

Based on systems biology viewpoints, biological molecules including genes, RNAs, proteins, and metabolites may interact with each other, which contributes to the development of complex phenotypes. Hence, the construction and analysis of biological networks, e.g., PPI network, gene co-expression network, miRNA-mRNA regulatory network, competing endogenous RNA (ceRNA) network, would help the holistic discovery of key players in disease pathogenesis [57]. For example, Tang et al. identified several hub genes as candidate biomarkers for predicting the occurrence of HBV-related hepatocellular carcinoma (HCC) from gene expression and PPI data [58]. They found that the identified genes could well distinguish stage I HCC samples from the normal controls, which indicated the potential of the genes for early detection of HBV-induced HCC. Meanwhile, two gene signatures in the result set, i.e., TOP2A and KIF11, could also be used for overall survival (OS) stratification of patients with HBV-HCC [58]. Huang et al. integrated the associations of PPI and miRNA-mRNA network to mine key miRNA-mRNA axis in HBV-HCC prognosis and therapy [59]. First, they downloaded publicly available datasets from online databases and performed the differentially expressed analysis on genes. Then hub genes with significantly differentially expressed patterns were extracted from the background, and miRNAs targeting the identified hubs were identified as key factors based on relationships in miRNA-mRNA network. The downstream web-lab experiments using real-time PCR approach convinced the biomarker potential of miRNAs-mRNAs in prognostic management of HBV-HCC patients [59].

The similar research pipeline could also be applied to HCV-mediated liver disease analysis. Liu et al. constructed a HCV-HCC-related PPI network and selected the top ten genes with high degrees for identification of biomarkers in predicting HCV-HCC development [60]. Zhan et al. identified key genes, pathways, and therapeutic targets for liver fibrosis associated with HBV and HCV infection based on a combination of bioinformatics prediction and experimental validation [61]. They found that the immune-inflammatory response pathway was shared by both HBV and HCV datasets, which highlighted the significance and role of immune and inflammatory responses to HBV/HCV infection [61].

1.4.3 Influenza Virus

Influenza, caused by the infection of influenza viruses, commonly occurs in Winter and Spring in China. Although most influenza diseases could be cured after proper treatment, patients with serious symptoms may still be life-threatening. The integration of bioinformatics prediction and clinical validation is one of the current ways widely acknowledged for translational researches of influenza. Liu et al. identified hub genes from weighted gene co-expression network and found that these genes were highly associated with the processes of antimicrobial response and neutrophils activity [62]. Moreover, two significant genes, i.e., BPI and MMP8, tended to be overexpressed in severe and dead cases, indicating their roles in regulating the development of influenza [62]. In addition to biomarker discovery, bioinformatics is also helpful for vaccine design. For example, Hu et al. used an integrated bioinformatics approach to evaluate annual perspective changes in influenza viruses. Based on the computational framework, the most plausible vaccine epitopes were calculated and compared [63]. Kaewpongsri et al. designed a new bioinformatics method to characterize viral sequences of A/H5N1. They collected the H5N1 viral isolations and performed a combination of genotypic testing and bioinformatics tools to detect the variations of H5N1 for designing appropriate vaccines against influenza [64].

To sum up, bioinformatics is powerful for control of virus infection in the era of big data and AI. On the one hand, the sequence and structure of virus could be precisely deciphered based on genome annotation and analytical tools. On the other hand, candidate biomarkers and driven signatures ranging from single molecules including genes, RNAs, and proteins to integrated pathways could be screened for translational applications such as drug design, vaccine development, and infectious pathogenic understanding.

1.5 Virus Informatics: From Virus Surveillance to Health Promotion

1.5.1 Opportunities for Precision Virus Management and Systems Healthcare

Traditional studies for virus analysis are largely dependent on experimental techniques, which would be time-consuming and costly. With the accumulation of data resources and the progress in informatics techniques, computer-aided methods promote the flourishing of virus study into an interdisciplinary informatics-experiment mode. The paradigm of virus informatics, therefore, is proposed for precision virus management and infectious control. As shown in Fig. 1.2, databases and knowledgebases are foundations for systematical modeling of complex viral statuses. Using the newly developed techniques such as AI, 5G communication, cloud computing, and block chain, the basic information and evolutionary characteristics of viruses can be quickly calculated and comprehensively measured for precision virus surveillance,

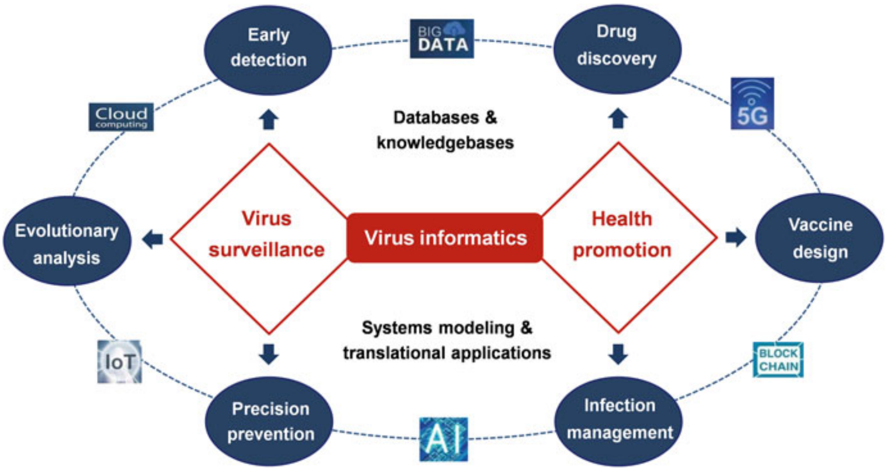

Fig. 1.2 The paradigm of virus informatics for virus surveillance and health promotion

including the early detection of viral development signatures, phylogenetic analysis on virus origin and evolution, and prevention from viral infection.

Virus informatics is also a strong weapon to fight against virus-induced diseases and promote systems health spectrum [65]. Currently, computer-aided drug and vaccine design have become a hot frontier for infection management and disease treatment. Compared with studies solely using experimental methods, the identification of key factors in viral infection process based on computational algorithms integrates a variety of biomedical knowledge to improve the efficiency and precision of data analysis, and it could further drive the discovery of novel clues and insights in virus–host pathogenesis for personalized therapeutics of virus-infected diseases.

1.5.2 Challenges and Perspectives

It is acknowledged that the term translational informatics brings unprecedented chances for precision virus surveillance and systems health promotion. The advances in informatics technologies have greatly changed the mode of computational modeling and intelligent computing, thereby creating the opportunity for systematical understanding of various biology at molecular, cellular, individual and population levels. Although the advantages are encouraging, limitations and challenges are still needed to be carefully considered and addressed.

Challenge and perspective 1: standardization and integration of multi-omics data and knowledge for population-based model construction and refining.

It should be admitted that data collection is an essential and the first step for computational modeling. However, the data from multiple omics sources tend to be highly heterogeneous. Hence, the development of ontology is urgently needed to provide standardized and normalized rules for data representation and integration. As the occurrence and progression of virus infection is a typical issue at the population level, bioinformatics models ought to be constructed and trained using population-based data to avoid overfitting and achieve enough sensitivity and specificity to assist clinical decision-making.

Challenge and perspective 2: Predicting dynamic variability and actionable signature alternations in virus evolution and infection.

Genomic mutation is one of the important features of viruses, and the interplay between virus and host cell is also a dynamic process. The informatics models, therefore, should have the ability to capture the changeable signatures in both virus evolution and infection, and it would not only help the understanding of viral origin but also be an effective way for the discovery of driven factors for disease control and treatment.

Challenge and perspective 3: combining computational strategies with biomedical experiments for translational pathogenesis understanding and anti-infective therapeutics development.

Translational informatics is not a substitution of traditional experimental researches, it aims at integrating innovative technologies of both informatics and experiments for systems-level viral studies. The combination of computational prediction with point-to-point experimental validation improves the flexibility and accuracy of developing personalized anti-infective therapeutic schemes and makes the designing of drugs and vaccines in a smart manner.

1.6 Conclusions

The advances of big data and translational informatics make the computational modeling of virus infection become reality. Databases and knowledgebases are well constructed to provide great resources for viral data sharing and analysis. Meanwhile novel bioinformatics and systems biology frameworks with AI-guided kernels contribute to systematical characterization of viruses in terms of the properties of genomic sequences, evolutionary patterns, and infectious pathogenesis. In the future work, population-level virus informatics studies with large-sample-based biomedical validations should be performed for health promotion of the translation from basic researches into clinical applications.

Acknowledgments This study was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the National Natural Science Foundation of China (grant nos. 32070671, and 31900490), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (grant no. 20KJB180010), and the regional innovation cooperation between Sichuan and Guangxi Provinces (grant no. 2020YFQ0019).

Competing Interests The authors declare no conflict of interest.

References

1. Hatano Y, Ideta T, Hirata A, Hatano K, Tomita H, Okada H et al (2021) Virus-driven carcinogenesis. Cancers (Basel) 13(11):2625
2. Windhaber S, Xin Q, Lozach PY (2021) Orthobunyaviruses: from virus binding to penetration into mammalian host cells. Viruses 13(5):872
3. Alnuqaydan AM, Almutary AG, Sukamaran A, Yang BTW, Lee XT, Lim WX et al (2021) Middle East respiratory syndrome (MERS) virus-pathophysiological axis and the current treatment strategies. AAPS PharmSciTech 22:173
4. Goyal M, Tewatia N, Vashisht H, Jain R, Kumar S (2021) Novel corona virus (COVID-19); global efforts and effective investigational medicines: a review. J Infect Public Health 14:910–921
5. Goettsch W, Beerenwinkel N, Deng L, Dolken L, Dutilh BE, Erhard F et al (2021) ITN-VIROINF: understanding (harmful) virus-host interactions by linking virology and bioinformatics. Viruses 13(5):766
6. Ramirez-Salinas GL, Garcia-Machorro J, Rojas-Hernandez S, Campos-Rodriguez R, de Oca AC, Gomez MM et al (2020) Bioinformatics design and experimental validation of influenza A virus multi-epitopes that induce neutralizing antibodies. Arch Virol 165:891–911
7. Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W (2021) Bioinformatics resources for SARS-CoV-2 discovery and surveillance. Brief Bioinform 22:631–641
8. Ibrahim B, McMahon DP, Hufsky F, Beer M, Deng L, Mercier PL et al (2018) A new era of virus bioinformatics. Virus Res 251:86–90
9. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V et al (2012) ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 40:D593–D598
10. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y et al (2017) Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res 45:D482–D490
11. Canakoglu A, Pinoli P, Bernasconi A, Alfonsi T, Melidis DP, Ceri S (2021) ViruSurf: an integrated database to investigate viral sequences. Nucleic Acids Res 49:D817–D824
12. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS (2018) A Reference Viral Database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. mSphere 3(2):e00069-18
13. Wang Y, Tong Y, Zhang Z, Zheng R, Huang D, Yang J et al (2021) ViMIC: a database of human disease-related virus mutations, integration sites and cis-effects. Nucleic Acids Res 50(D1):D918–D927
14. Yang X, Lian X, Fu C, Wuchty S, Yang S, Zhang Z (2021) HVIDB: a comprehensive database for human-virus protein-protein interactions. Brief Bioinform 22:832–844
15. Cook HV, Doncheva NT, Szklarczyk D, von Mering C, Jensen LJ (2018) Viruses.STRING: a virus-host protein-protein interaction database. Viruses 10(10):519
16. Xiang Y, Zou Q, Zhao L (2020) VPTMdb: a viral posttranslational modification database. Brief Bioinform 22(4):bbaa251
17. Cai Z, Fan Y, Zhang Z, Lu C, Zhu Z, Jiang T et al (2021) VirusCircBase: a database of virus circular RNAs. Brief Bioinform 22:2182–2190
18. Tang D, Li B, Xu T, Hu R, Tan D, Song X et al (2020) VISDB: a manually curated database of viral integration sites in the human genome. Nucleic Acids Res 48:D633–D641
19. Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK et al (2020) The 2019 novel coronavirus resource. Yi Chuan 42:212–221
20. Feng Z, Chen M, Liang T, Shen M, Chen H, Xie XQ (2021) Virus-CKB: an integrated bioinformatics platform and analysis resource for COVID-19 research. Brief Bioinform 22:882–895
21. Chen TF, Chang YC, Hsiao Y, Lee KH, Hsiao YC, Lin YH et al (2021) DockCoV2: a drug database against SARS-CoV-2. Nucleic Acids Res 49:D1152–D1159

22. Gowthaman R, Guest JD, Yin R, Adolf-Bryfogle J, Schief WR, Pierce BG (2021) CoV3D: a database of high resolution coronavirus protein structures. Nucleic Acids Res 49:D282–D287
23. Mahdi A, Blaszczyk P, Dlotko P, Salvi D, Chan TS, Harvey J et al (2021) OxCOVID19 Database, a multimodal data repository for better understanding the global impact of COVID-19. Sci Rep 11:9237
24. Shu Y, McCauley J (2017) GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill 22(13):30494
25. Squires RB, Noronha J, Hunt V, Garcia-Sastre A, Macken C, Baumgarth N et al (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respir Viruses 6:404–416
26. Ding X, Yuan X, Mao L, Wu A, Jiang T (2020) FluReassort: a database for the study of genomic reassortments among influenza viruses. Brief Bioinform 21:2126–2132
27. Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V et al (2008) BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. Nucleic Acids Res 36:D497–D503
28. Muthaiyan M, Naorem LD, Seenappa V, Pushan SS, Venkatesan A (2021) Ebolabase: Zaire ebolavirus-human protein interaction database for drug-repurposing. Int J Biol Macromol 182:1384–1391
29. Lathwal A, Kumar R, Raghava GPS (2020) OvirusTdb: a database of oncolytic viruses for the advancement of therapeutics in cancer. Virology 548:109–116
30. Usman Z, Velkov S, Protzer U, Roggendorf M, Frishman D, Karimzadeh H (2020) HDVdb: a comprehensive hepatitis D virus database. Viruses 12(5):538
31. Yan B, Zhang S, Yu S, Hussain S, Liu T, Wang B et al (2020) HRRD: a manually-curated database about the regulatory relationship between HPV and host RNA. Sci Rep 10:19586
32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
33. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340
34. Novak A, Miklos I, Lyngso R, Hein J (2008) StatAlign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. Bioinformatics 24:2403–2404
35. Troshin PV, Procter JB, Barton GJ (2011) Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA. Bioinformatics 27:2001–2002
36. Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J et al (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. Bioinformatics 32:3501–3503
37. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695
38. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274
39. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288
40. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D et al (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537
41. Zhang D, Gao F, Jakovlic I, Zou H, Zhang J, Li WX et al (2020) PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour 20:348–355
42. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
43. Abril JF, Guigo R (2000) gff2ps: visualizing genomic annotations. Bioinformatics 16:743–744
44. Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z et al (2015) IBS: an illustrator for the presentation and visualization of biological sequences. Bioinformatics 31:3359–3361

45. Zablocki O, Michelsen M, Burris M, Solonenko N, Warwick-Dugdale J, Ghosh R et al (2021) VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. PeerJ 9:e11088
46. Flageul A, Lucas P, Hirchaud E, Touzain F, Blanchard Y, Eterradossi N et al (2021) Viral variant visualizer (VVV): a novel bioinformatic tool for rapid and simple visualization of viral genetic diversity. Virus Res 291:198201
47. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B (2015) RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol 1:vev003
48. Alawi M, Burkhardt L, Indenbirken D, Reumann K, Christopeit M, Kroger N et al (2019) DAMIAN: an open source bioinformatics tool for fast, systematic and cohort based analysis of microorganisms in diagnostic samples. Sci Rep 9:16841
49. Borges V, Pinheiro M, Pechirra P, Guiomar R, Gomes JP (2018) INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. Genome Med 10:46
50. Li G, Ruan S, Zhao X, Liu Q, Dou Y, Mao F (2021) Transcriptomic signatures and repurposing drugs for COVID-19 patients: findings of bioinformatics analyses. Comput Struct Biotechnol J 19:1–15
51. Vastrad B, Vastrad C, Tengli A (2020) Bioinformatics analyses of significant genes, related pathways, and candidate diagnostic biomarkers and molecular targets in SARS-CoV-2/COVID-19. Gene Rep 21:100956
52. Xie TA, Han MY, Su XR, Li HH, Chen JC, Guo XG (2020) Identification of Hub genes associated with infection of three lung cell lines by SARS-CoV-2 with integrated bioinformatics analysis. J Cell Mol Med 24:12225–12230
53. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A (2020) A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe 27:671–80.e2
54. Min YQ, Mo Q, Wang J, Deng F, Wang H, Ning YJ (2020) SARS-CoV-2 nsp1: bioinformatics, potential structural and functional features, and implications for drug/vaccine designs. Front Microbiol 11:587317
55. Barker H, Parkkila S (2020) Bioinformatic characterization of angiotensin-converting enzyme 2, the entry receptor for SARS-CoV-2. PLoS One 15:e0240647
56. Teufel A (2015) Bioinformatics and database resources in hepatology. J Hepatol 62:712–719
57. Lin Y, Qian F, Shen L, Chen F, Chen J, Shen B (2019) Computer-aided biomarker discovery for precision medicine: data resources, models and applications. Brief Bioinform 20:952–975
58. Tang Y, Zhang Y, Hu X (2020) Identification of potential hub genes related to diagnosis and prognosis of hepatitis B virus-related hepatocellular carcinoma via integrated bioinformatics analysis. Biomed Res Int 2020:4251761
59. Huang DP, Zeng YH, Yuan WQ, Huang XF, Chen SQ, Wang MY et al (2021) Bioinformatics analyses of potential miRNA-mRNA regulatory axis in HBV-related hepatocellular carcinoma. Int J Med Sci 18:335–346
60. Liu J, Ma Z, Liu Y, Wu L, Hou Z, Li W (2019) Screening of potential biomarkers in hepatitis C virus-induced hepatocellular carcinoma using bioinformatic analysis. Oncol Lett 18:2500–2508
61. Zhan Z, Chen Y, Duan Y, Li L, Mew K, Hu P et al (2019) Identification of key genes, pathways and potential therapeutic agents for liver fibrosis using an integrated bioinformatics analysis. PeerJ 7:e6645
62. Liu S, Huang Z, Deng X, Zou X, Li H, Mu S et al (2021) Identification of key candidate biomarkers for severe influenza infection by integrated bioinformatical analysis and initial clinical validation. J Cell Mol Med 25:1725–1738
63. Hu YJ, Chow KC, Liu CC, Lin LJ, Wang SC, Wang SD (2015) Using combinatorial bioinformatics methods to analyze annual perspective changes of influenza viruses and to accelerate development of effective vaccines. J Formos Med Assoc 114:774–778

64. Kaewpongsri S, Sukasem C, Srichunrusami C, Pasomsub E, Zwang J, Pairoj W et al (2010) An integrated bioinformatics approach to the characterization of influenza A/H5N1 viral sequences by microarray data: implication for monitoring H5N1 emerging strains and designing appropriate influenza vaccines. Mol Cell Probes 24:387–395
65. Shen L, Ye B, Sun H, Lin Y, van Wietmarschen H, Shen B (2017) Systems Health: a transition from disease management toward health promotion. Adv Exp Med Biol 1028:149–164

Chapter 2
Detection and Prevention of Virus Infection

Ying Wang and Bairong Shen

Abstract The pathogenic mechanism of viral infection is a complex process involving viral mutation, viral integration, and various aspects of the interaction between the viral genome and the host. Moreover, the virus mutation will lead to the failure of related vaccines, leading to the increasing of vaccine development costs and difficulties in virus prevention. With the accumulation of various types of data, using bioinformatics methods to mine the potential viral characteristics of the pathogenic process can help virus detection and diagnosis, to take intervention measures to prevent disease development or develop effective antiviral therapies. In this chapter, we first outlined traditional approaches and emerging technologies of virus detection and prevention, and then summarized the latest developments in the bioinformatics methods application in different fields of virus researches. The emergence of artificial intelligence provides advanced analysis techniques for revealing key factors of virus infection and has been widely used in the virology community. In particular, we highlight machine learning and deep learning algorithms to identify factors/categories from complex multidimensional data and uncover novel patterns of virus or disease risk prediction.

Keywords Diagnosis · Prevention · Virus infection · Informatics · Machine learning

Y. Wang
Department of Laboratory Medicine, Shanghai Eastern Hepatobiliary Surgery Hospital, Shanghai, China

B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_2

2.1 Introduction

Viruses cause most of the infectious diseases, and more than 200 viruses have been confirmed to be pathogenic to humans [1]. In the past 50 years, many new emerging viruses are still being found worldwide, such as the Ebola virus in 1977, human immunodeficiency virus (HIV) in 1983, SARS virus in 2003, MERS in 2012, and the ongoing global pandemic of coronavirus disease 2019 (COVID-19). These emerging new infectious diseases seriously threaten human health and safety. Some are difficult to cure, due to lack of specific prevention and treatment methods, and the mortality rate is high. Some have caused and will continue to cause great harm to human beings as they are prone to becoming chronic diseases.

Nowadays, the rapid development of viral diagnosis technology has gradually formed an important branch of virology research. Since there is no effective treatment for most viruses, laboratory diagnosis has become an important strategy to control the epidemic and spread of the virus. With the continuous deepening of virology research, the diagnostic methods of viral infectious diseases have evolved from a single virus isolation at the beginning to more sensitive and specific diagnostic methods such as viral serology and molecular biology.

Vaccines are the most powerful weapon against viral infections. Although some progress has been made in the research of virus vaccines, there is still a certain gap in effective prevention currently. Many viruses show high mutation rates and can evolve rapidly to produce new mutant strains. For example, COVID-19 has evolved since the end of 2019. Different viral subtypes or branches are constantly being developed and spread globally. Worse still is the occurrence of virus antigenic drift, such as the continuous mutation of surface antigens of influenza virus neuraminidase (NA) and hemagglutinin (HA), leading to the emergence of antigenically distinct variants [2].

Therefore, it is still necessary to develop novel diagnostic methods, strengthen the research of virus vaccines, and carry out the mining and discovery of antiviral drugs from multiple levels and perspectives. With the emergence of a series of omics data (genomics, proteomics, metabolomics, etc.), sequencing technology has provided researchers with a large amount of virus and host genome information. Here, we aim to review the approaches and informatics for the detection and prevention of some virus infection, including influenza virus, coronavirus, HIV, human T-lymphotropic virus I (HTLV-1), human papillomavirus (HPV), herpes virus, hepatitis virus, arboviruses, filovirus, and rabies virus.

2.2 Virus Detection and Prevention

Viruses are classified based on their type of nucleic acid as genetic material, DNA or RNA. Compared with DNA viruses, RNA viruses are more likely to cause diseases and more deadly to the human host. The replication mode of RNA viruses

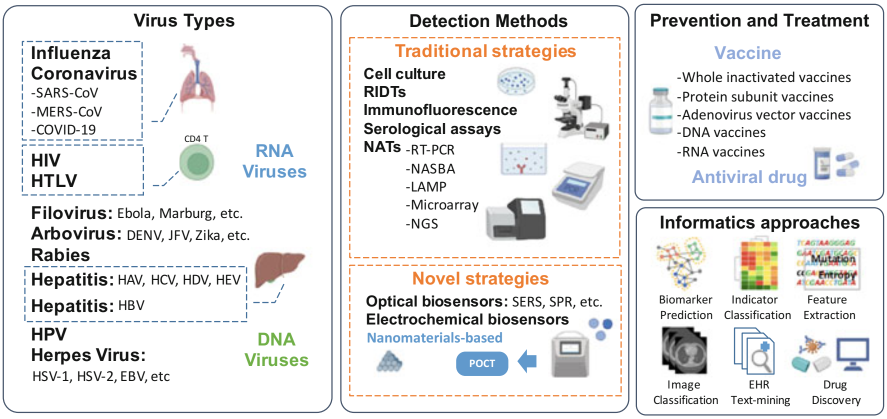

Fig. 2.1 Summary of detection, prevention, treatment, and informatics methods for DNA and RNA viruses

can be summarized as RNA self-replication and reverse transcription. Due to the low activity or even almost absent of the enzymes involved in the error repair mechanism during the replication process, RNA viruses are prone to mutations. From the Spanish flu in 1918 until the COVID-19 in 2019, RNA viruses may have constituted the greatest pandemic threat among all known pathogens in the past 100 years. Rostislav Bukasov et al. reviewed RNA viruses with pandemic tendency and common detection methods [3]. In this section, we summarized the traditional methods and novel biosensor diagnostic strategies, as well as the latest developments in the prevention and treatment for RNA and DNA viruses (Fig. 2.1).

2.2.1 Influenza Virus Detection and Prevention

Influenza is an acute respiratory infection caused by influenza virus. Influenza viruses include four types: A, B, C, and D. Detection methods used to identify influenza can be roughly divided into traditional and novel strategies. The traditional detection methods mainly include cell culture-based tests, rapid influenza diagnostic tests (RIDTs), immunofluorescence assays, serological assays, and nucleic acid-based tests (NATs).

2.2.1.1 Cell Culture-Based Detection

Influenza viruses are propagated in mammalian cells or embryonic eggs [4]. The cell culture lines used routinely include MDCK adherent cell line, VERO cell line, Hep-2 line, A549 cell line, and MRC5 cell line [5]. The sensitivity of this method is almost 100%. In influenza diagnostic research, it is considered to be the most traditional gold standard method, but results are usually not available within a few days [6]. Therefore, this method is currently rarely used for diagnostic purposes but is mainly used in the screening process of vaccine strains in epidemic seasons [5].

2.2.1.2 RIDTs

RIDTs are fast and easy of use, which are based on immunoassay to detect the nucleoprotein antigen of influenza virus in respiratory specimens [4]. Using colloidal gold tests, the results are usually obtained within 10–30 min. For commercially available RIDTs, several studies have carried out mutual evaluation or compared to reference methods and found that the sensitivity range of RIDTs is relatively wide [4, 6, 7]. Therefore, to improve the accuracy of RIDTs in the early diagnosis of influenza, it is still necessary to increase the sensitivity and specificity of their detection.

2.2.1.3 Immunofluorescence Assays

Immunofluorescence (IF) assays, including direct fluorescent antibody assay or indirect fluorescent antibody assay, mainly use fluorescent pigments to label antibody molecules and then bind to specific antigens in the specimen. IF assays are cheap, intuitive, and fast, with results within 2–4 h. Compared to RIDTs, IF assays have higher specificity and moderate sensitivity. However, the sensitivity is much lower than those of PCR-based methods, and well-trained researchers with expertise in fluorescence microscopy are required. Thus, it is still not a suitable test for clinical laboratory diagnosis of influenza virus but is valuable in confirming the results of RIDTs [7].

2.2.1.4 Serological Assays

Common serological assays include hemagglutination and hemagglutination inhibition (HAI) test, neuraminidase inhibition (NI) assay, complement fixation test (CFT), serum virus neutralization test (SVN), and enzyme-linked immunosorbent test (ELISA) [4, 7]. Because the influenza virus encodes HA and NA, HAI prevents the binding of HA on the virus surface to red blood cells by adding specific antibodies to bind to the virus. The titer of specific antibodies increasing by four times or more indicates positive and will reach the peak within 14 days [8]. Since this method usually requires two serum samples, it is mainly used as an auxiliary method for the detection of influenza but not conducive to the early diagnosis [7]. The NI assay is mainly used for the rapid determination and classification of neuraminidase subtypes (N1–N9) of the surface antigen of influenza A virus, recommended by the World Health Organization (WHO) [9]. CFT is a test for detecting antigens or antibodies that uses antigen–antibody complexes to combine with complement for the complement depletion in a solution of known concentration, which is presently used for the retrospective diagnosis of influenza [5]. The SVN test assesses the inhibitory effect of influenza virus infectivity by detecting neutralizing antibodies in human or animal serum. This method can achieve high sensitivity, but only if the antibodies match the antigens on the surface of the virus. The cumbersome procedure also limits its diagnostic application [10]. ELISA has been widely used to detect antigens or antibodies with high sensitivity and specificity and is routinely used for the rapid diagnosis of suspected influenza cases. Currently, ELISA has also been combined with microneutralization assays to determine the presence of influenza virus in microwell plates [10].

2.2.1.5 NATs

NATs is a series of technologies for direct detection of specific sequence of the virus. Compared with antigen-based tests, NATs have higher sensitivity and shorter time [4]. However, due to specific sequences depending on the aims of different researches, the specificity of NATs may vary widely. Currently, available NATs include reverse transcription PCR (RT-PCR), real-time RT-PCR, multiplex PCR, nucleic acid sequence-based amplification (NASBA), loop-mediated isothermal amplification (LAMP), microarray, and next-generation sequencing (NGS).

The PCR-based detection method is considered as a gold standard test refers to a diagnostic method with one of the highest sensitive for virus detection including influenza virus. Compared with the RT-PCR, real-time RT-PCR has the advantages of real-time monitoring of amplified products or treatment progress in patients, lower time-consuming as well as less human effort. Multiplex PCR detection method for multiple gene expressions can detect multiple respiratory pathogens simultaneously, including influenza virus. The multiplex PCR seems to have the highest diagnostic potential, characterized by high efficiency, systemicity and economic simplicity [5].

Similar to PCR-based detection methods, isothermal nucleic acid amplification method provides detection of a nucleic acid target sequence for pathogen in a high-sensitivity and less-stringent instrument requirements. This method does not rely on thermal cycling that would be very useful for high-throughput applications. To date, several methods of isothermal nucleic acid amplification are available, such as NASBA and LAMP [11–13].

Microarrays containing hundreds or thousands of probes can provide the potential for simultaneous detection of multiple pathogens. The sensitivity and specificity of most microarrays are comparable to that of other molecular diagnostic tests. For influenza virus, tested genes are usually focused on HA, NA, and M genes, and there are also some microarrays used to detect nucleoprotein (NP) and non-structural protein (NS) genes. However, due to the requirement of specialized instruments, microarrays are more appropriate for research and monitoring [14].

NGS is one of the most influential technologies in the field of genetics and medicine. To better develop novel and effective influenza inhibitors, Whitehead et al. [15] used the Illumina NGS platform to construct sequence function maps to optimize influenza binding proteins. Since NGS requires specialized equipment and takes a long time for bioinformatics analysis, this technology is more suitable for research purpose such as characterization of novel viral genome, comparative genome analysis, and genetic tracking [14].

Generally, traditional strategies for virus detection have been known for many years and can be performed under standard laboratory conditions. However, most traditional tests usually require special equipment or complex reaction control. Since it is difficult to achieve instrument miniaturization, these methods are just appropriate for clinical laboratory testing but not for point-of-care test (POCT) [16]. Biosensor is a kind of high-tech, developed from the mutual penetration of biology, chemistry, physics, medicine, electronic technology, and other disciplines. In recent years, with the continuous development of biological science, information science, and material science, biosensor technology has grown rapidly and vigorously. It is found that most sensors have high sensitivity, good selectivity, low cost, high degree of automation, miniaturization, and online continuous monitoring system. The current technical developments in the field of biosensors is based on the selection of signal output systems, mainly optical and electrochemical detection systems.

Optical biosensors commonly include the Surface-Enhanced Raman Scattering (SERS) and Surface Plasmon Resonance (SPR). Yang Sun et al. [17] have used SERS technology to develop a novel magnetic immunosensor to detect bird flu. Skilled operators and Raman spectrometers are still required for this method. SPR allows to track the binding kinetics of important molecules, as well as for quantitative detection of analytes [3]. Electrochemical biosensors contain voltammetry, current method, impedance method, and conductivity biosensor. The main benefits of using electrochemical biosensors are: fast, simple, fewer samples, and low cost. The results are generally available between 5 and 20 min. At present, use of nanomaterials in improving the performance of electrochemical biosensors is also being carried out [3, 18]. Karolina Dziabowska et al. also listed other novel ideas for biosensors [4].

Overall, taking into account the time efficiency, sensitivity, portability, and cost, biosensors have the potential for future clinical applications and POCT. Their high precision and detection rate opens up new horizons for the area of designing portable and lab-on-a-chip devices to solve many difficult problems in decades.

2.2.1.6 Influenza Virus Prevention

Inactivated influenza virus vaccine (IVV) is regarded as one of the main means to protect the host from influenza virus infection. The quadrivalent inactivated split vaccine covers the four types of currently circulating seasonal influenza viruses, H1, H3, Bv, and By. However, due to the vaccination rate, immunization failure, or other reasons, there is always a virus outbreak during the flu season. The other approach to prevention against influenza is anti-influenza therapy. Commonly used anti-influenza virus drugs are neuraminidase inhibitors (NAIs) and ion channel inhibitors based on NA-neuraminidase and M2-ion channel design. The licensed NAIs are oseltamivir (trade name Damivir), zanamivir, peramivir, and lanimivir [19]. However, resistant mutations can affect the NA catalytic site. Mutations in the catalytic site and the surrounding region were determined to be related to resistance or reduction in sensitivity to NAIs [20]. Amantadine drugs include amantadine and rimantadine. Due to the prevailing drug resistance, especially the amino acid substitution mutations at residues 26–34 of the transmembrane domain of M2 protein, these antiviral drugs are only effective if given in the early stages of infection and are currently phased out in the clinical use [21].

2.2.2 Coronavirus Detection and Prevention

Coronavirus is an important pathogen of many domestic animals and pets. In humans, coronavirus can cause respiratory infections. The pathogenicities of human coronavirus 229E (HCoV-229E), human coronavirus OC43 (HCoV-OC43), human coronavirus HKU1 (HCoV-HKU1), and human coronavirus NL63 (HCoV-NL63) are low. On the contrary, severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) are highly pathogenic, leading to severe respiratory diseases and even fatal in infected patients [22].

2.2.2.1 Coronavirus Detection

As we all know, the ongoing outbreak of COVID-19 has accelerated threats to the global public health. In clinical practice, there are three primary methods to diagnose COVID-19, which are NATs, chest CT imaging, and immunoassays. NATs are still the gold test for the diagnosis of COVID-19. In addition, studies have shown that chest imaging is beneficial at some stage of COVID-19 infection [23]. However, because the CT imaging features of patients infected with COVID-19 may be similar to those of other viral infections (e.g., influenza or SARS-CoV), the Centers for Disease Control still does not recommend using CT imaging to diagnose COVID-19. Nevertheless, the combination of epidemiological history, clinical symptoms, and CT images can help identify COVID-19 infection, especially under the circumstances of lacking laboratory test kits.

The qualitative antibody detection of COVID-19 can employ the IgM/IgG rapid detection kit based on immunoassay. The sensitivity of this method is lower than that of NATs, but it has practical value and can be used for rapid screening of previous infection individuals and identification of potential people whose immune systems have a strong enough to the virus.

To combat the rapidly spreading pandemic, POCT is the best solution. The equipment needs to be economical, sensitive, selective, user-friendly, and fast, which should be suitable for on-site testing in clinics and other environments with less infrastructure or on-site use during a virus outbreak. The isothermal nucleic acid amplification methods, such as rolling circle amplification (RCA) and RT-LAMP, have been established. Although the isothermal nucleic acid amplification methods simplified the RT-PCR detection procedure including the thermal cycler and the turnaround time for amplification, sample processing steps still require professional training of operators. A paper lateral flow measurement (LFA) technology was further developed, which is characterized by low-cost, easy to manufacture, and test at home. There are other emerging detection techniques that have been established to improve various methodological performances. For example, the use of colloidal gold nanoparticles combined with probes to achieve signal amplification strategies; the use of single-molecule ELISA to provide the detection limit of subfemoral protein concentration in order to enhance the sensitivity; the development of proximity linkage analysis (PLA) to improve the specificity; and the use of NGS and DNA microarrays to increase the throughput, etc. [24].

2.2.2.2 Coronavirus Prevention

Although many forms of vaccines against SARS-CoV and MERS-CoV have been developed, no one has been approved by the FDA yet. As of now, there are more than 200 COVID-19 vaccine candidates under the development worldwide.

Specifically, whole inactivated vaccines in clinical trials include CoronaVac (also known as PiCoVacc) developed by SinoVac Inc., and other different COVID-19 inactivated virus vaccine made by Sinopharm and Wuhan Institute of Biological Products. Additionally, Sinopharm Inc. also collaborated with Beijing Institute of Biological Products to develop BBIBP-CorV. Protein subunit vaccines are developed through chemical decomposition or control of proteolysis, then to extract the special viral protein structure and screen immunologically active fragments. Generally, protein subunit vaccines have few side effects and high safety. However, they still need adjuvants and vaccine delivery systems to enhance the immune response, to ensure the correct formation of immune memory. In clinical trials, protein subunit vaccines of COVID-19 include NVX-CoV2373 vaccine developed by Novavax and recombinant COVID-19 vaccine made by Anhui Zhifei Longcom Biopharmaceutical Co. Nucleic acid vaccine (DNA or RNA) refers to the direct introduction of exogenous genes that encode viral antigen proteins into animal cells. The expression system of the host cell synthesizes antigen proteins and then induces the host's immune response. To date, DNA vaccines of MERS and COVID-19 are mainly developed by Inovio. RNA vaccines against COVID-19 are made by Moderna and BioNTech/Pfizer. Adenovirus vector vaccines are also popular, such as Gam-COVID-Vac (Gamaleya Research Institute), AZD1222 (AstraZeneca and Oxford University), Ad5 (CanSino Biological Inc. and Beijing Institute of Biotechnology), and Ad26 (Johnson & Johnson and Beth Israel Deaconess Medical Center) [22, 25].

To date, there is still no effective antiviral drug for coronavirus infection. For patients with early SARS-CoV infection, ribavirin and high-dose steroids are used. Advanced patients use INF-a, lopinavir, and ribavirin [26]. However, the current effects of these commonly drugs and treatments are limited. The effect of high-dose steroid therapy and complications are still unclear. In addition, peptides derived from certain viral proteins can be used to inhibit viruses, such as peptide inhibitors in the MERS-CoV S2 region [27]. Other small-molecule drugs can also inhibit the viral infection. Researchers have found that HIV inhibitor ADSJ1 and 3-hydroxyphthalic anhydride-modified human serum albumin can inhibit the invasion of MERS-CoV, but their mechanisms need further study [28]. Nowadays, the pandemic of COVID-19 has attracted global attention. Remdesivir, chloroquine phosphate, and Lianhua Qingwen have been found to have certain clinical effects, but more comprehensive clinical trials are still in progress [29].

2.2.3 HIV and HTLV-1 Detection and Prevention

HIV is a retrovirus that causes defects in the human immune system. Similar to HIV, one of the key types in human T-lymphotropic virus, HTLV-1, is a tumorigenic RNA virus that causes adult T-cell leukemia through infecting CD4+ T cells.

2.2.3.1 HIV and HTLV-1 Detection

Until now, many traditional detection methods, such as ELISA, immunoassay, Western blot, radioimmunoprecipitation assay, NATs, have been developed to detect HIV and HTLV-1. Moreover, diverse technologies based on nanomaterials also have been employed for the construction of emerging diagnostic methods. Nanosensor technology uses biochemical reactions mediated by enzymes, immune components, cells, and tissues to provide corresponding information about particle behavior and characteristics, and then these reactions are converted into interpretable signals through the nano-to-micron technology. Nanosensors provide reproducible and fast results with a high degree of specificity and sensitivity for quantitative and qualitative detection. For example, Sarthak Nandi et al. reported that the nanosensor technology can be used to track infant infections during mother-to-child transmission, the latent pool in HIV-positive, and monitor patients with HIV receiving antiviral therapy [30].

So far, the electrochemical sensor detections for HIV and HTLV-1 gene based on nanomaterials include electrochemical impedance spectroscopy, square wave voltammetry, differential pulse voltammetry, and hybrid methods. Optical biosensors include quantum dots-based fluorescence analysis, metal nanoparticles, nanoclusters or nanosheets, fluorescence polarization measurement, chemiluminescence detection, nanoplasma analysis, dynamic light scattering detection, etc. These emerging technologies will help accelerate the promotion of early diagnosis for

HIV-1 and HTLV-1 infections, as well as the improvement of clinical treatment and prevention of viruses [31].

2.2.3.2 HIV and HTLV-1 Prevention

Presently, vaccines fail to prevent HIV or HTLV-1 infection. HIV vaccines can be divided into four research stages: neutralizing antibodies for humoral immunity, cellular immunity mediated by stimulating CD8 T cells, combined immunization applications for various vaccines, as well as the antigen modification and vector replication to induce stronger humoral and cellular immune response. The discovery and application of ultra-broad-spectrum neutralizing antibodies provide new ideas for HIV vaccine research. Using technologies such as next-generation sequencing and single cell sorting, scientists have discovered broadly neutralizing antibodies (VRC01) in a small number of HIV-infected individuals [32]. Immunogenic replication virus vectors are also increasingly used in the development of HIV vaccines, among which Ad26 and Ad35 combined immunization with adenovirus as a carrier has entered the human phase I trial [33].

Traditional HIV antiviral therapy includes antiretroviral therapy (ART) and antiviral drug targets. For instance, ibalizumab inhibits the viral activity by binding to the extracellular domain of CD4 [34]; Maraviroc is a small molecule chemokine receptor antagonist that can inhibit the replication of HIV by preventing the binding of gp120 and CCR5 [35]; The fusion inhibitor T-20 blocks the invasion of HIV by affecting the formation of two heptameric repeats in the extracellular region of gp41 [36]. Other types of drugs include nucleotide reverse transcriptase inhibitors (NRTIs), non-nucleotide reverse transcriptase inhibitors (NNRTIs), protease inhibitors, and integrase inhibitors. Novel HIV antiviral treatments include nucleic acid-based gene therapy strategies, protein or polypeptide-based gene therapy strategies, and antigen non-specific immunotherapy including RNA interference, gene editing, type I interferon, etc.

Although ART has made great progress in reducing the risk of AIDS, people infected with HTLV-1 have not yet benefited from such effective treatments. In recent years, virus-derived lentiviral vectors have shown great application potential in the field of vaccine development. A HTLV-1 lentiviral vector vaccine containing unique polypeptides that encode Tax, HBZ, p12, and p30 viral proteins, has been proven to safely and effectively induce immune responses in mouse models. However, these related vaccines still need to be verified in clinical trials [37].

2.2.4 HPV Detection and Prevention

HPV is a spherical DNA virus. The main infected area is human epidermis and mucosal squamous epithelium, resulting in the proliferation of squamous epithelium of human skin and mucosa. HPV infection can lead to common warts, cervical cancer, etc.

2.2.4.1 HPV Detection

Benign lesions caused by HPV infection can be divided into two categories according to the site of infection, namely whether they cause genital mucosal lesions and whether they cause skin lesions. Clinically, the forms of genital mucosal lesions are condyloma acuminatum, papular warts, and flat lesions. The most common type is condyloma acuminatum. Skin lesions caused by HPV are also frequent, manifesting as common warts (HPV-2), plantar warts (HPV-1), or flat warts (HPV-3). Generally, if the lesion is initially confirmed to be HPV-related benign lesions based on clinical manifestations, laboratory tests are not performed.

In addition to the above-mentioned benign lesions, traditional methods for HPV infection are based on PCR and probe hybridization. However, these methods cannot detect fragments that do not specifically bind to the designed primers and probes, so that different HPV genotypes may not be detected. Nevertheless, this limitation can be circumvented by unbiased high-throughput sequencing of the total nucleic acid of the sample. Furthermore, data will show whether there is viral transcription activity by cDNA sequencing, which is usually essential for the initiation and maintenance of the HPV malignant phenotype.

Among cancers related to anogenital HPV infection, cervical cancer is the most popular. Interestingly, Arroyo Mühr et al. reported the use of Novaseq 6000 to perform unbiased deep sequencing of HPV-negative patients with invasive cervical cancer, which can provide more comprehensive data for HPV screening [38]. In addition, HPV infection has also been identified as one of the predictive markers of head and neck squamous cell carcinoma (HNSCC). Aldo Venuti et al. reviewed the application of various detection techniques to HPV-infected HNSCC and summarized the use of serum and plasma to measure HPV methods, such as serum antibodies against HPV antigen, HPV DNA in plasma or saliva. The circulating HPV DNA has a potential clinical application value in the HPV-infected cancers [39]. R.B. Capone et al. used a combined detection method of conventional PCR, Southern blot hybridization, and qPCR to evaluate the level of circulating HPV DNA in patients with HNSCC [40]. It is expected that HPV detection combined with the high-throughput technology may help determine the molecular profile of any specific HPV+ or HPV− associated cancers and assist in diagnosing the risk of related diseases.

2.2.4.2 HPV Prevention

Currently, there are three licensed HPV vaccines, including Bivalent (Cervarix®, GlaxoSmithKline), Quadrivalent (Gardasil®, Merck), and Nonavalent (Gardasil9®, Merck). The Quadrivalent Gardasil was first licensed in 2006 and used to prevent HPV6, HPV11, HPV16, and HPV18, followed by the Bivalent Gardasil (Types 16 and 18) in 2007. In 2015, the Gardasil9 was approved which protects against HPV (Types 6, 11, 16, 18, 31, 33, 45, 52, and 58) [41].

2.2.5 Herpes Virus Detection

The herpes virus is a type of virus with an envelope and a double-stranded DNA genome. Most types of herpes viruses that can infect humans are herpes simplex virus type 1 (HSV-1), herpes simplex virus type 2 (HSV-2), Epstein–Barr virus (EBV), human cytomegalovirus (HCMV), Kaposi's sarcoma-associated herpesvirus (KSHV), varicella zoster virus (VZV), etc.

2.2.5.1 Herpes Virus Detection

Virus isolation and culture is a reliable basis for the clinical diagnosis of herpes virus infection. Commonly used methods for antibody detection include the complement fixation, enzyme-linked immunosorbent test, immunofluorescence, and neutralization test. HSV-1-specific PCR amplification or DNA probes used to identify this virus or DNA restriction endonuclease map can be used for the analysis of typing. HCMV is the largest member of the genome of the herpes virus family and can encode more than 200 proteins, which can cause severe complications in immunosuppressed patients and infected newborns. Laboratory testing of HCMV is mainly through the isolation of viruses from tissues or secretions. For rapid diagnosis of HCMV, the infected cells can be fixed for 24 h, and DNA probes can be used for in situ hybridization detection. Other detection methods include ELISA to detect IgM antibodies and IgG antibodies, Western blotting, and molecular hybridization techniques.

In addition, due to the overlapping manifestations of herpes viruses, clinical identification is difficult. The multiplex PCR is an attractive choice for molecular virology laboratories. Chang Ho Shin et al. has developed a quadruplex PCR approach that can quickly and accurately detect and type HSV-1, HSV-2, CMV, and EBV [42]. Cyril C.Y. Yip et al. established a multiplex PCR technique for detecting and distinguishing HSV-1, HSV-2, and VZV and further compared the performance with the commercially available RealStar alpha Herpesvirus PCR Kit 1.0 [43].

2.2.5.2 Herpes Virus Prevention

At present, there is still no preventable vaccine or available therapeutic vaccine against herpes virus infections. Although the research on several herpes viruses such as HSV-1 vaccines including wild strain vaccines, inactivated vaccines, recombinant vaccines, and DNA vaccines have been started and developed, there is no clinically effective vaccine to prevent HSV-1 infection. Drugs for the treatment of HSV-1 and HCMV can only inhibit virus replication during infection and cannot completely eliminate the virus. Generally, drugs based on HSV-1-encoded protein kinase that inhibit DNA replication during the lytic phase are used, such as the nucleoside drug acyclovir as well as its derivatives valacyclovir, penciclovir, and famciclovir. Other non-nucleoside drugs include foscarnet and other drugs [44]. For HCMV infection, the first choice for the treatment is ganciclovir, which can competitively inhibit the synthesis of HCMV DNA polymerase and directly prevent the extension of viral DNA [45]. For Kaposi sarcoma (KS)-associated herpesvirus (KSHV), patients are treated with antiviral drugs such as valganciclovir [46]. At present, there is no particularly good method to treat EBV-related diseases. Commonly used drugs include acyclovir, ganciclovir, arginine butyrate, foscarnet, cidofovir, as well as conventional radiotherapy and chemotherapy. Immunotherapy for EBV-related diseases is also under development. Studies have shown that the star molecule, PD-L1, is also highly expressed in various tumor diseases related to EBV. Inhibitors targeting PD-L1 have also shown positive therapeutic effects in the research of EBV-related diseases. The combination of multiple programs is still the future development direction of the treatment of EBV-related diseases [47, 48].

2.2.6 Hepatitis Virus Detection and Prevention

Hepatitis virus refers to the pathogen that causes viral hepatitis, mainly including hepatitis A virus (HAV), hepatitis B virus (HBV), hepatitis C virus (HCV), hepatitis D virus (HDV), and hepatitis E virus (HEV). Except for type A and E viruses which are infected through the intestinal tract, other types of viruses are spread through close contact, blood, and injection. The serological method of antigen antibody detection can be used for routine diagnosis of hepatitis virus.

2.2.6.1 Hepatitis Virus Detection

HAV is an RNA virus, and its diagnosis is usually confirmed by serological evidence of recent infection, namely detection of IgM antibodies against HAV. HAV RNA can also be detected in the feces and saliva of infected hosts, but the concentration is much lower than that in serum. The NAT detection methods for HAV include real-time PCR or nested PCR [49]. HBV is a DNA virus. The antigens of HBV are complex, including HBsAg, HBeAg, and HBcAg. HBsAg positive for more than 6 months in the blood is an important indicator for the diagnosis of chronic HBV infection (CHB). HBeAg is a serological marker molecule for active virus replication. The core protein HBcAg is highly immunogenic. Correspondingly, anti-HB is a protective neutralizing antibody. The appearance of anti-HBc IgM in the blood is usually the first immunological indication of HBV infection. Anti-HBc

IgG can survive for life after clinical cure and is a sign of HBV infection or previous infection. The gold standard for detecting the presence of HBV virus is to quantify the viral DNA load [50]. HCV is an RNA virus with a lipid shell, and its detection can be divided into the detection of anti-HCV antibodies and the detection of HCV antigens. The former methods include enzyme immunoassay (EIA), microparticle EIA, and chemiluminescence immunoassay. The latter have developed assays for the detection of HCV core antigen, as well as several approaches for the simultaneous detection of both [51]. Since the co-infection of HBV and HCV is also frequent, Shantanu Prakash et al. established a single-step multiple real-time PCR method that can simultaneously detect HBV and HCV [52]. HDV is a defective hepatotropic single-stranded RNA virus that needs the assistance of HBV to replicate. Therefore, HDV may occur as a simultaneous co-infection or superinfection in HBV-infected patients. HEV is a picornavirus. One of the principal diagnostic methods for HEV is molecular detection of RNA and specific antibodies against ORF2 using serum samples [53].

2.2.6.2 Hepatitis Virus Prevention

Vaccination is an important way to control and prevent the spread of hepatitis virus infection. Specifically, many countries have adopted universal HAV vaccinations in their children such as Havrix, Vaqta, and Twinrix [49]. There are two types of HBV vaccine. One is a plasma-derived vaccine, and the other is a recombinant vaccine manufacturing by expressing the *HBsAg* gene in *Saccharomyces cerevisiae*. These two generation vaccines are both safe and effective [54]. At present, the only commercially available vaccine against HEV is HEV 239 (Hecolin, Xiamen Chuangxin Biotechnology, China), which was registered in China in 2011 but has not yet been approved in other countries [53].

Previously, for the antiviral treatment of HAV, interferon has been evaluated for the acute HAV infection and shown to be effective in cell cultures; however, due to the limited case reports, the effectiveness is still unclear [49]. Antiviral treatments for HBV mainly include interferon alpha, long-acting interferons (PegINF-alpha), and nucleoside analogs [nucleoside (tide) analogs, NUCs], etc. Because the treatment of interferon combined with nucleoside analogs cannot target cccDNA, it is difficult to completely remove the HBV replication template-cccDNA in liver cells. To achieve the goal of completely curing chronic HBV infection, many potential target drugs for different stages of the HBV life cycle have been developed, which are mainly summarized into three categories: direct antiviral drugs targeting the components of the virus itself, host-targeted drugs targeting factors related to the virus life cycle, and targeting factors related to immune regulation [55, 56]. In approximately 50% of treated HCV patients, the peg-IFN/RBV combination can eliminate the HCV virus, which has become a standard for chronic HCV infection. In addition, direct-acting antivirals have also become another treatment option for HCV infection [51]. Currently, HDV treatment is still based on the interferon alpha. Other three treatment strategies, Myrcludex B, Lonafarnib, and REP 2139, are under evaluation [57]. For the antiviral treatment of HEV, Juliana Gil Melgaço et al. have shown that ribavirin can be used in patients with chronic HEV infections who are immunocompromised [53].

2.2.7 Arbovirus Detection and Prevention

Arboviruses are a group of arthropod-borne viruses. More than 130 arboviruses causing human diseases have been discovered worldwide, e.g., dengue fever virus (DENV), yellow fever virus (JFV), Japanese encephalitis virus (JEV), tick-borne encephalitis virus, Rift Valley fever virus (RVFV), West Nile virus (WNF), Crimean-Congo hemorrhagic fever virus (CCHFV), Chikungunya fever virus (CHIKV), and Zika virus (ZIKV).

2.2.7.1 Arbovirus Detection

The diagnosis of arboviruses generally includes the clinical features and epidemiological analysis of the diseases. Several traditional methods for arbovirus detection are cell culture virus isolation, serological analysis, molecular technology, etc. Laboratory diagnosis mainly includes serological analysis (e.g., hemagglutination inhibition test, complement fixation test, and neutralization test) and direct virus isolation. At present, many researchers have developed techniques to detect multiple arboviruses. For instance, Christian Drosten et al. established a one-step RT-PCR system that can simultaneously detect CCHFV, RVFV, DENV, and YFV [58]. José A. Boga et al. developed an approach based on multiplex real-time PCR that can detect DENV, CHIKV, ZIKV, YFV, and WNV, concurrently [59].

Metagenomics takes the entire microbial community genome in specific environmental samples as the research object for high-throughput sequencing. For clinical purposes, metagenomics NGS can accurately analyze all microorganisms in patient samples, which has an extremely high application value for pathogen research of infectious diseases. Based on Oxford nanopore technology, MinION is a low-cost handheld sequencer that can generate long reads up to 233 kb in real-time and has been used to detect various viruses, including DENV and ZIKV [60]. In addition, biosensors based on optics, electrochemistry, microfluidics, ELISA, and smartphones are the main methods used in detecting different biomarkers and serotypes of viruses including DENV [61]. Luo Lianghui et al. developed a magnetic surface molecularly imprinted resonance light scattering sensor that can detect JEV quickly and with high sensitivity [62]. In future, these emerging technologies will open up new prospects for the improvement of commercial biosensors in the early diagnosis of arbovirus infections.

2.2.7.2 Arboviruses Prevention

At present, there are only a few vaccines to prevent arboviruses. For instance, the 17D live attenuated vaccine produced from chicken embryos has been widely used in yellow fever endemic areas [62]. There are two vaccines for Japanese encephalitis, an inactivated vaccine derived from inactivated Japanese encephalitis virus cultured in hamster kidney cells and a live attenuated SA14-14-2 virus vaccine [63].

2.2.8 Filovirus Detection and Prevention

The filovirus is a single-stranded anti-strand RNA virus, including Ebola virus and Marburg virus. Ebola hemorrhagic fever caused by the Ebola virus is extremely fatal. The structure of Marburg virus is almost the same as that of Ebola, but their antigenic responses are different.

2.2.8.1 Filovirus Detection

Detection methods basically rely on the direct identification of virus particles, proteins, or specific RNA in suspected cases from whole blood, serum, or plasma. Importantly, because Ebola and MARV viruses are highly dangerous pathogens, virus isolation and identification must be performed in a special laboratory facility. Usually, the preferred method is via the NATs to directly detect the viral RNA. The common target genes for the detection of these two viruses include NP, L, and GP genes. Reverse transcription is required before PCR. The primary diagnostic methods include RT-PCR, qRT-PCR, and RT-LAMP [64]. In addition, sequencing technology is used to track the spread of pathogens and monitor virulence and potential drug resistance. Portable systems have been developed for on-site sequencing and analysis of EBOV samples, such as the MinION (Oxford Nanopore Technology) sequencing equipment used in Guinea during the 2014–2016 Ebola virus outbreak. Other novel nucleic acid means similar to RT-qPCR include FilmArray Biothreat E and have obtained the emergency use right during the EBOV outbreak (2014–2016). To achieve the goal of POCT diagnosis, the FILODIAG consortium developed a laser-based ultra-fast PCR device. The Mofina consortium established a POCT device for detecting Ebola and Marburg viruses. This equipment is small, rapid, and portable [64].

2.2.8.2 Filovirus Prevention

So far, there is still no effective treatment for filovirus infection. Ebola virus disease (EVD) and Mofina virus disease (MVD) have limited treatment options.

Conservative treatment is generally adopted, mainly maintenance treatment, including intravenous infusion to maintain the patient's blood oxygen concentration, blood pressure and electrolyte balance, and the treatment of secondary infections. Although there is no proven effective drugs treatment for EVD or MVD, some specific small molecule drugs are during the research and development process or in clinical trials, but their safety and effectiveness have yet to be confirmed. Maryam Keshtkar-Jahromi et al. reported that candidate therapies such as ZMapp, IFN-β, TKM-130803, Fabiravir, brincidofovir, and GS-5734 were entered into clinical trials [65]. In terms of vaccines, although none of the Ebola virus vaccines has been approved by the FDA, the 2014 West Africa epidemic quickly pushed a variety of Ebola vaccines into the clinical research stage. Two most promising candidate EBOV vaccines, rVSV-EBOV and ChAd3-EBO-Z, have been successfully verified to have a good protective effect on the West African population [66]. However, the MARV vaccine did not see an accelerated development similar to EBOV.

2.2.9 Rabies Virus Detection and Prevention

Rabies virus is a ribonucleic acid type rhabdovirus. As RABV is not resistant, freshly collected, refrigerated, or frozen samples should be sent to a professional laboratory for diagnosis in the fastest way.

2.2.9.1 Rabies Virus Detection

The most commonly used method for the antigen diagnosis of rabies virus is fluorescent antibody test. For the antibodies diagnosis of rabies virus, it is mainly used to determine whether the neutralizing antibody in the serum after vaccine immunization is positive and the titer is enough. Methods to measure the level of neutralizing antibodies are the fluorescent antibody virus neutralization (FAVN) test, the rapid fluorescent focus inhibition test (RFFIT), and the indirect ELISA test [67]. Molecular methods based on RT-PCR are increasingly being used for rabies diagnosis, and further combined with nucleotide sequencing for epidemiological investigations [68].

2.2.9.2 Rabies Virus Prevention

Rabies is a highly fatal infectious disease. Although there is still no effective treatment, rabies is 100% preventable. Injection or oral rabies vaccine can effectively prevent the occurrence of this disease. Animal rabies prevention products include inactivated vaccines, live attenuated vaccines, recombinant vaccines, and subunit vaccines. A recent work of Venice Du Pont et al. studied practical strategies for the mechanical characterization and resistance analysis of RABV drug candidates, and found a new molecular probe chemical type GRP-60367 performing well on specifically targeting RABV G protein and preventing G-mediated virus entry [69].

2.3 Informatics for Detection and Prevention of Virus Infection

In recent years, both the fields of genomics and computational science have undergone revolutions. The sequencing and computing capabilities have increased exponentially. The impact of these nonlinear developments on virology research is also multifaceted. The medical interest of viruses is mainly focused on pathogens and their infections, and the infections will further cause the host immune response. Advances in genomics and computational science help researchers to better understand the molecular mechanism of host immune response, to determine more effective prevention and treatment strategies against viral infections. Here, the application of bioinformatics methods to the detection and prevention of viral infections is mainly summarized in the following sections.

2.3.1 Gene Regulatory Network Modeling and Biomarker Prediction Based on Multi-omics Data

Bioinformatics has been widely used to mine microarray and deep sequencing data to reveal the heterogeneity between infection-related diseases and non-infectious disease samples and to identify potential biomarkers and signaling pathways related to viruses. Zhi-Ping Liu et al. proposed a novel and systematic transcription and posttranscriptional regulation framework based on the curated knowledge and time course expression data of H1N1 virus-infected human lung epithelial cells for in-depth analyzing the regulatory relationship among transcription factors, miRNAs, and genes [70]. To further understand the flavivirus pathogenesis, George Savidis et al. used functional genomics to identify Zika virus- and dengue virus-dependent factors [71]. To predict the potential biomarkers for personalized treatment of HBV-related HCC patients, several studies selected multiple datasets from Gene Expression Omnibus (GEO) database, screened out differentially expressed genes, and further analyzed their related biological functions, pathways, interaction networks and prognosis [72–75]. Abdul Arif Khan et al. identified host target genes and investigated the host–pathogen protein–protein interactions among all recent coronavirus outbreaks including MERS, SARS, and COVID-19 [76]. Other virus-related diseases for the screening and identification of potential biomarkers include HIV-associated heart diseases [77], EBV-related gastric cancer [78], etc.

2.3.2 Integrative Analysis and Classification Prediction Based on Clinical Indicators and Demographic Information

To provide clinical decision support for personalized treatment in precision medicine, the advanced bioinformatics analysis of big data is an emerging technology to have been widely used in the field of healthcare. The current paradigm for exploring virus-infected patients is still mainly based on the clinically relevant information. For example, Ali Mohammad Mokhtari et al. performed a study in India to investigate and analyze relationship among multi-level variables [79]. Alan J. Mueller-Breckenridge et al. used a series of virological and clinical factors to assess HBV activity and liver damage and combined histopathological analysis of liver biopsy to estimate fibrosis [80].

Moreover, the emergence of artificial intelligence (AI) provides advanced analysis techniques for revealing host and viral factors. As a subcategory of AI, machine learning (ML) algorithms recognize data patterns and simultaneously train multiple variables to build predictive models. In clinical studies, Yi Yin et al. conducted the univariate logistic regression combined with the AdaBoost algorithm to infer the risk factors for patients coinfected with HBV and HIV [81]. Na Wang et al. used the support vector machine (SVM) model to distinguish serum peptide profiles of hepatocellular carcinoma (HCC) from liver cirrhosis (LC) and found new noninvasive specific serum biomarkers for the discrimination of HBV-related HCC and LC [82]. Ying Wang et al. collected 33 indicators (i.e., demographic characteristics, blood routine indicators, and liver function) and used 4 ML models including extreme gradient boosting (XGBoost), random forest (RF), decision tree (DT), and logistic regression (LR) to predict and evaluate the population at high risk of HBV surface antigen detection [83]. Haochen Yao et al. used an ML model based on blood and urine tests to predict whether patients with COVID-19 would be at risk of severe symptoms and screened out 28 features for the potential severity of COVID-19-related biomarkers involved in the COVID-19 infection [84]. A.S. Albahri et al. summarized the application frequency of ML methods in COVID-19 detection and diagnosis, among which DT algorithms are the most frequently used, followed by naive Bayes, SVM, *k*-nearest neighbors, etc. [85]. In addition, Raman spectroscopy has received extensive attention in medical diagnosis and biomedical research. Saranjam Khan et al. applied Raman spectroscopy combined with SVM to predict HBV infection in human serum based on the spectral features [86].

Deep learning, as a branch field of ML, is an algorithm based on characterization learning of data, and mimics the mechanism of the human brain to interpret data, such as texts, images, and sounds. To date, the medical field has attracted more and more attention on the application of deep learning. Patrick Luckett et al. applied a deep learning model of cerebral blood flow to classify the cognitive impairment and weakness of HIV-infected patients [87]. Sebastian Klein et al. performed deep learning to predict prognosis of HPV related oropharyngeal squamous cell carcinoma based on H&E staining data or combine p16 status, and the performance is better than those of HPV-DNA combined with p16 status [88].

Interestingly, several studies focus on the estimation of viral infections risk based on the social network information, especially HIV. Tyler B. Wray et al. applied the ML approach to the ecological transient assessment data of HIV risk behavior to classify and predict the most important risk factors [89]. Cheng Zheng et al. used an iterative deep learning method to automatically identify online HIV influencers to increase the impact of HIV prevention activities [90]. Convolutional Neural Network (CNN) is one of the representative algorithms of deep learning. It is a type of feed forward neural network that characterized by the convolution calculation and deep structure. To increase the prediction performance of HIV status in the social network effectively, Yang Xiang et al. used a graph convolutional network model to train multiple social network data and provided a useful tool for detecting possible unknown HIV infections [91].

Actually, in a complex spectrum of viral diseases, a single laboratory test is unable to fully understand the medical history and progress of patients. Presently, many tools or pipelines have been developed to quickly analysis and integrate virus information retrieved from public domains for a better risk evaluation. Chin-Rur Yang et al. developed a set of integrated tools, FluConvert and IniFlu, which combines public available virological, epidemiological, and clinical information to discover new risk features of emerging influenza viruses [92].

2.3.3 Features Extraction and Analysis Based on Viral Genome Sequences

Mining potential virus-causing risk factors in the viral genome sequences will help formulate effective prevention and control measures to minimize the threat of future pandemics. With the in-depth understanding of the structure and function of viral genome, studies regarding the virus genetics and evolution have become one of the hotspots in the virology research. Chun Yu et al. and Liang Cai et al. explored the molecular characteristics and performed phylogenetic analysis of the virus N gene of rabies virus [93, 94]. Olivo Miotto et al. identified a set of key factors in the influenza A PB2 protein involved in the human-to-human transmission via the mutual information analysis [95].

It is known that vaccination is the main strategy to reduce the impact of virus outbreaks. Therefore, detecting the sequence variations of viral pathogens related to diseases is essential for the development of vaccines and therapies. On the one hand, it is a challenging task to identify genetic variants among rapidly evolving pathogens that adapt to the selection pressure of each host. Alexander G. Holman et al. used an ML model to construct novel methods for rapidly analyzing the genetics of evolutionary pathogens to identify amino acid features in the HIV env gene for dementia prediction [96]. On the other hand, virus subtypes or cross subtypes analysis can help ease the design and development process of vaccines and therapeutic interventions. In recent years, decoding the features that drive the biological functions from the main structure of virus subtypes has become the in-depth research direction. Norbert Nwankwo et al. and Charalambos Chrysostomou et al. used bioinformatics methods based on the digital signal processing to determine the origin of HIV-1 non-B subtype and [97] influenza A virus subtypes in the neuraminidase gene [98], respectively. Susanne Fischer et al. applied a novel affinity propagation clustering algorithm to construct a standardized subspecies classification on the basis of the RABV whole-genome sequence [99]. To identify, assemble, and classify coronavirus genomes accurately and quickly, Sara Cleemput et al. developed a coronavirus typing tool and provided a free web service that allows tracking of new viral mutations via a novel dynamic aligner [100].

Furthermore, the accumulation of mutations in the viral antigen recognition site can lead to antigen drift or antigen transfer, causing new virus strains that may lack human anti-heterologous immunity such as the H1N1 pandemic in 2009. Since data for the genome sequence can be obtained directly from clinical samples now, it is both efficient and economical for researchers to make some attempts in identifying antigenic variants based on the viral genome sequence. Lei Han et al. used multi-source serological data to develop a graph-guided multi-task sparse learning model that learns virus antigenicity-related mutations to infer antigenic variant of H3N2 virus [101]. Moreover, the reassortant virus strain poses a great risk of epidemics to human and animal health. Aaron T.L. Lun et al. developed protein typing methods, FluShuffle and FluResort, to correctly identify the source of virus proteins and the number of reconfiguration events required to produce influenza strains by the high-resolution mass spectrometry [102].

Compared with Sanger sequencing, researchers can in-depth explore HBV quasispecies (QS) characteristics based on GB-level viral sequencing data generated by NGS. In clinical practice, feature extraction from these data and convert them into indicators has important significance of research and application value. To reveal novel virus QS patterns of disease progression and risk prediction, the ML algorithms that allow identifying classification of factors from complex multidimensional data (hundreds to thousands of covariates) have been widely used in the feature analysis of virus sequences. Mingjie Wang et al. employed ML-assisted quantitative analysis of viral QS to accurately identify the immune tolerant stage of HBeAg-positive patients and develop an automatic QS analysis package [103]. Alan J. Mueller-Breckenridge et al. applied a RF model to determine the HBeAg status for patients with chronic HBV infection [80]. Shipeng Chen et al. calculated the QS pattern of the HBV rt region and provided the evidence the first time that HBV rt sequence contains important QS features for HCC risk prediction [104]. Haiyan Lei et al. mapped the original sequence with an average read length of 175 bps with human, bacterial, fungal, and viral genomes DNA databases to investigate the characterization of EBV genome in human peripheral blood B lymphocytes via NGS technology [105]. Rohan J. Meshram et al. conducted phylogenetic, mutation variability, sequence entropy, and mutation analysis strategies to predict the HCV NS5B protein epitope [106]. With increasing application of NGS technology in virus detection, the shortcomings associated with the bioinformatics pipelines have not yet been thoroughly developed, including the algorithm design challenge of

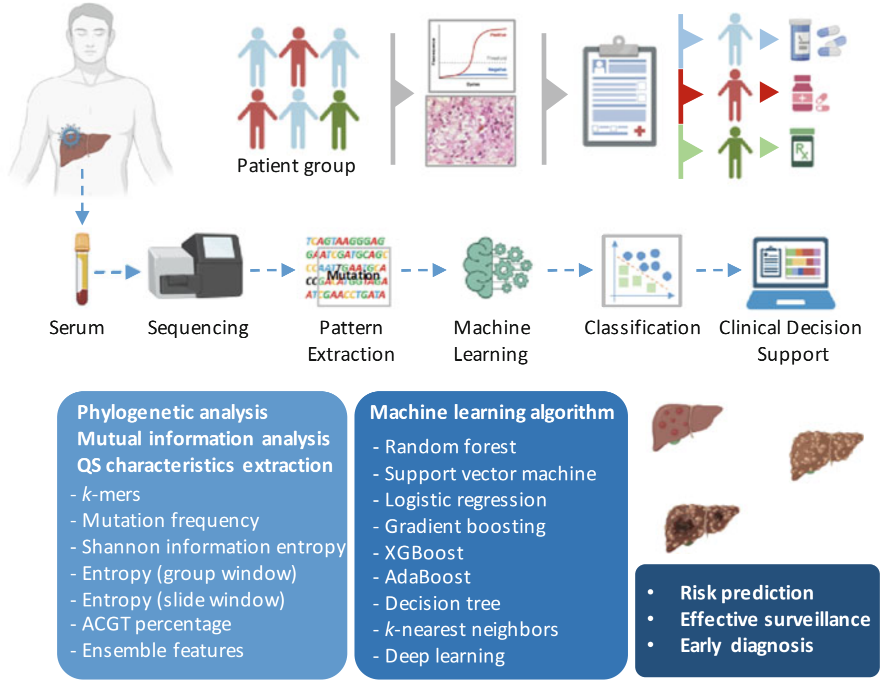

Fig. 2.2 ML-based clinical profile classification of HBV-infected patients [80] and sequence-based risk prediction for early detection

de novo assembly of viral genome. To evaluate the read loss caused by fragment alignment, Joel A. Southgate et al. provided a graph-based classifier for the reference genomes selection, assembly verification, and non-human strain detection [107]. Figure 2.2 showed the summary of features extraction and analysis based on viral genome sequences.

2.3.4 Image Classification Strategies Based on the Deep Learning Model

Regarding the medical image processing, the ML and deep learning model framework is widely used to extract image features or directly complete tasks such as classification and detection. For instance, a number of studies have built supervised ML models to distinguish X-rays of COVID-19 and other lung diseases [108–112]. Using CT scans data, Xiaoguo Zhang et al. utilized feature selection and four ML to construct the radiomics models and integrated deep learning to identify COVID-19 [113]. Ahmad Waleed Salehi et al. reviewed the deep learning models used to detect and predict coronavirus and summarized deep learning architectures that can classify chest CT and X-ray images into pneumonia and disease-free categories, including baseline CNN, DenseNet201, VGG16, VGG19, Inception_ResNet_V2, Inception_V3, Xception, Resnet50 and MobileNet_V2 [114]. Mohammed Chachan Younis et al. assessed deep learning models for predicting different coronavirus species and time series based on convolutional neural network models such as LetNet-5, AlexNet, VGG-16 net, Resnet-50, and Long Short-Term Memory [115]. Table 2.1 listed the detail description of recent coronavirus-related works.

2.3.5 Knowledge Discovery via Text Mining in Electronic Medical Records

Text Mining refers to an AI technology that extracts valuable information and knowledge from text data. With the rapid development of medical technology, massive, distributed, and heterogeneous medical data will be generated and stored in multiple medical IT systems including electronic health record (EHR), picture archiving and communications systems (PACS), hospital information systems (HIS), laboratory information systems (LIS), etc. Among them, the data content of EHR is rich which records detailed diagnosis and treatment of individuals. Therefore, the development of EHR-based text mining is of great significance for the prevention and control of viruses. At present, natural language processing (NLP) and ML classifiers are increasingly used to detect influenza cases from free text reports. Ye Ye et al. measured the feature selection and discrimination capabilities of NLP and Bayesian Network classifiers on key factors influencing the influenza detection [136]. Arturo López Pineda et al. used ML models combined with NLP technology to recognize and detect influenza from free text reports of the emergency department [137]. Julia L. Marcus et al. developed a HIV prediction model based on a large healthcare system to identify and improve pre-exposure prevention [138].

2.3.6 Drug Discovery

Molecular docking is a drug design tool through the characteristics of receptors and the interaction between receptors and drug molecules and has become a key technology in the computer-aided drug discovery research. With a large amount of protein–ligand complex structure data accumulated in public domains, researchers have identified a series of potential NA inhibitors through this approach. Li Zhang et al. developed a NA-specific scoring approach and used the RF algorithm to effectively screen NA inhibitors [139]. Shen Chang et al. filtered potential drug targets using two-side RNA-seq data and utilized the pre-trained deep learning drug target interaction model for providing a systematic drug discovery and relocation program [140]. N.R. Tomar et al. applied the molecular docking to simulate rabies virus neutralizing antibodies and provided theoretical support for the improvement

Table 2.1 Detail description of recent 20 coronavirus-related works (after 2020.10)

Reference	Method	Dataset	Brief summary
Aversano et al. [116]	Ensemble-based approach	CT scan images	VGG, Xception, and ResNet evolved with a genetic algorithm
Balaha et al. [117]	CovH2SD	CT scan images	Harris Hawks optimization and stacked deep learning
Banerjee et al. [118]	COFE-Net	Chest X-rays and CT scans	Fuzzy ensemble network (Inception V3, Inception ResNet V2 and DenseNet 201)
Verma et al. [119]	CovXmlc	Chest X-rays	SVM + last layer of VGG16 convolution network
Elharrouss et al. [120]	Encoder-decoder-based method	CT scan images	Structure and texture component extraction + encoder (VGG-16)-decoder architecture
Kumar et al. [121]	SARS-Net	Chest X-rays	Graph convolutional networks + Convolutional neural networks
Aviles-Rivero et al. [122]	GraphXCOVID	Chest X-rays	Deep graph diffusion pseudo-labelling
Liu et al. [123]	Weakly supervised segmentation	CT scan images	Uncertainty-aware self-ensembling and transformation-consistent mean teacher model with scribble-level annotation
Barshooi and Amirkhani [124]	Novel data augmentation method	Chest X-rays	Gabor filter and convolutional deep learning
Ghosh and Ghosh [125]	ENResNet	Chest X-rays	A modified residual network based enhancement
Nikolaou et al. [126]	Neural network	Chest X-rays	A dense layer on top of a pre-trained baseline CNN (EfficientNetB0)
Abdel-Basset et al. [127]	Two-stage deep learning framework	CT scan images	GR-U-Net redesigns + EfficientNet-B7
Li et al. [128]	A deep-learning-based framework	CT scan images	DNN (U-net++) + Res2Net + Clinical metadata embedding
Verma et al. [129]	Wavelet and deep learning-based detection	Thoracic X-ray images	A wavelet-based convolution neural network
Morís et al. [130]	Data augmentation approaches	Chest X-rays	Cycle generative adversarial networks
Guarrasi et al. [131]	Pareto optimization of deep networks	Chest X-rays	Ensemble CNNs

(continued)

Table 2.1 (continued)

Reference	Method	Dataset	Brief summary
Toğaçar et al. [132]	Local interpretable model-agnostic explanations method	CT scan images	Grad-CAM + CNNs (ResNet-18, ResNet-50, ResNet-101) + LIME
Bhattacharyya et al. [133]	A deep learning based approach	Chest X-rays	C-GAN + DNN (VGG-19) + ML models
Chakraborty et al. [134]	Transfer learning-based approach	Chest X-rays	Transfer learning approach on the pre-trained VGG-19
Malhotra et al. [135]	COMiT-Net	Chest X-rays	Multi-task networks (VGG16 Encoder-Decoder)

of novel therapies in the future [141]. Alexander M. Andrianov et al. combined the deep learning algorithm and molecular modeling method to identify small drug compounds that can be used as novel virus entry inhibitors [142].

2.4 Conclusions and Perspective

In this chapter, we introduced recent progress on detection and prevention approaches of viruses, including traditional and novel strategies, that have been developed to diagnose, prevent, and treat virus infections, covering cell culture-based tests, RIDTs, immunofluorescence assays, serological assays, NATs, biosensors, antiviral treatment, and vaccines. Currently, more and more different types of virus-related data are being available, bioinformatics has become a powerful methodology for the detection and prevention of virus infections and has been widely used in the virology community. Among them, the emergence of AI technology has greatly improved our ability to detect and prevent the risk of virus outbreaks, especially using ML and deep learning algorithms to train clinical experience data, build a virus infection recognition or classification model with high accuracy which can assist clinicians in rapid clinical diagnosis. With the accumulation of various types of viral data, it is expected that more efficient bioinformatics approach/tool/pipeline/databases will arise to realize global viral data integration, process analysis, and mining via the combination of virus genome and human genome data and help researches on virus mutation, evolution, traceability, and treatment.

References

1. Woolhouse M et al (2012) Human viruses: discovery and emergence. Philos Trans R Soc Lond Ser B Biol Sci 367(1604):2864–2871
2. Roubidoux EK, Schultz-Cherry S (2021) Animal models utilized for the development of influenza virus vaccines. Vaccines (Basel) 9(7):787
3. Bukasov R, Dossym D, Filchakova O (2021) Detection of RNA viruses from influenza and HIV to Ebola and SARS-CoV-2: a review. Anal Methods 13(1):34–55
4. Dziabowska K, Czaczyk E, Nidzworski D (2018) Detection methods of human and animal influenza virus-current trends. Biosensors (Basel) 8(4):94
5. Wozniak-Kosek A, Kempinska-Miroslawska B, Hoser G (2014) Detection of the influenza virus yesterday and now. Acta Biochim Pol 61(3):465–470
6. Koski RR, Klepser ME (2017) A systematic review of rapid diagnostic tests for influenza: considerations for the community pharmacist. J Am Pharm Assoc (2003) 57(1):13–19
7. Kim DK, Poudel B (2013) Tools to detect influenza virus. Yonsei Med J 54(3):560–566
8. Cox NJ, Subbarao K (1999) Influenza. Lancet 354(9186):1277–1282
9. Pedersen JC (2008) Neuraminidase-inhibition assay for the identification of influenza A virus neuraminidase subtype or neuraminidase antibody specificity. Methods Mol Biol 436:67–75
10. Zhang H, Miller BL (2019) Immunosensor-based label-free and multiplex detection of influenza viruses: state of the art. Biosens Bioelectron 141:111476
11. Poon LL et al (2005) Detection of human influenza A viruses by loop-mediated isothermal amplification. J Clin Microbiol 43(1):427–430
12. McMullen AR et al (2016) Pathology consultation on influenza diagnostics. Am J Clin Pathol 145(4):440–448
13. Lau LT, Fung YW, Yu AC (2006) Detection of animal viruses using nucleic acid sequence-based amplification (NASBA). Dev Biol (Basel) 126:7–15; discussion 323
14. Malanoski AP, Lin B (2013) Evolving gene targets and technology in influenza detection. Mol Diagn Ther 17(5):273–286
15. Whitehead TA et al (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 30(6):543–548
16. Quesada-Gonzalez D, Merkoci A (2018) Nanomaterial-based devices for point-of-care diagnostic applications. Chem Soc Rev 47(13):4697–4709
17. Sun Y et al (2017) A promising magnetic SERS immunosensor for sensitive detection of avian influenza virus. Biosens Bioelectron 89(Pt 2):906–912
18. Kim SM et al (2020) Recent development of aptasensor for influenza virus detection. Biochip J 14:327–339
19. Beck CR et al (2013) Neuraminidase inhibitors for influenza: a review and public health perspective in the aftermath of the 2009 pandemic. Influenza Other Respir Viruses 7(Suppl 1):14–24
20. Ferraris O, Lina B (2008) Mutations of neuraminidase implicated in neuraminidase inhibitors resistance. J Clin Virol 41(1):13–19
21. Musharrafieh R et al (2019) The L46P mutant confers a novel allosteric mechanism of resistance toward the influenza A virus M2 S31N proton channel blockers. Mol Pharmacol 96(2):148–157
22. Li YD et al (2020) Coronavirus vaccine development: from SARS and MERS to COVID-19. J Biomed Sci 27(1):104
23. Islam N et al (2021) Thoracic imaging tests for the diagnosis of COVID-19. Cochrane Database Syst Rev 3:CD013639
24. Pokhrel P, Hu C, Mao H (2020) Detecting the coronavirus (COVID-19). ACS Sens 5(8):2283–2296
25. Zhao J et al (2020) COVID-19: coronavirus vaccine development updates. Front Immunol 11:602256

26. Stockman LJ, Bellamy R, Garner P (2006) SARS: systematic review of treatment effects. PLoS Med 3(9):e343
27. Wang C et al (2019) Combining a fusion inhibitory peptide targeting the MERS-CoV S2 protein HR1 domain and a neutralizing antibody specific for the S1 protein receptor-binding domain (RBD) showed potent synergism against pseudotyped MERS-CoV with or without mutations in RBD. Viruses 11(1):31
28. Zhao G et al (2013) A safe and convenient pseudovirus-based inhibition assay to detect neutralizing antibodies and screen for viral entry inhibitors against the novel human coronavirus MERS-CoV. Virol J 10:266
29. Li H et al (2020) Overview of therapeutic drug research for COVID-19 in China. Acta Pharmacol Sin 41(9):1133–1140
30. Nandi S et al (2020) Biosensor platforms for rapid HIV detection. Adv Clin Chem 98:1–34
31. Mozhgani SH et al (2020) Nanotechnology based strategies for HIV-1 and HTLV-1 retroviruses gene detection. Heliyon 6(5):e04048
32. Wu X (2018) HIV broadly neutralizing antibodies: VRC01 and beyond. Adv Exp Med Biol 1075:53–72
33. Baden LR et al (2016) Assessment of the safety and immunogenicity of 2 novel vaccine platforms for HIV-1 prevention: a randomized trial. Ann Intern Med 164(5):313–322
34. Chahine EB, Durham SH (2021) Ibalizumab: the first monoclonal antibody for the treatment of HIV-1 infection. Ann Pharmacother 55(2):230–239
35. Giraudy I et al (2021) In vitro inhibitory effect of maraviroc on the association of the simian immunodeficiency virus envelope glycoprotein with CCR5. Virus Genes 57(1):106–110
36. Poveda E et al (2002) Evolution of the gp41 env region in HIV-infected patients receiving T-20, a fusion inhibitor. AIDS 16(14):1959–1961
37. Futsch N, Mahieux R, Dutartre H (2017) HTLV-1, the other pathogenic yet neglected human retrovirus: from transmission to therapeutic treatment. Viruses 10(1):1
38. Arroyo Muhr LS et al (2020) Deep sequencing detects human papillomavirus (HPV) in cervical cancers negative for HPV by PCR. Br J Cancer 123(12):1790–1795
39. Venuti A, Paolini F (2012) HPV detection methods in head and neck cancer. Head Neck Pathol 6(Suppl 1):S63–S74
40. Capone RB et al (2000) Detection and quantitation of human papillomavirus (HPV) DNA in the sera of patients with HPV-associated head and neck squamous cell carcinoma. Clin Cancer Res 6(11):4171–4175
41. Athanasiou A et al (2020) HPV vaccination and cancer prevention. Best Pract Res Clin Obstet Gynaecol 65:109–124
42. Shin CH et al (2003) Detection and typing of HSV-1, HSV-2, CMV and EBV by quadruplex PCR. Yonsei Med J 44(6):1001–1007
43. Yip CCY et al (2019) Evaluation of RealStar(R) alpha herpesvirus PCR kit for detection of HSV-1, HSV-2, and VZV in clinical specimens. Biomed Res Int 2019:5715180
44. Kimberlin DW, Whitley RJ (2007) Chapter 64: Antiviral therapy of HSV-1 and -2. In: Human herpesviruses: biology, therapy, and immunoprophylaxis. Cambridge University Press, Cambridge
45. Grossi P, Baldanti F (1997) Treatment of ganciclovir-resistant human cytomegalovirus infection. J Nephrol 10(3):146–151
46. Lurain K, Yarchoan R, Uldrick TS (2018) Treatment of Kaposi sarcoma herpesvirus-associated multicentric Castleman disease. Hematol Oncol Clin North Am 32(1):75–88
47. Ma SD et al (2016) PD-1/CTLA-4 blockade inhibits Epstein-Barr virus-induced lymphoma growth in a cord blood humanized-mouse model. PLoS Pathog 12(5):e1005642
48. Fang W et al (2015) PD-L1 is remarkably over-expressed in EBV-associated pulmonary lymphoepithelioma-like carcinoma and related to poor disease-free survival. Oncotarget 6(32):33019–33032
49. Abutaleb A, Kottilil S (2020) Hepatitis A: epidemiology, natural history, unusual clinical manifestations, and prevention. Gastroenterol Clin N Am 49(2):191–199
50. Song JE, Kim DY (2016) Diagnosis of hepatitis B. Ann Transl Med 4(18):338

51. Ansaldi F et al (2014) Hepatitis C virus in the new era: perspectives in epidemiology, prevention, diagnostics and predictors of response to therapy. World J Gastroenterol 20(29):9633–9652
52. Prakash S, Jain A, Jain B (2016) Development of novel triplex single-step real-time PCR assay for detection of Hepatitis Virus B and C simultaneously. Virology 492:101–107
53. Melgaco JG et al (2018) Hepatitis E: update on prevention and control. Biomed Res Int 2018:5769201
54. Chang MH, Chen DS (2015) Prevention of hepatitis B. Cold Spring Harb Perspect Med 5(3):a021493
55. Vigano M et al (2018) Treatment of hepatitis B: is there still a role for interferon? Liver Int 38(Suppl 1):79–83
56. Koumbi L (2015) Current and future antiviral drug therapies of hepatitis B chronic infection. World J Hepatol 7(8):1030–1040
57. Caviglia GP, Rizzetto M (2020) Treatment of hepatitis D: an unmet medical need. Clin Microbiol Infect 26(7):824–827
58. Drosten C et al (2002) Rapid detection and quantification of RNA of Ebola and Marburg viruses, Lassa virus, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, dengue virus, and yellow fever virus by real-time reverse transcription-PCR. J Clin Microbiol 40(7):2323–2330
59. Boga JA et al (2019) Simultaneous detection of Dengue virus, Chikungunya virus, Zika virus, Yellow fever virus and West Nile virus. J Virol Methods 268:53–55
60. Batovska J et al (2017) Metagenomic arbovirus detection using MinION nanopore sequencing. J Virol Methods 249:79–84
61. Basso CR et al (2018) An easy way to detect dengue virus using nanoparticle-antibody conjugates. Virology 513:85–90
62. Luo L et al (2019) Fast and sensitive detection of Japanese encephalitis virus based on a magnetic molecular imprinted polymer-resonance light scattering sensor. Talanta 202:21–26
63. Li X et al (2014) Immunogenicity and safety of currently available Japanese encephalitis vaccines: a systematic review. Hum Vaccin Immunother 10(12):3579–3593
64. Emperador DM et al (2019) Diagnostics for filovirus detection: impact of recent outbreaks on the diagnostic landscape. BMJ Glob Health 4(Suppl 2):e001112
65. Keshtkar-Jahromi M et al (2018) Treatment-focused Ebola trials, supportive care and future of filovirus care. Expert Rev Anti-Infect Ther 16(1):67–76
66. Wang Y et al (2017) Ebola vaccines in clinical trial: the promising candidates. Hum Vaccin Immunother 13(1):153–168
67. Realegeno S et al (2018) An ELISA-based method for detection of rabies virus nucleoprotein-specific antibodies in human antemortem samples. PLoS One 13(11):e0207009
68. Woldehiwet Z (2005) Clinical laboratory advances in the detection of rabies virus. Clin Chim Acta 351(1–2):49–63
69. Du Pont V et al (2020) Identification and characterization of a small-molecule rabies virus entry inhibitor. J Virol 94(13):e00321–e00320
70. Liu ZP et al (2014) Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. BMC Bioinformatics 15:336
71. Savidis G et al (2016) Identification of Zika virus and Dengue virus dependency factors using functional genomics. Cell Rep 16(1):232–246
72. Zhang X, Wang L, Yan Y (2020) Identification of potential key genes and pathways in hepatitis B virus-associated hepatocellular carcinoma by bioinformatics analyses. Oncol Lett 19(5):3477–3486
73. Zeng XC et al (2020) Screening and identification of potential biomarkers in hepatitis B virus-related hepatocellular carcinoma by bioinformatics analysis. Front Genet 11:555537
74. Tang Y, Zhang Y, Hu X (2020) Identification of potential hub genes related to diagnosis and prognosis of hepatitis B virus-related hepatocellular carcinoma via integrated bioinformatics analysis. Biomed Res Int 2020:4251761

75. Chen Z et al (2019) Identification of potential key genes for hepatitis B virus-associated hepatocellular carcinoma by bioinformatics analysis. J Comput Biol 26(5):485–494
76. Khan AA, Khan Z (2021) Comparative host-pathogen protein-protein interaction analysis of recent coronavirus outbreaks and important host targets identification. Brief Bioinform 22(2):1206–1214
77. Rasheed S, Hashim R, Yan JS (2015) Possible biomarkers for the early detection of HIV-associated heart diseases: a proteomics and bioinformatics prediction. Comput Struct Biotechnol J 13:145–152
78. Wang H et al (2021) Screening and identification of key genes in EBV-associated gastric carcinoma based on bioinformatics analysis. Pathol Res Pract 222:153439
79. Mokhtari AM et al (2021) Association of routine hepatitis B vaccination and other effective factors with hepatitis B virus infection: 25 years since the introduction of National Hepatitis B Vaccination in Iran. Iran J Med Sci 46(2):93–102
80. Mueller-Breckenridge AJ et al (2019) Machine-learning based patient classification using hepatitis B virus full-length genome quasispecies from Asian and European cohorts. Sci Rep 9(1):18892
81. Yin Y et al (2021) A noninvasive prediction model for hepatitis B virus disease in patients with HIV: based on the population of Jiangsu, China. Biomed Res Int 2021:6696041
82. Wang N et al (2014) Serum peptide pattern that differentially diagnoses hepatitis B virus-related hepatocellular carcinoma from liver cirrhosis. J Gastroenterol Hepatol 29(7):1544–1550
83. Wang Y et al (2019) Predicting hepatitis B virus infection based on health examination data of community population. Int J Environ Res Public Health 16(23):4842
84. Yao H et al (2020) Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. Front Cell Dev Biol 8:683
85. Albahri AS et al (2020) Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. J Med Syst 44(7):122
86. Khan S et al (2018) Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. Photodiagn Photodyn Ther 23:89–93
87. Luckett P et al (2019) Deep learning analysis of cerebral blood flow to identify cognitive impairment and frailty in persons living with HIV. J Acquir Immune Defic Syndr 82(5):496–502
88. Klein S et al (2021) Deep learning predicts HPV association in oropharyngeal squamous cell carcinomas and identifies patients with a favorable prognosis using regular H&E stains. Clin Cancer Res 27(4):1131–1138
89. Wray TB et al (2019) Using smartphone survey data and machine learning to identify situational and contextual risk factors for HIV risk behavior among men who have sex with men who are not on PrEP. Prev Sci 20(6):904–913
90. Zheng C, Wang W, Young SD (2021) Identifying HIV-related digital social influencers using an iterative deep learning approach. AIDS 35(Suppl 1):S85–S89
91. Xiang Y et al (2019) Network context matters: graph convolutional network model over social networks improves the detection of unknown HIV infections among young men who have sex with men. J Am Med Inform Assoc 26(11):1263–1271
92. Yang CR et al (2020) FluConvert and IniFlu: a suite of integrated software to identify novel signatures of emerging influenza viruses with increasing risk. BMC Bioinformatics 21(1):316
93. Yu C et al (2011) [Analysis on nucleoprotein gene sequence of 25 rabies virus isolates in Guizhou Province, China]. Bing Du Xue Bao 27(6):549–556
94. Cai L et al (2011) Molecular characteristics and phylogenetic analysis of N gene of human derived rabies virus. Biomed Environ Sci 24(4):431–437
95. Miotto O et al (2008) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. BMC Bioinformatics 9(Suppl 1):S18

96. Holman AG, Gabuzda D (2012) A machine learning approach for identifying amino acid signatures in the HIV env gene predictive of dementia. PLoS One 7(11):e49538
97. Nwankwo N (2013) A digital signal processing-based bioinformatics approach to identifying the origins of HIV-1 non B subtypes infecting US Army personnel serving abroad. Curr HIV Res 11(4):271–280
98. Chrysostomou C, Seker H (2013) Signal-processing-based bioinformatics approach for the identification of influenza A virus subtypes in neuraminidase genes. Annu Int Conf IEEE Eng Med Biol Soc 2013:3066–3069
99. Fischer S et al (2018) Defining objective clusters for rabies virus sequences using affinity propagation clustering. PLoS Negl Trop Dis 12(1):e0006182
100. Cleemput S et al (2020) Genome Detective Coronavirus Typing Tool for rapid identification and characterization of novel coronavirus genomes. Bioinformatics 36(11):3552–3555
101. Han L et al (2019) Graph-guided multi-task sparse learning model: a method for identifying antigenic variants of influenza A(H3N2) virus. Bioinformatics 35(1):77–87
102. Lun AT, Wong JW, Downard KM (2012) FluShuffle and FluResort: new algorithms to identify reassorted strains of the influenza virus by mass spectrometry. BMC Bioinformatics 13:208
103. Wang M et al (2021) Viral quasispecies quantitative analysis: a novel approach for appraising the immune tolerant phase of chronic hepatitis B virus infection. Emerg Microbes Infect 10(1):842–851
104. Chen S et al (2021) Using quasispecies patterns of hepatitis B virus to predict hepatocellular carcinoma with deep sequencing and machine learning. J Infect Dis 223(11):1887–1896
105. Lei H et al (2013) Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. BMC Genomics 14:804
106. Meshram RJ, Gacche RN (2015) Effective epitope identification employing phylogenetic, mutational variability, sequence entropy, and correlated mutation analysis targeting NS5B protein of hepatitis C virus: from bioinformatics to therapeutics. J Mol Recognit 28(8):492–505
107. Southgate JA et al (2020) Influenza classification from short reads with VAPOR facilitates robust mapping pipelines and zoonotic strain detection for routine surveillance applications. Bioinformatics 36(6):1681–1688
108. Jain G et al (2020) A deep learning approach to detect Covid-19 coronavirus with X-ray images. Biocybern Biomed Eng 40(4):1391–1405
109. El Asnaoui K, Chawki Y (2021) Using X-ray images and deep learning for automated detection of coronavirus disease. J Biomol Struct Dyn 39(10):3615–3626
110. Brunese L et al (2020) Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. Comput Methods Prog Biomed 196:105608
111. Brunese L et al (2020) Machine learning for coronavirus covid-19 detection from chest x-rays. Procedia Comput Sci 176:2212–2221
112. Albahli S, Albattah W (2020) Detection of coronavirus disease from X-ray images using deep learning and transfer learning algorithms. J Xray Sci Technol 28(5):841–850
113. Zhang X et al (2021) A deep learning integrated radiomics model for identification of coronavirus disease 2019 using computed tomography. Sci Rep 11(1):3938
114. Waleed Salehi A, Baglat P, Gupta G (2020) Review on machine and deep learning models for the detection and prediction of Coronavirus. Mater Today Proc 33:3896–3901
115. Younis MC (2021) Evaluation of deep learning approaches for identification of different corona-virus species and time series prediction. Comput Med Imaging Graph 90:101921
116. Aversano L et al (2021) Deep neural networks ensemble to detect COVID-19 from CT scans. Pattern Recogn 120:108135
117. Balaha HM, El-Gendy EM, Saafan MM (2021) CovH2SD: a COVID-19 detection approach based on Harris Hawks Optimization and stacked deep learning. Expert Syst Appl 186:115805
118. Banerjee A et al (2022) COFE-Net: an ensemble strategy for computer-aided detection for COVID-19. Measurement (Lond) 187:110289

119. Verma SS, Prasad A, Kumar A (2022) CovXmlc: high performance COVID-19 detection on X-ray images using Multi-Model classification. Biomed Signal Process Control 71:103272
120. Elharrouss O, Subramanian N, Al-Maadeed S (2022) An encoder-decoder-based method for segmentation of COVID-19 lung infection in CT images. SN Comput Sci 3(1):13
121. Kumar A et al (2022) SARS-Net: COVID-19 detection from chest x-rays by combining graph convolutional network and convolutional neural network. Pattern Recogn 122:108255
122. Aviles-Rivero AI et al (2022) GraphXCOVID: explainable deep graph diffusion pseudo-labelling for identifying COVID-19 on chest X-rays. Pattern Recogn 122:108274
123. Liu X et al (2022) Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images. Pattern Recogn 122:108341
124. Barshooi AH, Amirkhani A (2022) A novel data augmentation based on Gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-ray images. Biomed Signal Process Control 72:103326
125. Ghosh SK, Ghosh A (2022) ENResNet: a novel residual neural network for chest X-ray enhancement based COVID-19 detection. Biomed Signal Process Control 72:103286
126. Nikolaou V et al (2021) COVID-19 diagnosis from chest x-rays: developing a simple, fast, and accurate neural network. Health Inf Sci Syst 9(1):36
127. Abdel-Basset M et al (2021) Two-stage deep learning framework for discrimination between COVID-19 and community-acquired pneumonia from chest CT scans. Pattern Recogn Lett 152:311–319
128. Li Z et al (2021) A deep-learning-based framework for severity assessment of COVID-19 with CT images. Expert Syst Appl 185:115616
129. Verma AK et al (2021) Wavelet and deep learning-based detection of SARS-nCoV from thoracic X-ray images for rapid and efficient testing. Expert Syst Appl 185:115650
130. Moris DI et al (2021) Data augmentation approaches using cycle-consistent adversarial networks for improving COVID-19 screening in portable chest X-ray images. Expert Syst Appl 185:115681
131. Guarrasi V et al (2022) Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays. Pattern Recogn 121:108242
132. Togacar M et al (2022) Detection of COVID-19 findings by the local interpretable model-agnostic explanations method of types-based activations extracted from CNNs. Biomed Signal Process Control 71:103128
133. Bhattacharyya A et al (2022) A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images. Biomed Signal Process Control 71:103182
134. Chakraborty S, Paul S, Hasan KMA (2022) A transfer learning-based approach with deep CNN for COVID-19- and pneumonia-affected chest X-ray image classification. SN Comput Sci 3(1):17
135. Malhotra A et al (2022) Multi-task driven explainable diagnosis of COVID-19 using chest X-ray images. Pattern Recogn 122:108243
136. Ye Y et al (2014) Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc 21(5):815–823
137. Lopez Pineda A et al (2015) Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J Biomed Inform 58:60–69
138. Marcus JL et al (2019) Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. Lancet HIV 6(10):e688–e695
139. Zhang L et al (2017) Virtual screening approach to identifying influenza virus neuraminidase inhibitors using molecular docking combined with machine-learning-based scoring function. Oncotarget 8(47):83142–83154
140. Chang S, Wang LH, Chen BS (2020) Investigating core signaling pathways of hepatitis B virus pathogenesis for biomarkers identification and drug discovery via systems biology and deep learning method. Biomedicine 8(9):320

141. Tomar NR et al (2010) Molecular docking studies with rabies virus glycoprotein to design viral therapeutics. Indian J Pharm Sci 72(4):486–490
142. Andrianov AM et al (2021) Application of deep learning and molecular modeling to identify small drug-like compounds as potential HIV-1 entry inhibitors. J Biomol Struct Dyn 2021:1–19

Chapter 3
Bioinformatics for the Origin and Evolution of Viruses

Jiajia Chen, Yuxin Zhang, and Bairong Shen

Abstract The ongoing pandemic of coronavirus disease 2019 (COVID19) caused by infection with human SARS-CoV-2 is a global threat to the human population. World effort has arisen toward the characterization of the origin and evolutionary features of this devastating virus. The development of high-throughput sequencing platforms has facilitated the surveillance of viral sequence diversity in both human and animal populations. Bioinformatics pipelines are readily available for ongoing virus tracking on a global level. In this chapter, we summarize the bioinformatics tools in the origin tracing and evolutionary analyses of viruses with a highlight in their application in SARS-CoV-2, which will facilitate the prevention of the pandemic as well as custom-designed antiviral strategies.

Keywords SARS-CoV-2 · Virus · Evolution · Origin · Bioinformatics

3.1 Introduction

SARS-CoV-2, a novel coronavirus which causes the COVID-19 disease in humans, outbroke in late 2019 leading to a pandemic. With 196 million confirmed cases and four million deaths [1], COVID-19 is quickly becoming the most important health concern in the world.

SARS‒CoV‒2 belongs to the β genus of coronaviruses (CoVs) along with MERS-CoV and SARS-CoV. As the global COVID-19 pandemic continues to rage, knowledge about the origin of this virus and its mechanisms of dissemination is of great importance for future epidemic control. Although there has been much

J. Chen (✉)
School of Chemistry and Life Science, Suzhou University of Science and Technology, Suzhou, Jiangsu, China

Y. Zhang · B. Shen
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_3

speculation over the origin of the causative virus, no definite conclusion has been given so far.

SARS-CoV-2 is an RNA virus that lacks proofreading capability to correct replication errors. Therefore it keeps evolving actively with novel mutations. The geographical and temporal distributions of these mutations and their impact on the virulence and infectivity of virus require in-depth evaluation and investigation.

Recent revolutionary developments in sequencing technology have provided an unprecedented amount of sequence data for structural, phylogenetic, and mutational studies of viruses [2]. The integration of viral sequence data with spatial, temporal, and other metadata in a bioinformatics framework will facilitate the inference of the origin and evolutionary dynamics of the viral epidemic.

In this chapter, we introduce the bioinformatics tools currently available for the origin tracing and evolutionary analyses of viruses and highlight their updated application in SARS-CoV-2, which will facilitate the prevention of the COVID-19 pandemic as well as custom-designed antiviral strategies.

3.2 Informatics for Tracing the Origin of Viruses

3.2.1 Mechanism of Virus Dissemination

The potential dissemination route of SARS-CoV-2 is controversial. Molecular phylogeny of viral genomes has provided some clues about the possible paths of dissemination (Fig. 3.1). It is widely accepted that zoonotic dissemination of CoVs occurs via an intermediate host species, in which viruses are selected for better adaption to human receptors, promoting the species barrier crossing. The bat CoV (BtRaTG13) is the closest viral strain to SARS-CoV-2 characterized by now, sharing 96.1% genome identity. Given this genetic proximity, it can be postulated that SARS-CoV-2 possibly originated from bats [3]. However, there is no definite evidence that bats are the direct progenitor to humans. Recent analysis showed that Malayan pangolin is a potential animal reservoir and intermediate host, which shares <90% genome identity with SARS-CoV-2 [3–5]. SARS-CoV-2 invades the cell by binding with the host cellular angiotensin-converting enzyme (ACE2) receptors through the Spike glycoprotein (S protein). Some pangolin coronaviruses bear striking similarities to SARS-CoV-2 in the receptor binding domain (RBD) of S protein, including all six critical residues for ACE2 binding [4]. In the phylogenic tree of S protein (Fig. 3.2a), the PnGX-P2V and BtRaTG13 strains are the closest to that of human SARS-CoV-2, indicating a possible recombination between pangolin- and bat-derived CoVs [6].

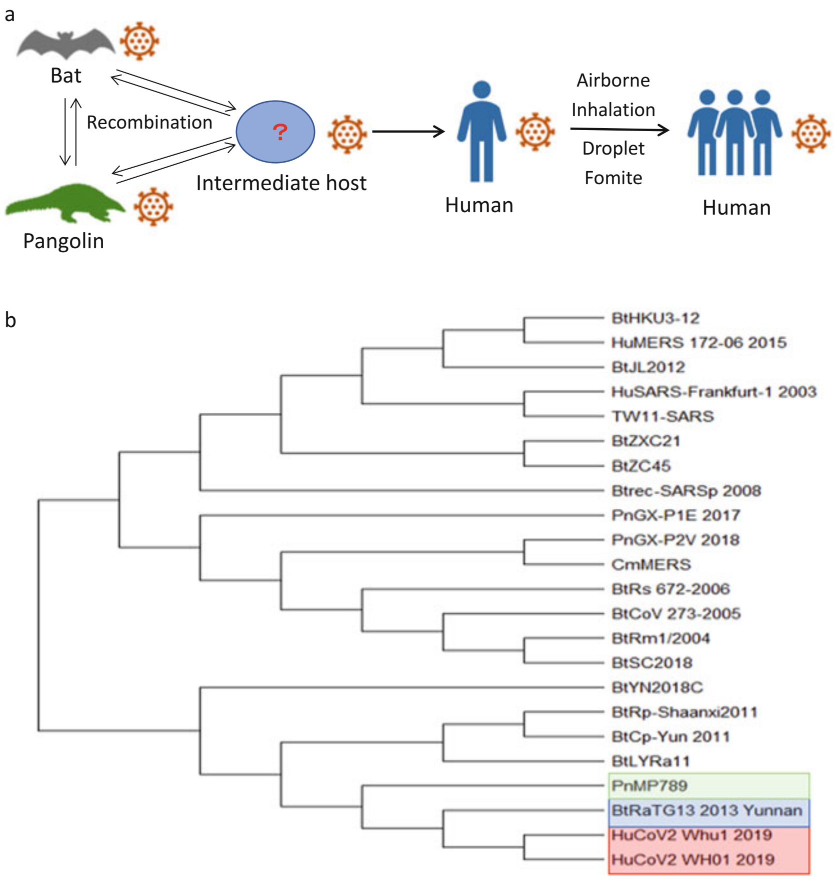

Fig. 3.1 Origin and phylogeny of SARS-CoV-2. (**a**) Putative dissemination routes of SARS-CoV-2. (**b**) Phylogenetic tree inferred from complete genomes from SARS-CoV-2 and other CoVs in bats, pangolins, camels, and humans

3.2.2 Zoonotic Origin or Accidental Laboratory Escape?

There have been several speculations that HCoV-19 was artificially manipulated and accidentally escaped from the laboratory [7]. The laboratory-origin stories are based on the observation of the insertion of a furin proteolytic cleavage site between the S1 and S2 subunits of the S protein (Fig. 3.2b). The furin cleavage site is unique in the S protein of SARS-CoV-2 and absent in other β coronaviruses. Given the uniqueness, it has been suggested that this insertion must be recent and might be artificially generated in experiments to humanize the bat RaTG13 virus. This hypothesis was supported by a recent analysis of codon usage of SARS-CoV-2 and other β coronaviruses [8]. Significant codon usage bias was revealed on spike and membrane genes between SARS-CoV-2 and its phylogenetic relatives, implying

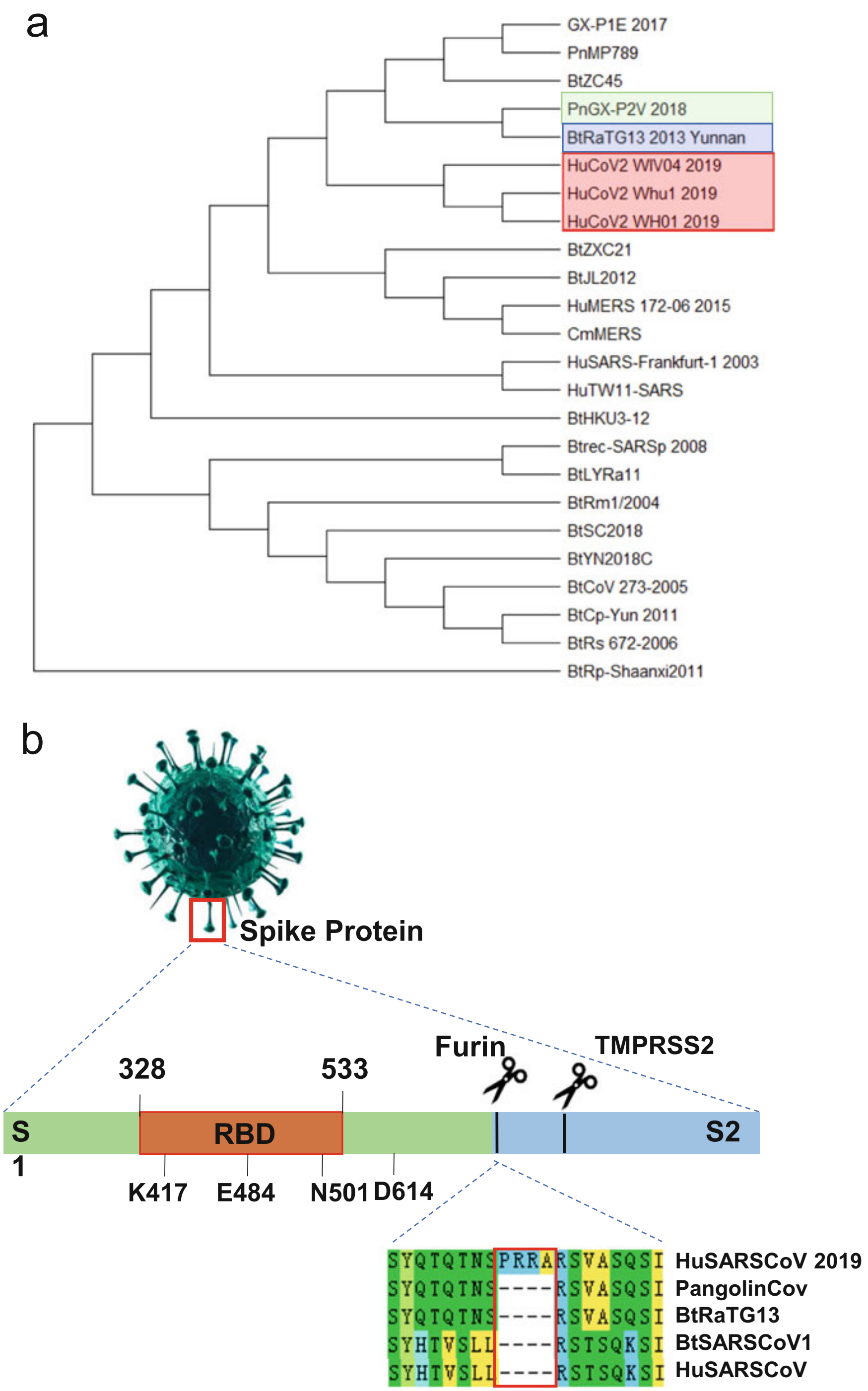

Fig. 3.2 Structure and evolution of the spike protein. (**a**) Phylogenetic tree of S proteins among different CoV species. (**b**) Structure and prevalent mutations of the SARS-CoV-2 spike proteins

these two genes are under different selection pressures and might originate from different evolutionary backgrounds.

Amid claims of deliberate manipulation, bioinformatics and molecular phylogeny approaches have provided strong pieces of evidence that argue for natural selection. Phylogenetic inferences performed around the furin cleavage sites showed that the appearance of insertion occurs in multiple coronavirus species [9]. Computational prediction revealed that the RBD sequence is not optimal for ACE2 receptor binding, lending further support to the natural selection hypothesis.

However, the current knowledge is far from sufficient to conclude the origin of SARS-CoV-2. Since this issue is critical to formulating future prevention and biosafety policies, transparent and open investigation is urgently needed.

3.2.3 Phylogenetic Inference of Virus Origin

Phylogenetic inference is a widely used method in virology and a key element of virus origin and epidemiological investigation. Molecular phylogeny analyzes the molecular sequence information, including DNA and protein mutations, codon usage, etc. through statistical methods to calculate the similarity between sequences and estimate the rate of molecular evolution, the time of sequence divergence, and the phylogenetic position of a species or gene.

Phylogenetic tree is the most widespread visualization of virus phylogeny. It can be constructed using multiple methods, e.g., distance-based methods neighbor joining (NJ), character-based methods maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference (BI).

NJ, ML, and BI approaches are based on substitution models of nucleotide or amino acid. NJ method reconstructs new neighbors by continuously connecting the nearest neighbors, and finally forms a tree topology. NJ method treats all positions on the sequence equally, and is suitable for short sequences with small evolutionary distance and few information sites. ML method is the best tree-building algorithm when the evolutionary model is selected reasonably, but it is computationally intensive and time-consuming. BI method is a new systematic evolutionary analysis method that uses Bayesian deduction to predict phylogenetic history. It not only retains the basic principles of ML method, but also introduces the Markov chain Monte Carlo method (MCMC) to simulate the later probability distribution of the evolutionary tree. MP method does not employ substitution models. It calculates all possible correct topological structures and selects the topological structure with the smallest number of substitutions as the optimal phylogenetic tree.

Software implementing these phylogenetic methods have been developed to realize phylogenetic inference on a genome scale (Table 3.1), e.g., BIONJ [10] implementing NJ, PhyML [11], IQ-TREE [12] and RAxML [13] implementing ML and MrBayes [14], PhyloBayes [15], BEAST2.5 [16] implementing BI. Various alignment-free approaches also exist for phylogenetic tree-building, e.g., the average common subsequence [17], CVTree [18], and k-mer [19]. Recently, several

Table 3.1 Bioinformatic tools for phylogenetic inference of viruses

Methodology	URL	Reference
Phylogenetic analysis		
BIONJ	http://www.atgc-montpellier.fr/bionj/	[10]
PhyML	http://www.atgc-montpellier.fr/phyml/	[11]
IQ-TREE	http://www.iqtree.org/	[12]
RAxML	http://www.exelixis-lab.org/software.html	[13]
MrBayes	http://nbisweden.github.io/MrBayes/	[14]
PhyloBayes	http://www.atgc-montpellier.fr/phylobayes/	[15]
BEAST2.5	http://www.beast2.org/	[16]
Phylosuite	http://phylosuite.jushengwu.com/	[20]
PAUP	http://paup.csit.fsu.edu/	[21]
MEGA	https://www.megasoftware.net/	[22]
Nextstrain	https://nextstrain.org/	[23]
Phylogeographic analysis		
GeoBoost2	https://zodo.asu.edu/geoboost2/	[24]
SpreaD3	https://github.com/phylogeography/SPREAD	[25]
ZooPhy	https://github.com/ZooPhy	[26]
CoVerage	https://sarscoverage.org/	[27]

integrative platforms, e.g., PhyloSuite [20], PAUP [21], and MEGA [22], have been proposed to streamline phylogenetic inference and data management.

Phylogenetic methods have found wide application in studying the evolution of viruses. For example, Nextstrain [23], a popular viral sequence analysis platform, focuses on visualized phylogenetic and enables real-time tracking of the transmission and evolution of the latest SARS-CoV-2 genomes. New computational tools are required to solve some special issues of virus phylogeny, such as recombination, virus-host interaction, and horizontal gene transfer during evolution.

3.2.4 Virus Originated from Different Geographical Locations

Phylogeography is an emerging subdiscipline in public health that explores the geographic pattern of taxa from the perspective of phylogeny, and estimates the gene flow, historical or ecological barriers that affect the spatial pattern. In this field geographical information is accounted for when analyzing the evolutionary lineage of those species. Phylogeography of viruses allows researchers to estimate the origin and migration path of viruses over time within a special context, thus better predicting the risk to humans in specific geographic areas.

Nucleotide sequences repositories such as GenBank have provided a wealth of sequence data for phylogeographic research. However, the amount of geographical information is scarce in GenBank entries, which may hinder the analysis and even give erroneous conclusions. The ever growing number of virus sequences needs matching information extraction tools.

To address this issue, a combination of Named Entity Recognition (NER) based informatics tools have been developed to extract and disambiguate geographic locations in the free-text content of articles, in order to include more geographical data for phylogeographic research. For example, GeoBoost2 [24] is a natural language processing (NLP) based tool for automate geographic locations extraction of infected hosts from the literature using supervised and distant supervision methods. Advanced Web technology (Web 2.0 and the Semantic Web) can also integrate diverse geography data within evolutionary models. SpreaD3 [25] is a software that implements a flexible Bayesian statistical framework that incorporates spatial diffusion by mapping phylogenies onto both discrete and continuous spatial information. Another example is ZooPhy [26], an automated phylogeographic pipeline of zoonotic RNA viruses. ZooPhy retrieves genetic, geographical, and public-health data through data-mining methods, performs Bayesian phylogeographic analysis via BEAST and visualizes virus migration via SpreaD3. CoVerage [27] is an online interactive dashboard to visualize and track COVID-19 cases in real time. This web resource calculates country/region-wise frequency over time and identifies SARS-CoV-2 lineages with a selective advantage in individual countries/regions based on viral genome sequence data shared via GISAID.

3.3 Prediction of SARS-CoV-2 Evolution

3.3.1 Data Resources for Virus Mutations

Virus genome sequences are being generated and shared at an unprecedented rate in virus-centered nucleotide sequences repositories such as GenBank [28], Global Initiative on Sharing All Influenza Data (GISAID) [29], CNCB/NGDC database [30], and the Virus Pathogen Resource (ViPR) [31] (Table 3.2). These databases play important roles in sequence archiving, phylogenic inference, and mutation detection. Among them, GISAID provides the largest number of SARS-CoV-2 genome sequences. At the time of writing, it has more than one million SARS-CoV-2 sequences available.

Recently, databases of nucleotide or amino acid variations have been built for SARS-CoV-2, which could track the epidemic trend of mutation in the viral genomes of interest. GESS [32] is a single nucleotide variants (SNVs) database of over two million SARS-CoV-2 genomes, while CoV-GLUE [33] is an amino acid variation database that contains replacements, insertions, and deletions in GISAID SARS-CoV-2 sequences.

Table 3.2 Bioinformatics resources for SARS-CoV-2 evolution analysis

Recourses	URL	Reference
Database		
NCBI	https://www.ncbi.nlm.nih.gov/genome/viruses/	[28]
GISAID	https://www.gisaid.org/	[29]
CNCB/NGDC	https://bigd.big.ac.cn/ncov/	[30]
ViPR	https://www.viprbrc.org/	[31]
GESS	https://wan-bioinfo.shinyapps.io/GESS/	[32]
CoV-GLUE	http://cov-glue.cvr.gla.ac.uk/#/home	[33]
Mutation identification and visualization		
CoV-Spectrum	https://cov-spectrum.ethz.ch/	[34]
CoVariants	https://covariants.org/	[35]
CoVizu	http://filogeneti.ca/covizu/#	[36]
Genomic Signature Analysis	https://covid19genomes.csiro.au/index.html#	[37]
Geographic Mutation Tracker	https://www.cbrc.kaust.edu.sa/covmt/	[38]
EACoV server	http://cov.lichtargelab.org/	[39]
MicroGMT	https://github.com/qunfengdong/MicroGMT	[40]
AutoVEM	https://github.com/Dulab2020/AutoVEM	[41]
Recombination analysis		
RDP4	http://web.cbio.uct.ac.za/∼darren/rdp.html	[42]
SimPlot	https://sray.med.som.jhmi.edu/SCRoftware/simplot/	[43]
Mutation detection at structural level		
COVID-3D	http://biosig.unimelb.edu.au/covid3d/	[44]
Coronavirus3D	https://coronavirus3d.org/index.html	[45]
VIStEDD	https://prokoplab.com/vistedd/	[46]
mCSM-PPI2	http://biosig.unimelb.edu.au/mcsm_ppi2/	[47]
CATH resource	http://funvar.cathdb.info/	[48]
EVCouplings	https://marks.hms.harvard.edu/sars-cov-2/	[49]
Spike protein mutation analysis		
Sequence Analysis Pipeline	https://cov.lanl.gov/content/index	[50]
MutationAnalyzer	https://weilab.math.msu.edu/MutationAnalyzer/	[51]
Spike Protein Mutations Monitoring	https://www.molnac.unisa.it/BioTools/cov2smt/index.php	[52]
Covid-Miner	https://covid-miner.ifo.gov.it/	[53]
Conservation analysis		
GERP	http://mendel.stanford.edu/SidowLab/downloads/gerp/	[54]
PhastCons	http://compgen.cshl.edu/phast/	[55]
PhyloP	https://ccg.epfl.ch/mga/hg19/phylop/phylop.html	[56]
Selection pressure calculation		
EasyCodeML	https://github.com/BioEasy/EasyCodeML	[57]
datamonkey	https://www.datamonkey.org/	[58]

3.3.2 Characterization of Virus Mutations

Viruses evolve through mutations and genetic recombination to adapt to environmental changes or evade host immunity. World effort has arisen toward the characterization of the genetic variation and evolutionary characteristics of SARS-CoV-2.

3.3.2.1 Mutation Detection at Sequence Level

Mutation sites, e.g., nucleotide substitutions and indels, could be readily identified through the alignment of full-length viral genomes. Enabled by CoV-19 data shared via GISAID, a variety of tools are now available for tracking the key variations in viral genome as well as epidemic trends with customizable visualizations, e.g., CoV-Spectrum [36], CoVariants [37], CoVizu [38], Genomic Signature Analysis [39], and Geographic Mutation Tracker [40]. With these tools one is able to find out mutation determinants of a variant, the functional impact as well as to visualize the diversity of SARS-CoV-2 genomes.

In addition to presenting the pre-defined mutations, platforms are also available for early identification of new variants. EACoV server [39] identifies variants and epitopes from SARS-CoV-2 proteome using Evolutionary Trace (ET) method. MicroGMT is a Python-based mutation tracker which allows for sequence mapping and indel and SNV calling [40] in microbial genomes. AutoVEM [41] is an combinatorial tool for haplotypes classification, mutations deletion and analysis in SARS-CoV-2.

In addition to mutations in the viral genome, several human mutation databases, e.g., dbSNP [59], gnomAD [60], 1KGP [61], Topmed [62], UK10K [63], and CHINAMAP [64] also provide comprehensive landscape of variation on human host proteins, e.g., ACEs that may interact with SARS-CoV-2. The annotation of SARS-CoV-2 variant data in human genome can help identify the human mutations that determine virus susceptibility and disease outcomes.

3.3.2.2 Recombination Analysis

Gene recombination is an important mechanism of virus evolution. Viruses can produce a large number of genetic mutations through genetic recombination, which is much faster than mutations caused by mutations alone. In order to reveal the role of gene recombination in the evolution of certain genes, it is necessary to obtain and validate possible recombination signals. Multiple algorithms exist for recombination analysis and have been implemented in RDP4 [42], a popular Windows program that detects recombination events amongst a group of aligned sequences. SimPlot [43] is another popular recombination analysis tool, which detects recombination signals by assessing the sequence similarity between target sequence and the reference sequence through pattern changes of the dot diagram.

3.3.2.3 Mutation Detection at Structural Level

Various structure-based resources have been established that map mutation data onto protein structures to interpret the impacts of the mutation on protein function and target binding. COVID-3D [44] and Coronavirus3D [45] rely on the spatial mapping of variant information onto experimental and predicted protein structures of SAR-CoV-2 to determine the mutation location within functional sites as well as the impact on function of the protein. In addition to SARS-CoV-2 proteins, VIStEDD [46], mCSM-PPI2 [47], and CATH [48] also map variants onto human receptors to predict the molecular impact of mutations on virus-host protein interactions.

Most methods above are based solely on evolutionary sequence conservation. Recently, an unsupervised statistical method EVCoupling [50] was developed which considers the residue dependencies between mutation positions to predict the effects of mutations and make inferences on protein function.

3.3.3 Evolutionary Analysis of Virus Spike Protein and Host Cell Receptor ACE2

The spike protein on the viral envelope is a key player in virus evolution and species barrier crossing. It binds to ACE2 through the RBD to mediate the virus entry into host cells. As shown in Fig. 3.2b, the spike protein contains two functional subunits S1 and S2, where S1 is responsible for binding to host cell receptors, and the S2 subunit is responsible for the membrane fusion between virus and host cells. During the infection process, the S protein is cleaved by the host protease (e.g., TMPRSS2) into the N-terminal S1 subunit and the C-terminal S2 subunit, and changes from the pre-fusion state to the post-fusion state. S1 and S2 are composed of an extracellular domain and a single transmembrane helix, which mediate receptor binding and membrane fusion, respectively. S1 consists of an N-terminal domain and a receptor binding domain (RBD), which is essential for determining tissue tropism and host range [65].

Some bioinformatic tools are specifically designed for analysis of S protein mutations, e.g., COVID-19 Viral Genome Analysis Pipeline [50], MutationAnalyzer [51], Spike Protein Mutations Monitoring [52], and Covid-Miner [53]. Among the observed S protein mutations, the D614G mutation is the most prevalent and its significance has been functionally characterized. The D614G mutation increases the viral load in the upper respiratory tract of COVID-19 patients and may facilitate transmission [66]. However, D614G does not seem to affect the interaction domain with ACE2, which is responsible for the viral entry into epithelial cells [67]. Apart from D614G, other mutations have been identified in RBD as host range determinants and may significantly alter the receptor binding capability. Recent reports [68, 69] performed the multiple sequence alignment to analyze the mutational dynamics in RBD and found that L455, F486, Q493, S494, N501, and Y505 residues on RBD are crucial for ACE2 recognition. RBD is the major target of the neutralizing antibodies currently under development which aim to block the binding of ACE2 to the viral receptor binding domain, thus inhibiting the membrane fusion capacity of virus. It should be noted that some of the RBD point mutations have little effect on the ability of the virus to infect, but have developed broad-spectrum resistance to neutralizing antibodies. For example, the RBD mutants A475V, L452R, V483A, and F490L are resistant to certain neutralizing antibodies [70]. A recent study [71] conducted a phylogenetic analysis of ACE2 orthologous genes from 410 vertebrate species to score the binding ability of SARS-CoV-2 Spike protein and ACE2. Species with high scores are likely to be infected with the SARS-CoV-2 virus through the ACE2 receptor, while species with low scores have a lower probability of infection. However, these predictions are only based on computer simulation and need to be confirmed by direct experimental data.

3.3.4 Mutation and Virulence Analysis

Key mutations in the RBD of SARS-CoV-2 create new inter-protein contacts which may alter the binding affinity and eventually the infectivity. Most variants with D614G and its combination with other mutations, e.g., D614G + K458R, D614G + I472V, have generally increased infectivity. There are also some mutations that reduced infectivity, e.g., V341I, D405V, V503F, P521S.

N501Y found in the B.1.1.7 variant potentially increased virulence and transmissibility [72, 73]. Co-occurrence of mutations in the RBD, e.g., K417N + E484K + N501Y, K417T + E484K + N501Y variants is more lethal with reduced antigenicity [74] than N501Y mutation alone. N439K [75] in S protein has increased binding affinity to the human ACE2 receptor and may contribute to immunity evasion.

3.3.5 Mutation Constraints and Drug/Vaccine resistance

The spike protein is the major target for vaccine design. The emergence of variants with mutations in spike protein may confer resistance to the neutralizing monoclonal antibodies and reduce the activity of convalescent serum [76, 77].

Variants with L452R, E484K, N501Y, Y508H, and combinations, e.g., D614G + A435S, D614G + I472V, K417N-E484K-N501Y, H69/V70 deletion +N501Y + D614G and E484K + N501Y + D614G [78–81] exhibited attenuated sensitivity to vaccine-induced and monoclonal antibodies [82, 83]. Therefore, immune evasion is likely to occur after acquisition of these mutations, and design strategies of COVID-19 vaccine against challenges from the variants for antibody escape are needed.

Greaney et al. [84] scanned the mutations in the RBD that affect antibody binding and generated a complete escape mutation map to predict the mutations under positive selection in the presence of antibodies and enable rational design of antibodies. A number of other interesting polymorphisms outside the spike protein have been described throughout the rest of the genome. For example, a deletion of 382 bp in the ORF8 showed significantly higher replicative fitness in vitro [85] and may assist with host immune evasion.

Genes with strong evolutionary conservation may have more important functions. As the central functional motif involved in ACE2 binding, it was earlier believed that the RBD residues are also subject to strong evolutionary constraints [86, 87]. A conservation tracks program [88] has been established for identifying evolutionarily conserved elements on surface residues of antibody epitopes that bind the SARS-CoV-2 RBD. The program discovered multiple highly conserved footprints on the RBD surface with a stronger antibody neutralizing effect. Antibodies targeting conserved conformational-epitopes on RBDs can cross-react with a range of related viruses and are expected to be a major target for therapeutic exploitation.

The conserved elements during viral evolution can be identified with a variety of bioinformatics tools. GERP (Genomic Evolutionary Rate Profiling) [54] uses the maximum likelihood method to estimate the evolution rate of a specific site. It identifies constrained motifs in multiple alignments by quantifying substitution deficits referred to as “Rejected Substitutions,” which represents the strength of purifying selection on the site. The score reflects the conservativeness of the site. The higher the score, the more conservative, and the more deleterious the substitutions. PhastCons [55] and PhyloP [56] identify conserved elements based on alignment and a model of neutral evolution called phylo-HMM, and then compute conservation scores to estimate the likelihood of aligned DNA sequences under purifying selection. PhyloP only considers the current column of the comparison and PhastCons also considers the adjacent columns of the comparison column, which makes PhastCons more sensitive to the conservative segments, while PhyloP is more accurate in the definition of conservative segments.

These tools provide a framework to inform the formulation of antibody cocktails against multiple conservation sites, in order to prevent mutation escape from individual antibodies.

3.3.6 Prediction of the Fitness of the Virus Mutations

Whether the virus develops resistance against environmental pressure during the epidemic is an important factor that affects the infectivity. The detection and quantification of evolutionary pressure have been a hot area of research in recent years. Genes containing adaptive mutations in the genome are constantly increasing due to the predominance of positive natural selection. These genes with adaptive mutations are generally potential drugs targets.

As synonymous mutations are largely invisible to natural selection, while nonsynonymous mutations can be under strong selective pressure, the ratio of synonymous/nonsynonymous substitution rate ($\omega = \mathrm{d}N/\mathrm{d}S$) is an important indicator of selective pressure at the protein level, in which $\omega = 1$ means neutral mutation, $\omega < 1$ represents purifying selection, and $\omega > 1$ means diversifying positive selection.

CodeML is one of the most widely used programs searching for genomic signatures of positive selection and has been implicated in bioinformatics tools such as EasyCodeML [57]. Datamonkey [58] is another web-server for evolutionary pressure analysis which provides a collection of methods for interrogating coding-sequence alignments for imprints of natural selection. Users can run different algorithms such as SLAC, FEL, or REL to detect sites undergoing positive (adaptive) evolution or negative evolution.

3.3.7 Long-Term Evolution and Herd Immunity

Herd immunity refers to immunization of a large proportion of the population to protect the susceptible individuals by reducing the proportion of vulnerable hosts to a level below the transmission threshold [89]. A high level of herd immunity means a high percentage of hosts in the group that are resistant to infection.

Induction of herd immunity by mass vaccination is the only ethical way for epidemic prevention. The decision whether to introduce herd immunity artificially by immunization against a particular disease will depend on several epidemiological principles. The disease must carry a substantial risk; the risk of contracting the disease must be considerable; and the vaccine must be effective and safe.

Effective herd immunity depends on several factors, including the proportion of the immunized population, the duration and efficacy of the immune response, and the stability of the viral epitopes.

The threshold proportion of the population that needs to be immunized to achieve herd immunity depends on the basic reproduction number R0, i.e., the average number of people who can be infected by an infected person in a fully susceptible and well-mixed population. The formula for calculating the herd immunity threshold is 1-1/R0, that is, the more people each infected person can infect, the higher the proportion of population immunity is needed to achieve herd immunity. For example, measles is very contagious, R0 is generally between 12 and 18, and the calculated herd immunity threshold is 92–94% of the total population. The lower the infectivity of the virus, the lower the reproduction number, the lower the threshold. For SARS-CoV-2, global immunization coverage of 50–66% of population is required to achieve herd immunity. Globally, till August 2021, a total of 4.3 billion vaccine doses have been administered. Given the current infection rates, it still represents a massive challenge. In addition, clinical data suggest that the length of immunity response against SARS-CoV-2 vaccines may not be significant and the vaccine may not be efficacious in all patients. In conclusion, development and manufacture of more effective vaccines are needed to provide a safer possible way to reach COVID-19 herd immunity.

3.4 Conclusions

Bioinformatics analysis has shed light on the possible origins and evolution trends of multiple viruses including SARS-CoV-2. The future of virus bioinformatics will depend on the development of specific bioinformatic tools, establishment of virus-specific databases as well as the collaboration of interdisciplinary research projects, in order to better understand the molecular epidemiology of viruses.

Acknowledgement This work is supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019), the National Natural Science Foundation of China (Grant No. 32070671, 31770903), Technology R&D Program of Suzhou (Grant SNG201905), as well as the Qinglan Project of Jiangsu Province.

References

1. World Health Organization. https://covid19.who.int/ (2021). Accessed August 11 2021
2. Dolja VV, Koonin EV (2018) Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res 244:36–52. https://doi.org/10.1016/j.virusres.2017.10.020
3. 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. https://doi.org/10.1038/s41586-020-2012-7
4. Lam TT, Jia N, Zhang YW, Shum MH, Jiang JF, Zhu HC et al (2020) Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583(7815):282–285. https://doi.org/10.1038/s41586-020-2169-0
5. Zhang T, Wu Q, Zhang Z (2020) Probable Pangolin origin of SARS-CoV-2 associated with the COVID-19 Outbreak. Curr Biol 30(8):1578. https://doi.org/10.1016/j.cub.2020.03.063
6. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou JJ et al (2020) Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature 583(7815):286–289. https://doi.org/10.1038/s41586-020-2313-x
7. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nat Med 26(4):450–452. https://doi.org/10.1038/s41591-020-0820-9
8. Gu H, Chu DKW, Peiris M, Poon LLM (2020) Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. Virus Evol 6(1):veaa032. https://doi.org/10.1093/ve/veaa032
9. Sallard E, Halloy J, Casane D, Decroly E, van Helden J (2021) Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. Environ Chem Lett:1–17. https://doi.org/10.1007/s10311-020-01151-1
10. Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14(7):685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808

11. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33(Web Server issue):W557–W559. https://doi.org/10.1093/nar/gki352
12. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274. https://doi.org/10.1093/molbev/msu300
13. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35(21):4453–4455. https://doi.org/10.1093/bioinformatics/btz305
14. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S et al (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61(3):539–542. https://doi.org/10.1093/sysbio/sys029
15. Rodrigue N, Lartillot N (2014) Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics 30(7):1020–1021. https://doi.org/10.1093/bioinformatics/btt729
16. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchene S, Fourment M, Gavryushkina A et al (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 15(4):e1006650. https://doi.org/10.1371/journal.pcbi.1006650
17. Zhu QH, Stephen S, Taylor J, Helliwell CA, Wang MB (2014) Long noncoding RNAs responsive to Fusarium oxysporum infection in Arabidopsis thaliana. New Phytol 201(2):574–584. https://doi.org/10.1111/nph.12537
18. Sun J, Xu Z, Hao B (2010) Whole-genome based Archaea phylogeny and taxonomy: a composition vector approach. Chin Sci Bull 55(22):2323–2328. https://doi.org/10.1007/s11434-010-3008-8
19. Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517. https://doi.org/10.1186/1471-2164-9-517
20. Zhang D, Gao F, Jakovlic I, Zou H, Zhang J, Li WX et al (2020) PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour 20(1):348–355. https://doi.org/10.1111/1755-0998.13096
21. Wilgenbusch JC, Swofford D (2003.;Chapter 6:Unit 64) Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics. https://doi.org/10.1002/0471250953.bi0604s00
22. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739. https://doi.org/10.1093/molbev/msr121
23. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C et al (2018) Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34(23):4121–4123. https://doi.org/10.1093/bioinformatics/bty407
24. Magge A, Weissenbacher D, O'Connor K, Tahsin T, Gonzalez-Hernandez G, Scotch M (2020) GeoBoost2: a natural languageprocessing pipeline for GenBank metadata enrichment for virus phylogeography. Bioinformatics 36(20):5120–5121. https://doi.org/10.1093/bioinformatics/btaa647
25. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P (2016) SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol 33(8):2167–2169. https://doi.org/10.1093/molbev/msw082
26. Scotch M, Mei C, Brandt C, Sarkar IN, Cheung K (2010) At the intersection of public-health informatics and bioinformatics: using advanced Web technologies for phylogeography. Epidemiology 21(6):764–768. https://doi.org/10.1097/EDE.0b013e3181f534dd
27. Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20(5):533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

28. Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M et al (2004) National center for biotechnology information viral genomes project. J Virol 78(14):7291–7298. https://doi.org/10.1128/JVI.78.14.7291-7298.2004
29. Shu Y, McCauley J (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill 22(13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
30. Song S, Ma L, Zou D, Tian D, Li C, Zhu J et al (2020) The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics Proteomics Bioinformatics. https://doi.org/10.1016/j.gpb.2020.09.001
31. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V et al (2012) ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 40(Database issue):D593–D598. https://doi.org/10.1093/nar/gkr859
32. Fang S, Li K, Shen J, Liu S, Liu J, Yang L et al (2021) GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences. Nucleic Acids Res 49(D1):D706–DD14. https://doi.org/10.1093/nar/gkaa808
33. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: a web application for tracking sARS-CoV-2 genomic variation. Preprint at https://www.preprints.org/manuscript/202006.0225/v1, 2021
34. Chen C, Nadeau S, Yared M, Voinov P, Stadler T. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Preprint at https://arxiv.org/abs/2106.08106 (2021)
35. Hodcroft E: CoVariants: SARS-CoV-2 mutations and variants of interest. https://covariants.org/ (2021). Accessed August 12, 2021.
36. Rambaut A, Holmes EC, O'Toole A, Hill V, McCrone JT, Ruis C et al (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5(11):1403–1407. https://doi.org/10.1038/s41564-020-0770-5
37. Bauer DC, Tay AP, Wilson LOW, Reti D, Hosking C, McAuley AJ et al (2020) Supporting pandemic response using genomics and bioinformatics: a case study on the emergent SARS-CoV-2 outbreak. Transbound Emerg Dis 67(4):1453–1462. https://doi.org/10.1111/tbed.13588
38. Alam I, Radovanovic A, Incitti R, Kamau AA, Alarawi M, Azhar EI et al (2021) CovMT: an interactive SARS-CoV-2 mutation tracker, with a focus on critical variants. Lancet Infect Dis 21(5):602. https://doi.org/10.1016/S1473-3099(21)00078-5
39. Wang C, Konecki D, Marciano D, Govindarajan H, Williams A, Wastuwidyaningtyas B, et al. Identification of evolutionarily stable functional and immunogenic sites across the SARS-CoV-2 proteome and the greater coronavirus family. Preprint at https://www.researchsquare.com/article/rs-95030/v3 (2021)
40. Xing Y, Li X, Gao X, Dong Q (2020) MicroGMT: a mutation tracker for SARS-CoV-2 and other microbial genome sequences. Front Microbiol 11:1502. https://doi.org/10.3389/fmicb.2020.01502
41. Xi B, Jiang D, Li S, Lon JR, Bai Y, Lin S et al (2021) AutoVEM: an automated tool to real-time monitor epidemic trends and key mutations in SARS-CoV-2 evolution. Comput Struct Biotechnol J 19:1976–1985. https://doi.org/10.1016/j.csbj.2021.04.002
42. Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16(6):562–563. https://doi.org/10.1093/bioinformatics/16.6.562
43. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG et al (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 73(1):152–160. https://doi.org/10.1128/JVI.73.1.152-160.1999
44. Portelli S, Olshansky M, Rodrigues CHM, D'Souza EN, Myung Y, Silk M et al (2020) Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. Nat Genet 52(10):999–1001. https://doi.org/10.1038/s41588-020-0693-3
45. Sedova M, Jaroszewski L, Alisoltani A, Godzik A (2020) Coronavirus3D: 3D structural visualization of COVID-19 genomic divergence. Bioinformatics 36(15):4360–4362. https://doi.org/10.1093/bioinformatics/btaa550

46. Gupta R, Charron J, Stenger CL, Painter J, Steward H, Cook TW et al (2020) SARS-CoV2 (COVID-19) structural/evolution dynamicome: insights into functional evolution and human genomics. bioRxiv. https://doi.org/10.1101/2020.05.15.098616
47. Rodrigues CHM, Myung Y, Pires DEV, Ascher DB (2019) mCSM-PPI2: predicting the effects of mutations on protein-protein interactions. Nucleic Acids Res 47(W1):W338–WW44. https://doi.org/10.1093/nar/gkz383
48. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P et al (2019) CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res 47(D1):D280–D2D4. https://doi.org/10.1093/nar/gky1097
49. Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C et al (2017) Mutation effects predicted from sequence co-variation. Nat Biotechnol 35(2):128–135. https://doi.org/10.1038/nbt.3769
50. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W et al (2020) Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 Virus. Cell 182(4):812–27 e19. https://doi.org/10.1016/j.cell.2020.06.043
51. Chen J, Gao K, Wang R, Wei GW (2021) Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. Chem Sci 12(20):6929–6948. https://doi.org/10.1039/d1sc01203g
52. Oliva R, Shaikh AR, Petta A, Vangone A, Cavallo L (2021) D936Y and other mutations in the fusion core of the SARS-CoV-2 Spike protein heptad repeat 1: frequency, geographical distribution, and structural effect. Molecules 26(9). https://doi.org/10.3390/molecules26092622
53. Massacci A, Sperandio E, D’Ambrosio L, Maffei M, Palombo F, Aurisicchio L et al (2020) Design of a companion bioinformatic tool to detect the emergence and geographical distribution of SARS-CoV-2 Spike protein genetic variants. J Transl Med 18(1):494. https://doi.org/10.1186/s12967-020-02675-4
54. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6(12):e1001025. https://doi.org/10.1371/journal.pcbi.1001025
55. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15(8):1034–1050. https://doi.org/10.1101/gr.3715005
56. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20(1):110–121. https://doi.org/10.1101/gr.097857.109
57. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW (2019) EasyCodeML: a visual tool for analysis of selection using CodeML. Ecol Evol 9(7):3891–3898. https://doi.org/10.1002/ece3.5015
58. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL (2018) Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol 35(3):773–777. https://doi.org/10.1093/molbev/msx335
59. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311. https://doi.org/10.1093/nar/29.1.308
60. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q et al (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581(7809):434–443. https://doi.org/10.1038/s41586-020-2308-7
61. Kuehn BM (2008) 1000 Genomes Project promises closer look at variation in human genome. JAMA 300(23):2715. https://doi.org/10.1001/jama.2008.823
62. Burgess DJ (2021) The TOPMed genomic resource for human health. Nat Rev Genet 22(4):200. https://doi.org/10.1038/s41576-021-00343-x
63. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y et al (2015) The UK10K project identifies rare variants in health and disease. Nature 526(7571):82–90. https://doi.org/10.1038/nature14962

64. Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J et al (2020) The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res 30(9):717–731. https://doi.org/10.1038/s41422-020-0322-9
65. Khan A, Zia T, Suleman M, Khan T, Ali SS, Abbasi AA et al (2021) Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: An insight from structural data. J Cell Physiol. https://doi.org/10.1002/jcp.30367
66. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z et al (2020) Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181(4):894–904 e9. https://doi.org/10.1016/j.cell.2020.03.045
67. Guzzi PH, Mercatelli D, Ceraolo C, Giorgi FM (2020) Master regulator analysis of the SARS-CoV-2/human interactome. J Clin Med 9(4). https://doi.org/10.3390/jcm9040982
68. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O et al (2020) Cryo-EM Structure of the 2019-nCoV spike in the prefusion conformation. bioRxiv. https://doi.org/10.1101/2020.02.11.944462
69. Kadam SB, Sukhramani GS, Bishnoi P, Pable AA, Barvkar VT (2021) SARS-CoV-2, the pandemic coronavirus: molecular and structural insights. J Basic Microbiol 61(3):180–202. https://doi.org/10.1002/jobm.202000537
70. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S et al (2020) The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 182(5):1284–94 e9. https://doi.org/10.1016/j.cell.2020.07.012
71. Damas J, Hughes GM, Keough KC, Painter CA, Persky NS, Corbo M et al (2020) Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates. bioRxiv. https://doi.org/10.1101/2020.04.16.045302
72. Zhao S, Lou J, Cao L, Zheng H, Chong MKC, Chen Z et al (2021) Quantifying the transmission advantage associated with N501Y substitution of SARS-CoV-2 in the UK: an early data-driven analysis. J Travel Med 28(2). https://doi.org/10.1093/jtm/taab011
73. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L et al (2021) Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 593(7858):266–269. https://doi.org/10.1038/s41586-021-03470-x
74. Collier DA, De Marco A, Ferreira I, Meng B, Datir RP, Walls AC et al (2021) Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. Nature 593(7857):136–141. https://doi.org/10.1038/s41586-021-03412-7
75. Thomson EC, Rosen LE, Shepherd JG, Spreafico R, da Silva FA, Wojcechowskyj JA et al (2021) Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell 184(5):1171–87 e20. https://doi.org/10.1016/j.cell.2021.01.037
76. Wang P, Nair MS, Liu L, Iketani S, Luo Y, Guo Y et al (2021) Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. Nature 593(7857):130–135. https://doi.org/10.1038/s41586-021-03398-2
77. Wang P, Casner RG, Nair MS, Wang M, Yu J, Cerutti G et al (2021) Increased Resistance of SARS-CoV-2 Variant P.1 to Antibody Neutralization. bioRxiv. https://doi.org/10.1101/2021.03.01.433466
78. Jangra S, Ye C, Rathnasinghe R, Stadlbauer D, Krammer F, Simon V et al (2021) The E484K mutation in the SARS-CoV-2 spike protein reduces but does not abolish neutralizing activity of human convalescent and post-vaccination sera. medRxiv. https://doi.org/10.1101/2021.01.26.21250543
79. Wang Z, Schmidt F, Weisblum Y, Muecksch F, Barnes CO, Finkin S et al (2021) mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. Nature 592(7855):616–622. https://doi.org/10.1038/s41586-021-03324-6
80. Xie X, Liu Y, Liu J, Zhang X, Zou J, Fontes-Garfias CR et al (2021) Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. Nat Med 27(4):620–621. https://doi.org/10.1038/s41591-021-01270-4

81. Xie X, Zou J, Fontes-Garfias CR, Xia H, Swanson KA, Cutler M et al (2021) Neutralization of N501Y mutant SARS-CoV-2 by BNT162b2 vaccine-elicited sera. bioRxiv. https://doi.org/10.1101/2021.01.07.425740
82. Collier DA, De Marco A, Ferreira I, Meng B, Datir R, Walls AC et al (2021) SARS-CoV-2 B.1.1.7 sensitivity to mRNA vaccine-elicited, convalescent and monoclonal antibodies. medRxiv. https://doi.org/10.1101/2021.01.19.21249840
83. Greaney AJ, Loes AN, Crawford KHD, Starr TN, Malone KD, Chu HY et al (2021) Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. Cell Host Microbe 29(3):463–76 e6. https://doi.org/10.1016/j.chom.2021.02.003
84. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN et al (2021) Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. Cell Host Microbe 29(1):44–57 e9. https://doi.org/10.1016/j.chom.2020.11.007
85. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J et al (2020) Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. MBio 11(4). https://doi.org/10.1128/mBio.01610-20
86. Letko M, Marzi A, Munster V (2020) Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat Microbiol 5(4):562–569. https://doi.org/10.1038/s41564-020-0688-y
87. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science 367(6485):1444–1448. https://doi.org/10.1126/science.abb2762
88. Zhou D, Duyvesteyn HME, Chen CP, Huang CG, Chen TH, Shih SR et al (2020) Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. Nat Struct Mol Biol 27(10):950–958. https://doi.org/10.1038/s41594-020-0480-y
89. Papachristodoulou E, Kakoullis L, Parperis K, Panos G (2020) Long-term and herd immunity against SARS-CoV-2: implications from current and past knowledge. Pathog Dis 78(3). https://doi.org/10.1093/femspd/ftaa025

Chapter 4
In Silico Drug Discovery for Treatment of Virus Diseases

Shikha Joon, Rajeev K. Singla, and Bairong Shen

Abstract Viral infections have remained a serious public health burden despite significant improvements in medical and pharmaceutical research in recent years. In silico approaches for drug discovery and design are fruitful for the management of a plethora of viral diseases. Virtual screening of libraries is performed using various computational tools to search for potential antiviral compounds. For this, a rational approach is used that comprises filtration of the screened compounds using docking, ligand- or pharmacophore-based similarity searches. The selected candidates are then tested in vitro to ascertain their biological activity. This minimizes the overall cost and time incurred in conventional drug designing methods. In this book chapter, we have discussed various methods of drug discovery and design, and their applications for the development of effective antiviral compounds. A descriptive methodology for the management of some common and notorious viral diseases is also outlined.

Keywords In silico drug discovery and design · Viral diseases · Virtual screening · Docking · Ligand · Pharmacophore · Antiviral compounds

Shikha Joon and Rajeev K. Singla have contributed equally to this work and share the first authorship.

S. Joon · R. K. Singla
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China

iGlobal Research and Publishing Foundation, New Delhi, India

B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_4

4.1 Introduction

Management of viral infections has been often onerous as it demands a longer treatment regime, which increases the economic burden, especially on the developing nations. Besides causing acute and chronic illnesses, these notorious bugs have the potential to give way to the pandemics affecting global health and economic stability [1]. The occurrence of high variations in the viral genomes is a major barrier to the development of efficient antiviral therapeutics. This genome variability is indispensable for the survival of the viruses as it confers adaptability and resistance to them against myriad antivirals. This necessitates a prompt and efficient drug discovery approach.

Drug discovery and development against various diseases (the viral diseases) is an arduous and cost-extensive process. To find a promising candidate, conventional drug discovery methods rely on the synthesis of various compounds usually large in number followed by their screening. In recent decades, in silico techniques such as virtual screening (VS) have accelerated and economized the process of drug discovery and development [2]. VS has a wide range of applications for the discovery and development of potential antivirals. This involves the identification of a target and lead structure followed by the development of a potential drug compound and its optimization.

The past couple of decades have witnessed an increased interest and efforts to use a combinational approach focused on in silico, chemical, biological approaches to expedite antiviral drug discovery and development [3]. In silico approaches have proven to be pivotal in deciphering the underlying molecular mechanisms, reveal potential drug targets, and their inhibitory compounds. These candidate molecules or lead compounds pave the way to the designing of novel and efficient anti-infectious drugs [4]. In silico approaches can either be used for de novo synthesis of novel drug candidates or optimization of ADMET (absorption, distribution, metabolism, excretion, and toxicity) profile of known drugs from various databases for a particular disease. In this chapter, we have discussed various methods of drug discovery and design, and their applications for the development of effective antiviral compounds. A descriptive methodology for the management of some common and notorious viral diseases is also outlined along with the most recent in silico-based antiviral drug discovery and development.

4.2 In Silico Drug Designing: Concepts and Methods

In silico drug designing is also known as virtual screening. This exploits computational models for describing the interactions between the macromolecules (most commonly these are proteins) and their respective ligands. This is accomplished by employing several computer-based methods. These fall under two main categories known as the two- and the three-dimensional approaches (2D and 3D methods).

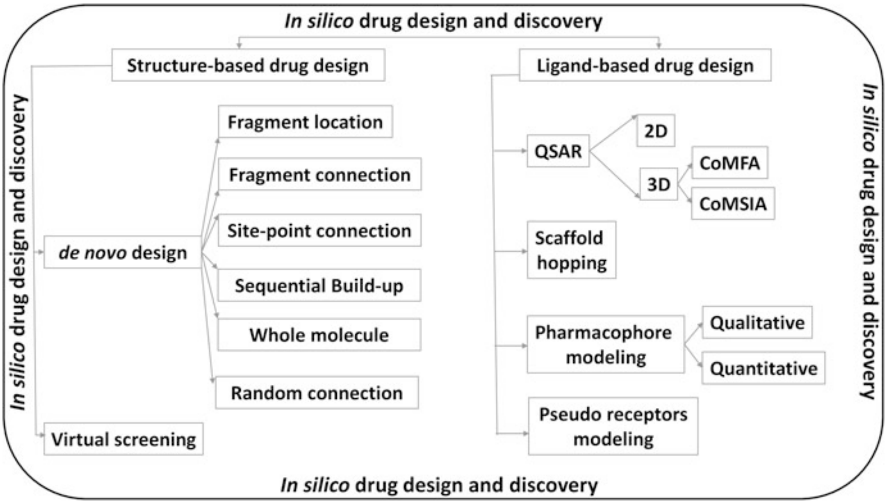

Fig. 4.1 An outline of principles underlying in silico drug discovery and design

These have different computational performance or efficiency and applications. Amongst these, the 2D approaches are most popularly exploited as "molecular filters" to minimize the potential molecules for downstream screening. This is credited to their much-reduced calculation times over 3D approaches. These are discussed in detail in the following sections and illustrated in Fig. 4.1.

4.2.1 2D/Descriptor-Based Approach

The methods based on this approach calculate and compare the scalar molecular properties. These are then used to identify molecules with similar molecular attributes (calculated). Machine-learning tools and techniques, for example, the linear correlation of experimentally determined biological activity and neural networks are used to select and weigh the relevant molecular descriptors for specific target binding.

4.2.2 3D/Conformation-Based Approach

The methods based on this approach are aimed at describing the chemical and steric complementarity of the target (macromolecule)-ligand binding for their 3D conformations. In recent times, 3D or conformation-based computational approaches are the methods of choice for antiviral drug discovery. These 3D models are developed using either the structure-based drug design or ligand-based drug design approaches.

The structure-based drug design (SBDD) requires information on the 3D structures of potential drug targets (macromolecules, usually, the proteins) whose binding pocket is known, as in the target-ligand co-crystallized structure. This information on the 3D structure of biological targets is determined experimentally using the technique, such as the X-ray crystallography, and archived in the Protein Data Bank (PDB) [5]. On the contrary, the ligand-based drug design (LBDD) approach works well when the 3D structural data of the biological targets are unavailable. Therefore, it relies on a set of known biomolecules or ligands that binds with the potential drug target for the disease under study. However, the predicted binding site (on the target) for these ligands mandates experimental validation using mutagenesis studies. This has an advantage as it may reveal a set of novel small molecule inhibitors for the target of interest (for a specific disease).

Both SBDD and LBDD (3D) approaches disseminate crucial information for lead optimization. The putative ligand (active) is positioned in the respective binding site on the target molecule using SBDD approaches. This enables the chemical optimization of the said compound. On the other hand, LBDD offers a range of compounds whose inhibitory action against similar targets is known. This allows analysis of critical aspects related to lead optimization. These approaches, their methodology, and their types are discussed in detail in the following sections.

4.2.3 Structure-Based Drug Design (SBDD)

As mentioned in the previous section, the methods that rely on this approach exploit the structural information of the intended drug target for a particular disease. This, in turn, allows for the development of potential inhibitors against that specific disease target. Here, the disease targets are usually the cellular receptors whose structure is to be known for developing potential inhibitors. The structural data are most commonly acquired using X-ray crystallography and/or Nuclear Magnetic Resonance (NMR).

4.2.3.1 Threading and Homology Modeling for SBDD

Threading and homology modeling is a pair of useful computational modeling techniques when the structural details of the intended drug target are absent. Threading enables modeling those target proteins that lack their counterparts (homologous proteins). In the absence of the homologous counterparts, the structural details are completely unknown. Here, a compatibility search is made for a specific amino acid sequence against the structures in a database. These structures have folds that are known, which, in turn, are used to build the structure of the protein in question (query).

Homology modeling is a comparative approach that exploits the sequence (amino acid sequence) homology between the query protein and its counterpart (homologous protein or template) whose structure is known. Here, one known structure of the homologous protein is sufficient [6]. Homology modeling is performed sequentially as follows: Firstly, the homologous protein in its 3D structural form is identified. This becomes the template. Secondly, the sequence alignment of the query (target protein) and the template is done. Next, the model for the query protein is generated followed by its refining and final validation using various computational tools [7, 8]. In recent years, this approach has gained much popularity for generating 3D models of the intended target proteins when the crystal structures are not available.

4.2.3.2 De Novo SBDD

This approach is based on the notion "from the beginning." Here, the active site of the potential drug target is characterized structurally. This includes the revelation of amino acid sequences that together form the active site and their orientation that confer ligand specificity to the binding site. This, in turn, is crucial for designing ligands that specifically bind to their respective drugs targets. The structural data for the active site of the target protein can be acquired and its analysis can be done using various computational approaches. These computational tools known as computer-aided ligand design (CALD) are of immense help in deciphering potential small molecule inhibitors (or ligands) from the available databases or de novo designing [9]. Reportedly, there are six main classes of CALD methods that comprise fragment connection methods, fragment location methods, random connection methods, sequential build-up methods, site point connection methods, and whole molecule methods [9]. A description of CALD methods is outlined in Table 4.1.

Table 4.1 Classification of CADD methods for SBDD

Method	Description
Fragment connection methods	Positioning the fragments and connecting them using either the "linkers" or "scaffolds". This is done to achieve a desired orientation of the positioned fragments.
Fragment location methods	The desired or favourable atom(s) or small fragment(s) locations (within the active site) are determined.
Sequential build-up methods	The ligand is constructed either atom by atom or fragment by fragment.
Random connection methods	This is a combination of sequential build-up and fragment connection methods, which introduces randomness using bond disconnection strategy.
Site-point connection methods	At first, the locations or site points are determined, followed by fragment placement in the active site. This is done to ensure that appropriate atoms occupy the determined locations.
Whole molecule methods	Placement of compounds (in different conformations) into the target's active site. This is done to assess target-ligand complementarity, in terms of shape and electrostatic force.

4.2.3.3 Structure-Based Virtual Screening (VS) for SBDD

This is the most sought-after SBDD approach (3D) for lead identification that complements high-throughput screening (HTS). It comprises molecular docking and scoring functions. Molecular docking predicts the binding mode of ligand (assumed to be flexible) to its specific binding site (assumed to be highly rigid) on the target whose 3D structure is known while target-ligand binding affinity is measured using scoring. Here, the database search for plausible target binding compounds is carried out. These compounds are then ranked to obtain a potential target-specific subset, which is then proceeded for further biological assays [6, 10, 11]. For this, various docking tools have been developed with different algorithms for ligand placement and scoring [12]. These include AutoDock [13], DOCK [14], FlexX [15], FRED [16], GLIDE [17], GOLD [18], LigandFit [19], Surflex [20], and GRIP docking [21, 22] (Table 4.2). There is considerable data on the application of structure-based VS that relies on molecular docking [23–29]. Structure-based VS has accelerated the process of drug discovery and development with improved efficiency [30]. Ligand placement is largely reliable for assessing various possibilities for the binding of ligand to the intended target (and their geometries). But, the scoring functions cannot be relied upon for all the targets. This could occur due to its incapacity to correctly estimate the effects of entropy [31]. This could, however, be overcome by developing manual and customized scoring functions (empirical) for specific applications. This approach is considered amongst the most promising methods for rational lead optimization as it allows to visualize the hypothetical target-ligand binding (docking). Even though there are crude energy calculations involved, it offers ready availability of the compounds from various libraries. This makes in vitro biological assays less challenging and generates fewer false positives [11].

Table 4.2 Molecular docking tools for structure-based virtual screening

Tool	Description/algorithm used	Reference
FlexX	Based on ligand fragmentation and incremental reconstruction	Rarey et al. [15]
DOCK FRED	Based on molecular shape algorithms	Ewing et al. [14] McGann et al. [16]
GOLD AutoDock	Based on genetic algorithms	Jones et al. [18] Morris et al. [13]
Glide	Based on systematic search algorithms	Halgren et al. [17]
LigandFit	Based on monte Carlo optimization	Venkatachalam et al. [19]
Surflex	Based on surface-based molecular similarity	Jain et al. [20]
GRIP docking	Based on PLP scoring function	Singla et al. [21] Singla and Dubey [22]

4.2.4 Ligand-Based Drug Design (LBDD)

As mentioned in the previous sections, the LBDD approach works well when the 3D structural data of the biological targets are unavailable. It, therefore, solely relies on a set of known biomolecules or ligands that binds with the potential drug target for the disease under study. This is accomplished using 3D quantitative structure-activity relationships (3D QSAR) and pharmacophore modeling tools. Here, predictive models are generated that enable lead identification and optimization [32].

4.2.4.1 Quantitative Structure-Activity Relationship (QSAR) for LBDD

The process of quantitatively correlating molecular descriptors with functions for a group of similar compounds is known as QSAR. Here, the structural properties of a molecule are referred to as the "molecular descriptors" or "descriptors" while "functions" is the term used for the biological activities, physicochemical attributes, toxicity, etc. [33]. QSAR relies on the notion that a molecule's structure possesses the attributes (electronic, geometric, and steric properties) that confer specific activities to it (physical, chemical, and biological properties) [34]. This approach strives to study a series of molecules with varied structures and properties to deduce a structure-property relationship empirically. As mentioned in the earlier sections, this is a promising alternative approach for LBDD when the structural data is not available. The QSAR approach comprises methods that correlate the structure of a molecule with experimental data concerning its biological activities. These methods have distinct nomenclature based on the data collected. For example, the quantitative structure-property relationship (QSPR) includes methods that model physicochemical properties while the quantitative structure-toxicity relationship (QSTR) is used to correlate molecular structure with toxicological data [35].

4.2.4.2 Pseudoreceptor Modeling for LBDD

Here, the structures of known bioactive or ligands are exploited for reconstructing the 3D structure of an unknown target. This is a novel concept in computer-aided drug discovery and development (CADD) that generates a precise and explicit model of a target or receptor. The receptor model so generated can be utilized for affinity predictions and to generate other models [36]. These are classified as 3D QSAR methods that link LBDD with the receptor or target-based drug design [37].

4.2.4.3 Pharmacophore Mapping/Modeling for LBDD

The concept of "pharmacophore" was introduced by Paul Ehrlich in the 1900s, which is an amalgamation of *phoros* (the carrier or bearer) and *pharmacon* (drug) [38]. So, pharmacophore is a part of a molecular framework bearing essential attributes or structural features that confer specific biological or pharmacological properties to a drug, such as receptor recognition and binding [38, 39].

It strives to achieve the following:

(a) To find the essential features or attributes needed to confer a specific biological activity.
(b) To determine the bioactive or the molecular conformation needed.
(c) To develop a rule for the alignment or superposition for the compound series.

These are achieved through a sequential computational approach that comprises a selection of the drug target, preparation of database, generation of the pharmacophore model, and 3D screening [40–44]. Some automated programs have been developed for pharmacophore mapping, such as Catalyst [41, 45], DISCO [46], GASP [47], LigandScout [48–50], MOE [51, 52], and Phase [53]. Table 4.3 describes these automated pharmacophore mapping programs. In particular, these automated programs recognize interaction sites that include hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic (H), ionizable/negative ionic (N), ionizable/positive ionic (P), and aromatic rings (R). These develop the pharmacophore models by exploiting the "activity" attribute to differentiate the actives from the non-actives. The pharmacophore models that exhibit inactivity are discarded. In recent times, the pharmacophore-based VS has become an essential tool for hit identification in drug designing studies in the absence of the target's 3D structural information. Also, it offers prompt screening to identify potential compounds from a library of numerous compounds. Moreover, these generate descriptive and highly transparent models (or 3D pharmacophore representations) that are amenable to modifications with easy interpretation of results. However, these fail to completely

Table 4.3 Automated pharmacophore mapping programs

Tool	Description	Reference
LigandScout	The pharmacophore is derived from complexes formed by the protein and the ligand. It interprets geometries of the ligand. The accurate hybridization states are assigned. The possible protein-ligand interactions are classified by applying a defined set of rules.	Wolber and Langer [49], Wolber et al. [50], Seidel et al. [48]
Catalyst, GASP, Phase, MOE	The superpositioning of pharmacophore molecules is done using cascading n-point pharmacophore fingerprints.	Hecker et al. [45], Güner et al. [41], Jones et al. [47], Dixon et al. [53], Halgren [52], Cheong et al. [51]

address the biophysical interactions of the drug in question, which is a limitation of this approach [54–60].

4.2.4.4 Scaffold Hopping for LBDD

Using this approach, novel compounds with structural diversity, but shared specific biological attributes are found [61]. Here, the structural skeleton or the chemical core structure of a scaffold is altered while the biological activity (or interaction properties in the 3D space) is preserved [62]. This technique aids the identification of novel and structurally diverse classes of compounds as potential inhibitors of the target protein [63, 64]. The automated scaffold hopping methods include Recore [65] (http://www.biosolveit.de/), BROOD (http://www.eyesopen.com/), SHOP (Scaffold Hopping) [62] (http://www.moldiscovery.com/), and MOE (Molecular Operating Environment) (http://www.chemcomp.com/).

4.3 The Journey of In Silico Drug Discovery and Design for the Treatment of Viral Diseases

In silico drug discovery and designing methods had a remarkable journey for the treatment and management of viral diseases. These include viruses for AIDS (acquired human deficiency syndrome), dengue, hepatitis C, influenza, mononucleosis, severe acute respiratory syndrome (SARS-CoV1 and SARS-CoV2), West Nile fever, and yellow fever. The genome of these viruses encodes structural and non-structural proteins. There are three structural proteins, namely the core and envelope proteins (E1 and E2, fusion) while the non-structural or enzymatic components include NS1, NS2A, and NS2B, NS3, NS4A, and NS4B, and NS5 proteins. Together, these facilitate replication and assembly of the virion [66]. As these participate in crucial activities needed for viral fusion, replication, and virion assembly, they make up attractive antiviral targets. Different databases are screened to discover potential antiviral compounds whose antiviral activity was previously not known. These antiviral compounds are specific to the known or novel targets of the causative viral agent in question (Table 4.4). The proceeding literature provides an overview of the journey of in silico drug discovery and design for the treatment of viral diseases. For example, the neuraminidase (NA) enzyme of the influenza virus is a potential drug target. This is known to be involved in the viral transportation into the host [67]. To this end, the shape-based virtual screening (VS) approach was adopted by Kirchmair et al., wherein five potential katsumadain A (NA inhibitor) analogs were identified from the National Cancer Institute (NCI) compound database (http://cactus.nci.nih.gov/download/nci/) [68, 69]. These were further validated by in vitro chemiluminescence and cytotoxicity assays. Another potential antiviral drug target is reverse transcriptase (RT) of the Human Immunodeficiency Virus (HIV-1). RT

Table 4.4 An overview of in silico drug discovery and design for the treatment of viral diseases

Disease	Causative organism	Family	Potential target	Antiviral identified	In silico intervention		In vitro assays	PMID
					Methodology and software	Database Screened		
Influenza	Influenza viruses	Orthomyxoviridae	Neuraminidase (NA)	Katsumadain A analogs	Shape-based VS: ROCS, Protein-ligand docking: GOLD	NCI	Chemiluminescence-based NA inhibition & cell cytotoxicity assay	**21452980**
AIDS	HIV	Retroviridae	Reverse transcriptase (RT)	RT inhibitor, NNRTIs	High throughput docking with physicochemical filters: FRED & FILTER, VS: GLIDE	NCI, ZINC, Maybridge screening collection	Non-radioactive colorimetric assay, cell cytotoxicity assay	**19703027, 19374380**
			Reverse transcriptase	DHBNH analogs	Shape-based VS: ROCS, 2D-fingerprint similarity: ECFP method, ligand-based pharmacophore model & screening: LigandScout, Catalyst, 3D-pharmacophore VS: GRID	NCI	RNase H polymerase-independent cleavage assay & RdDp activity assay	**22361685**
			HIV-1 Integrase	BAS-044249 (Asinex code)	Receptor-based pharmacophore search	Asinex Gold Collection	Integrase & antiviral activity assays	**19447621**
			gp120-CD4 binding	gp120-CD4 inhibitors	Docking: GOLD, shape-based similarity search: ROCS	ZINC version 7	Cell-based infectivity assays	**21169023**

			CTD of Capsid protein	NYAD-1: peptide-based CTD inhibitor	Docking-based VS and analog search: GLIDE	ZINC	Cell-based antiviral activity & cytotoxicity assays	**21168336**
			CXCR4 receptor (Human anti-HIV target)	CXCR4 receptor inhibitor	Library construction, Ligand-Based Pharmacophore Modeling: MOE & Discovery Studio, Ligand-based shape matching VS: PARAFIT, ROCS & HEX, Receptor-based VS: AUTODOCK, GOLD, FRED, & HEX 4.8	Maybridge	Not available	**19358515**
Hepatitis C	Hepatitis C virus	Flaviviridae	NS3-4A serine protease	Novel NS3-4A inhibitors	GENIUS: an induced-fit docking program was developed	MDL available chemical directory 2005 (ACD)	Protease inhibition activity, replicon & ATP assays	**21992802**
			NS5 polymerase	NS5 inhibitors	Pharmacophore modeling (3D): Cerius2	Asinex commercial database	RdRp activity, replicon & cytotoxicity assays	**21531135**
			NS5B polymerase	Isoxazole analogs	HTS & ligand-based docking: LigandFit	Maybridge screening collection	RdRp activity, replicon & MTS cell cytotoxicity assays	**21281131**
			NS5B polymerase: AP-1 pocket (binds tetracyclic indole inhibitor)	Rhodanine & imidazo-coumarin analogs	Docking, structure-based VS & HTVS: GLIDE	ChemBridge	RdRp inhibition assay	**20627595**

(continued)

Table 4.4 (continued)

Disease	Causative organism	Family	Potential target	Antiviral identified	In silico intervention		In vitro assays	PMID
					Methodology and software	Database Screened		
			NS5B polymerase	NS5B inhibitors	3D QSAR with ligand- & structure-based approach: Surflex-Sim & Autodock 4	NCI Diversity Set	NS5B RdRp assay & binding mode analysis	**20225870**
Dengue	Dengue virus	Flaviviridae	E-protein: ligand binding pocket	Commercially-purchased compounds	Docking-based VS: GOLD (v.2.1)	Maybridge chemical database	MTT cell cytotoxicity, antiviral inhibition, & fusion inhibition assays	**19781577**
			E-protein: hydrophobic pocket	E-protein inhibitor	High-throughput docking VS: GLIDE (v.2.7)	–	CFI, cell viability & PRA assays, time-of-addition & immunofluorescence studies, endosomal pH evaluation & compound virus binding assay	**19223625**
			E-protein: βOG-binding pocket	E-protein inhibitor	Focused library construction using VS: generation of Novartis in-house compound docking set, docking: GLIDE SP	Novartis in-house natural product collection	Virus-liposome fusion, plaque & antiviral activity assays, quantitative real-time RT-PCR, cell cytotoxicity assay	**19800368**
			E-protein	E-protein inhibitor	VS: DOCK v5.1.1, GOLD v2.1.2, FlexX v2.2.0	Commercial databases (13)	MTT cell viability & plaque reduction assays	**19241120**

			NS5 RNA methyltransferase	NS5 inhibitors	Molecular docking: Chimera, ligand preparation: LigPrep, VS & ligand docking: GLIDE v.4.5, generation of pharmacophore & database searching: Phase v.2.5	ZINC v.5, CACDB (commercial database)	Methyltransferase activity assays, CF-1 & cytotoxicity assays	**20108931**
Yellow fever	Yellow fever virus	Flaviviridae	E-protein: βOG-binding pocket	E-protein inhibitors	HTVS: Ligandinfo, Structure-based virtual screening: GLIDE v5.5	Ligandinfo (a meta-database)	Not available	**21369890**
West Nile fever	West Nile virus	Flaviviridae	NS3: NS2B cofactor binding site	NS3 inhibitors	Docking-based VS	NCI	Fluorescence-based proteinase, cell cytotoxicity & replicon assay	**21050032**
Mononucleosis	*Epstein-Barr virus*	Herpes virus family (herpesviridae)	EBNA1: DNA binding activity inhibition	EBNA1-DNA binding inhibitors	HTVS: DOCK v.4.0	SPECS, ZINC	FP assay & EMSA	**20405039**
Severe acute respiratory syndrome	SARS-CoV&	Coronaviridae	3CLpro	3CLpro inhibitors	Structure-based VS: AutoDock version 3.0.5, ligand and structure-based VS	Chembridge, Asinex Platinum collection	*In vitro* enzyme inhibition assays, whole-cell CPE assay	**21470860, 21604711**

(continued)

Table 4.4 (continued)

Disease	Causative organism	Family	Potential target	Antiviral identified	In silico intervention		In vitro assays	PMID
					Methodology and software	Database Screened		
			RNA pseudoknot in the -1 RF site	Small molecule RNA pseudoknot inhibitors and -1 RF disruptors	Construction of a 3D-SARS-pseudoknotstructural model & VS: DOCK 4.0	LeadQuest	Cell-based -1 RF assay	**21591761**
	SARS-CoV2	Coronaviridae	3CLpro	3CLpro inhibitors	Molecular docking-based VS: AutoDock Vina	ZINC & FDA approved drugs	–	**33392261**

Abbreviations: 3D, three-dimensional; βOG, n-octyl-beta-D-glucoside; CF-1, cell-based flavivirus immunodetection; CLpro, chymotrypsin-like protease; CPE, cytopathic effects; CTD, carboxy-terminal domain; DHBNH, dihydroxy benzoyl naphthyl hydrazone; EBNA1, Epstein-Barr nuclear antigen 1; ECFP, extended-connectivity fingerprints; EMSA, electrophoresis mobility shift assay; E-protein, envelope protein; FP, fluorescence polarization; HIV-1, human immunodeficiency virus-1; HTVS, high-throughput virtual screening; MTT, methyl thiazole tetrazolium; MTS, 3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophe-nil)-2H-tetrazolium; NCI, national cancer institute database; NNRTIs, non-nucleoside inhibitors of HIV-1 reverse transcriptase; PRA, plaque reduction assay; QSAR, quantitative structure-activity relationship; RdDp, RNA-dependent DNA polymerase; RdRp, RNA-dependent RNA polymerase; RF, ribosomal frameshifting; ROCS, rapid overlay of chemical structures; RT-PCR, reverse transcription-polymerase chain reaction; SARS-COV, Severe acute respiratory syndrome-coronavirus is involved in the reverse transcription mechanism, that is, from RNA to DNA. Various in silico approaches have been adopted to search for or design potential RT inhibitors. These include high-throughput docking with physicochemical filters (FRED and FILTER) from the NCI database. In this study, four best hits from amongst 2800 compounds were found to possess anti-RT activity in vitro [70]. Further, docking-based VS (by GLIDE) of ZINC [71] and Maybridge screening collection (https://www.thermofisher.com) revealed nine-hit compounds. Amongst these, three compounds demonstrated anti-RT activity in vitro at low-micromolar concentrations [72]. Further, a shape-based VS screening of the NCI database for dihydroxy benzoyl naphthyl hydrazine (DHBNH) analogs was performed. A consecutive ligand-based VS was conducted for the best hits using ligand-based pharmacophore screening and 3D-, 2D-similarity searches. In vitro, anti-RT activity evaluation of the compounds bearing novel scaffolds revealed remarkable antiviral activity [54, 55]. In another study, novel thiohydantoin, thiobarbituric acid, and rhodamine-based analogs were explored from the Asinex gold collection (http://www.asinex.com/, a commercial database) against the integrase (IN) enzyme of HIV-1. IN is known to aid the integration of viral DNA with the host cell genome. The receptor-based pharmacophore search revealed BAS-044249 (Asinex code), which demonstrated significant anti-IN inhibitory activity in vitro [73]. The entry of HIV-1 into the host system can also be abrogated via inhibition of binding of glycoprotein 120 (gp120) to the CXCR4 or CCR5 chemokine co-receptors and CD4 receptor. Accordingly, novel gp120 inhibitors were identified using a combination of docking and shape-based similarity search (ZINC database) approaches [74]. Furthermore, novel small molecule anti-capsid protein (CTD, C-terminal domain) inhibitors were identified using a docking-based VS (ZINC database) and analog search (GLIDE). Capsid protein is crucial for the assembly and maturation of HIV-1. The best hit compounds demonstrated significant anti-HIV-1 activity in vitro [75]. Finally, human targets were also explored as a promising antiviral therapeutic strategy. This strategy is based on the notion that viral escape mutants are capable of rendering the antiviral drugs inactive, which can be overcome by including those host protein targets that are essential for the replication of the virus. These, however, are nonessential for the host. The human CXCR4 receptor is a promising target to prohibit virus entry into the host system. For this, a combined structure and ligand-based VS workflow were developed and validated that targeted the CXCR4 receptor. Five potential hits were generated from the Maybridge database that required synthesis and experimental testing [76].

The Flaviviridae family comprises viruses that majorly cause hepatitis C, dengue, yellow fever, and West Nile fever [77]. The potential drug targets include NS3-4A serine protease, NS5 or NS5B polymerase, and envelope- or E-protein. GENIUS, a novel induced-fit docking method was developed by Takaya et al., which generated 13 new inhibitors of hepatitis C virus (HCV) NS3-4A protease. Their HCV inhibitory activity was ascertained by in vitro assays [78]. Further, pharmacophore modeling (3D)-based VS was performed using Cerius2 that sought NS5 polymerase as the key target for the discovery of novel anti-HCV inhibitors. This yielded a potential anti-NS5 polymerase inhibitor in the activity assays in vitro [79]. Furthermore, the structure-based VS of the Maybridge database and in vitro revealed an isoxazole analog as potential anti-NS5B polymerase inhibitors [80]. Using a structure-based VS approach followed by compound synthesis and structure-activity relationship, the rhodanine, and imidazocoumarin analogs were identified as promising anti-HCV NS5B polymerase inhibitors from the ChemBridge database. These identified compounds led to allosteric inhibition of the NS5B polymerase upon binding to the enzyme's AP-1 pocket as confirmed by the in vitro RdRp inhibitory assay [81]. In another interesting study, a combined 3D QSAR approach that comprised ligand- and structure-based VS was used to search for novel HCV NS5B polymerase inhibitory compounds from the NCI database. This revealed a potential HCV NS5B polymerase inhibitor [82]. There is considerable evidence from the docking studies on the potential antiviral candidature of the E-protein of the dengue virus (DNV). These studies screened the commercial databases or built their customized library for discovering novel DNV inhibitors that targeted ligand-binding, hydrophobic, or the βOG pocket of E-protein [77, 83–85]. In an interesting study, novel DENV NS5 polymerase inhibitors were discovered using in vitro-driven VS of commercial databases (ZINC and CACDB) [86]. Novel anti-E protein compounds against the yellow fever virus (YFV) were explored using docking studies by Umamaheswari et al. [87]. Ligand-based VS of NCI revealed potential allosteric inhibitors of NS3 proteinase of the West Nile virus (WNV) [88]. A high-throughput VS for anti-Epstein-Barr nuclear antigen 1 (EBNA1) inhibitors against Epstein-Barr virus (EBV) led to the discovery of four selective inhibitory compounds [89].

The spike protein (S protein) and 3-chymotrypsin-like protease (3CLpro) are attractive drug targets against severe acute respiratory syndrome caused by SARS-CoV (coronavirus) and SARS-CoV-2 [68, 69]. Structure- and ligand-based VS identified potential 3CLpro inhibitors against SARS from the commercial databases. In vitro assays confirmed their candidature as promising anti-SARS inhibitory compounds [90–92]. In a study by Park and coworkers, the structure-based VS identified RNA pseudoknot-binding ligands that potentially inhibited −1 ribosomal frameshifting (RF) of SARS-CoV [93].

4.4 Conclusion

In silico methods have proved their merit in the discovery and design of novel drugs, particularly, for the treatment of viral diseases. In silico approaches comprise rapid and efficient tools that score over the conventional drug discovery methods for the management of viral infections that are accompanied by resistance to almost any commercial drug. VS, for example, is an essential component of computer-assisted drug discovery and development that is mainly employed for early lead discovery to expedite the process. Despite the several advantages being offered by in silico approaches, the drug development process demands improvisations, particularly, to combat the emerging and re-emerging viral diseases. Moreover, efforts are needed to enhance the accuracy of these in silico methods, which in turn are expected to improve the quality of the procured experimental data. To this end, an amalgamation of in silico and chemical-biological methods can be sought as an efficient approach.

Acknowledgement This work was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019), and the National Natural Science Foundation of China (32070671).

References

1. Neumann G, Noda T, Kawaoka Y (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature 459:931–939
2. McInnes C (2007) Virtual screening strategies in drug discovery. Curr Opin Chem Biol 11:494–502
3. Kapetanovic IM (2008) Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact 171:165–176
4. Shaikh SA, Jain T, Sandhu G et al (2007) From drug target to leads–sketching a physicochemical pathway for lead molecule design in silico. Curr Pharm Des 13:3454–3470
5. Berman HM, Battistuz T, Bhat TN et al (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899–907
6. Kroemer RT (2007) Structure-based drug design: docking and scoring. Curr Protein Pept Sci 8:312–328
7. Cavasotto CN, Phatak SS (2009) Homology modeling in drug discovery: current trends and applications. Drug Discov Today 14:676–683
8. Krieger E, Nabuurs SB, Vriend G (2003) Homology modeling. Methods Biochem Anal 44:509–523
9. Murcko MA (2007) In: Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry, vol 11. John Wiley & Sons, Inc., Hoboken, NJ, pp 1–67
10. Lyne PD (2002) Structure-based virtual screening: an overview. Drug Discov Today 7:1047–1055
11. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432:862–865
12. Kalyaanamoorthy S, Chen YP (2011) Structure-based drug design to augment hit discovery. Drug Discov Today 16:831–839
13. Morris GM, Goodsell DS, Halliday RS et al (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J. Comput Chem 19:1639–1662
14. Ewing TJ, Makino S, Skillman AG et al (2001) DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 15:411–428
15. Rarey M, Kramer B, Lengauer T et al (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
16. McGann MR, Almond HR, Nicholls A et al (2003) Gaussian docking functions. Biopolymers 68:76–90
17. Halgren TA, Murphy RB, Friesner RA et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47:1750–1759
18. Jones G, Willett P, Glen RC et al (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
19. Venkatachalam CM, Jiang X, Oldfield T et al (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. J Mol Graph Model 21:289–307

20. Jain AN (2003) Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 46:499–511
21. Singla RK, Ali M, Kamal MA, Dubey AK (2018) Isolation and characterization of nuciferoic acid, a novel keto fatty acid with hyaluronidase inhibitory activity from *Cocos nucifera* Linn. endocarp. Curr Top Med Chem 18(27):2367–2378
22. Singla RK, Dubey AK (2019) Phytochemical profiling, GC-MS analysis and α-amylase inhibitory potential of ethanolic extract of *Cocos nucifera* Linn. endocarp. Endocr Metab Immune Disord Drug Targets 19:419–442
23. Finn J (2012) Application of SBDD to the discovery of new antibacterial drugs. Methods Mol Biol 841:291–319
24. Jenwitheesuk E, Samudrala R (2005) Virtual screening of HIV-1 protease inhibitors against human cytomegalovirus protease using docking and molecular dynamics. AIDS 19:529–531
25. Kuck D, Singh N, Lyko F et al (2010) Novel and selective DNA methyltransferase inhibitors: docking-based virtual screening and experimental evaluation. Bioorg Med Chem 18:822–829
26. Pierri CL, Parisi G, Porcelli V (2010) Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. Biochim Biophys Acta 1804:1695–1712
27. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. J Comput Aided Mol Des 16:151–166
28. Wang D, Wang F, Tan Y et al (2012) Discovery of potent small molecule inhibitors of DYRK1A by structure-based virtual screening and bioassay. Bioorg Med Chem Lett 22:168–171
29. Waszkowycz B (2002) Structure-based approaches to drug design and virtual screening. Curr Opin Drug Discov Devel 5:407–413
30. Ghosh S, Nie A, An J, Huang Z (2006) Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol 10:194–202
31. Warren GL, Andrews CW, Capelli AM et al (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
32. Acharya C, Coop A, Polli JE et al (2011) Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. Curr Comput Aided Drug Des 7:10–22
33. Merz KM Jr, Ringe D et al (2010) Drug design. Cambridge University Press
34. Eleni P, Dimitra HL (2003) Review in quantitative structure activity relationships on lipoxygenase inhibitors. Mini Rev Med Chem 3:487–499
35. Winkler DA (2002) The role of quantitative structure–activity relationships (QSAR) in biomolecular discovery. Brief Bioinform 3:73–86
36. Vedani A (1994) Pseudoreceptor modeling - a tool in the pharmacological screening process. ALTEX 11:11–21
37. Tanrikulu Y, Schneider G (2008) Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening. Nat Rev Drug Discov 7:667–677
38. Ehrlich P (1909) Present status of chemotherapy. Chem Ber 42:17–47
39. Gund P (1977) Three-dimensional pharmacophoric pattern searching. Prog Mol Subcell Biol 11:117–143
40. Caporuscio F, Tafi A (2011) Pharmacophore modelling: a forty year old approach and its modern synergies. Curr Med Chem 18:2543–2553
41. Güner O, Clement O, Kurogi Y (2004) Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. Curr Med Chem 11:2991–3005
42. Kim KH, Kim ND, Seong BL (2010) Pharmacophore-based virtual screening: a review of recent applications. Expert Opin Drug Discov 5:205–222
43. Melagraki G, Afantitis A (2011) Ligand and structure based virtual screening strategies for hit-finding and optimization of hepatitis C virus (HCV) inhibitors. Curr Med Chem 18:2612–2619
44. Sun H (2008) Pharmacophore-based virtual screening. Curr Med Chem 15:1018–1024

45. Hecker EA, Duraiswami C, Andrea TA et al (2002) Use of catalyst pharmacophore models for screening of large combinatorial libraries. J Chem Inf Comput Sci 42:1204–1211
46. Lin SK (2000) Pharmacophore perception, development and use in drug design. Edited by Osman F. Güner. Molecules 5:987–989
47. Jones G, Willett P, Glen RC (2000) In: Güner OF (ed) Pharmacophore perception, development and use in drug design. International University Line, La Jolla, CA, pp 85–86
48. Seidel T, Ibis G, Bendix F et al (2010) Strategies for 3D pharmacophore-based virtual screening. Drug Discov Today Technol 7:e221–e228
49. Wolber G, Langer T (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160–169
50. Wolber G, Dornhofer AA, Langer T (2006) Efficient overlay of small organic molecules using 3D pharmacophores. J Comput Aided Mol Des 20:773–788
51. Cheong SL, Federico S, Venkatesan G et al (2011) Pharmacophore elucidation for a new series of 2-aryl-pyrazolo-triazolo-pyrimidines as potent human A3 adenosine receptor antagonists. Bioorg Med Chem Lett 21:2898–2905
52. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17:490–519
53. Dixon SL, Smondyrev AM, Rao SN (2006) PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem Biol Drug Des 67:370–372
54. Distinto S, Yáñez M, Alcaro S et al (2012a) Synthesis and biological assessment of novel 2-thiazolylhydrazones and computational analysis of their recognition by monoamine oxidase B. Eur J Med Chem 48:284–295
55. Distinto S, Esposito F, Kirchmair J et al (2012b) Identification of HIV-1 reverse transcriptase dual inhibitors by a combined shape-, 2D-fingerprint- and pharmacophore-based virtual screening approach. Eur J Med Chem 50:216–229
56. Noha SM, Atanasov AG, Schuster D et al (2011) Discovery of a novel IKK-β inhibitor by ligand-based virtual screening techniques. Bioorg Med Chem Lett 21:577–583
57. Noha SM, Jazzar B, Kuehnl S et al (2012) Pharmacophore-based discovery of a novel cytosolic phospholipase A(2)α inhibitor. Bioorg Med Chem Lett 22:1202–1207
58. Schuster D, Markt P, Grienke U et al (2011a) Pharmacophore-based discovery of FXR agonists. Part I: Model development and experimental validation. Bioorg Med Chem 19:7168–7180
59. Schuster D, Kowalik D, Kirchmair J et al (2011b) Identification of chemically diverse, novel inhibitors of 17β-hydroxysteroid dehydrogenase type 3 and 5 by pharmacophore-based virtual screening. J Steroid Biochem Mol Biol 125:148–161
60. Waltenberger B, Wiechmann K, Bauer J et al (2011) Pharmacophore modeling and virtual screening for novel acidic inhibitors of microsomal prostaglandin E2 synthase-1 (mPGES-1). J Med Chem 54:3163–3174
61. Tsunoyama K, Amini A, Sternberg MJ et al (2008) Scaffold hopping in drug discovery using inductive logic programming. J Chem Inf Model 48:949–957
62. Bergmann R, Linusson A, Zamora I (2007) SHOP: scaffold HOPping by GRID-based similarity searches. J Med Chem 50:2708–2717
63. Renner S, Schneider G (2006) Scaffold-hopping potential of ligand-based similarity concepts. ChemMedChem 1:181–185
64. Zhao H (2007) Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug Discov Today 12:149–155
65. Maass P, Schulz-Gasch T, Stahl M et al (2007) Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. J Chem Inf Model 47:390–399
66. Fusco DN, Chung RT (2012) Novel therapies for hepatitis C: insights from the structure of the virus. Annu Rev Med 63:373–387
67. Gong J, Xu W, Zhang J (2007) Structure and functions of influenza virus neuraminidase. Curr Med Chem 14:113–122

68. Kirchmair J, Rollinger JM, Liedl KR et al (2011a) Novel neuraminidase inhibitors: identification, biological evaluation and investigations of the binding mode. Future Med Chem 3:437–450
69. Kirchmair J, Distinto S, Liedl KR et al (2011b) Development of anti-viral agents using molecular modeling and virtual screening techniques. Infect Disord Drug Targets 11:64–93
70. Bustanji Y, Al-Masri IM, Qasem A et al (2009) In silico screening for non-nucleoside HIV-1 reverse transcriptase inhibitors using physicochemical filters and high-throughput docking followed by in vitro evaluation. Chem Biol Drug Des 74:258–265
71. Irwin JJ, Shoichet BK (2005) ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
72. Nichols SE, Domaoal RA, Thakur VV et al (2009) Discovery of wild-type and Y181C mutant non-nucleoside HIV-1 reverse transcriptase inhibitors using virtual screening with multiple protein structures. J Chem Inf Model 49:1272–1279
73. Rajamaki S, Innitzer A, Falciani C et al (2009) Exploration of novel thiobarbituric acid-, rhodanine- and thiohydantoin-based HIV-1 integrase inhibitors. Bioorg Med Chem Lett 19:3615–3618
74. Lalonde JM, Elban MA, Courter JR et al (2011) Design, synthesis and biological evaluation of small molecule inhibitors of CD4-gp120 binding based on virtual screening. Bioorg Med Chem 19:91–101
75. Curreli F, Zhang H, Zhang X et al (2011) Virtual screening based identification of novel small-molecule inhibitors targeted to the HIV-1 capsid. Bioorg Med Chem 19:77–90
76. Pérez-Nueno VI, Pettersson S, Ritchie DW et al (2009) Discovery of novel HIV entry inhibitors for the CXCR4 receptor by prospective virtual screening. J Chem Inf Model 49:810–823
77. Yennamalli R, Subbarao N, Kampmann T et al (2009) Identification of novel target sites and an inhibitor of the dengue virus E protein. J Comput Aided Mol Des 23:333–341
78. Takaya D, Yamashita A, Kamijo K et al (2011) A new method for induced fit docking (GENIUS) and its application to virtual screening of novel HCV NS3-4A protease inhibitors. Bioorg Med Chem 19:6892–6905
79. Kim ND, Chun H, Park SJ et al (2011) Discovery of novel HCV polymerase inhibitors using pharmacophore-based virtual screening. Bioorg Med Chem Lett 21:3329–3334
80. Lin YT, Huang KJ, Tseng CK et al (2011) Efficient in silico assay of inhibitors of hepatitis C Virus RNA-dependent RNA polymerase by structure-based virtual screening and in vitro evaluation. Assay Drug Dev Technol 9:290–298
81. Talele TT, Arora P, Kulkarni SS et al (2010) Structure-based virtual screening, synthesis and SAR of novel inhibitors of hepatitis C virus NS5B polymerase. Bioorg Med Chem 18:4630–4638
82. Musmuca I, Caroli A, Mai A et al (2010) Combining 3-D quantitative structure-activity relationship with ligand based and structure based alignment procedures for in silico screening of new hepatitis C virus NS5B polymerase inhibitors. J Chem Inf Model 50:662–676
83. Kampmann T, Yennamalli R, Campbell P et al (2009) In silico screening of small molecule libraries using the dengue virus envelope E protein has identified compounds with antiviral activity against multiple flaviviruses. Antivir Res 84:234–241
84. Poh MK, Yip A, Zhang S et al (2009) A small molecule fusion inhibitor of dengue virus. Antivir Res 84:260–266
85. Wang QY, Patel SJ, Vangrevelinghe E et al (2009) A small-molecule dengue virus entry inhibitor. Antimicrob Agents Chemother 53:1823–1831
86. Podvinec M, Lim SP, Schmidt T et al (2010) Novel inhibitors of dengue virus methyltransferase: discovery by in vitro-driven virtual screening on a desktop computer grid. J Med Chem 53:1483–1495
87. Umamaheswari A, Kumar MM, Pradhan D et al (2011) Docking studies towards exploring antiviral compounds against envelope protein of yellow fever virus. Interdiscip Sci 3:64–77
88. Shiryaev SA, Cheltsov AV, Gawlik K et al (2011) Virtual ligand screening of the National Cancer Institute (NCI) compound library leads to the allosteric inhibitory scaffolds of the West Nile Virus NS3 proteinase. Assay Drug Dev Technol 9:69–78

89. Li N, Thompson S, Schultz DC et al (2010) Discovery of selective inhibitors against EBNA1 via high throughput in silico virtual screening. PLoS One 5:e10126
90. Abdusalam AAA, Murugaiyah V (2020) Identification of potential inhibitors of 3CL protease of SARS-CoV-2 from ZINC database by molecular docking-based virtual screening. Front Mol Biosci 7:603037
91. Mukherjee P, Shah F, Desai P et al (2011) Inhibitors of SARS-3CLpro: virtual screening, biological evaluation, and molecular dynamics simulation studies. J Chem Inf Model 51:1376–1392
92. Nguyen TT, Ryu HJ, Lee SH et al (2011) Virtual screening identification of novel severe acute respiratory syndrome 3C-like protease inhibitors and in vitro confirmation. Bioorg Med Chem Lett 21:3088–3091
93. Park SJ, Kim YG, Park HJ (2011) Identification of RNA pseudoknot-binding ligand that inhibits the −1 ribosomal frameshifting of SARS-coronavirus by structure-based virtual screening. J Am Chem Soc 133:10094–10100

Chapter 5
Vaccines and Immunoinformatics for Vaccine Design

Shikha Joon, Rajeev K. Singla, and Bairong Shen

Abstract The host immune system recognizes and responds to the selective antigens or epitopes (immunome) of the intruding pathogen over an entire organism. The immune response so generated is ample to confer the desired immunity and protection to the host. This led to the conception of immunome-derived vaccines that exploit selective genome-derived antigens or epitopes from the pathogen's immunome and not its entire genome or proteome. These are designed to elicit the required immune response and confer protection against future invasions by the same pathogen. Immunoinformatics through its epitope mapping tools allows direct selection of antigens from a pathogen's genome or proteome, which is critical for the generation of an effective vaccine. This paved way for novel vaccine design strategies based on the mapped epitopes for translational applications that includes prophylactic, therapeutic, and personalized vaccines. In this chapter, various Immunoinformatics tools for epitope mapping are presented along with their applications. The methodology for immunoinformatics-assisted vaccine design is also outlined.

Keywords Antigens · Epitopes · Immunome · Immunome-derived vaccines · Immunoinformatics · Epitope mapping

Shikha Joon and Rajeev K. Singla contributed equally to this work.

S. Joon · R. K. Singla · B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

S. Joon · R. K. Singla
iGlobal Research and Publishing Foundation, New Delhi, India

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_5

5.1 Introduction

Vaccination has unquestionably contributed to the maintenance of a healthy population worldwide. It has enhanced the quality of life and curtailed medical expenses while the disease burden and associated morbidity and mortality are minimized [1]. However, vaccine development is presented with various challenges concerning emerging and re-emerging infectious diseases, complicated life cycles and antigen diversity associated with the causative pathogens, need for personalized or tailored vaccines together with the cost and time invested [2, 3]. To conquer these roadblocks, it is pertinent to strategize the process of vaccine development and devise novel approaches that are efficient, economical, and swift.

Breakthroughs in the field of immunology together with the development of advanced Bioinformatics tools paved the way to novel platforms for vaccine design. The advent of various immunome-mining tools embarked on a new journey of immunology and led to the conception of novel vaccine designing strategies. These comprise "genome-derived vaccines" or "immunome-derived vaccines," "reverse vaccinology," and "vaccinomics" [4–10]. The principle behind immunome-derived vaccines is that recognition of specific antigens or a set of epitopes (B-cell or T-cell epitopes) induces competent and protective host immune responses against the intruding pathogen. For instance, a lock (antibody) can be opened with the key (pathogen) bearing the matching strings (epitopes) at its tip (antigen). In this case, the entire key is not required to open the lock, but the critical peptides or epitopes at its tip. Upon recognizing their specific key strings or epitopes, the immune locks or cells, such as T cells are activated. These, in turn, raise an alarm to other arsenals of the immune system regarding the trespassing pathogen. This laid the foundation of immunome-derived vaccines that comprise specific epitopes from the pathogen's genome or proteome using various Immunoinformatics tools.

Several vaccines have been designed against myriad infectious diseases using Immunoinformatics tools. These include immunome-derived vaccines for *Staphylococcus aureus* [11, 12], *Streptococcus pneumoniae* [13], *Neisseria meningitidis* [10], group B streptococcus [14], *Porphyromonas gingivalis* [15], and *Chlamydia pneumoniae* [16], which are under development. Immunoinformatics tools have also been exploited to develop anticancer vaccines and others that confers protection against autoimmune diseases [17, 18]. In this chapter, the feasibility of Immunoinformatics tools in unraveling critical immune determinants for vaccine design is discussed together with a methodology for designing immunome-derived vaccines.

5.2 Concept of Immunome-Derived Vaccines

T cells orchestrate the host's protective immunological response to a disease-causing pathogen. These responses are, in particular, targeted to the specific epitopes or short peptide sequences present on the antigens of the attacking pathogen. In

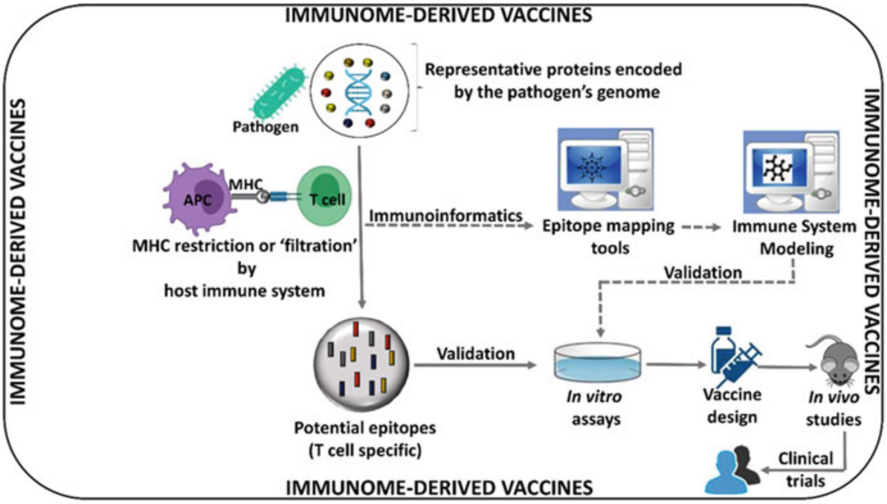
IMMUNOME-DERIVED VACCINES
Pathogen
Representative proteins encoded by the pathogen's genome
MHC
APC
T cell
MHC restriction or 'filtration' by host immune system
Immunoinformatics
Epitope mapping tools
Immune System Modeling
Validation
Validation
Potential epitopes (T cell specific)
In vitro assays
Vaccine design
In vivo studies
Clinical trials

Fig. 5.1 Conceptualization of immunome-derived vaccines using in silico epitope mapping tools that mimic host immune restriction

brief, the attacking pathogen is encountered by specific immune cells called antigen-presenting cells (APCs), which undertake to process all the antigens encoded by the entire genome of the pathogen. At this stage, MHC restriction or "immune system filter" allows the presentation of only selected processed proteins (fragmented proteins) or peptides to the T cells. These are recognized as the MHC ligands while those that fail to undergo MHC binding are sieved out in the screening process. Advanced epitope mapping tools have now made it possible to mimic this immune filtration considerably. A set of epitopes (immunogenic), specifically the T-cell epitopes, that elicit protective T-cell-mediated immunological responses can be identified from the immunome of a pathogen. Upon identification of potential T-cell epitopes, their immunogenicity can be confirmed with in vitro and in vivo studies. Finally, vaccine design and delivery can be commenced when desirable protection is obtained. As already stated, a pathogen's immunome comprises epitopes from all the proteins (cellular and non-cellular) (Fig. 5.1) [19].

5.3 The Journey of Vaccine Development Through Genome to Immunome

Choosing an appropriate antigen(s) is a major hurdle for the development of an effective vaccine. This roadblock in vaccine development has largely been overcome with the advent of novel genome analysis tools. These advanced tools are based on Bioinformatics and Immunoinformatics that aid in the selection of antigenic proteins from the pathogens' (in question) genomes directly. These, in turn, paved way for novel vaccine designing strategies based on the genome or immunome of the pathogen. The “immunome” refers to a set of immune-specific genes and proteins (antigens and epitopes) [20].

5.3.1 Antigen Presentation, and Activation and Generation of T- and B-Cell Immune Responses

Upon encountering an intruder or a pathogen, the host immune system responds through its various arsenals that include pathogen attacking antibodies (produced by B cells) and T cells. Here, both T helper cells (THCs) and cytotoxic T-lymphocytes (CTLs or killer T cells) participate in the T-cell responses. THCs trigger the antibody responses while CTLs are responsible for eliminating the pathogen-hijacked (intracellular pathogens) host cells or antigen-presenting cells (APCs). APCs display pathogen-specific epitopes via antigen processing, which are bound to the major histocompatibility complex (MHC: MHC I and MHC II) molecules on their surface. These, in turn, are recognized by their cognate T cells and stimulate the T-cell responses. “Epitopes” refers to the short peptide sequences present on the surface of an antigen (Fig. 5.2) [21].

Induction of the immunological memory and its quality, as well as magnitude, critically effectuates immune system activation. This, in turn, has a conspicuous effect on the efficacy of the developed vaccine. Besides, factors such as the category of memory cells induced and the longevity of antibodies generated markedly affect

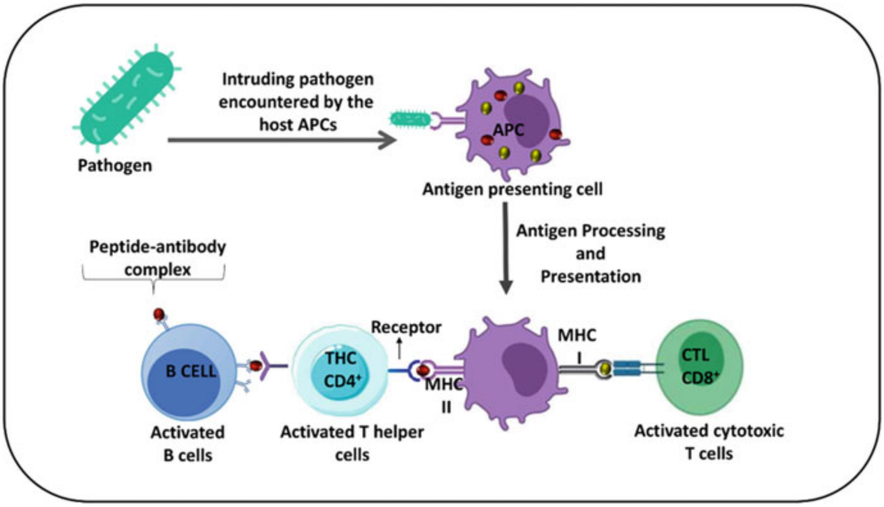

Fig. 5.2 Illustration of antigen presentation, and activation and generation of T- and B-cell immune responses in the human host upon pathogen exposure [21]

a vaccine's efficacy [22]. Previously, it was thought that the key immunological effectors induced upon pathogen entry are the antibody-producing B cells (humoral response), and the vaccines that induce B cells are more potent. However, accumulating shreds of evidence confirmed that vaccines based on the T-cell epitopes trigger protective immunological responses. These T cells generated protective immunological responses need the development of immune memory (B- and T-cell memory) only to a set of immunodominant epitopes from key pathogenic antigen(s) instead of every conceivable peptide from the whole pathogen. Essentially, these are amongst those few critical peptides that are capable of stimulating a T-cell response upon clearing the antigen processing and binding stages. This can be well understood by the following example. Smallpox (Variola) infection was averted by immunization with the cowpox virus vaccine (Vaccinia) even though their causative virus culprits are non-identical but related. A plausible explanation for this could be the conserved or shared B- and T-cell epitopes amongst the etiological agents of their etiological agents. These shared epitopes could successfully drive protective immunological responses via the generation of B and T memory cells in the host system against conserved Variola epitopes as well upon immunization with cowpox vaccine. Other examples include the HBV (Hepatitis B Virus) vaccine developed against the Hepatitis B virus. This vaccine contains Hepatitis B surface antigen (HBsAg) as the sole virus protein and provides sufficient protection against HBV. Further, the BCG (Bacillus Calmette–Guérin) vaccine against tuberculosis (TB) comprising only a subset of potent mycobacterial antigens drives host immunological responses against *Mycobacterium tuberculosis* bacterium. These examples strengthen the fact that vaccines based on a conserved set of immunodominant epitopes are proficient in thwarting the intruding bugs at par or even more in comparison to their counterparts that contain the entire pathogens or their whole proteins. As T cells orchestrate both humoral and cell-mediated immunological memory, the T-cell epitopes are now considered as the prime immune targets for developing an effective vaccine. For this, the critical T-cell epitopes can be shortlisted from the pathogen's immunome by making a comparison of their genome sequences and Immunoinformatics tools can then be exploited for epitope mapping.

5.4 Comparison of the Genome Sequences from Pathogens for Vaccine Development: A Novel Approach

Although an organism's immunome is obtained from its proteome (through its genome), the whole proteome does not serve as the starting point for epitome mapping. The reason being that the proteome comprises all the proteins of a pathogen. Several of these proteins are frequently conserved through diverse microbial species and participate in myriad activities including those that are immune-specific. These are, therefore, undesirable candidates for developing a vaccine. However, there are several ways to limit the candidate proteins from an organism's vast proteome for

Table 5.1 Strategies to exploit genome sequencing data for discovering vaccine candidates

Strategy	Outcome
Identification and elimination of genes other than those that are quintessential for immune response generation	Only immune-specific candidate antigens will be focused minimizing the undesirable self-immune reactions or cross-reactivity from vaccines comprising the screened candidates
Comparison of genome sequences of virulent and non-virulent organisms	Identification of virulence-associated genes for screening protein counterparts
Comparison of genome sequences from pathogens and uncovering their conserved genes, which are either used in existing vaccines or are potential future candidates	Identification of established antigens against a known pathogen which can be exploited for generating protective immunity against another infectious agent Examples: BCG vaccine against TB, and cowpox virus vaccine (vaccinia) against Smallpox (variola)
Employ DNA-microarray approach to screen pathogen-specific essential genes (adaptation, survival, and virulence genes)	Unveil genes, which in turn code for proteins (secretory proteins) that undergo upregulation post host immune interactions

epitope mapping. For instance, the genome sequences of virulent and non-virulent organisms can be compared. This way the genes that are indispensable for the pathogenesis of virulent counterparts along with those that generate host immune responses can be shortlisted for further analysis. Another strategy could be to uncover the conserved genes in the organisms that are proven vaccine candidates and their target pathogen using comparative genome sequencing. This will disseminate crucial information on antigen selection for vaccines (Table 5.1). Eventually, the pathogen-specific proteins, which are indispensable for its survival and virulence can be screened as potential vaccine targets. These include various upregulated and secretory proteins, which facilitate its adaption to the host environment post-intrusion. Thus, a comparison of genome sequences can reveal potential targets that can be subjected to epitope mapping for discovering critical T-cell epitopes using suitable Immunoinformatics tools outlined in the following section

5.5 Designing Vaccines Using Immunoinformatics Tools

Once the protein targets (potential antigens) from the pathogen in question are shortlisted for vaccine design, the immunodominant regions are determined using appropriate Immunoinformatics tools. In vitro testing of these potential immunogens is then undertaken to ascertain their immunogenicity. As stated in the previous section, Immunoinformatics tools aid in discovering the critical T-cell epitopes on the selected antigens that interact with the host T cells. These Immunoinformatics

Table 5.2 A list of T and B cell epitope prediction tools and databases

Name	Website/URLs	PMID	References
T cell epitope prediction tools and databases			
TEPITOPE	www.vaccinome.com	15542373	[24]
ProPred	http://www.imtech.res.in/raghava/proped	11751237	[29]
MULTIPRED	http://antigen.i2r.a-star.edu.sg/multipred/	15980449	[30]
MHCPred 2.0	http://www.darrenflower.info/mhcpred/	16539539	[31]
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	12717023	[32]
NetCTL 1.2	http://www.cbs.dtu.dk/services/NetCTL/	17973982	[33]
IEDB	http://www.immuneepitope.org/	15760272	[34]
IMGT®	http://imgt.cines.fr	18978023	[35]
EpiToolKit	http://www.epitoolkit.org	25712691	[36]
MMBPred	http://www.imtech.res.in/raghava/mmbpred/	14511568	[37]
SYFPEITHI	http://www.syfpeithi.de	10602881	[38]
ElliPro	http://tools.immuneepitope.org/tools/ElliPro	19055730	[39]
B cell Epitope Prediction Tools			
ABCpred	http://www.imtech.res.in/raghava/abcpred	16894596	[40]
COBEpro	http://scratch.proteomics.ics.uci.edu/	12807816	[41]
BEPITOPE	http://bepitope.ibs.fr/	**12557235**	[42]
DiscoTope	http://www.cbs.dtu.dk/services/DiscoTope/	**23300419**	[43]
BepiPred	http://www.cbs.dtu.dk/services/BepiPred	**28472356**	[44]
Pepitope	http://pepitope.tau.ac.il/	**17977889**	[45]
BCPREDS	http://ailab.cs.iastate.edu/bcpreds/	19642274	[46]

tools for epitope mapping utilize algorithms that exploit threading, non-linear functions, and neural networks [23–28]. These tools allow to scan of the sequences from protein targets and predict potential T-cell epitopes. Table 5.2 summarizes the Immunoinformatics tools employed in epitope mapping (both B- and T-cell epitopes).

5.6 Advanced Immunoinformatics Tools for Epitope Mapping

Before an MHC-bound peptide (antigen) is presented to the T cells, it must be excised or processed from their native proteins within the APCs of the host. This involves enzyme-mediated proteolytic cleavage or proteasomal processing and their transportation by ancillary proteins such as the transporter associated with antigen processing (TAP) to the endoplasmic reticulum (ER). All these steps are indispensable for peptide binding to MHCs, their presentation of MHC-bound peptides to the T cells, and elicitation of T-cell-dependent immune stimulation [47]. There exist several Immunoinformatics tools that have improvised epitope mapping in terms of its quality for shortlisting potential epitopes as vaccine

Table 5.3 A list of antigen-processing tools for vaccine design

Name	Website	PMID	References
Proteasomal cleavage prediction tools			
NetChop 20S	http://www.cbs.dtu.dk/Services/NetChop/	11983929	[48]
PAProC	http://www.paproc.de/	11345595	[49]
MAPPP	http://www.mpiib-berlin.mpg.de/MAPPP/	10047495	[53]
PCPS	http://imed.med.ucm.es/Tools/pcps/	32162269	[54]
Pcleavage	https://webs.iiitd.edu.in/raghava/pcleavage/	15988831	[55]
TAP processing prediction tools			
PRED^{TAP}	http://antigen.i2r.a-star.edu.sg/predTAP	16719926	[52]
SVMTAP	http://www-bs.informatik.uni-tuebingen.de/WAPP	**15987883**	[51]
TAPPred	http://bioinformatics.uams.edu/mirror/tappred/	**14978300**	[50]

components. For instance, Immunoinformatics tools such as NetChop 20S [48] and PAProC (Prediction Algorithm for Proteasomal Cleavages) [49] can predict proteasomal processing for several proteins intended for vaccine development. These online tools are based on artificial neural networks and have revamped Immunoinformatics screening for epitope mapping (T-cell epitopes). Then there are other epitope mapping tools, which can precisely predict the fate of shortlisted peptides concerning their processing and differentiate processed peptides from those that will not be amenable to processing. For instance, TAPPred [50], SVMTAP [51], and PRED^{TAP} [52] can predict which peptides will successfully pass through the TAP during their processing and has enhanced the precision of epitope mapping [53]. Table 5.3 summarizes Immunoinformatics tools for predicting proteasomal cleavage sites and TAP processing of antigens.

5.7 What Makes Immunoinformatics Tools so Proficient at Identifying Critical Antigens/T-Cell Epitopes for Vaccines?

The Immunoinformatics tools have proven to be incredibly valuable as they considerably minimize the research efforts, cost, and time needed for epitope mapping over traditional vaccine design methods. The candidate protein sequences can be selected and Immunoinformatics tools can be exploited to determine the immunogenic regions in the protein. As stated earlier, these immunogenic regions or short peptides drive appropriate immune responses in the host. Alternatively, the Open Reading Frames can be selected when the name and function of the gene of interest have not been assigned yet. The synthesis of these shortlisted immunogenic peptides that comprise potent epitopes can be undertaken followed by their screening in vitro. For this purpose, the T cells that are primed with these immunogens are obtained from the patients. Next, in vitro immunological assays, such as enzyme-linked immunospot (ELISpot) and cytokine assays by flow cytometry can be exploited. When the synthesized peptides yield a positive immune response in these assays these can be considered as potent immunogens for further analysis. These immunogens are capable of interacting with the immune system of the host and generate an appropriate immune response when the host encounters the associated disease or infection. These peptides undergo cellular processing and presentation within the APCs of the host. The presented peptides (or antigen in this case) display epitopes on their surfaces that elicit appropriate host immune responses. Once the epitopes are confirmed, their peptide counterparts that partake in eliciting T-cell-specific immune responses are the antigens of choice for the vaccine. Alternately, the whole protein counterpart of these peptides can be selected for developing a subunit vaccine.

5.8 Immunoinformatics: A Boon for Vaccine Design and Development

This section discusses applications of Immunoinformatics for in silico vaccine design and development and modeling and simulation of immunological responses as two broad categories.

5.8.1 In Silico Vaccine Design and Development

With the advent of genome sequencing and proteomics (comparative) together with immunoinformatics tools, it is now possible to implement new vaccine design methodologies.

A novel concept in vaccine design was introduced and termed, “reverse vaccinology,” which identifies potential immunogens (extracellular) from the pathogen’s entire genome. This indeed is an economic and time-intensive approach. This approach was foremost used for developing a conjugate vaccine against *Neisseria meningitides*, a causative agent for meningococcal meningitides and sepsis [56, 57].

5.8.1.1 Microarray-Based Vaccine Design

The microarray technology allows the screening of pathogenic genes from various bugs at different phases of their growth and conditions. This considerably minimizes the candidate genes in a pathogen’s genome for vaccine design. It is worth mentioning here that immunogenicity, structural motifs, and signal peptides obtained from sequencing of the given genome are vital ingredients for vaccine design [19].

5.8.1.2 Epitope-Based Vaccine Design

The vaccines that are based on the entire antigen or protein of a pathogen often present the risk of undesirable immunogenicity to the host. This is greatly eliminated by designing epitope-based vaccines. These vaccines comprise immunodominant epitopes that induce the desired immunological responses and confer protective immunity to the host [58]. Vaccines designed using this approach comprise a start codon (single) and an epitope inserted in a vaccine construct consecutively [59]. Subunit vaccines are now designed using this approach and require the prediction of ligands that have binding promiscuity [60].

5.8.1.3 Peptide-Based Approach for Vaccine Design

Peptide-based vaccines comprise small peptides obtained from epitopes, which are engineered to augment the immunological response. The prediction of potential peptides that bind to MHC molecules on APCs is quintessential as these peptides undergo MHC class I recognition. The MHC binding peptides can be predicted using any of the three approaches mentioned below and then a voting scheme can be applied for their integration to obtain better outcomes. At first, qualitative and quantitative immunological data is procured followed by quadratic programming. Another method is based on linear programming while the last method exploits sequence profiles. For this, the known epitopes are clustered and candidate peptides are then scored to yield sequence profiles. This is a more convenient approach over sequence-based methods to identify peptides that bind to MHC [61].

5.8.1.4 Non-alignment-Based Vaccine Design

Previous approaches for identifying antigens relied on sequence alignment and had several limitations. For instance, there exist proteins that are structurally and biologically similar, albeit they may not have the same sequence. A novel non-alignment-based approach for vaccine design was introduced to overcome this constraint. At first, three datasets were designed for viruses, bacteria, and tumors, respectively. Validation of the developed models was then done employing the leave-one-out cross-validation (LOO-CV) method on these sets. Tests sets were used for external validation. The outcome was the webserver VaxiJen (http://www.darrenflower.info/VaxiJen/), which comprises these validated models [62] (Table 5.4).

Table 5.4 In silico DNA vaccine design tools

Tool/server	Website/URLs	PMID	References
DyNAVacS	http://miracle.igib.res.in/dynavac/	16845007	[63]
NERVE	http://www.bio.unipd.it/molbinfo	**16848907**	[64]
VaxiJen	http://www.jenner.ac.uk/VaxiJen	**17207271**	[62]
Vaxign	http://www.violinet.org/vaxign/	20671958	[65]
VIOLIN	http://www.violinet.org	24259431	[66]

5.8.1.5 Designing DNA Vaccines

It is a known fact that DNA vaccines elicit both humoral and cell-mediated immunological responses. These are particularly helpful in combating intracellular pathogens. For this, various.

Webservers and softwares have been developed, which include DYNAVACS [63], NERVE [64], and VIOLIN [66]. DYNAVACS has various modules such as mapping sites for restriction enzymes, optimization of codons for heterologous genes, insertion of Kozak sequences, designing primers, genetic engineering for gene therapy, and customized sequence insertion [63]. The subunit vaccines can be designed against bacterial bugs using the automated NERVE software [64]. VIOLIN is a web-based database that executes curation, storage, and analysis of published vaccine data [66]. It comprises integrated programs for data mining and searches such as VAXPRESSO, VAXLERT, VAXMESH, and LITSEARCH. Now, potential vaccine targets can be predicted with the aid of VAXIGN, which is a vaccine design Webserver. In addition to predicting immunogenic vaccine targets, it can also predict the transmembrane domain, subcellular location, probability of adhesion, conserved sequences in distinct genomes, immunogen and host proteome sequence, and binding of the predicted epitope to MHCs [65] (Table 5.4).

5.8.2 Modeling and Simulation of Immunological Responses

Modeling and simulation of immunological responses give a qualitative and quantitative understanding of the immune system on the whole. The generated immune models are capable of testing the antigen–antibody interactions and deduce immunological responses for the antigen in question. These are particularly useful in evaluating the efficacy of a vaccine candidate or drug administration. Several infection models have been developed to date including Hepatitis C and HIV (human immunodeficiency virus) [34, 67]. These models disseminate information on adaptive immune responses and the virus infection cycle within the host. The immune system is simulated to gain an insight into the complex immune responses by exploiting a combinational approach comprising experimental and computational data. IMMUNOGRID (http://www.immunogrid.org) and VIROLAB

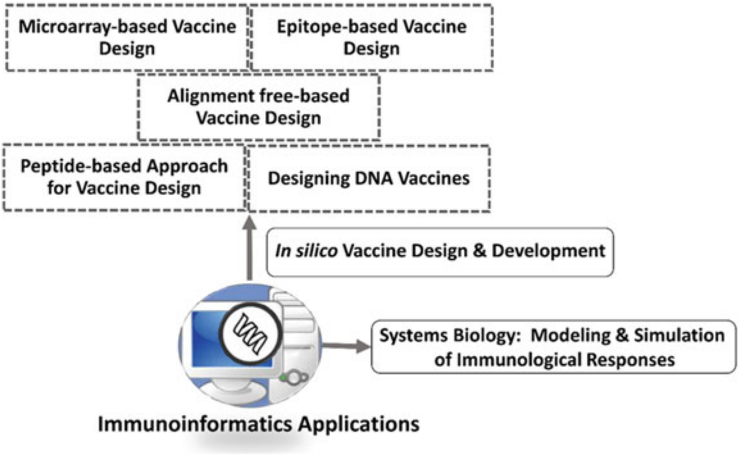
Microarray-based Vaccine Design
Epitope-based Vaccine Design
Alignment free-based Vaccine Design
Peptide-based Approach for Vaccine Design
Designing DNA Vaccines
In silico Vaccine Design & Development
Systems Biology: Modeling & Simulation of Immunological Responses
Immunoinformatics Applications

Fig. 5.3 Various immunology-related applications of Immunoinformatics

(http://www.virolab.org:080/virolab) are immune simulation projects based on this approach that aims at developing an in silico library for myriad infections [68]. Another example includes SIMISYS 0.3, which simulates healthy and diseased states of the host based on its interpretations of immune cell interactions and the pathogen [56, 69]. Various immunology-related applications of Immunoinformatics are shown in Fig. 5.3.

5.9 Designing Vaccines Using Immunoinformatics Tools: Pitfalls and Future Interventions

Although immune-derived vaccines save the time, cost, and labor invested in designing vaccines using conventional wet laboratory techniques, they are accompanied by certain limitations as follows.

- Immunoinformatics tools rely on in vitro, in vivo, and clinical findings to generate substantial data for further analysis. Therefore, these can expedite the process of vaccine design and development but, not replace the wet laboratory immunology research.
- Data quality and robustness of the algorithm exploited determine the prediction quality of Immunoinformatics tools. For instance, sequencing data acquired from high-throughput sequencing many times is flawed. The incorrect sequencing or annotations affect the quality of analysis as they can yield ambiguous results [70].
- The existing tools for genome mapping are not efficient in identifying the immunogenic epitopes from the pathogen's non-protein antigenic entities, such as lipids and carbohydrates.

- To date, only a few Toll-receptor agonists have been discovered, which are required in innate immune responses and in turn in generating protective immune responses [71]. Immunoinformatics tools that can model and select pathogen-associated molecular patterns (PAMPs) are underway though.
- Further, the available Immunoinformatics tools are inefficient in generating reliable predictions for B-cell epitopes (conformational epitopes). These interact with the host antibodies [72].
- Finally, the identified immunogens or their predicted epitopes mandate in vitro and in vivo efficacy evaluation with appropriate challenge animal models before clinical trials [19].

5.10 Conclusion

In recent years, research in the field of genome-based vaccine development has gained momentum. This can be credited to the enormous and expanding information on microbial genome complemented with novel genome sequencing tools that allow the comparison of genomes from distinct microbial species. Immunoinformatics with advanced antigen selection approaches and tools have expedited vaccine research and development. The validation of the predicted antigens for vaccine development relies on highly efficient in vitro assays that measure antigen elicited T-cell response. Howbeit, the time and cost invested in preliminary antigen selection procedures are greatly minimized with genome scanning and Immunoinformatics over classic in vitro methods. As these in silico methods allow scanning of the genome sequences illuminating potential vaccine targets, in-depth knowledge of protein (of the pathogen) structure and function is not needed anymore. Also, these have eliminated the need for cloning of genes and protein (of the pathogen) isolation before the target screening. By scanning the antigen or peptide sequences in the target pathogen, potential epitopes for the vaccine can be discovered circumventing long-drawn-out in vitro cloning procedures. An amalgamated approach employing bench-research and Bioinformatics tools, such as the confirmation of T-cell (in vitro) and genome sequencing can expedite the process of vaccine development over traditional vaccine design strategies. In conclusion, Immunoinformatics has emerged as a promising in silico approach, which might revolutionize vaccine design and development.

Acknowledgments This work was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019) and the National Natural Science Foundation of China (32070671).

References

1. Terry FE, Moise L, Martin RF et al (2015) Time for T? Immunoinformatics addresses vaccine design for neglected tropical and emerging infectious diseases. Expert Rev Vaccines 14:21–35
2. Poland GA, Whitaker JA, Poland CM et al (2016) Vaccinology in the third millennium: scientific and social challenges. Curr Opin Virol 17:116–125
3. Servín-Blanco R, Zamora-Alvarado R, Gevorkian G et al (2016) Antigenic variability: obstacles on the road to vaccines against traditionally difficult targets. Hum Vaccin Immunother 12:2640–2648
4. De Groot AS, Martin W (2003) From immunome to vaccine: epitope mapping and vaccine design tools. Novartis Found Symp 254:57–72
5. Doytchinova IA, Taylor P, Flower DR (2003) Proteomics in vaccinology and immunobiology: an informatics perspective of the immunone. J Biomed Biotechnol 2003:267–290
6. Jongeneel V (2001) Towards a cancer immunome database. Cancer Immun 1:3
7. Klade CS (2002) Proteomics approaches towards antigen discovery and vaccine development. Curr Opin Mol Ther 4:216–223
8. Pederson T (1999) The immunome. Mol Immunol 36:1127–1128
9. Petrovsky N, Brusic V (2002) Computational immunology: the coming of age. Immunol Cell Biol 80:248–254
10. Rappuoli R (2001) Reverse vaccinology, a genome-based approach to vaccine development. Vaccine 19:2688–2691
11. Etz H, Minh DB, Henics T et al (2002) Identification of in vivo expressed vaccine candidate antigens from Staphylococcus aureus. Proc Natl Acad Sci U S A 99:6573–6578
12. Weichhart T, Horky M, Söllner J et al (2003) Functional selection of vaccine candidate peptides from Staphylococcus aureus whole-genome expression libraries in vitro. Infect Immun 71:4633–4641
13. Adamou JE, Heinrichs JH, Erwin AL et al (2001) Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis. Infect Immun 69:949–958
14. Tettelin H, Masignani V, Cieslewicz MJ et al (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proc Natl Acad Sci U S A 99:12391–12396
15. Ross BC, Czajkowski L, Hocking D et al (2001) Identification of vaccine candidate antigens from a genomic analysis of Porphyromonas gingivalis. Vaccine 19:4135–4142
16. Montigiani S, Falugi F, Scarselli M et al (2002) Genomic approach for analysis of surface proteins in Chlamydia pneumoniae. Infect Immun 70:368–379
17. Mathiassen S, Lauemøller SL, Ruhwald M et al (2001) Tumor-associated antigens identified by mRNA expression profiling induce protective anti-tumor immunity. Eur J Immunol 31:1239–1246
18. Robinson WH, Garren H, Utz PJ et al (2002) Millennium award. Proteomics for the development of DNA tolerizing vaccines to treat autoimmune disease. Clin Immunol 103:7–12
19. De Groot AS (2004) Immunome-derived vaccines. Expert Opin Biol Ther 4:767–772
20. Ortutay C, Vihinen M (2009) Immunome knowledge base (IKB): an integrated service for immunome research. BMC Immunol 10:3
21. Chaplin DD (2010) Overview of the immune response. J Allergy Clin Immunol 125:S3–S23
22. Sarkander J, Hojyo S, Tokoyoda K (2016) Vaccination to gain humoral immune memory. Clin Transl Immunology 5:e120
23. Altuvia Y, Margalit H (2004) A structure-based approach for prediction of MHC-binding peptides. Methods 34:454–459
24. Bian H, Hammer J (2004) Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. Methods 34:468–475
25. Brusic V, Bajic VB, Petrovsky N (2004) Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications. Methods 34:436–443

26. De Groot AS, Bishop EA, Khan B et al (2004) Engineering immunogenic consensus T helper epitopes for a cross-clade HIV vaccine. Methods 34:476–487
27. Doytchinova IA, Guan P, Flower DR (2004) Quantitative structure-activity relationships and the prediction of MHC supermotifs. Methods 34:444–453
28. Sung MH, Simon R (2004) Candidate epitope identification using peptide property models: application to cancer immunotherapy. Methods 34:460–467
29. Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17:1236–1237
30. Zhang GL, Khan AM, Srinivasan KN et al (2005) MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. Nucleic Acids Res 33:W172–W179
31. Guan P, Hattotuwagama CK, Doytchinova IA et al (2006) MHCPred 2.0: an updated quantitative T-cell epitope prediction server. Appl Bioinforma 5:55–61
32. Nielsen M, Lundegaard C, Worning P et al (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci 12:1007–1017
33. Larsen MV, Lundegaard C, Lamberth K et al (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics 8:424
34. Peters B, Sidney J, Bourne P et al (2005) The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 3:e91
35. Lefranc MP, Giudicelli V, Ginestoux C et al (2009) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 37:D1006–D1012
36. Schubert B, Brachvogel HP, Jürges C et al (2015) EpiToolKit—a web-based workbench for vaccine design. Bioinformatics 31:2211–2213
37. Bhasin M, Raghava GP (2003) Prediction of promiscuous and high-affinity mutated MHC binders. Hybrid Hybridomics 22:229–234
38. Rammensee H, Bachmann J, Emmerich NP et al (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219
39. Ponomarenko J, Bui HH, Li W et al (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 9:514
40. Saha S, Raghava GP (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65:40–48
41. Saxová P, Buus S, Brunak S et al (2003) Predicting proteasomal cleavage sites: a comparison of available methods. Int Immunol 15:781–787
42. Odorico M, Pellequer JL (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. J Mol Recognit 16:20–22
43. Kringelum JV, Lundegaard C, Lund O et al (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol 8:e1002829
44. Jespersen MC, Peters B, Nielsen M et al (2017) BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res 45:W24–W29
45. Mayrose I, Penn O, Erez E et al (2007) Pepitope: epitope mapping from affinity-selected peptides. Bioinformatics 23:3244–3246
46. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting flexible length linear B-cell epitopes. Comput Syst Bioinformatics Conf 7:121–132
47. van Endert PM, Tampé R, Meyer TH et al (1994) A sequential model for peptide binding and transport by the transporters associated with antigen processing. Immunity 1:491–500
48. Keşmir C, Nussbaum AK, Schild H et al (2002) Prediction of proteasome cleavage motifs by neural networks. Protein Eng 15:287–296
49. Nussbaum AK, Kuttler C, Hadeler KP et al (2001) PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. Immunogenetics 53:87–94
50. Bhasin M, Raghava GP (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci 13:596–607
51. Dönnes P, Kohlbacher O (2005) Integrated modeling of the major events in the MHC class I antigen processing pathway. Protein Sci 14:2132–2140
52. Zhang GL, Petrovsky N, Kwoh CK et al (2006) PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. Immunome Res 2:3

53. Petrovsky N, Brusic V (2004) Virtual models of the HLA class I antigen processing pathway. Methods 34:429–435
54. Gomez-Perosanz M, Ras-Carmona A, Reche PA (2020) PCPS: a web server to predict proteasomal cleavage sites. Methods Mol Biol 2131:399–406
55. Bhasin M, Raghava GP (2005) Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. Nucleic Acids Res 33:W202–W207
56. Kalita JK, Chandrashekar K, Hans R et al (2006) Computational modelling and simulation of the immune system. Int J Bioinforma Res Appl 2:63–88
57. Pizza M, Scarlato V, Masignani V et al (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287:1816–1820
58. Gallimore A, Hengartner H, Zinkernagel R (1998) Hierarchies of antigen-specific cytotoxic T-cell responses. Immunol Rev 164:29–36
59. Morris S, Kelley C, Howard A et al (2000) The immunogenicity of single and combination DNA vaccines against tuberculosis. Vaccine 18:2155–2163
60. Zhao B, Sakharkar KR, Lim CS et al (2007) MHC–peptide binding prediction for epitope based vaccine design. Int J Integr Biol 1:127–140
61. Florea L, Halldórsson B, Kohlbacher O et al (2003) Epitope prediction algorithms for peptide-based vaccine design. Proc IEEE Comput Soc Bioinform Conf 2:17–26
62. Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 8:4
63. Harish N, Gupta R, Agarwal P et al (2006) DyNAVacS: an integrative tool for optimized DNA vaccine design. Nucleic Acids Res 34:W264–W266
64. Vivona S, Bernante F, Filippini F (2006) NERVE: new enhanced reverse vaccinology environment. BMC Biotechnol 6:35
65. He Y, Xiang Z, Mobley HL (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010:297505
66. He Y, Racz R, Sayers S et al (2014) Updates on the web-based VIOLIN vaccine database and analysis system. Nucleic Acids Res 42:D1124–D1132
67. Gong T, Cai Z (2005) Visual modeling and simulation of adaptive immune system. Conf Proc IEEE Eng Med Biol Soc 2005:6116–6119
68. De Groot AS, Rappuoli R (2004) Genome-derived vaccines. Expert Rev Vaccines 3:59–76
69. Castiglione F, Liso A (2005) The role of computational models of the immune system in designing vaccination strategies. Immunopharmacol Immunotoxicol 27:417–432
70. Bahrami AA, Payandeh Z, Khalili S et al (2019) Immunoinformatics: in silico approaches and computational design of a multi-epitope, immunogenic protein. Int Rev Immunol 38:307–322
71. Imler JL, Hoffmann JA (2001) Toll receptors in innate immunity. Trends Cell Biol 11:304–311
72. Enshell-Seijffers D, Denisov D, Groisman B et al (2003) The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1. J Mol Biol 334:87–101

Chapter 6
Predicting the Disease Severity of Virus Infection

Xin Qi, Li Shen, Jiajia Chen, Manhong Shi, and Bairong Shen

Abstract The COVID-19 pandemic has resulted in unprecedented burden on global health and economic systems, promoting worldwide efforts to understand, control, and fight the disease. Due to the wide spectrum of clinical severity, effective risk factors, biomarkers, and models for predicting disease severity and mortality in COVID-19 patients are urgently needed to provide guidance for clinical intervention and management. In this chapter, we first describe the infection features of different COVID-19 strains and the potential of clinical features, cytokine storm and biomarkers in predicting the severity of COVID-19 patients. We focus on how scoring systems, mathematical models and artificial intelligence (AI)-based models can promote the classification of COVID-19 severity at the population or individual level. Moreover, the development perspective of biomarkers and models for predicting the severity of COVID-19 is prospected. Therefore, this chapter highlights the clinical significance of biomarkers and models related to COVID-19 severity and provides important clues for improving the outcomes of COVID-19 patients, thereby facilitating timely disease assessment and precision medicine for individual COVID-19 patients.

Xin Qi and Li Shen contributed equally to this work.

X. Qi (✉) · J. Chen
School of Chemistry and Life Sciences, Suzhou University of Science and Technology, Suzhou, China
e-mail: qixin@usts.edu.cn

L. Shen · B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

M. Shi
College of Information and Network Engineering, Anhui Science and Technology University, Fengyang, China

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_6

Keywords COVID-19 severity · Risk factor · Cytokine storm · Biomarker · Scoring system · Mathematical model · AI-based model

6.1 Introduction

The recent occurrence and spread of COVID-19 have posed a huge threat to healthcare, economic, and social systems around the world. It is caused by infection with a novel coronavirus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Increasing evidence showed that the majority of patients with COVID-19 exhibit mild or moderate clinical symptoms, such as cough, fever, fatigue, and pneumonia, while 19% of patients still develop to severe illness or even organ failure [1]. Without timely diagnosis and treatment, severe cases of COVID-19 possess a high risk of poor prognosis [2]. Therefore, it is urgent to identify risk factors or develop effective models for predicting severity and providing guidance for individualized intervention and treatment of COVID-19 patients.

Biomarkers are measurable alterations in the composition of tissues or body fluids, which can objectively indicate physiological or pathological processes with high specificity and sensitivity. Multiple types of biomarkers including molecular alteration, biochemical index, hematological parameters, and clinical features have been extensively utilized as effective tools for the diagnosis, staging, and treatment of different diseases. Especially, biomarkers possess huge clinical significance if they could track real-time disease progression and severity. At present, due to high mortality of severe patients with COVID-19, scientists and medical workers from all over the world are devoted to the development and identification of biomarkers that can predict the severity of the disease. For example, expression levels of plasma IFN-γ-induced protein 10 (IP-10), monocyte chemotactic protein-3 (MCP-3), interleukin-6 (IL-6), and C-reactive protein (CRP) have been demonstrated to be promising biomarkers that were highly associated with COVID-19 severity and progression [3, 4]. Furthermore, recent progresses in multi-omics technologies have spurred considerable efforts to survey biomolecules for predicting the severity and prognosis of COVID-19 patients. Overmyer et al. have successfully identified 219 biomolecules with high relevance to COVID-19 severity based on large-scale transcriptomic and proteomic analysis [5].

Predictive models that integrate key variables or clinical parameters can also benefit the evaluation of COVID-19 severity. Currently, methods ranging from scoring systems to mathematical models, machine learning models and deep learning models have been proposed to predict COVID-19 severity, thereby helping patients to receive prompt and effective therapeutic regimens at early stages. For example, using univariate and multivariate logistic regression analysis, Zhang et al. 6 have established a novel predictive scoring system for COVID-19 severity, which was composed of factors including age, white blood cell count (WBC), neutrophil (NEU), glomerular filtration rate (GFR), and myoglobin. Its excellent prediction value was confirmed by the area under the curve (AUC) in the ROC curve analysis

[6]. Based on three key indicators (e.g., lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP)) for COVID-19 prognostic prediction, a machine learning-based model was developed that can predict the mortality risk of severe patients with over 90% accuracy [7]. Therefore, the development of COVID-19 severity models is of great significance for customizing personalized treatment regimens and even reducing mortality.

To gain a deep understanding of the role and application potential of the identified predictive factors and models, we here firstly introduce the infection features of different COVID-19 strains and summarize the key clinical indicators highly associated with COVID-19 severity. Then, we describe the impact of cytokine storm in COVID-19 disease process and severity. Finally, we systematically review and appraise the available predictive biomarkers and models of COVID-19 severity, highlighting the importance of disease severity prediction for precision medicine and healthcare for COVID-19 patients (Fig. 6.1).

6.2 Informatics for Diagnosis of COVID-19 with Varying Severity

6.2.1 Different SARS-CoV-2 Strains and Infection Severity

Recent researches suggest that the emergence of multiple highly transmissible variants of SARS-CoV-2 has not only exacerbated the COVID-19 pandemic, but also brought considerable difficulties and challenges for the design of an effective vaccine on a global scale. Currently, several new SARS-CoV-2 genetic variants including Alpha variant (B.1.1.7), Beta variant (B.1.351), Gamma variant (P.1), and Delta variant (B.1.617.2), are reported to possess increased transmissibility and pathogenicity compared with the original strain, arousing health concerns around the world.

The first SARS-CoV-2 variant of concern with an amino acid mutation from an aspartate to a glycine at position 614 (D614G) in the spike protein was identified in early March 2020 and spread quickly to global dominance by April 2020 [8], attaching considerable attention. The spike protein composed of S1 and S2 subunits is a vital structural region of the coronavirus for receptor recognition and membrane fusion [9]. The D614G mutation is located in the S1 subunit, which plays a crucial role in SARS-CoV-2 entry into host cell [10]. Emerging studies indicated that SARS-CoV-2 harboring D614G mutation has higher infectivity [11, 12].

The B.1.1.7 variant (N501Y.V1) was initially detected in the United Kingdom (UK) in September 2020 [13], and rapidly became the dominant SARS-CoV-2 strain in the southeast and east of England [14]. Notably, the strain had spread to 160 countries by 3 June, 2021 [15]. Genomic surveillance revealed that the strain has 23 mutations across the virus genome, especially including the N501Y (Asn501Tyr) mutation in the spike protein, which might enhance angiotensin-converting enzyme

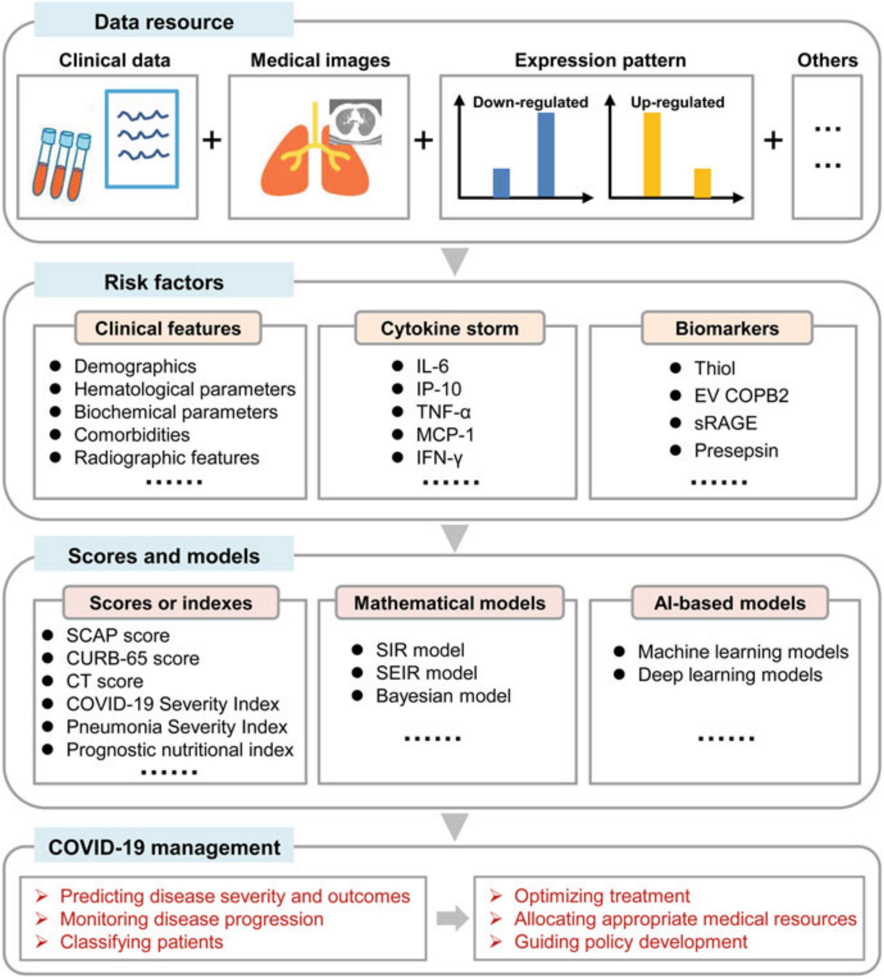

Fig. 6.1 Paradigm of predicting the disease severity of COVID-19. Based on large-scale clinical data, medical images and expression profiling data, a variety of risk factors, scores, and models can be used to predict disease severity and mortality, thereby providing guidance for clinical intervention and management of patients with COVID-19. *EV* extracellular vesicles; *sRAGE* soluble receptor for advanced glycation end product; *SIR* susceptible-infectious-recovered; *SEIR* susceptible-exposed-infectious-recovered; *AI* artificial intelligence

2 (ACE2) receptor-binding affinity [16]. Preliminary evidence indicated that the B.1.1.7 strain is more transmissible than previously reported variants. Davies et al. [17] found that the UK variant B.1.1.7 is 43 to 90% (95% credible intervals (CI): 38–130%) more transmissible than preexisting lineages. Volz et al. [18] and Graham et al. [19] independently discovered that B.1.1.7 has a significant transmission advantage over preexisting variants, with a multiplicative increase in the effective

reproduction number, R_t, by a factor of 1.5–2.0 or 1.35. Besides, Bager et al. [20] found that there is an increased risk of hospitalization for people infected with B.1.1.7 relative to other lineages in Denmark.

The B.1.351 variant (501Y.V2), first identified in October 2020 [13], rapidly caused the outbreak of COVID-19 in South Africa [21]. It harbors three pivotal mutations in the receptor-binding domain of the spike protein, including N501Y, K417N, and E484K, thereby enhancing the binding affinity of spike protein to the ACE2 receptor [22]. By analyzing the global spread of SARS-CoV-2 variants, Campbell et al. [15] found a marked increase in effective reproduction numbers of B.1.351 strain at 25% (95% CI: 20–30%), highlighting its increased transmissibility. Besides, it is estimated that the transmission rate of the 501Y.V2 variant is 50% higher than the preexisting variant in South Africa [21].

The P.1 variant (501Y.V3) first detected in Manaus in the Brazilian Northern region, was alleged as the main cause of the second COVID-19 wave that occurred in Amazonas in November 2020 [23]. The P.1 strain has 17 mutations, including the N501Y, K417N, and E484K in the spike protein that shared with the B.1.351 variant [24]. Epidemiological data showed that prevalence of P.1 variant increased sharply to 73% in January 2021 and replaced the original strain in less than 2 months [25]. Evidence indicated that the P.1 variant has caused a striking increase in intensive care unit (ICU) admission, mechanical ventilation (MV) need, and mortality of young adults [26], confirming its high transmissibility and lethality.

The B.1.617 lineage was first identified in India in October 2020 [27]. It has since then contributed to the surge of COVID-19 cases in India and the UK and further rapidly spread to 43 countries [28]. The lineage has three main subtypes including B.1.617.1, B.1.617.2, and B.1.617.3, which contain diverse mutations in the spike protein. Among them, B.1.617.2 (Delta variant) is deemed to spread faster than other preexisting variants [27]. Emerging evidence showed that the growth rate of B.1.617.2 was higher than that of B.1.1.7, and its doubling time was between 5–14 days [29]. Currently, B.1.617.2 with increased proportion in sequenced lineages is the prevailing variant in UK, indicating a competitive advantage. Moreover, Salvatore et al. found that B.1.617.2 exhibited higher transmissibility than previously circulating strains [30].

Collectively, given the critical role of spike-ACE2 binding affinity in the initial stage of SARS-CoV-2 infection, accumulating studies have uncovered that several SARS-CoV-2 variants with significantly increased transmissibility and disease severity including B.1.1.7, B.1.351, P.1, and B.1.617.2 strain, harbor different number of mutations in the spike protein, which could enhance the spike-ACE2 binding affinity. Therefore, characterization of new variants is critical for monitoring the extent of the COVID-19 pandemic and maintaining the effectiveness of vaccination.

6.2.2 Clinical Features Associated with COVID-19 Severity

Since the outbreak of COVID-19 epidemic, considerable efforts have been made to discover clinical characteristics and risk factors related to the disease severity, thereby supporting decision-making in clinical practice. As listed in Table 6.1, clinical features associated with COVID-19 severity were mainly classified into four types: demographics, hematological and biochemical parameters, comorbidities, and radiographic features.

6.2.2.1 Demographics

Male gender and older age are the most frequently reported demographic factors that can increase the severity and mortality of COVID-19, and are therefore included as key variables in multiple risk scores or indexes that predict the disease severity [31]. Barek and colleagues found that male patients and patients older than 50 years are significantly associated with the higher risk of cases severity [32]. The meta-analyses on 36,470 patients by Pijls et al. also showed that male patients and patients with age $\geq$ 70 have a higher risk for severe COVID-19 [33].

6.2.2.2 Hematological and Biochemical Parameters

Through laboratory biochemistry tests, increasing evidence demonstrated that the involvement of hematological abnormalities is prominent in severe patients with COVID-19 [34]. As listed in Table 6.1, the hematological markers, such as neutrophil-to-lymphocyte ratio (NLR) [35–39], platelet-to-lymphocyte ratio (PLR) [35, 37, 39], serum amyloid A (SAA) [40, 41], C-reactive protein (CRP) [38, 40, 42–44], D-dimer [44–47], ferritin [42, 48], ALT [42], AST [42], albumin [42], immunoglobulin deficiency [49], blood urea nitrogen (BUN)/creatinine (Cr) ratio [50], troponin I (cTnI) [44], high sensitivity C-reactive protein-prealbumin ratio (HsCPAR) [51], and high-sensitivity C-reactive protein-albumin ratio (HsCAR) [51], can play a predictive role in the stratification of COVID-19 severity. Especially, NLR, PLR, and BUN/Cr ratios were reported to be associated with COVID-19 severity as well as mortality, and mean platelet volume (MPV) and D-dimer are predictive makers of hospitalization and severity in children with COVID-19. Therefore, these hematological and biochemical parameters altered significantly in severe cases may serve as promising biomarkers for identifying patients needing hospitalization and intensive care.

Table 6.1 Clinical features for predicting COVID-19 severity and mortality

Clinical feature	Function	Number of patients	Performance	Type	Study
Male sex	Risk factor associated with COVID-19 severity	10,014		Demographics	[32]
Male sex	Risk factor associated with COVID-19 severity	36,470		Demographics	[33]
Male sex	Risk factor associated with COVID-19 in-hospital mortality	200		Demographics	[65]
Male sex	Risk factor associated with COVID-19 severity	548		Demographics	[66]
Older age	Risk factor associated with COVID-19 severity	10,014		Demographics	[32]
Older age	Risk factor associated with COVID-19 severity	36,470		Demographics	[33]
Older age	Risk factor associated with COVID-19 in-hospital mortality	200		Demographics	[65]
Older age	Risk factor associated with COVID-19 severity	548		Demographics	[66]
NLR	Predictor of severity and mortality in COVID-19 patients of Pakistan	191	AUC of 0.841 for ICU patients; AUC of 0.860 for deceased patients	Hematological and biochemical parameters	[35]
NLR	Predictor of severity and mortality in COVID-19 patients	4546	AUC of 0.85 for predicting severity, AUC of 0.90 for predicting mortality	Hematological and biochemical parameters	[36]
NLR	Prognostic and risk stratifying factor of COVID-19 severity	233		Hematological and biochemical parameters	[37]

(continued)

Table 6.1 (continued)

Clinical feature	Function	Number of patients	Performance	Type	Study
NLR	Predictor of COVID-19 severity	443	AUC of 0.737	Hematological and biochemical parameters	[38]
NLR	Indicator of COVID-19 severity	101		Hematological and biochemical parameters	[39]
PLR	Predictor of the severity and mortality in COVID-19 patients of Pakistan	191	AUC of 0.703 for ICU patients; AUC of 0.677 for deceased patients	Hematological and biochemical parameters	[35]
PLR	Prognostic and risk stratifying factor of COVID-19 severity	233		Hematological and biochemical parameters	[37]
PLR	Indicator of COVID-19 severity	101		Hematological and biochemical parameters	[39]
SAA	Predictor of the severity and recovery of COVID-19	35	AUC of 0.818 for predicting severity; AUC of 0.923 for predicting recovery	Hematological and biochemical parameters	[40]
SAA	Predictor of COVID-19 progression and severity	150	AUC of 0.89	Hematological and biochemical parameters	[41]
CRP	Predictor of COVID-19 severity	443	AUC of 0.734	Hematological and biochemical parameters	[38]
CRP	Predictor of the severity of COVID-19	35	AUC of 0.804	Hematological and biochemical parameters	[40]
CRP	Indicator of COVID-19 severity	66		Hematological and biochemical parameters	[42]
CRP	Predictor of COVID-19 severity	144	AUC of 0.85	Hematological and biochemical parameters	[43]
CRP	Indicator of COVID-19 severity	77	AUC of 0.83	Hematological and biochemical parameters	[44]

D-dimer	Indicator of COVID-19 severity	77	AUC of 0.78	Hematological and biochemical parameters	[44]
D-dimer	Predictor of hospitalization due to COVID-19 in children	251	AUC of 0.776	Hematological and biochemical parameters	[45]
D-dimer	Early identification of severe cases in children	171		Hematological and biochemical parameters	[46]
D-dimer	Predictive marker for COVID-19 severity	972		Hematological and biochemical parameters	[47]
Ferritin	Indicator of COVID-19 severity	66		Hematological and biochemical parameters	[42]
Ferritin	Predictor of severity and mortality in COVID-19	157	AUC of 0.69	Hematological and biochemical parameters	[48]
ALT	Indicator of COVID-19 severity	66		Hematological and biochemical parameters	[42]
AST	Indicator of COVID-19 severity	66		Hematological and biochemical parameters	[42]
Albumin	Indicator of COVID-19 severity	66		Hematological and biochemical parameters	[42]
Immunoglobulin deficiency	Indicator of COVID-19 severity	62		Hematological and biochemical parameters	[49]
BUN/Cr ratio	Predictor of disease severity and survival of COVID-19 patients	139	AUC of 0.98 for predicting severity, AUC of 0.95 for predicting mortality	Hematological and biochemical parameters	[50]
cTnI	Indicator of COVID-19 severity	77	AUC of 0.76	Hematological and biochemical parameters	[44]
HsCPAR	Predictor of COVID-19 severity	114	AUC of 0.80	Hematological and biochemical parameters	[51]
HsCAR	Predictor of COVID-19 severity	114	AUC of 0.81	Hematological and biochemical parameters	[51]

(continued)

Table 6.1 (continued)

Clinical feature	Function	Number of patients	Performance	Type	Study
Lymphocyte percentage (LYM%)	Predictor of prognosis in COVID-19 patients	5		Hematological and biochemical parameters	[67]
Lymphocyte-monocyte ratio (LMR)	Indicator of COVID-19 severity	101		Hematological and biochemical parameters	[39]
Platelet	Predictor of COVID-19 severity	443	AUC of 0.634	Hematological and biochemical parameters	[38]
Hypertension	Risk factor associated with COVID-19 severity	2893		Comorbidities	[55]
Hypertension	Risk factor associated with COVID-19 severity and mortality	6560		Comorbidities	[68]
Diabetes	Risk factor associated with COVID-19 severity and in-hospital death	78,874		Comorbidities	[56]
Diabetes	Risk factor associated with COVID-19 severity	6451		Comorbidities	[69]
Diabetes	Risk factor associated with COVID-19 mortality	1122		Comorbidities	[70]
Cardiovascular disease	Risk factor associated with COVID-19 severity	1527		Comorbidities	[57]
Cardiovascular disease	Risk factor associated with COVID-19 severity	54		Comorbidities	[71]
Chronic kidney disease	Risk factor associated with COVID-19 severity and mortality	8932		Comorbidities	[58]

Cancer	Risk factor associated with COVID-19 severity	32,404		Comorbidities	[59]
Non-thyroidal illness syndrome (NTIS)	Predictor of adverse outcomes in COVID-19 patients predominantly of mild-to-moderate severity	367		Comorbidities	[72]
Chest CT examination	Predictor of COVID-19 severity	1078	AUC of 0.91	Radiographic features	[64]
Quantitative CT imaging	Early prediction of COVID-19 severity	74		Radiographic features	[63]

6.2.2.3 Comorbidities

Patients with COVID-19 present a wide range of clinical phenotypes that can be used to help identify high risk and critically ill patients. Primarily transmitted by respiratory droplets, the most common clinical feature of symptomatic individuals with COVID-19 is the acute respiratory manifestations, including cough, fever, dyspnea, and fatigue [52, 53]. Besides, the presence of comorbidities is also tightly associated with the severity and clinical outcomes of COVID-19. Hypertension, diabetes, and cardiovascular disease have been identified as important risk factors for COVID-19 severity since the first description of the virus disease [54]. For example, by performing a meta-analysis, Lippi et al. [55] found that hypertension was associated with an approximately 2.5-fold increase in the risk of severe and poor prognosis for COVID-19, especially in the elderly; Mantovani and colleagues [56] found that preexisting diabetes was associated with a nearly twofold increased risk of severe illness and an approximate threefold higher risk of in-hospital death of COVID-19 patients; Li et al. [57] discovered that acute cardiac injury was associated with the severity and prognosis of patients with COVID-19. Furthermore, chronic kidney diseases and cancer have recently emerged as frequent comorbidities associated with severity of COVID-19 patients. For example, a meta-analysis of 42 studies involving 8932 patients showed that COVID-19 patients had a strikingly increased risk of developing severe illness or death [58]; Ofori-Asenso et al. [59] found that COVID-19 patients with cancer face higher risk of severity than those without cancer. Therefore, the presence of preexisting comorbidity should be regarded as a critical factor in risk stratification of COVID-19 severity and outcomes.

As mentioned above, SARS-CoV-2 infection of host cells is triggered by the binding of the virus spike protein to the ACE2 receptor, which also has a vital role in the development of cardiovascular diseases, hypertension, and diabetes. Thus, one possible mechanism by which comorbidity increases the risk of severity and death is that the increased ACE2 expression promoted virus entry [60].

6.2.2.4 Radiographic Features

The clinical manifestations of COVID-19 range from asymptomatic infection to acute respiratory illness and respiratory failure. Chest radiography (CXR) and computed tomography (CT) scan have been widely used to detect the distribution of lung abnormalities. Hui and colleagues found that the degree of CXR abnormalities can distinguish patients with severe COVID-19, and its performance (AUC $= 0.987$) was comparable to or better than that of well-characterized laboratory markers [61]. Liu et al. found that the alteration of radiographic features examined by high-resolution computed tomography (HRCT) scans, including ground-glass opacity, nodular opacities, consolidation, air bronchogram, and pleural effusion, were valuable in assessing disease severity and viral clearance for COVID-19 [62]. Li et al. [63] reported that 5 days after the appearance of initial symptoms, CT could predict the COVID-19 patients who progressed to severe symptoms later with 95% confidence. Similarly, a study enrolling 1078 patients with COVID-19 pneumonia showed that CT opacity scores can accurately distinguish the critical patients with excellent performance (AUC = 0.91) [64]. Therefore, CXR and CT examinations possess key clinical value for the follow-up of COVID-19 pneumonia.

6.2.3 Cytokine Storm for Classification of COVID-19 Severity Prediction

The dysregulation of cytokines in COVID-19 patients is one of the main contributors to death, which is technically named cytokine storm syndrome (CSS). In fact, CSS is not a new syndrome from this pandemic but was proposed about one century ago [73]. Several recent worldwide viral pandemics, such as SARS, H1N1, and MERS, arise scientists' attention to CSS. General mechanisms of CSS can be described as the over-expression of several pro-inflammatory cytokines like IFN-γ that break the balance of immune system and cause the excessive immune response, leading to a systematic over-inflammation [74]. Due to its tight association with high lethality, a precise evaluation and classification of CSS are highly demanded.

During the COVID-19 pandemic, numerous investigations have noticed positive correlations between the severities of CSS and COVID-19. As early as the beginning of this outbreak, Huang et al. have been aware of the existence of cytokine storm in patients requiring ICU admission [75]. Following studies from Horby et al., Zhu et al., Del Valle et al., Mathew et al. and so on gave further detailed illustration of the positive correlations from both biological and clinical views [76–79]. Besides, serum concentration of several pro-inflammatory cytokines (e.g., IL-3 and IL-6) was confirmed to be an excellent indicator of COVID-19 severity (Table 6.2) [80, 81]. Especially, due to the rapid increases in the number of patients with COVID-19, overloading of the capacity of public health services may result in a shortage in doctors and computational resources for medical image analysis, compared to which cytokine testing is much time-saving. Therefore, these findings raise a potential classification strategy that the measurement of CSS severity may serve as one of the grading standards for COVID-19.

6.2.4 Biomarkers for Prediction of COVID-19 Severity

To ensure timely treatment, it is imperative to identify effective biomarkers that can stratify the severity of COVID-19 patients. Currently, besides the clinical features and cytokine storm mentioned above, disordered expression pattern of proteins involved in oxidative stress, extracellular vesicles (EVs), and immune response has

Table 6.2 Biomarkers for prediction of COVID-19 severity

Biomarker	Function	Number of patients	Performance	Study
Native thiol	Biomarker for predicting COVID-19 severity	144	AUC of 0.83	[43]
Thiol	Biomarker for predicting COVID-19 severity	517	AUC of 0.949	[86]
Native thiol	Predictor for disease severity in both children and adults with COVID-19	79 children and 74 adults	AUC of 0.614 for distinguishing pediatric patients; AUC of 0.701 for distinguishing adult patients	[87]
Total thiol	Predictor for disease severity in both children and adults with COVID-19	79 children and 74 adults	AUC of 0.618 for distinguishing pediatric patients; AUC of 0.699 for distinguishing adult patients	[87]
EV COPB2	Early prediction of COVID-19 severity	31	AUC of 0.85	[89]
sRAGE	Predictor of MV need in COVID-19 inpatients and mortality	164	AUC of 0.871 for predicting MV; AUC of 0.903 for predicting mortality	[93]
sRAGE	Risk factor associated with mortality of COVID-19 patients	50		[94]
PSP	Predictor of mortality in COVID-19 patients	75	AUC of 0.75	[96]
PSP	Biomarker for predicting COVID-19 severity	6		[97]
IL-3	Prognostic marker for the outcome of severe COVID-19	105		[81]
IL-6	Indicator of COVID-19 severity	901	AUC of 0.97 in predicting in-hospital death	[80]
IL-6	Indicator of COVID-19 severity	77	AUC of 0.95	[44]
IL-6	Predictor of COVID-19 severity in adults	60	AUC of 0.70	[98]
IP-10	Predictor of COVID-19 progression and severity	150	AUC of 0.90	[41]
IP-10	Predictor of COVID-19 severity in children	60	AUC of 0.77	[98]

been identified as promising biomarkers for predicting the severity and mortality of COVID-19 (Table 6.2).

Oxidative stress is defined as the disproportion in the oxidant-antioxidant balance marked by excessive production of reactive oxygen species (ROS) [82]. It has been demonstrated that various viral infections are accompanied by oxidative stress, which further affects the disease pathogenesis, including immune and inflammatory responses, as well as cell and tissue damage [83]. Thiols are the main component of the antioxidant defense system involved in oxidative stress and a key indicator of the redox state of cells [84]. The maintenance of thiol-disulfide homeostasis is critical for viral entry and fusion into target cells, and can be impacted by oxidative stress [85]. Emerging evidence suggests that oxidation of thiols to disulfides modulated by oxidative stress may elevate the binding affinity of the spike protein to the ACE2 receptor, thereby exacerbating the severity of COVID-19 [85]. The research results of Kalem et al. [43] confirmed the important role of oxidative stress in COVID-19 pathogenesis, and showed that the level of native thiol is a highly sensitive biomarker for COVID-19 severity stratification (AUC = 0.83). Similarly, by comparing the thiol levels of patients with mild, moderate, severe, and critical COVID-19, Erel et al. [86] found that the thiol level was negatively correlated with the disease severity and could serve as an independent risk factor. Aykac et al. [87] also identified serum levels of native thiol and total thiol as independent predictors of COVID-19 severity in children and adults.

Extracellular vesicles (EVs) (e.g., exosomes and microvesicles) are responsible for transferring information to target cells, playing a pivotal role in regulating cell communication under physiological and pathological processes. A variety of RNAs and proteins in serum exosomes have been identified as potent and highly specific disease biomarkers. Especially, exosomes derived from virus-infected cells could spread viral infection by delivering viral proteins to normal cells [88], highlighting its potential role in mediating SARS-CoV-2 infection. Fujita et al. [89] found that the abundance of EV COPB2 protein in serum of mild COVID-19 patients was higher than that of severe COVID-19 patients. Moreover, EV COPB2 expression level proved to be an important predictor of COVID-19 severity with high accuracy (AUC = 0.85). These findings suggest that EV protein may be a key biomarker for monitoring the course of COVID-19 and assessing the severity of the disease.

Receptor for advanced glycation end product (RAGE), a member of the immunoglobulin superfamily, has been reported to play an important role in lung inflammation and pathogen-induced pneumonia [90]. In particular, the serum level of soluble RAGE (sRAGE) is an indicator of the status of bacterial infection, inflammatory response and lung epithelial injury as well as a predictor of the progression of acute respiratory distress syndrome (ARDS) patients without COVID-19 [91, 92]. Recently, Lim et al. [93] found that sRAGE level in serum positively correlated with COVID-19 severity and could serve as an excellent biomarker for predicting the mortality (AUC = 0.903) and the need for MV (AUC = 0.871) in patients with COVID-19. Kapandji et al. [94] discovered that plasma sRAGE levels in ARDS patients with COVID-19 were higher than that in ARDS patients without COVID-19.

Presepsin (PSP), a soluble CD14 subtype, is known to modulate immune responses through interactions with T and B cells [95]. Currently, increasing evidence indicated that PSP could serve as a powerful biomarker for early diagnosis and prognosis prediction in patients with pneumonia. The data from Zaninotto et al. [96] showed that COVID-19 patients with PSP values higher than 250 ng/L spent significantly longer in the ICU than patients with lower PSP values. And ROC curve analysis showed that the AUC value of PSP was 0.72 in predicting mortality, revealing its prognostic role in COVID-19. Besides, Fukada and colleagues [97] also identified PSP as a potential biomarker for predicting COVID-19 severity.

6.3 Models for Classification and Prediction of COVID-19 Severity

6.3.1 Scores or Indexes for COVID-19 Severity Measurement at the Individual Level

Early identification of COVID-19 patients who are at high risk of severe illness and/or mortality is of great significance for optimizing clinical decision-making and allocating medical resources. Recently, based on outcome-related variables, a variety of prognostic scores or indexes have been developed to facilitate the detection of COVID-19 patients at high risk of severe or critical illness (Table 6.3).

Various conventional pneumonia severity scoring systems proposed to determine the outcome of community-acquired pneumonia (CAP), such as the Pneumonia Severity Index/Pneumonia Outcome Study Trial (PSI/PORT), CURB-65, and Severe Community-acquired pneumonia (SCAP), have proven to be strong indicators of COVID-19 severity and mortality. For example, Anurag et al. [99] found that the SCAP score could serve as a useful screening tool for distinguishing severe cases of COVID-19 with high sensitivity (0.905) and specificity (0.842), and SCAP, PSI/PORT, CURB-65 scores were all accurate predictors of 14-day mortality. Moreover, when comparing the performance in predicting mortality in COVID-19 patients, the SCAP score was superior to the PSI/PORT and CURB-65 scores [99], and the accuracy of the PSI/PORT score was higher than that of the CURB-65 score [99, 100].

The COVID-GRAM score was firstly established to accurately predict the risk of developing critical illness in Chinese COVID-19 patients based on 10 variables frequently measured upon hospital admission [101]. Armiñanzas et al. [102] further explored the ability of COVID-GRAM score in predicting severity among Caucasian patients with COVID-19. They found that the COVID-GRAM score exhibited excellent accuracy for evaluating disease severity, and it was identified as an independent indicator of critical illness (AUC $=$ 0.779). In terms of predicting 30-day mortality, the COVID-GRAM score (AUC $=$ 0.88) shows a higher distinguishing performance than CURB-65 score (AUC $=$ 0.83) [102].

Table 6.3 Scores or indexes for COVID-19 Severity measurement at the individual level

Scores or indexes	Function	Included parameters	Number of patients	Performance	Study
SCAP score	Predicting disease severity and 14-day mortality of COVID-19	Arterial pH, systolic pressure, confusion, BUN, respiratory rate, X-ray multilobar bilateral, PaO_2, and age	122	AUC of 0.873 for predicting severity; AUC of 0.963 for predicting 14-day mortality	[99]
PSI/PORT score	Predicting disease severity and 14-day mortality of COVID-19	Age, sex, nursing home resident, neoplastic disease, liver disease, congestive heart failure, cerebrovascular disease, renal disease, altered mental status, respiratory rate, systolic blood pressure, temperature, pulse, arterial pH, BUN, sodium, glucose, hematocrit, PaO_2, and pleural effusion on X-ray	122	AUC of 0.713 for predicting severity; AUC of 0.953 for predicting the 14-day mortality	[99]
CURB-65 score	Predicting disease severity and 14-day mortality of COVID-19	Confusion, urea, respiratory rate, blood pressure and age	122	AUC of 0.643 for predicting severity; AUC of 0.950 for predicting the 14-day mortality	[99]
CURB-65 score	Predicting COVID-19 severity		523	AUC of 0.727 for predicting critical illness; AUC of 0.835 for predicting 30-day mortality	[102]
COVID-GRAM score	Predicting COVID-19 severity	CXR abnormality, age, hemoptysis, dyspnea, unconsciousness, number of comorbidities, cancer history, NLR ratio, lactate dehydrogenase and direct bilirubin	523	AUC of 0.779 for predicting critical illness; AUC of 0.88 for predicting 30-day mortality	[102]

(continued)

Table 6.3 (continued)

Scores or indexes	Function	Included parameters	Number of patients	Performance	Study
Bennouar et al.'s risk score	Early prediction of COVID-19 severity and in-hospital mortality	Age, blood urea nitrogen, LDH, NLR, CRP, albumin, and natremia	570	AUC of 0.74 and 0.90, respectively for severity and mortality prediction	[112]
COVID-19 index	Predicting COVID-19 severity	D-dimer, ESR, and lymphocyte	147	AUC of 0.843	[104]
COVID-19 severity index	Predicting COVID-19 severity	Age, male gender, heart failure, COPD, diabetes with end-organ damage, chest X-ray, respiratory rate, SpO_2, SpO_2 in COPD, supplemental O_2, systolic BP, pulse, temperature, dyspnea, D-dimer, lymphocytes and platelets	220	AUC of 0.94 when applied 24 h prior to transferal; AUC of 0.88 when applied 48 h prior to ICU admission	[105]
PSI	Predicting 30-day mortality in patients with COVID-19	Age, gender, long-term care facility resident, accompanying disease, symptoms at diagnosis and laboratory measurements	681	AUC of 0.91	[100]
KPI score	Predicting COVID-19 severity	Age, CRP, procalcitonin, lymphocyte percentage, monocyte percentage and serum albumin	581	AUC of 0.843 in the training dataset; AUC of 0.794 in the validation cohort	[106]
SII	Determining COVID-19 severity	Platelet count and NLR	233		[37]
Zhang et al.'s scoring system	Predicting COVID-19 severity	Age, white blood cell count, neutrophil, glomerular filtration rate, and myoglobin	80	AUC of 0.906	[6]
PNI	Discriminating COVID-19 severity	Serum albumin and total lymphocyte count	101	AUC of 0.790	[103]
PNI	Predicting COVID-19 severity	Albumin and LYM counts	114	AUC of 0.76	[51]

EPI-SCORE	Predicting COVID-19 severity	WHO severity classification, acute renal failure, age, LDH levels, lymphocyte count, and aPTT at admission	295	AUC of 0.91	[113]
CT severity score	Predicting clinical severity and outcome in COVID-19	Defined by summing up individual scores from 20 lung subsegments visualized on CT chest scan	67	AUC of 0.87	[107]
CT severity score	Predicting COVID-19 severity		500	AUC of 0.960	[108]
Total CT score	Predicting the prognosis of COVID-19 patients.		134	AUC of 0.817	[109]
Visual-coronary artery calcification score (V-CACS)	Predicting clinical severity and outcome in COVID-19	Calculated based on 2016 society of cardiovascular computed tomography (SCCT)/society of thoracic radiology (STR) guidelines on coronary artery calcium scoring of non-cardiac chest CT scans	67	AUC of 0.75	[107]
CXR score	Predicting clinical outcomes in hospitalized patients with COVID-19		240	AUC of 0.685	[110]
LUS score	Predicting the likelihood of COVID-19 patients progressing to severe disease		31	AUC of 0.910	[111]

The prognostic nutritional index (PNI) calculated based on the serum albumin level and total lymphocyte count, is known to be a risk-stratified tool reflecting the immune and nutritional status of patients. Wang et al. [103] found that the PNI was markedly lower in critical patients with COVID-19 than that in non-critical controls, and demonstrated to be a good factor for distinguishing COVID-19 severity (AUC = 0.790). Xue et al. [51] also found PNI was closely associated with the risk of severe COVID-19. Similarly, multiple score systems based on clinical and biochemical parameters have been developed with satisfying performance in predicting the severity and mortality of COVID-19, such as COVID-19 index [104], COVID-19 Severity Index [105], Kuwait prognosis indicator (KPI) score [106], and systematic immune-inflammation (SII) index [51]. Besides, the role of scoring systems based on lung detection techniques, such as CT severity score [107, 108], total CT score [109], CXR score [110] and lung ultrasound (LUS) score [111], have been demonstrated for predicting disease severity and outcome in COVID-19 patients.

6.3.2 Mathematical Models for Population Level Prediction of the COVID-19 Severity

In epidemiology, it is critical to estimate the transmissibility of pathogens and the severity of the outbreak at a population level, which is commonly determined by parameters, such as basic reproduction number (R_0) [114] and effective reproduction number (R_t or R_e) [115]. Mathematical models have long been used as key tools to estimate the transmission dynamics of infectious diseases via epidemiological parameters, which can provide important clues for understanding the epidemiological situation and assessing whether the control measures taken are having an obvious effect.

Susceptible-infectious-recovered (SIR) model and susceptible-exposed-infectious-recovered (SEIR) model are typical infection expansion models that have recently been established to evaluate the transmission dynamics of COVID-19 around different regions. For example, SIR models have been employed to assess the transmissibility and severity of the first-wave COVID-19 in major cities of China outside Hubei province, estimate the potential impact of relaxing interventions in a possible second wave [116], and simulate the spread of COVID-19 within different communities [117]. SEIR models have been developed to evaluate the scale and time of the epidemic peak as well as the eventual size of the outbreak under different intervention strategies in China [118], simulate COVID-19 scenarios at state level in the United States [119], and forecast the median rate of symptom onset in the first 12 months across African countries [120].

Bayesian models are increasingly being used to estimate the epidemiology of COVID-19 with the advantage of integrating prior information with current information and more fully considering the uncertainties associated with models and parameter values [121]. Saqib et al. [121] introduced a hybrid polynomial-Bayesian ridge regression (PBRR) model that can forecast the progression of COVID-19 outbreak with high accuracy and reliability. This is a typical example of assessing the ability of a modified Bayesian model to predict the progression of the COVID-19 pandemic. Moreover, Bayesian models have contributed to quantify the impact of COVID-19 interventions [122], evaluate biases in the prevalence and severity of COVID-19 reported in Wuhan, China, based on international traveler case data [123], and identify high-mortality risk in hospitalized patients with COVID-19 [124].

Totally, various mathematical models have been developed to estimate the dynamic transmissibility and severity of the COVID-19 pandemic at population level, determine the effectiveness of implemented public health interventions, and further guide policy development.

6.3.3 Artificial Intelligence (Machine Learning or Deep Learning) Models for Classification of COVID-19 Severity

Artificial intelligence (AI) is an innovative computer technology that uses computational methods to simulate human intelligent behavior and critical thinking. Given the outstanding ability to analyze mountains of complex medical data, AI-based techniques, especially machine learning and deep learning models, have been recognized as a helpful tool to fight the COVID-19 by identifying high-risk patients, evaluating disease severity and mortality, and predicting the scale of disease outbreaks in different regions. Most of these AI models have been developed based on information or variables, such as biochemical indicators, medical images, and clinical data.

Rapid biochemical tests and protein profiling in blood enables monitoring of multiple critical biomarkers, which indicate changes in cell/organ functions and are frequently used to predict the severity of diverse diseases. Increasing evidence suggest that machine learning or deep learning approaches based on those available biochemical parameters can improve the quality of clinical decision-making for patients with COVID-19. For example, Aktar et al. [125] developed predictive machine learning models using a large number of routine blood parameters and proved that these methods can predict COVID-19 disease severity with high accuracy. Cobre et al. [126] implemented four machine learning-based models including artificial neural networks (ANN), decision trees (DT), discriminant analysis by partial least squares (PLS-DA), and k-nearest neighbors (KNN), for COVID-19 diagnosis and severity prediction based on biochemical, hematological, and urinary parameters. They found that all of the four models can effectively diagnose COVID-19 patients and predict COVID-19 severity with high accuracy (>84%) and ferritin was identified as the most critical variable in all models. Furthermore, based on the key proteins determined by blood protein profiling, Yaşar et al. [127] compared the accuracy of machine learning (gradient boosted trees (GBTs) and random forest (RF)) and deep learning approaches in the prediction of COVID-19 severity. Their results indicated that the proposed GBTs model achieved the highest accuracy in predicting COVID-19 disease severity compared with other models.

CT, CXR, and LUS are noninvasive tools for monitoring the progression and severity of lung diseases. Especially, automated severity assessment of COVID-19 from medical images through AI-based models plays a pivotal role in identifying patients that are in urgent need of intensive care. The CT scan is capable of monitoring the manifestations of COVID-19 during the disease progression with high sensitivity by providing a three-dimensional display of the pulmonary vessels. Agarwal and colleagues [128] developed a novel block imaging approach for effectively detecting COVID-19 severity based on CT images and demonstrated that deep learning exhibited superior performance compared with machine learning models. Yu et al. [129] found that the deep learning model DenseNet-201 combined with cubic SVM classifiers achieved high accuracy for rapid discrimination of COVID-19 severity from CT scans. The CXR is another attractive radiological imaging method due to its flexibility and low cost. Cohen et al. [130] proposed a deep learning model for monitoring the severity of COVID-19 as well as treatment efficacy using CXR images. In addition, as a powerful visual-inspection based approach, the LUS provides real-time and high-resolution views of the lungs without the risk of radiation. Dastider et al. [131] developed an effective deep learning model that integrated convolutional neural network (CNN) with Long Short-Term Memory (LSTM) for predicting COVID-19 severity from LUS images.

Given the important role of medical imaging (especially CT) and clinical features (including symptoms and laboratory findings) in the diagnosis and prognosis of COVID-19, AI models integrated with CT features and clinical variables have also been demonstrated to predict COVID-19 severity and progression risk to critical illness with high accuracy [132]. For example, the machine learning model extracted features from CT images and clinical laboratory measurements can identify severe cases of COVID-19 with a cross-validated AUC value of 0.93, indicating its role in predicting COVID-19 severity [133]. The open database containing CT images, clinical features and laboratory-confirmed status was able to discriminate severe cases via deep learning algorithm with high accuracy (AUC = 0.884) [134]. Shiri et al. [135] proved that combination of radiomic and clinical features can effectively improve the prognostic performance of the machine learning model in predicting survival outcomes.

Furthermore, a super learner ensemble of 14 statistical and machine learning models were developed and proved to be more powerful for predicting disease severity of COVID-19 patients with cardiovascular conditions, highlighting the great potential of super learning ensembles in improving predictive performance [136]. Therefore, AI-based models contribute greatly to the classification and prediction of COVID-19 severity.

6.4 Summary and Perspectives

The outbreak of COVID-19 continues to cause high mortality, pose a huge threat to global public healthcare systems and bring profound economic implications due to the emergence of novel SARS-CoV-2 variants. Especially, the Delta variant with increased transmissibility triggers new waves of COVID-19 pandemic throughout the world. As patients with COVID-19 are presenting with a spectrum of clinical severity, ranging from asymptomatic disease to severe, life-threatening infections requiring ICU admission or MV, the most urgent issue in the management and intervention of COVID-19 patients is to monitor disease severity and distinguish individuals at high risk of mortality.

At present, considerable efforts have been made to identify risk factors and biomarkers associated with the severity and mortality of COVID-19 disease and develop effective predictive models for detecting COVID-19 severity from the individual level or population level. On the one hand, increasing evidence indicates that a variety of clinical features, including demographic, biochemical, hematological, inflammatory and radiographic findings, are closely related to the severity of COVID-19, and that the cytokine storm is a major driver of critical illness in COVID-19. On the other hand, based on the clinical risk factors and medical images, a growing number of prognostic scores/indexes and AI models have been developed for individual level prediction of COVID-19 severity, and multiple mathematical models have been utilized to predict the local and global transmissibility, infectivity and severity of COVID-19 at the population level.

However, current methods for predicting or classifying the severity of COVID-19 still have some key issues that need to be addressed. First, considering the impact of new variants on COVID-19 outbreak and effectiveness of vaccination, increasing effort need to be put on the mechanism of SARS-CoV-2 mutation to effectively fight with COVID-19. Second, only physiological index, CSS grading system or image analysis is not enough for precise classification, a combined model or standard, which includes both image, and clinical and physiological indexes, should be constructed for COVID-19 grading. Moreover, to achieve superior predictive performance, novel scoring systems and AI models are required to be developed based on large-scale clinical data, comprehensively taking population, region, and environment into account.

In summary, it is of great significance to identify reliable severity predictors and develop effective models for monitoring disease progression, classifying patients, optimizing treatment, and allocating appropriate medical resources during the COVID-19 pandemic.

Acknowledgments This work was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019), National Natural Science Foundation of China (Grant Nos. 32070671, 31900490, 31770903), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 19KJB180027), 2021 Anhui

Provincial Universities Excellent Top-notch Talents Training Program (gxyq2021200), Anhui Science and Technology University Stabilization and Introduction of Talents (XWWD202101).

References

1. Wu Z, McGoogan JM (2020) Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA 323(13):1239–1242
2. Arabi YM, Murthy S, Webb S (2020) COVID-19: a novel coronavirus and a novel challenge for critical care. Intensive Care Med 46(5):833–836
3. Yang Y et al (2020) Plasma IP-10 and MCP-3 levels are highly associated with disease severity and predict the progression of COVID-19. J Allergy Clin Immunol 146(1):119–127. e4.
4. Broman N et al (2021) IL-6 and other biomarkers as predictors of severity in COVID-19. Ann Med 53(1):410–412
5. Overmyer KA et al (2021) Large-scale multi-omic analysis of COVID-19 severity. Cell Syst 12(1):23–40 e7
6. Zhang C et al (2020) A novel scoring system for prediction of disease severity in COVID-19. Front Cell Infect Microbiol 10:318
7. Yan L et al (2020) Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. MedRxiv
8. Lauring AS, Hodcroft EB (2021) Genetic variants of SARS-CoV-2—what do they mean? JAMA 325(6):529–531
9. Huang Y et al (2020) Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin 41(9):1141–1149
10. Zhang Y, Xi H, Juhas M (2020) Biosensing detection of the SARS-CoV-2 D614G mutation. Trends Genet
11. Grubaugh ND et al (2021) Public health actions to control new SARS-CoV-2 variants. Cell 184(5):1127–1132
12. Zhang L et al (2020) SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun 11(1):1–9
13. Saha S et al (2021) COVID-19 rise in Bangladesh correlates with increasing detection of B. 1.351 variant, 2021. BMJ Glob Health 6(5):e006012
14. Emary KR et al (2021) Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 variant of concern 202012/01 (B. 1.1. 7): an exploratory analysis of a randomised controlled trial. Lancet 397(10282):1351–1362
15. Campbell F et al (2021) Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Eur Secur 26(24):2100509
16. Starr TN et al (2020) Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 182(5):1295–1310. e20
17. Davies NG, et al 2021 Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1.7 in England. Science. 372(6538)
18. Volz E et al (2021) Assessing transmissibility of SARS-CoV-2 lineage B. 1.1. 7 in England. Nature 593(7858):266–269
19. Graham MS et al (2021) Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B. 1.1. 7: an ecological study. Lancet Public Health 6(5):e335–e345
20. Bager P, et al (2021) Increased risk of hospitalisation associated with infection with SARS-CoV-2 lineage B. 1.1.7 in Denmark

21. Abdool Karim SS, de Oliveira T (2021) New SARS-CoV-2 variants-clinical, public health, and vaccine implications. N Engl J Med
22. Ramanathan M, et al (2021) SARS-CoV-2 B. 1.1. 7 and B. 1.351 spike variants bind human ACE2 with increased affinity. Lancet Infectious Diseases
23. Barbosa G et al (2021) Rapid spread and high impact of the variant of concern P. 1 in the largest city of Brazil. J Infect 83(1):119–145
24. Faria NR et al (2021) Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. Science 372(6544):815–821
25. Coutinho RM et al (2021) Model-based evaluation of transmissibility and reinfection for the P. 1 variant of the SARS-CoV-2. MedRxiv
26. de Souza FSH et al (2021) Second wave of COVID-19 in Brazil: younger at higher risk. Eur J Epidemiol 36(4):441–443
27. Planas D et al (2021) Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature:1–7
28. Bernal JL, et al (2021) Effectiveness of COVID-19 vaccines against the B. 1.617.2 variant. MedRxiv
29. Challen R, et al (2021) Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B. 1.617.2 in England. MedRxiv
30. Salvatore M, et al (2021) Resurgence of SARS-CoV-2 in India: potential role of the B. 1.617.2 (Delta) variant and delayed interventions. MedRxiv
31. Gallo Marin B et al (2021) Predictors of COVID-19 severity: a literature review. Rev Med Virol 31(1):1–10
32. Barek MA, Aziz MA, Islam MS (2020) Impact of age, sex, comorbidities and clinical symptoms on the severity of COVID-19 cases: a meta-analysis with 55 studies and 10014 cases. Heliyon 6(12):e05684
33. Pijls BG et al (2021) Demographic risk factors for COVID-19 infection, severity, ICU admission and death: a meta-analysis of 59 studies. BMJ Open 11(1):e044640
34. Mina A, Van Besien K, Platanias LC (2020) Hematological manifestations of COVID-19. Leuk Lymphoma 61(12):2790–2798
35. Asghar MS et al (2020) Hematological parameters predicting severity and mortality in COVID-19 patients of Pakistan: a retrospective comparative analysis. J Commun Hosp Intern Med Perspect 10(6):514–520
36. Li X et al (2020) Predictive values of neutrophil-to-lymphocyte ratio on disease severity and mortality in COVID-19 patients: a systematic review and meta-analysis. Crit Care 24(1):647
37. Rokni M et al (2020) Comparison of clinical, para-clinical and laboratory findings in survived and deceased patients with COVID-19: diagnostic role of inflammatory indications in determining the severity of illness. BMC Infect Dis 20(1):869
38. Shang W et al (2020) The value of clinical parameters in predicting the severity of COVID-19. J Med Virol 92(10):2188–2192
39. Waris A et al (2021) Evaluation of hematological parameters as an indicator of disease severity in Covid-19 patients: Pakistan's experience. J Clin Lab Anal 35(6):e23809
40. Fu J et al (2020) The value of serum amyloid A for predicting the severity and recovery of COVID-19. Exp Ther Med 20(4):3571–3577
41. Haroun RA, Osman WH, Eessa AM (2021) Interferon-γ-induced protein 10 (IP-10) and serum amyloid A (SAA) are excellent biomarkers for the prediction of COVID-19 progression and severity. Life Sci 269:119019
42. Ghweil AA et al (2020) Characteristics, outcomes and indicators of severity for COVID-19 among sample of ESNA quarantine Hospital's patients, Egypt: a retrospective study. Infect Drug Resist 13:2375–2383
43. Kalem AK et al (2021) A useful and sensitive marker in the prediction of COVID-19 and disease severity: thiol. Free Radic Biol Med 166:11–17
44. Wang M et al (2020) Differences of inflammatory and non-inflammatory indicators in coronavirus disease-19 (COVID-19) with different severity. Infect Genet Evol 85:104511

45. Guner Ozenen G et al (2021) Demographic, clinical, and laboratory features of COVID-19 in children: the role of mean platelet volume in predicting hospitalization and severity. J Med Virol 93(5):3227–3237
46. Hançerli Törün S, et al (2021) Plasma D-dimer: a promising indicator of COVID-19 infection severity or only an acute phase reactant. Minerva Pediatr (Torino)
47. Wungu CDK et al (2021) Meta-analysis of cardiac markers for predictive factors on severity and mortality of COVID-19. Int J Infect Dis 105:551–559
48. Ahmed S et al (2021) Evaluation of serum ferritin for prediction of severity and mortality in COVID-19—a cross sectional study. Ann Med Surg (Lond) 63:102163
49. Husain-Syed F et al (2021) Immunoglobulin deficiency as an indicator of disease severity in patients with COVID-19. Am J Physiol Lung Cell Mol Physiol 320(4):L590–L599
50. Ok F et al (2021) Predictive values of blood urea nitrogen/creatinine ratio and other routine blood parameters on disease severity and survival of COVID-19 patients. J Med Virol 93(2):786–793
51. Xue G et al (2020) Novel serological biomarkers for inflammation in predicting disease severity in patients with COVID-19. Int Immunopharmacol 89(Pt A):107065
52. Lai C-C et al (2020) Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): facts and myths. J Microbiol Immunol Infect 53(3):404–412
53. Berlin DA, Gulick RM, Martinez FJ (2020) Severe covid-19. N Engl J Med 383(25):2451–2460
54. ERA-EDTA Council; ERACODA Working Group (2021) Chronic kidney disease is a key risk factor for severe COVID-19: a call to action by the ERA-EDTA. Nephrol Dial Transplant 36(1):87–94
55. Lippi G, Wong J, Henry BM (2020) Hypertension and its severity or mortality in coronavirus disease 2019 (COVID-19): a pooled analysis. Pol Arch Intern Med 130(4):304–309
56. Mantovani A et al (2020) Diabetes as a risk factor for greater COVID-19 severity and in-hospital death: a meta-analysis of observational studies. Nutr Metab Cardiovasc Dis 30(8):1236–1248
57. Li B et al (2020) Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. Clin Res Cardiol 109(5):531–538
58. Wang B et al (2021) The involvement of chronic kidney disease and acute kidney injury in disease severity and mortality in patients with COVID-19: a meta-analysis. Kidney Blood Press Res 46(1):17–30
59. Ofori-Asenso R et al (2020) Cancer is associated with severe disease in COVID-19 patients: a systematic review and meta-analysis. Ecancermedicalscience 14:1047
60. Zheng Y-Y et al (2020) COVID-19 and the cardiovascular system. Nat Rev Cardiol 17(5):259–260
61. Hui TC et al (2020) Clinical utility of chest radiography for severe COVID-19. Quant Imaging Med Surg 10(7):1540
62. Liu X et al (2020) Temporal radiographic changes in COVID-19 patients: relationship to disease severity and viral clearance. Sci Rep 10(1):1–9
63. Li K et al (2021) Early prediction of severity in coronavirus disease (COVID-19) using quantitative CT imaging. Clin Imaging 78:223–229
64. Jafari R et al (2021) Identification, monitoring, and prediction of disease severity in patients with COVID-19 pneumonia based on chest computed tomography scans: a retrospective study. Adv Exp Med Biol 1321:265–275
65. Palaiodimos L et al (2020) Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx. New York Metabolism 108:154262
66. Li X et al (2020) Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol 146(1):110–118
67. Tan L et al (2020) Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. Signal Transduct Target Ther 5(1):33

68. Pranata R et al (2020) Hypertension is associated with increased mortality and severity of disease in COVID-19 pneumonia: a systematic review, meta-analysis and meta-regression. J Renin-Angiotensin-Aldosterone Syst 21(2):1470320320926899
69. Gregory JM et al (2021) COVID-19 severity is tripled in the diabetes community: a prospective analysis of the pandemic's impact in type 1 and type 2 diabetes. Diabetes Care 44(2):526–532
70. Bode B et al (2020) Glycemic characteristics and clinical outcomes of COVID-19 patients hospitalized in the United States. J Diabetes Sci Technol 14(4):813–821
71. Chen Q et al (2020) Cardiovascular manifestations in severe and critical patients with COVID-19. Clin Cardiol 43(7):796–802
72. Lui DTW et al (2021) Role of non-thyroidal illness syndrome in predicting adverse outcomes in COVID-19 patients predominantly of mild-to-moderate severity. Clin Endocrinol
73. Yongzhi X (2021) COVID-19-associated cytokine storm syndrome and diagnostic principles: an old and new issue. Emerg Microbes Infect 10(1):266–276
74. Fajgenbaum DC, June CH (2020) Cytokine storm. N Engl J Med 383(23):2255–2273
75. Huang C et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506
76. RECOVERY Collaborative Group et al (2021) Dexamethasone in hospitalized patients with Covid-19. N Engl J Med 384(8):693–704
77. Zhu Z et al (2020) Clinical value of immune-inflammatory parameters to assess the severity of coronavirus disease 2019. Int J Infect Dis 95:332–339
78. Del Valle DM et al (2020) An inflammatory cytokine signature predicts COVID-19 severity and survival. Nat Med 26(10):1636–1643
79. Mathew D et al (2020) Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science 369(6508)
80. Zhang J et al (2020) Serum interleukin-6 is an indicator for severity in 901 patients with SARS-CoV-2 infection: a cohort study. J Transl Med 18(1):406
81. Bénard A et al (2021) Interleukin-3 is a predictive marker for severity and outcome during SARS-CoV-2 infections. Nat Commun 12(1):1–8
82. Yoshikawa T, Naito Y (2002) What is oxidative stress? Jpn Med Assoc J 45(7):271–276
83. Chernyak B et al (2020) COVID-19 and oxidative stress. Biochem Mosc 85(12):1543–1553
84. Rodosskaia N, Chernousova G (2010) Immune system and thiols: some peculiarities of thiol exchange. Comp Immunol Microbiol Infect Dis 33(1):65–71
85. Suhail S et al (2020) Role of oxidative stress on SARS-CoV (SARS) and SARS-CoV-2 (COVID-19) infection: a review. Protein J:1–13
86. Erel Ö et al (2021) A sensitive indicator for the severity of COVID-19: thiol. Turk J Med Sci 51(3):921–928
87. Aykac K et al (2021) Oxidant and antioxidant balance in patients with COVID-19. Pediatr Pulmonol
88. Pegtel DM et al (2010) Functional delivery of viral miRNAs via exosomes. Proc Natl Acad Sci 107(14):6328–6333
89. Fujita Y et al (2021) Early prediction of COVID-19 severity using extracellular vesicle COPB2. J Extracell Vesicles 10(8):e12092
90. Dozio E et al (2020) Soluble receptor for advanced glycation end products and its forms in COVID-19 patients with and without diabetes mellitus: a pilot study on their role as disease biomarkers. J Clin Med 9(11):3785
91. Jabaudon M et al (2018) Receptor for advanced glycation end-products and ARDS prediction: a multicentre observational study. Sci Rep 8(1):1–11
92. Jabaudon M et al (2021) Changes in plasma soluble receptor for advanced glycation end-products are associated with survival in patients with acute respiratory distress syndrome. J Clin Med 10(10):2076
93. Lim A et al (2021) Soluble receptor for advanced glycation end products (sRAGE) as a biomarker of COVID-19 disease severity and indicator of the need for mechanical ventilation, ARDS and mortality. Ann Intensive Care 11(1):50

94. Kapandji N et al (2021) Importance of lung epithelial injury in COVID-19 associated acute respiratory distress syndrome: value of plasma sRAGE. Am J Respir Crit Care Med 204(3):359–362
95. Chenevier-Gobeaux C et al (2015) Presepsin (sCD14-ST), an innate immune response marker in sepsis. Clin Chim Acta 450:97–103
96. Zaninotto M et al (2020) Presepsin in risk stratification of SARS-CoV-2 patients. Clin Chim Acta 507:161–163
97. Fukada A et al (2021) Presepsin as a predictive biomarker of severity in COVID-19: a case series. J Med Virol 93(1):99–101
98. Ozsurekci Y et al (2021) Predictive value of cytokine/chemokine responses for the disease severity and management in children and adult cases with COVID-19. J Med Virol 93(5):2828–2837
99. Anurag A, Preetam M (2021) Validation of PSI/PORT, CURB-65 and SCAP scoring system in COVID-19 pneumonia for prediction of disease severity and 14-day mortality. Clin Respir J 15(5):467–471
100. Satici C et al (2020) Performance of pneumonia severity index and CURB-65 in predicting 30-day mortality in patients with COVID-19. Int J Infect Dis 98:84–89
101. Liang W et al (2020) Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. JAMA Intern Med 180(8):1081–1089
102. Armiñanzas C et al (2021) Usefulness of the COVID-GRAM and CURB-65 scores for predicting severity in patients with COVID-19. Int J Infect Dis 108:282–288
103. Wang ZH et al (2020) Predictive value of prognostic nutritional index on COVID-19 severity. Front Nutr 7:582736
104. Dong Y et al (2020) A novel simple scoring model for predicting severity of patients with SARS-CoV-2 infection. Transbound Emerg Dis 67(6):2823–2829
105. Huespe I et al (2020) COVID-19 severity index: a predictive score for hospitalized patients. Med Intensiva (Engl Ed)
106. Jamal MH et al (2020) A biomarker based severity progression indicator for COVID-19: the Kuwait prognosis indicator score. Biomarkers 25(8):641–648
107. Nair AV et al (2021) Utility of visual coronary artery calcification on non-cardiac gated thoracic CT in predicting clinical severity and outcome in COVID-19. Clin Imaging 74:123–130
108. Palwa AR et al (2021) Chest CT severity score as an auxiliary grading tool to COVID-19 pneumonia imaging classification: a tertiary care experience in Pakistan. J Coll Physicians Surg Pak 31(1):14–20
109. Zhou S et al (2020) Chest CT imaging features and severity scores as biomarkers for prognostic prediction in patients with COVID-19. Ann Transl Med 8(21):1449
110. Reeves RA et al (2020) Performance of a severity score on admission chest radiograph in predicting clinical outcomes in hospitalized patients with coronavirus disease (COVID-19). AJR Am J Roentgenol
111. Xian J et al (2021) The clinical value of bedside ultrasound in predicting the severity of coronavirus disease-19 (COVID-19). Ann Transl Med 9(4):336
112. Bennouar S et al (2021) Development and validation of a laboratory risk score for the early prediction of COVID-19 severity and in-hospital mortality. Intensive Crit Care Nurs 64:103012
113. de Terwangne C et al (2020) Predictive accuracy of COVID-19 World Health Organization (WHO) severity classification and comparison with a Bayesian-Method-Based Severity Score (EPI-SCORE). Pathogens 9(11)
114. Dietz K (1993) The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res 2(1):23–41
115. Shim E et al (2020) Transmission potential and severity of COVID-19 in South Korea. Int J Infect Dis 93:339–344

116. Leung K et al (2020) First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. Lancet 395(10233):1382–1393
117. Cooper I, Mondal A, Antonopoulos CG (2020) A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons Fractals 139:110057
118. Yang Z et al (2020) Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 12(3):165
119. IHME COVID-19 Forecasting Team (2020) Modeling COVID-19 scenarios for the United States. Nat Med 27:94–105
120. Van Zandvoort K et al (2020) Response strategies for COVID-19 epidemics in African settings: a mathematical modelling study. BMC Med 18(1):1–19
121. Saqib M (2021) Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. Appl Intell 51(5):2703–2713
122. Anderson SC et al (2020) Quantifying the impact of COVID-19 control measures using a Bayesian model of physical distancing. PLoS Comput Biol 16(12):e1008274
123. Niehus R, et al (2020) Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. MedRxiv
124. Momeni-Boroujeni A et al (2021) A dynamic Bayesian model for identifying high-mortality risk in hospitalized COVID-19 patients. Infect Dis Rep 13(1):239–250
125. Aktar S et al (2021) Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: statistical analysis and model development. JMIR Med Inform 9(4):e25884
126. Cobre AF et al (2021) Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? Comput Biol Med 134:104531
127. Yaşar Ş, Çolak C, Yoloğlu S (2021) Artificial intelligence-based prediction of Covid-19 severity on the results of protein profiling. Comput Methods Prog Biomed 202:105996
128. Agarwal M et al (2021) A novel block imaging technique using nine artificial intelligence models for COVID-19 disease classification, characterization and severity measurement in lung computed tomography scans on an Italian Cohort. J Med Syst 45(3):28
129. Yu Z et al (2020) Rapid identification of COVID-19 severity in CT scans through classification of deep features. Biomed Eng Online 19(1):63
130. Cohen JP et al (2020) Predicting COVID-19 pneumonia severity on chest X-ray with deep learning. Cureus 12(7):e9448
131. Dastider AG, Sadik F, Fattah SA (2021) An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound. Comput Biol Med 132:104296
132. Purkayastha S et al (2021) Machine learning-based prediction of COVID-19 severity and progression to critical illness using CT imaging and clinical data. Korean J Radiol 22(7):1213–1224
133. Li D et al (2020) Prediction of COVID-19 severity using chest computed tomography and laboratory measurements: evaluation using a machine learning approach. JMIR Med Inform 8(11):e21604
134. Ning W et al (2020) Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat Biomed Eng 4(12):1197–1207
135. Shiri I et al (2021) Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients. Comput Biol Med 132:104304
136. Ehwerhemuepha L et al (2021) A super learner ensemble of 14 statistical learning models for predicting COVID-19 severity among patients with cardiovascular conditions. Intell Based Med 5:100030

Chapter 7
Modeling the Virus Infection at the Population Level

Cong Wu, Xuemeng Fan, Tong Tang, and Bairong Shen

Abstract As pointed out by many researchers in the last few decades, differential equations with fractional (non-integer) order differential operators, in comparison with classical integer order ones, have apparent advantages in modeling. A Caputo fractional order system of ordinary differential equations is introduced to model the virus infection at the population level in this chapter. As well known, there are two main methods to study the dynamics of a model: qualitative analysis and numerical modeling. Here the qualitative analysis, including uniqueness, invariant set, and stability, is first presented with intuitive derivation. Then the famous genetic algorithm is introduced to numerically model the dynamics of virus infection, i.e. to adjust the parameters of the Caputo fractional model such that its solution can properly fit real data and predict future.

Keywords Virus infection at the population level; Caputo fractional model; Qualitative analysis; Numerical modeling

7.1 Introduction

The coronavirus disease (COVID-19) broke out in Wuhan, China at the very beginning of 2020. As reported in [1], this disease is caused by the novel coronavirus of zoonotic origin. Currently, the spread of COVID-19 is controlled very well in China. Only a small amount of confirmed cases are reported seldomly. However, outside of China, such as United States of America, Russia, and India, it seems that those people are suffering a serious and deadly situation. For the whole world, the COVID-19 epidemic is still far away from its end. Therefore, it has been significant to predict the trend of epidemic up to this moment.

C. Wu · X. Fan · T. Tang · B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: congwu@wchscu.cn; fanxuemeng@wchscu.cn; tangtong@wchscu.cn; bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_7

Epidemic models prove to be an effective method for prediction. Especially, Caputo fractional order models have shown their advantages in modeling virus infection at the population level. Fractional order operators (derivatives and integrals), as explained in Remark 6.4 in [2], are a very natural tool to model memory-dependent phenomena. They provide an excellent instrument for the description of memory and hereditary properties of various processes, which endows fractional order models, in comparison with classical integer order ones, apparent advantages in modeling, see Preface of [3]. In 2013, Diethelm used a Caputo fractional order model to simulate an outbreak of dengue fever in [4] and pointed out that the parameters at the right hand side of the model must satisfy the dimension requirement there. In the same year, Gonzalez-Parra, Arenas, and Chen-Charpentier applied a Caputo fractional order epidemic model for the simulation of outbreaks of influenza A(H1N1) [5]. In 2019, Almeida, Brito da Cruz, Martins, and Monteiro built a Caputo fractional order epidemiological model for the varicella outbreak among Shenzhen school children, China [6]. In 2021, Xu, Yu, Chen, and Lu adopted a Caputo fractional order model to forecast the epidemic trend of COVID-19 in United States of America [7].

There are two main methods to study the dynamics of a model: qualitative analysis and numerical modeling. This is also reflected in [5–7]. Specifically, in [5], the uniqueness and stability were studied, and the least squares method and the Nelder-Mead algorithm were adopted to adjust the parameters of the Caputo fractional order model; in [6], the stability was also investigated, and the routine $fminsearch$ from the MATLAB Optimization Toolbox was used for fitting the model's parameters; in [7], the invariant set and stability were analyzed, and the internal function $lsqcurvefit$ of MATLAB was applied for numerical modeling. As well known, qualitative analysis may help us to figure out dynamics of models prior to numerically solving them. However, in practice, real models perhaps do not possess those qualitative properties. In this situation, the method other than qualitative analysis—numerical modeling may be taken into consideration. It can be used to adjust parameters such that models fit real data and perform precise prediction.

In terms of their importance in modeling, both the qualitative analysis and numerical modeling will be introduced in this chapter. The former is to include uniqueness, invariant set, and stability, while the latter is to be implemented by a heuristic algorithm—genetic algorithm. Compared to those routine ones, the genetic algorithm [8] has advantages in robustness, convergence, parallelism, and scalability.

7.2 Model Form

The Caputo fractional order ordinary differential equations will be considered as the form of the virus infection model at the population level owning to their apparent advantages, in comparison with classical integer order ones, in modeling.

7.2.1 Caputo Fractional Order Ordinary Differential Equation

At first, referring to a book by Diethelm [2], we introduce definitions for fractional integrals and derivatives. A fractional integral on $L_1[a, b]$ is given by

$$ {}_aI_t^{\gamma} f(t) = \frac{1}{\Gamma(\gamma)} \int_a^t \frac{f(\tau)}{(t-\tau)^{1-\gamma}} \mathrm{d}\tau, \ a \le t \le b, $$

where $\gamma > 0$ and $\Gamma(\cdot)$ is the Gamma function. For an arbitrary non-integer number $p > 0$, the Riemann-Liouville and Caputo fractional derivatives are, respectively, defined by

$$ {}_a^RD_t^p f(t) = D^{[p]+1}[{}_aI_t^{[p]-p+1} f(t)] $$

and

$$ {}_a^CD_t^p f(t) = {}_aI_t^{[p]-p+1}[D^{[p]+1} f(t)], $$

where $[p]$ represents p's integer part; D, RD, ${}^CD$, respectively, denotes the first-order derivative, Riemann-Liouville fractional derivative, Caputo fractional derivative. If $f \in AC^{[p]+1}[a, b]$ (the set of functions with absolutely continuous derivative of order $[p]$), then the fractional derivatives ${}_a^RD_t^q f$ and ${}_a^CD_t^p f$ exist almost everywhere on $[a, b]$ [2]. In particular, for $0 < p < 1$, ${}_a^RD_t^q f$ and ${}_a^CD_t^p f$ exist almost everywhere on $[a, b]$, if $f \in AC[a, b]$ (the set of absolutely continuous functions).

Now it becomes ready to consider the Caputo fractional order ordinary differential equation

$$ \begin{cases} {}_{t_0}^CD_t^{\alpha} x = f(t, x) \\ x(t_0) = x_0, \end{cases} \tag{7.1} $$

where $\alpha \in (0, 1)$; $f : \mathbb{R}_+ \times B(\rho) \to \mathbb{R}^n$, for some $\rho > 0$, is the given vector field function; $t_0 \in \mathbb{R}_+$ is an initial time; and $x_0 \in \mathbb{R}^n$ is an initial value vector. Here $B(\rho) := \{x \in \mathbb{R}^n : ||x|| < \rho\}$, where $|| \centerdot ||$ denotes the Euclidean norm.

Lemma 1 (Existence) *Assume that f is continuous on the closed set $\bar{S} = \{(t, x) : t \in [t_0, t_0 + a], ||x - x_0||_1 \le b\}$, for some $a > 0$, $b > 0$ such that $\bar{S} \subset \mathbb{R}_+ \times B(\rho)$. Then (7.1) has a solution $x(t) \in C[t_0, t_0+h]$, where $h = \min\{a, [b\Gamma(\alpha+1)/M]^{1/\alpha}\}$ and $M = \max_{(t,x)\in\bar{S}} ||f(t, x)||_1$.*

Note that $|| \centerdot ||_1$ denotes the norm given by $||x||_1 = \sum_{i=1}^n |x_i|$, for $x \in \mathbb{R}^n$.

Lemma 2 (Uniqueness) *Assume that f is continuous in t and Lipschitz in x on $\bar{S}$. Then (7.1) has a unique solution $x(t) \in C[t_0, t_0 + h]$.*

Lemma 3 (Global Uniqueness) *Assume that f is continuous in t and Lipschitz in x on $G = [t_0, t_0 + a] \times \mathbb{R}^n$. Then (7.1) has a unique solution $x(t) \in C[t_0, t_0 + a]$.*

Remark 1 As $a = \infty$, i.e. $G = [t_0, \infty] \times \mathbb{R}^n$, (7.1) has a unique solution $x(t) \in C[t_0, \infty)$.

The lemmas above Lemma 1, 2, and 3 are, respectively, extended from Theorem 6.1, Theorem 6.5, and Theorem 6.8 in [2], according to Remark 6.1 in [2].

Definition 1 The constant x^* is an equilibrium point of the Caputo fractional order nonautonomous system in (7.1), if and only if $f(t, x^*) = 0$, for all $t \geq t_0$.

Definition 2 Assume $f(t, 0) \equiv 0$, and let $x(t) = x(t, t_0, x_0)$ denote the solution of (7.1). Then the trivial solution to the Caputo fractional order nonautonomous system in (7.1) is said to be stable, if for any $\epsilon > 0$, $t_0 \geq 0$, there exists a $\delta(t_0, \epsilon) > 0$ such that $||x_0|| < \delta$ implies $||x(t)|| < \epsilon$, for all $t \geq t_0$.

Lemma 4 (Lyapunov's Indirect Method [9]) *Assume $f(t, x) \equiv f(x) = Ax + g(x)$, where $A \in \mathbb{R}^n \times \mathbb{R}^n$ is the Jacobian matrix of f at $\mathbf{0}$, and $g : \mathbb{R}^n \to \mathbb{R}^n$ is of order higher than the linearization in $B(r)$, a neighborhood of $\mathbf{0}$, i.e. $g(0) = 0$ and*

$$\lim_{r \to 0} \sup_{x, y \in B(r), x \neq y} \frac{||g(x) - g(y)||}{||x - y||} = 0.$$

If one of A's nonzero eigenvalues λ satisfies

$$|arg(\lambda)| < \frac{\alpha\pi}{2}$$

and g is Lipschitz on $\mathbb{R}^n$, then the trivial solution of (7.1) is unstable.

7.2.2 Caputo Fractional Order Model of the Virus Infection at the Population Level

In 1995, a model of the virus infection at the population level was proposed in [10] as

$$\begin{cases} \dot{S}(t) = \Lambda - \beta S(t)I(t) - \alpha S(t) \\ \dot{E}(t) = \beta S(t)I(t) - (\alpha + \sigma)E \\ \dot{I}(t) = \sigma E(t) - (\alpha + \gamma)I(t) \\ \dot{R}(t) = \gamma I(t) - \alpha R(t), \end{cases} \tag{7.2}$$

where $S(t)$, $E(t)$, $I(t)$, and $R(t)$ are, respectively, the number of susceptible, exposed, infective, and recovered individuals at time t; Λ, α, β, γ, and σ is the birth rate, natural death rate, transmission rate, recovery rate, and incubation rate, respectively. See (1.1) with $p = q = 1$ in [10]. In 2010, the model above was updated with a general nonlinear incidence to

$$\begin{cases} \dot{S}(t) = \Lambda - \beta F(S(t))G(I(t)) - \alpha S(t) \\ \dot{E}(t) = \beta F(S(t))G(I(t)) - (\alpha + \sigma)E(t) \\ \dot{I}(t) = \sigma E(t) - (\alpha + \gamma)I(t) \\ \dot{R}(t) = \gamma I(t) - \alpha R(t), \end{cases} \tag{7.3}$$

where $F, G : \mathbb{R} \to \mathbb{R}_+$, $F(0) = 0$ and $G(0) = 0$, see the epidemic model in the case of delay $\omega = 0$ above (15) in [11]. Here the product $F(S)G(I)$ is the so-called incidence rate of general form.

In 2013, a Caputo fractional model was first derived in [12] by Diethelm, based on some model similar to (7.2) and a rule that the time dimension of both sides of models must be consistent, see (9) in [12]. In 2020, a Caputo fractional model with consistent time dimension in both sides and a general nonlinear incidence rate was given as

$$\begin{cases} {}_{t_0}^{C}D_t^q S(t) = \Lambda^q - \beta^q F(S(t))G(I(t)) - \alpha^q S(t) \\ {}_{t_0}^{C}D_t^q E(t) = \beta^q F(S(t))G(I(t)) - (\alpha^q + \sigma^q)E(t) \\ {}_{t_0}^{C}D_t^q I(t) = \sigma^q E(t) - (\alpha^q + \gamma^q)I(t) \\ {}_{t_0}^{C}D_t^q R(t) = \gamma^q I(t) - \alpha^q R(t), \end{cases} \tag{7.4}$$

in [13], where $0 < q \leq 1$. This Caputo fractional model will be the object of discussion in the chapter.

7.3 Qualitative Analysis

Qualitative analysis, including uniqueness, invariant set, and stability, helps to figure out dynamics of models prior to numerically solving them. Specifically, the uniqueness of solutions tells us that the integral curves for a model cannot cross; the invariant set sketches the scope where solutions cannot go out once entering; and the stability implies the eventual tendency of solutions once their initial conditions are given sufficiently close to the origin.

7.3.1 Uniqueness

Let $x = [S\ E\ I\ R]^T$ and f denote the right hand side function (vector field function) of (7.4), then

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \frac{\partial f_1}{\partial x_4} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \frac{\partial f_2}{\partial x_4} \\ \frac{\partial f_3}{\partial x_1} & \frac{\partial f_3}{\partial x_2} & \frac{\partial f_3}{\partial x_3} & \frac{\partial f_3}{\partial x_4} \\ \frac{\partial f_4}{\partial x_1} & \frac{\partial f_4}{\partial x_2} & \frac{\partial f_4}{\partial x_3} & \frac{\partial f_4}{\partial x_4} \end{bmatrix} = \begin{bmatrix} -\beta^q G(I)\frac{dF}{dS} - \alpha^q & 0 & -\beta^q F(S)\frac{dG}{dI} & 0 \\ \beta^q G(I)\frac{dF}{dS} & -(\alpha^q + \sigma^q) & \beta^q F(S)\frac{dG}{dI} & 0 \\ 0 & \sigma^q & -(\alpha^q + \gamma^q) & 0 \\ 0 & 0 & \gamma^q & -\alpha^q \end{bmatrix}.$$

By the Mean Value Theorem for vector-valued functions, if $F, G \in C^1(\mathbb{R})$, and $F(S), G(I), F'(S)$ and $G'(I)$ are all bounded, then f is (continuous in t and) Lipschitz in x on $[0, \infty) \times \mathbb{R}^4$. By Lemma 3, we have the following uniqueness result for (7.4).

Theorem 1 *If $F, G \in C^1(\mathbb{R})$, and $F(S), G(I), F'(S)$ and $G'(I)$ are all bounded, then the initial value problem of (7.4) has a unique solution on $[t_0, \infty)$.*

7.3.2 Invariant Set

Theorem 2 *The set $\Omega = \{(S, E, I, R) \in \mathbb{R}^4 : 0 \leq S + E + I + R \leq \Lambda^q/\alpha^q\}$ is an invariant set of (7.4).*

Proof Let $N = S + E + I + R$, then by adding the four Caputo fractional order ordinary differential equations of (7.4), we derive

$${}_{t_0}^{C}D_t^q N(t) = \Lambda^q - \alpha^q N(t).$$

Let $N(t_0) = N_0$, where $0 \leq N_0 \leq \Lambda^q/\alpha^q$, then by Lemma 2.1 in [14],

$$N(t) = N_0 \mathscr{E}_q[-\alpha^q (t - t_0)^q] + \int_{t_0}^{t} (t - \tau)^{q-1} \mathscr{E}_{q,q}[-\alpha^q (t - \tau)^q] \Lambda^q d\tau,$$

where $\mathscr{E}_q$ and $\mathscr{E}_{q,q}$ denote the Mittag-Leffler functions with one parameter q and two the same parameters q, respectively.

By [15], these two Mittag-Leffler functions are nonnegative. Thus, $N(t) \geq 0$ for all $t \geq t_0$.

As $N(t_0) = \Lambda^q/\alpha^q$, $N(t) \equiv \Lambda^q/\alpha^q$ for all $t \geq t_0$. That is, for $t \geq t_0$,

$$\frac{\Lambda^q}{\alpha^q} \mathscr{E}_q[-\alpha^q (t - t_0)^q] + \int_{t_0}^{t} (t - \tau)^{q-1} \mathscr{E}_{q,q}[-\alpha^q (t - \tau)^q] \Lambda^q d\tau \equiv \frac{\Lambda^q}{\alpha^q}.$$

It follows, as $0 \leq N_0 < \Lambda^q/\alpha^q$,

$$N_0 \mathscr{E}_q[-\alpha^q(t-t_0)^q] + \int_{t_0}^{t} (t-\tau)^{q-1} \mathscr{E}_{q,q}[-\alpha^q(t-\tau)^q]\Lambda^q d\tau < \frac{\Lambda^q}{\alpha^q},$$

i.e. $N(t) < \Lambda^q/\alpha^q$. Therefore, for any $(S_0, E_0, I_0, R_0) \in \Omega$, $(S(t), E(t), I(t), R(t)) \in \Omega$, for all $t \geq t_0$. □

7.3.3 Stability

Let us find the possible equilibrium points for (7.4). Let the right hand side of (7.4) be zero, i.e.

$$\begin{cases} \Lambda^q - \beta^q F(S)G(I) - \alpha^q S = 0 \\ \beta^q F(S)G(I) - (\alpha^q + \sigma^q)E = 0 \\ \sigma^q E - (\alpha^q + \gamma^q)I = 0 \\ \gamma^q I - \alpha^q R = 0. \end{cases} \tag{7.5}$$

Obviously, $(\Lambda^q/\alpha^q, 0, 0, 0)$ is a constant solution to (7.5) so is an equilibrium of (7.4). As introduced in [13], there may be another equilibrium point. Let us go through the detailed derivation. Adding the first and second equation in (7.5) yields

$$\Lambda^q - \alpha^q S - (\alpha^q + \sigma^q)E = 0.$$

Substituting the third equation into the second equation in (7.5), for the replacement of I by E, gives

$$\beta^q F(S) G\left(\frac{\sigma^q E}{\alpha^q + \gamma^q}\right) - (\alpha^q + \sigma^q)E = 0.$$

Combining these two equations, we derive

$$\beta^q F\left(\frac{\Lambda^q - (\alpha^q + \sigma^q)E}{\alpha^q}\right) G\left(\frac{\sigma^q E}{\alpha^q + \gamma^q}\right) - (\alpha^q + \sigma^q)E = 0.$$

Let $\bar{F}$ denote the right hand side of this equation, then

$$\bar{F} = \bar{F}(E) = \beta^q F\left(\frac{\Lambda^q - (\alpha^q + \sigma^q)E}{\alpha^q}\right) G\left(\frac{\sigma^q E}{\alpha^q + \gamma^q}\right) - (\alpha^q + \sigma^q)E.$$

Since $G(0) = 0$, $\bar{F}(0) = 0$. Similarly, since $F(0) = 0$,

$$\bar{F}\left(\frac{\Lambda^q}{\alpha^q + \sigma^q}\right) = -\Lambda^q < 0.$$

Moreover, let

$$k = \beta^q F\left(\frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI}|_{I=0}\frac{\sigma^q}{(\alpha^q + \gamma^q)(\alpha^q + \sigma^q)},$$

then

$$\begin{aligned}\frac{d\bar{F}}{dE}|_{E=0} &= \beta^q \frac{dF}{dS}|_{S=\frac{\Lambda^q}{\alpha^q}} \frac{-(\alpha^q + \sigma^q)}{\alpha^q} G\left(\frac{\sigma^q E}{\alpha^q + \gamma^q}\right)|_{E=0} \\ &\quad + \beta^q F\left(\frac{\Lambda^q - (\alpha^q + \sigma^q)E}{\alpha^q}\right)|_{E=0}\frac{dG}{dI}|_{I=0}\frac{\sigma^q}{\alpha^q + \gamma^q} - (\alpha^q + \sigma^q) \\ &= \beta^q F\left(\frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI}|_{I=0}\frac{\sigma^q}{\alpha^q + \gamma^q} - (\alpha^q + \sigma^q) \\ &= (\alpha^q + \sigma^q)(k - 1).\end{aligned}$$

If $k > 1$, then it follows the continuity of $\bar{F}$, there exists $E_* \in (0, \Lambda^q/(\alpha^q + \sigma^q))$ such that $\bar{F}(E_*) = 0$. Correspondingly,

$$S_* = \frac{\Lambda^q - (\alpha^q + \sigma^q)E_*}{\alpha^q}, I_* = \frac{\sigma^q E_*}{\alpha^q + \gamma^q}, R_* = \frac{\gamma^q I_*}{\alpha^q}.$$

Moreover,

$$\begin{aligned}\frac{d\bar{F}}{dE}|_{E=E_*} &= \beta^q \frac{dF}{dS}|_{S=\frac{\Lambda^q - (\alpha^q + \sigma^q)E_*}{\alpha^q}} \frac{-(\alpha^q + \sigma^q)}{\alpha^q} G(\frac{\sigma^q E}{\alpha^q + \gamma^q})|_{E=E_*} \\ &\quad + \beta^q F\left(\frac{\Lambda^q - (\alpha^q + \sigma^q)E}{\alpha^q}\right)|_{E=E_*}\frac{dG}{dI}|_{I=\frac{\sigma^q E_*}{\alpha^q + \gamma^q}}\frac{\sigma^q}{\alpha^q + \gamma^q} - (\alpha^q + \sigma^q) \\ &= \beta^q \frac{dF}{dS}|_{S=S_*}\frac{-(\alpha^q + \sigma^q)}{\alpha^q} G(I_*)\end{aligned}$$

$$+ \beta^q F(S_*)\frac{dG}{dI}|_{I=I_*}\frac{\sigma^q}{\alpha^q + \gamma^q} - (\alpha^q + \sigma^q).$$

If $(d\bar{F}/dE)_{E=E_*} < 0$, then the positive E_* is unique, because for any other possible positive E_{**}, we must have

$$\frac{d\bar{F}}{dE}|_{E=E_{**}} > 0.$$

In the integer order case, the stability and instability of an equilibrium point for a nonlinear autonomous system may be investigated through that of an equilibrium point for its linearization, see pp.139 in [16]. There this linearization method is called Lyapunov's indirect method. It seems very natural to extend this idea for Caputo fractional order autonomous systems. However, there is only one proven result on instability [9], and none on stability despite a few applications, e.g. [13, 17], based on this idea. In this section, only the instability of $(\Lambda^q/\alpha^q, 0, 0, 0)$ is discussed. The steps for investigating the instability of (S_*, E_*, I_*, R_*) are the same.

We may first translate the nonzero equilibrium point $(\Lambda^q/\alpha^q, 0, 0, 0)$ to the origin. Let $\bar{S} = S - \Lambda^q/\alpha^q$, then (7.4) can be rewritten as

$$\begin{cases} {}_{t_0}^{C}D_t^q \bar{S} = -\beta^q F\left(\bar{S} + \frac{\Lambda^q}{\alpha^q}\right)G(I) - \alpha^q \bar{S} \\ {}_{t_0}^{C}D_t^q E = \beta^q F\left(\bar{S} + \frac{\Lambda^q}{\alpha^q}\right)G(I) - (\alpha^q + \sigma^q)E \\ {}_{t_0}^{C}D_t^q I = \sigma^q E - (\alpha^q + \gamma^q)I \\ {}_{t_0}^{C}D_t^q R = \gamma^q I - \alpha^q R. \end{cases} \tag{7.6}$$

Clearly, the origin $(0, 0, 0, 0)$ becomes an equilibrium point of the translated model (7.6). The Jacobian matrix of (7.6) is

$$J(\bar{S}, E, I, R) = \begin{bmatrix} -\beta^q \frac{dF}{dS}|_{S=(\bar{S}+\frac{\Lambda^q}{\alpha^q})} G(I) - \alpha^q & 0 & -\beta^q F\left(\bar{S} + \frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI} & 0 \\ \beta^q \frac{dF}{dS}|_{S=(\bar{S}+\frac{\Lambda^q}{\alpha^q})} G(I) & -(\alpha^q + \sigma^q) & \beta^q F\left(\bar{S} + \frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI} & 0 \\ 0 & \sigma^q & -(\alpha^q + \gamma^q) & 0 \\ 0 & 0 & \gamma^q & -\alpha^q \end{bmatrix}.$$

Thus,

$$J(\mathbf{0}) = \begin{bmatrix} -\alpha^q & 0 & -\beta^q F\left(\frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI} & 0 \\ 0 & -(\alpha^q + \sigma^q) & \beta^q F\left(\frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI} & 0 \\ 0 & \sigma^q & -(\alpha^q + \gamma^q) & 0 \\ 0 & 0 & \gamma^q & -\alpha^q \end{bmatrix}.$$

We may derive the eigenvalues of $J(\mathbf{0})$: $\lambda_1 = -\alpha^q$,

$$\lambda_{2,3} = \frac{-(2\alpha^q + \sigma^q + \gamma^q) \pm \sqrt{(\sigma^q - \gamma^q)^2 + 4\beta^q \sigma^q F\left(\frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI}|_{I=0}}}{2},$$

and $\lambda_4 = -\alpha^q$.

As $k > 1$, it can be concluded that

$$\frac{dG}{dI}|_{I=0} > 0.$$

In this case, $\lambda_{1,3,4} < 0$, we only need to focus on λ_2, which is calculated as follows

$$\begin{aligned}\lambda_2 &= \frac{-(2\alpha^q + \sigma^q + \gamma^q) + \sqrt{(\sigma^q - \gamma^q)^2 + 4\beta^q \sigma^q F\left(\frac{\Lambda^q}{\alpha^q}\right)\frac{dG}{dI}|_{I=0}}}{2} \\ &> \frac{-(2\alpha^q + \sigma^q + \gamma^q) + \sqrt{(\sigma^q - \gamma^q)^2 + 4(\alpha^q + \gamma^q)(\alpha^q + \sigma^q)}}{2} \\ &= \frac{-(2\alpha^q + \sigma^q + \gamma^q)}{2} \\ &\quad + \frac{\sqrt{[(\alpha^q + \gamma^q) - (\alpha^q + \sigma^q)]^2 + 4(\alpha^q + \gamma^q)(\alpha^q + \sigma^q)}}{2} \\ &= 0.\end{aligned}$$

Clearly, $|arg(\lambda_2)| < q\pi/2$. Applying Lemma 4, we have the following theorem.

Theorem 3 *If $F, G \in C^1(\mathbb{R})$, $F(0) = 0, G(0) = 0$, and $F(S), G(I), F'(S)$ and $G'(I)$ are all bounded, and moreover, $k > 1$, then the equilibrium point $(\Lambda^q/\alpha^q, 0, 0, 0)$ of (7.4) is unstable.*

Proof As analyzed in Section 3.1, $F, G \in C^1(\mathbb{R})$, and $F(S), G(I), F'(S)$ and $G'(I)$ are all bounded, then the right hand side function of (7.4), i.e. that of (7.6), is Lipschitz on $\mathbb{R}^4$. Moreover, $k > 1$ implies $|arg(\lambda_2)| < q\pi/2$. By Lemma 4, the zero equilibrium point of (7.6) is unstable. The translation relation implies that the equilibrium point $(\Lambda^q/\alpha^q, 0, 0, 0)$ of (7.4) is unstable. □

7.4 Numerical Modeling

In fact, the qualitative analysis is meaningless unless the model is sufficiently precise. Once an effective virus model is established, it may be used to predict the dynamics of virus infection (e.g. the number of infectious, dead, and recovered individuals) in the future through analyzing the properties of its solution (qualitative analysis) or directly computing its numerical solution. Compared to qualitative analysis, a rapid presentation of numerical solution seems more important in practice, because established models perhaps do not have expected properties. In this section, we shall introduce a numerical method (**genetic algorithm**) to build a model from available data (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series), and then use its numerical solution for prediction.

7.4.1 Genetic Algorithm

The genetic algorithm, motivated from the natural selection, is an adaptive algorithm for searching global optimal solution [18]. Based on the principles of heredity, variation, and survival of the fittest in Darwinian evolution theory, the genetic algorithm has inherent advantage in solving NP-hard problems, see [8] for more details about the NP-hard problem. In the genetic algorithm for a specific problem, at first, one population with a certain number of individuals is generated. Each individual represents a feasible solution in the search space of optimization and each individual will be given a particular fitness. In biology, the fitness of an individual is regarded as the individual's ability to survive in environment, while in minimization problems, it is the value of the optimization objective function of the feasible solution corresponding to this individual. All the individuals will be screened. The individuals with less fitness will be eliminated through selection or competition, while the individuals with higher fitness will mate with each other and produce new ones that will be added to the population. For the sake of realizing the mate process, binary strings are used to code each individual and so may be interpreted as chromosomes. As well known, in biological mate process, chromatids perform crossover, mutation, and inversion behaviors to ensure population diversity. The genetic algorithm performs the same operations on the binary strings representing chromosomes and then generates new individuals that are similar to but different from their parents, so as to realize the global search in the solution space. In the later stage of the algorithm, the mutation probability will be increased to avoid local optimal solutions. The algorithm continues to calculate the fitness of the newly formed population and continues to repeat the natural selection and reproduction process, till the optimal solution is found or the threshold number of iteration is reached. In the eventual population obtained, the feasible solution corresponding to the individual with the highest fitness will be regarded as the global optimal solution.

As introduced above, a genetic algorithm may be implemented by the following main steps.

1. Initialize parameters: iteration number "eranum," population size "N," cross probability "pcross," mutation probability "pmutation," inversion probability

"pinversion," precision "precision," upper and lower limits of solution space "UP" and "LP."
2. Determine the string length of binarily encoded feasible solutions.
3. Generate an initial population: arbitrarily generate "N" number of feasible solutions in the search space, and binarily encode them.
4. Implement the principle of survival of the fittest: decode the binary codes, then compute the fitness of all the individuals in the population, and then select the individuals with higher fitness as the parent generation by using the nonlinear method of roulette, and let them mate and produce new individuals.
5. Perform crossover: exchange segments of any two binary strings of the parent generation (chromosome) by the probability of "pcross," and produce new individuals.
6. Perform mutation: create mutations on single points from 0 to 1 or from 1 to 0 on the binary strings of the new individuals by probability "pmutation."
7. Perform inversion: interchange segments of the binary strings of the new individuals by probability "pinversion."
8. Add new individuals to the population.
9. Iterate: repeat steps 3–8 continuously till the optimal solution is found or the threshold number of iteration is reached.

Further, these main steps above may be described by the following pseudocode.

```
%Input: eranum, N, pcross, pmutation, pinversion, ...
    precision, UP and LP.
%output: global optimal solution.
parameter initialization: eranum, N, pcross, pmutation, ...
    pinversion, precision, UP, LP;
binary encoding length: bitslength;
initial population: Pop = InitPopGray(N, bitslength);
while i < eranum
     fitness: fit = fitness(Pop);
     selection: select = NonlinearRankSelect(fit,Pop);
     crossover: newpop = CrossOver(select, pcross);
     mutation: newpop = Mutation(newpop, pmutation);
     inversion: newpop = Inversion(newpop, pinversion);
     addition: Pop = select U newpop;
end
```

7.4.2 Setup

Here we set up the MATLAB code for the genetic algorithm to build the Caputo fractional order model of the virus infection at population level, i.e. to adjust the parameters of (7.4) such that its solution coincides the real data as possible. We select

$$F(S) = e^{\ln S}, G(I) = \frac{e^{\ln I}}{S(t_0) + E(t_0) + I(t_0) + R(t_0) + \Lambda}.$$

Obviously, $F, G : (0, \infty) \to (0, \infty)$. The initial condition is selected as

$$R(t_0) = 8474, I(t_0) = 224560-8474, E(t_0) = 0.25*I(t_0), S(t_0) = 1500*I(t_0)+\Lambda,$$

where $R(t_0)$ is the number of recovered cases and $I(t_0)$ is the number of confirmed cases with a subtraction of $R(t_0)$ on April 1st 2020, and $\Lambda = 433994$. Moreover, the factors (0.25 and 1500) in the expressions of $E(t_0)$ and $S(t_0)$ are selected such that the total population $S(t_0) + E(t_0) + I(t_0) + R(t_0)$ approximately equals the real population of the United States of America.

The goal of the genetic algorithm is to obtain

$$\arg\min_{\text{para}} a_1 \sum_{i=2}^{8} |\text{Confirmed}_i - \text{I}_i(\text{para}) - \text{R}_i(\text{para})| + a_2 \sum_{i=2}^{8} |\text{Recovered}_i - \text{R}_i(\text{para})|,$$

where $a_1 = 0.2$, $a_2 = 0.8$, and i means the ith day in April 2020. Further, Confirmed_i, Recovered_i denote the real number of confirmed, recovered cases, respectively, and I_i, R_i denote the derived number of confirmed, recovered cases from (7.4) with the selected F, G and initial condition, respectively, on the ith day in April 2020. Here, “para” denotes all the parameters of (7.4) including q, Λ, β, α, σ, and γ, whose first value is selected as 0.6, 433994, 0.2209, 5.2×10^{-7}, 0.2467, and 0.1031, respectively.

7.4.3 Implementation

The main file is presented as follows and all the function files are attached in Appendix. There are totally twelve functions used in the main file that are: fun, F, G, $InitPopGray$, $b2f$, $NonlinearRankSelect$, $CrossOver$, $EqualCrossOver$, $MultiPointCross$, $Mutation$, $Inversion$, $foms prediction$. The method used in function fun and foms_prediction to compute the numerical solution of the Caputo fractional model (7.4) is referred from [19]. All the other functions are written based on [8].

```
%  solves problems of the form:
%      min F(X)  subject to:  LB < =  X < =  UB
%  FUN           - objective function

%parameters
q=0.6;
Lambda=433994;
beta=0.2209;
```

```
alpha = 5.2 * 1e-7;
sigma=0.2467;
gama=0.1031;
para = [q, Lambda, beta, alpha, sigma, gama];
UB = [1-1e-3, 600000, 0.9,  0.1, 0.9,  0.9];
LB = [0.1, 300000, 0.001,  1e-10, 0.001, 0.001];
real_pop = [224560 256792 289087 321477 351354 382747 ...
    413516 444731; 8474 9001 9707 14652 17448 19581 ...
    21763 23559];
syms x
FUN = @(x) fun(x, real_pop);
eranum = 200;popsize = 100;pCross = 0.8;pMutation = ...
    0.1;pInversion = 0.15;options = [0 1e-4]; precision ...
    = [1e-3, 100,1e-3, 1e-13, 1e-3, 1e-3];

%GA
bounds = [LB;UB]';bits = [];VarNum = size(bounds,1);Pop= [];
bits = ceil(log2((bounds(:,2)-bounds(:,1))' ./ precision));
Pop = InitPopGray(popsize,bits);
bits0 = cumsum([0 bits]);
[m,n] = size(Pop);
NewPop = zeros(m,n);
children1 = zeros(1,n);
children2 = zeros(1,n);
pm0 = pMutation;
BestPop = zeros(eranum,n);
Trace = zeros(eranum,length(bits)+1);
i = 1;
while i ≤ eranum
    [m,n] = size(Pop);
    value = zeros(1,m);
    for j=1:m
        value(j)=feval(FUN,(b2f(Pop(j,:),bounds,bits)));
    end
    [MaxValue, Index] = min(value);
    BestPop(i,:)=Pop(Index,:);
    Trace(i,1)=MaxValue;
    Trace(i,(2:length(bits)+1))=b2f(BestPop(i,:),bounds, ...
        bits);
    [selectpop] = NonlinearRankSelect(value,Pop);
    [CrossOverPop] = CrossOver(selectpop,pCross,...
        round(unidrnd(eranum-i)/eranum), VarNum);
    [MutationPop] = Mutation(CrossOverPop,pMutation,VarNum);
    [InversionPop] = Inversion(MutationPop,pInversion);
    Pop=InversionPop;
    pMutation=pm0+(i^4)*(pCross/3-pm0)/(eranum^4);
    p(i)=pMutation;
    i=i+1;
end
t = 1:eranum;
plot(t,Trace(:,1)');
title('Genetic algorithm for ...
    optimization');xlabel('eranum');ylabel('maxfitness');
[MaxFval,I] = min(Trace(:,1));
```

```
X = Trace(I,(2:length(bits)+1));
hold on;  plot(I,MaxFval,'*');
text(I+5,MaxFval,['FMAX = ' num2str(MaxFval)]);
str1 = sprintf('GA seeks the optimal solution X = [%s], ...
    FUN(X)=%f\n in the %d th ...
    generation',num2str(X),MaxFval,I);
disp(str1);

%plot
real_pop_test = [224560 256792 289087 321477 351354 ...
    382747 413516 444731 480667 515081 544183 571440 ...
    598380 627205 652611 682626 715656 743588 769684 ...
    799512 825429 854288 887858 920185 950581 977082 ...
    1000785 1025362 1051800 1081020; 8474 9001 9707 ...
    14652 17448 19581 21763 23559 25410 28790 31270 ...
    32988 43482 47763 52096 54703 58545 64840 70337 ...
    72329 75204 77366 80203 99079 100372 106988 111424 ...
    115936 120720 153947
];
real_pop_test =  real_pop_test(:,1:12);
foms_prediction(X, 8, real_pop_test);
```

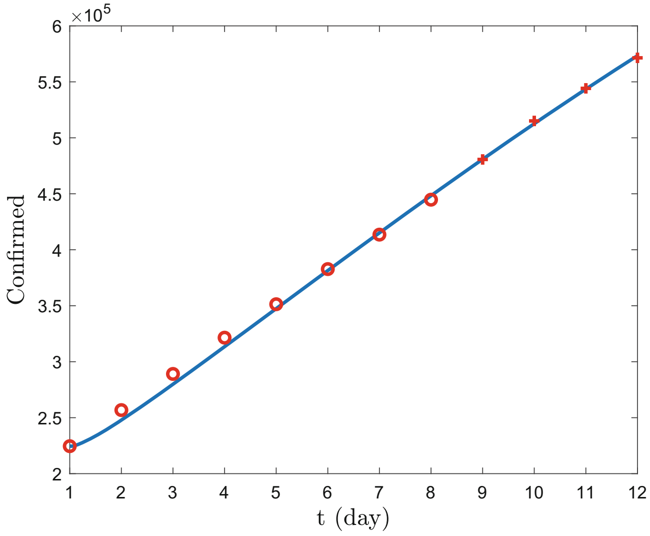

Fig. 7.1 Confirmed cases

The result of implementing the main file is shown in Figs. 7.1 and 7.2. In these two figures, the circles denote the real data reported from April 1 to 8 that is used to adjust the parameters of (7.4) such that the goal function given in Sect. 7.4.2 is minimized; the crosses denote the real data reported from April 9 to April 12 that is only used to compare with the predicted data (sketched by full lines) derived from

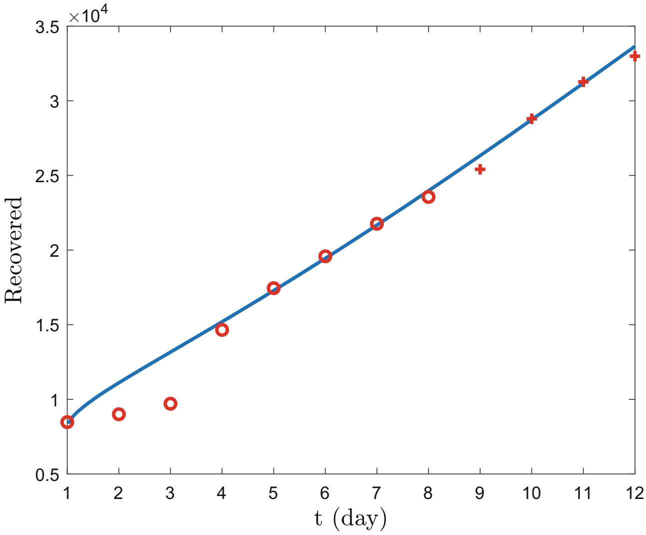

Fig. 7.2 Recovered cases

the built model (i.e. model (7.4) with adjusted parameters). As shown in these two figures, the genetic algorithm works well in adjusting the model to fit the real data, and to predict.

Acknowledgments This work was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019) and the National Natural Science Foundation of China (32070671,12101430).

Appendix

```
% function fun

function [erro] = fun( para, real_pop )
h=1/24;
T = size(real_pop, 2) - 1;
N=24*T;

%parameters
q=para(1);
Lambda=para(2);
beta=para(3);
alpha=para(4);
```

```
sigma=para(5);
gama=para(6);

%initial conditions
flag = true;
I0= 224560 - 8474;
S0= 1500*I0 + Lambda;
E0= 0.25*I0;
R0= 8474;
t0=0;
S = [];
S1 = [];
E = [];
E1 = [];
I = [];
I1 = [];
R = [];
R1 = [];

%end values
M1=0;
S(N+1)=[0];
S1(N+1)=[0];
M2=0;
E(N+1)=[0];
E1(N+1)=[0];
M3=0;
I(N+1)=[0];
I1(N+1)=[0];
M4=0;
R(N+1)=[0];
R1(N+1)=[0];

%iteration
S1(1)=S0+h^q*(Lambda^q-F(S0)*G(Lambda,I0)*beta^q-S0* ...
    alpha^q)/(gamma(q)*q);
E1(1)=E0+h^q*(F(S0)*G(Lambda,I0)*beta^q-E0*sigma^q-E0* ...
    alpha^q)/(gamma(q)*q);
I1(1)=I0+h^q*(E0*sigma^q-(gama^q+alpha^q)*I0)/(gamma(q)*q);
R1(1)=R0+h^q*(I0*gama^q-alpha^q*R0)/(gamma(q)*q);
S(1)=S0+h^q*(Lambda^q-F(S1(1))*G(Lambda,I1(1))*beta^q- ...
    S1(1)*alpha^q+q*(Lambda^q-F(S0)*G(Lambda,I0)*beta^q ...
    -S0*alpha^q))/gamma(q+2);
E(1)=E0+h^q*(F(S1(1))*G(Lambda,I1(1))*beta^q-E1(1)*sigma^q ...
    -E1(1)*alpha^q+q*(F(S0)*G(Lambda,I0)*beta^q-E0*sigma^q ...
    -E0*alpha^q))/gamma(q+2);
I(1)=I0+h^q*(E1(1)*sigma^q-(gama^q+alpha^q)*I1(1) ...
    +q*(E0*sigma^q-(gama^q+alpha^q)*I0))/gamma(q+2);
R(1)=R0+h^q*(I1(1)*gama^q-alpha^q*R1(1)+q*(I0*gama^q ...
    -alpha^q*R0))/gamma(q+2);

for n=1:N-1
    M1=(n^(q+1)-(n-q)*(n+1)^q)*(Lambda^q-F(S0) ...
        *G(Lambda,I0)*beta^q-S0*alpha^q);
    M2=(n^(q+1)-(n-q)*(n+1)^q)*(F(S0)*G(Lambda,I0)*beta^q ...
        -E0*sigma^q-E0*alpha^q);
    M3=(n^(q+1)-(n-q)*(n+1)^q)*(E0*sigma^q ...
        -(gama^q+alpha^q)*I0);
    M4=(n^(q+1)-(n-q)*(n+1)^q)*(I0*gama^q-alpha^q*R0);
    N1=((n+1)^q-n^q)*(Lambda^q-F(S0)*G(Lambda,I0)*beta^q ...
        -S0*alpha^q);
    N2=((n+1)^q-n^q)*(F(S0)*G(Lambda,I0)*beta^q-E0*sigma^q ...
        -E0*alpha^q);
    N3=((n+1)^q-n^q)*(E0*sigma^q-(gama^q+alpha^q)*I0);
    N4=((n+1)^q-n^q)*(I0*gama^q-alpha^q*R0);
    for j=1:n
        M1=M1+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (Lambda^q-F(S(j))*G(Lambda,I(j))*beta^q-S(j)* ...
            alpha^q);
        M2=M2+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (F(S(j))*G(Lambda,I(j))*beta^q-E(j)*sigma^q ...
            -E(j)*alpha^q);
        M3=M3+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (E(j)*sigma^q-(gama^q+alpha^q)*I(j));
        M4=M4+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (I(j)*gama^q-alpha^q*R(j));
        N1=N1+((n-j+1)^q-(n-j)^q)*(Lambda^q-F(S(j))* ...
            G(Lambda,I(j))*beta^q-S(j)*alpha^q);
        N2=N2+((n-j+1)^q-(n-j)^q)*(F(S(j))*G(Lambda,I(j))* ...
            beta^q-E(j)*sigma^q-E(j)*alpha^q);
        N3=N3+((n-j+1)^q-(n-j)^q)*(E(j)*sigma^q ...
            -(gama^q+alpha^q)*I(j));
        N4=N4+((n-j+1)^q-(n-j)^q)*(I(j)*gama^q-alpha^q*R(j));
    end

    S1(n+1)=S0+h^q*N1/(gamma(q)*q);
    E1(n+1)=E0+h^q*N2/(gamma(q)*q);
    I1(n+1)=I0+h^q*N3/(gamma(q)*q);
    R1(n+1)=R0+h^q*N4/(gamma(q)*q);
    S(n+1)=S0+h^q*((Lambda^q-F(S1(n+1))*G(Lambda,I1(n+1))* ...
        beta^q-S1(n+1)*alpha^q)+M1)/gamma(q+2);
    E(n+1)=E0+h^q*((F(S1(n+1))*G(Lambda,I1(n+1))*beta^q ...
        -E1(n+1)*sigma^q-E1(n+1)*alpha^q)+M2)/gamma(q+2);
    I(n+1)=I0+h^q*((E1(n+1)*sigma^q-(gama^q+alpha^q)* ...
        I1(n+1))+M3)/gamma(q+2);
    R(n+1)=R0+h^q*((I1(n+1)*gama^q-alpha^q*R1(n+1))+M4)/ ...
        gamma(q+2);
    if(¬(isreal(S(n+1)) && isreal(E(n+1)) && ...
        isreal(I(n+1)) && isreal(R(n+1))))
        flag = false;
        break
    end
end

%optimization
if flag
    S = [S0 S]; E = [E0 E]; I = [I0 I]; R = [R0 R];
    s = []; e = []; i = []; r = []; t = [];
    s(1)=S0;
    e(1)=E0;
    i(1)=I0;
    r(1)=R0;
    t(1)=t0;
    for n=2:N+1
        s(n)=S(n-1);
        e(n)=E(n-1);
        i(n)=I(n-1);
        r(n)=R(n-1);
        t(n)=t0+(n-1)*h;
    end

    erro =  sum([0.2, 0.8] * abs(real_pop(:,2:end) - [ ...
        I(24:24:(length(I)-1)) + R(24:24:(length(R)-1)); ...
        R(24:24:(length(R)-1))] ));
    if(isnan(erro) )
        erro = inf;
    end
else
    erro = inf;
end
```

```
% function F

function F=F(S)
F=exp(log(S));
end
```

```
% function G

function G = G(Lambda, I)
I0= 224560 - 8474;
S0= 1500*I0 + Lambda;
E0= 0.25*I0;
R0= 8474;
G=exp(log(I)) / (S0 + E0 + I0 + R0 + Lambda);
end
```

```
% function InitPopGray

function [initpop]=InitPopGray(popsize,bits)
len=sum(bits);
initpop=zeros(popsize,len);
for i=2:popsize-1
    pop=round(rand(1,len));
    pop=mod(([0 pop]+[pop 0]),2);
    initpop(i,:)=pop(1:end-1);
```

```
end
initpop(popsize,:)=ones(1,len);
```

```
% function b2f

function [fval] = b2f(bval,bounds,bits)
scale=(bounds(:,2)-bounds(:,1))'./(2.^bits-1); ...
    numV=size(bounds,1);
cs=[0 cumsum(bits)];
for i=1:numV
  a=bval((cs(i)+1):cs(i+1));
  fval(i)=sum(2.^(size(a,2)-1:-1:0).*a)*scale(i)+bounds(i,1);
end
```

```
% function NonlinearRankSelect

function [selectpop]=NonlinearRankSelect(value,Pop, m, n)
[m,n] = size(Pop);
selectpop = zeros(m,n);
fit  =  1./value;
selectprob = fit/sum(fit);
q = max(selectprob);
x = zeros(m,2);
x(:,1) = [m:-1:1]';
[y x(:,2)] = sort(selectprob);
r = q/(1-(1-q)^m);
newfit(x(:,2))=r*(1-q).^(x(:,1)-1);
newfit = cumsum(newfit);
rNums = sort(rand(m,1));
fitIn = 1;newIn = 1;
while newIn ≤ m
    if rNums(newIn)<newfit(fitIn)
        selectpop(newIn,:) = Pop(fitIn,:);
        newIn = newIn+1;
    else
        fitIn = fitIn+1;
    end
end
```

```
% function CrossOver

function [NewPop]=CrossOver(OldPop,pCross,opts, VarNum)
[m,n] = size(OldPop);
NewPop = [];
r=rand(1,m);
y1=find(r<pCross);
y2=find(r≥pCross);
len=length(y1);
```

```
if len>2&&mod(len,2)==1
    y2(length(y2)+1)=y1(len);
    y1(len)=[];
end
if length(y1)≥2
   for i=0:2:length(y1)-2
       if opts==0
           [NewPop(y1(i+1),:),NewPop(y1(i+2),:)] ...
               =EqualCrossOver(OldPop(y1(i+1),:), ...
               OldPop(y1(i+2),:));
       else
           [NewPop(y1(i+1),:),NewPop(y1(i+2),:)] ...
               =MultiPointCross(OldPop(y1(i+1),:), ...
               OldPop(y1(i+2),:), VarNum);
       end
   end
end
NewPop(y2,:)=OldPop(y2,:);
```

```
% function EqualCrossOver

function ...
    [children1,children2]=EqualCrossOver(parent1,parent2)
n = length(parent1);
children1 = parent1;
children2 = parent2;
hidecode=round(rand(1,n));
crossposition=find(hidecode==1);
holdposition=find(hidecode==0);
children1(crossposition)=parent1(crossposition);
children1(holdposition)=parent2(holdposition);
children2(crossposition)=parent2(crossposition);
children2(holdposition)=parent1(holdposition);
```

```
% function MultiPointCross

function ...
    [Children1,Children2]=MultiPointCross(Parent1,Parent2, ...
    VarNum)
n = length(Parent1);
Children1=Parent1;
Children2=Parent2;
Points=sort(unidrnd(n,1,2*VarNum));
for i=1:VarNum
    Children1(Points(2*i-1):Points(2*i)) ...
        =Parent2(Points(2*i-1):Points(2*i));
    Children2(Points(2*i-1):Points(2*i)) ...
        =Parent1(Points(2*i-1):Points(2*i));
end
```

```
% function Mutation

function [NewPop]=Mutation(OldPop,pMutation,VarNum)
[m,n] = size(OldPop);
NewPop = [];
r=rand(1,m);
position=find(r≤pMutation);
len=length(position);
if len≥1
   for i=1:len
       k=unidrnd(n,1,VarNum);
       for j=1:length(k)
           if OldPop(position(i),k(j))==1
              OldPop(position(i),k(j))=0;
           else
              OldPop(position(i),k(j))=1;
           end
       end
   end
end
NewPop=OldPop;
```

```
% function Inversion

function [NewPop]=Inversion(OldPop,pInversion)
[m,n] = size(OldPop);
NewPop=OldPop;
r=rand(1,m);
PopIn=find(r≤pInversion);
len=length(PopIn);
if len≥1
    for i=1:len
        d=sort(unidrnd(n,1,2));
        if d(1)≠1&d(2)≠n
           NewPop(PopIn(i),1:d(1)-1)=OldPop(PopIn(i), ...
               1:d(1)-1);
           NewPop(PopIn(i),d(1):d(2))=OldPop(PopIn(i), ...
               d(2):-1:d(1));
           NewPop(PopIn(i),d(2)+1:n)=OldPop(PopIn(i),d(2) ...
               +1:n);
       end
   end
end
```

```
% function foms_prediction

function [S, E, I, R, t] = foms_prediction( para, T, ...
    real_pop_pre)
h=1/24;
N=24*(size(real_pop_pre,2)-1);
```

```

%parameters
q=para(1);
Lambda=para(2);
beta=para(3);
alpha=para(4);
sigma=para(5);
gama=para(6);

%initial conditions
I0= 224560 - 8474;
S0= 1500*I0 + Lambda;
E0= 0.25*I0;
R0= 8474;
t0=0;
S = [];
S1 = [];
E = [];
E1 = [];
I = [];
I1 = [];
R = [];
R1 = [];

%end values
M1=0;
S(N+1)=[0];
S1(N+1)=[0];
M2=0;
E(N+1)=[0];
E1(N+1)=[0];
M3=0;
I(N+1)=[0];
I1(N+1)=[0];
M4=0;
R(N+1)=[0];
R1(N+1)=[0];

%iteration
S1(1)=S0+h^q*(Lambda^q-F(S0)*G(Lambda,I0)*beta^q-S0* ...
    alpha^q)/(gamma(q)*q);
E1(1)=E0+h^q*(F(S0)*G(Lambda,I0)*beta^q-E0*sigma^q-E0* ...
    alpha^q)/(gamma(q)*q);
I1(1)=I0+h^q*(E0*sigma^q-(gama^q+alpha^q)*I0)/(gamma(q)*q);
R1(1)=R0+h^q*(I0*gama^q-alpha^q*R0)/(gamma(q)*q);
S(1)=S0+h^q*(Lambda^q-F(S1(1))*G(Lambda,I1(1))*beta^q ...
    -S1(1)*alpha^q+q*(Lambda^q-F(S0)*G(Lambda,I0)*beta^q ...
    -S0*alpha^q))/gamma(q+2);
E(1)=E0+h^q*(F(S1(1))*G(Lambda,I1(1))*beta^q-E1(1)*sigma^q ...
    -E1(1)*alpha^q+q*(F(S0)*G(Lambda,I0)*beta^q-E0*sigma^q ...
    -E0*alpha^q))/gamma(q+2);
I(1)=I0+h^q*(E1(1)*sigma^q-(gama^q+alpha^q)*I1(1) ...
    +q*(E0*sigma^q-(gama^q+alpha^q)*I0))/gamma(q+2);
```

```
R(1)=R0+h^q*(I1(1)*gama^q-alpha^q*R1(1)+q*(I0*gama^q ...
    -alpha^q*R0))/gamma(q+2);

for n=1:N-1
    M1=(n^(q+1)-(n-q)*(n+1)^q)*(Lambda^q-F(S0) ...
        *G(Lambda,I0)*beta^q-S0*alpha^q);
    M2=(n^(q+1)-(n-q)*(n+1)^q)*(F(S0)*G(Lambda,I0)*beta^q ...
        -E0*sigma^q-E0*alpha^q);
    M3=(n^(q+1)-(n-q)*(n+1)^q)*(E0*sigma^q ...
        -(gama^q+alpha^q)*I0);
    M4=(n^(q+1)-(n-q)*(n+1)^q)*(I0*gama^q-alpha^q*R0);
    N1=((n+1)^q-n^q)*(Lambda^q-F(S0)*G(Lambda,I0)*beta^q ...
        -S0*alpha^q);
    N2=((n+1)^q-n^q)*(F(S0)*G(Lambda,I0)*beta^q-E0*sigma^q ...
        -E0*alpha^q);
    N3=((n+1)^q-n^q)*(E0*sigma^q-(gama^q+alpha^q)*I0);
    N4=((n+1)^q-n^q)*(I0*gama^q-alpha^q*R0);
    for j=1:n
        M1=M1+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (Lambda^q-F(S(j))*G(Lambda,I(j))*beta^q-S(j)* ...
            alpha^q);
        M2=M2+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (F(S(j))*G(Lambda,I(j))*beta^q-E(j)*sigma^q ...
            -E(j)*alpha^q);
        M3=M3+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (E(j)*sigma^q-(gama^q+alpha^q)*I(j));
        M4=M4+((n-j+2)^(q+1)+(n-j)^(q+1)-2*(n-j+1)^(q+1))* ...
            (I(j)*gama^q-alpha^q*R(j));
        N1=N1+((n-j+1)^q-(n-j)^q)*(Lambda^q-F(S(j))* ...
            G(Lambda,I(j))*beta^q-S(j)*alpha^q);
        N2=N2+((n-j+1)^q-(n-j)^q)*(F(S(j))*G(Lambda,I(j))* ...
            beta^q-E(j)*sigma^q-E(j)*alpha^q);
        N3=N3+((n-j+1)^q-(n-j)^q)*(E(j)*sigma^q ...
            -(gama^q+alpha^q)*I(j));
        N4=N4+((n-j+1)^q-(n-j)^q)*(I(j)*gama^q-alpha^q*R(j));
    end

    S1(n+1)=S0+h^q*N1/(gamma(q)*q);
    E1(n+1)=E0+h^q*N2/(gamma(q)*q);
    I1(n+1)=I0+h^q*N3/(gamma(q)*q);
    R1(n+1)=R0+h^q*N4/(gamma(q)*q);
    S(n+1)=S0+h^q*((Lambda^q-F(S1(n+1))*G(Lambda,I1(n+1))* ...
        beta^q-S1(n+1)*alpha^q)+M1)/gamma(q+2);
    E(n+1)=E0+h^q*((F(S1(n+1))*G(Lambda,I1(n+1))*beta^q ...
        -E1(n+1)*sigma^q-E1(n+1)*alpha^q)+M2)/gamma(q+2);
    I(n+1)=I0+h^q*((E1(n+1)*sigma^q-(gama^q+alpha^q)* ...
        I1(n+1))+M3)/gamma(q+2);
    R(n+1)=R0+h^q*((I1(n+1)*gama^q-alpha^q*R1(n+1))+M4)/ ...
        gamma(q+2);
end

%plot
S = [S0 S]; E = [E0 E]; I = [I0 I]; R = [R0 R];
s = []; e = []; i = []; r = []; t = [];
```

```
s(1)=S0;
e(1)=E0;
i(1)=I0;
r(1)=R0;
t(1)=t0;
for n=2:N+1
    s(n)=S(n-1);
    e(n)=E(n-1);
    i(n)=I(n-1);
    r(n)=R(n-1);
    t(n)=t0+(n-1)*h;
end
Tr = 1:size(real_pop_pre,2);%t(1:24:(length(t)-1))
Cr =  real_pop_pre(1,:);
Rr = real_pop_pre(2,:);

figure
plot(t+1,i + r,'lineWidth',2)
xlabel('t (day)','Interpreter','Latex','FontSize',14);
ylabel('Confirmed','Interpreter','Latex','FontSize',14);
hold on
plot(Tr(1:T),Cr(1:T),'ro','lineWidth',2)
xlabel('t (day)','Interpreter','Latex','FontSize',14);
hold on
plot(Tr((T+1):end),Cr((T+1):end),'r+','lineWidth',2)
xlabel('t (day)','Interpreter','Latex','FontSize',14);
set(gca,'xlim',[1,12],'xtick',[1:1:12]);

figure
plot(t+1,r,'lineWidth',2)
xlabel('t (day)','Interpreter','Latex','FontSize',14);
ylabel('Recovered','Interpreter','Latex','FontSize',14);
hold on
plot(Tr(1:T),Rr(1:T),'ro','lineWidth',2)
xlabel('t (day)','Interpreter','Latex','FontSize',14);
hold on
plot(Tr((T+1):end),Rr((T+1):end),'r+','lineWidth',2)
xlabel('t (day)','Interpreter','Latex','FontSize',14);
set(gca,'xlim',[1,12],'xtick',[1:1:12]);
```

References

1. Lin QY, Zhao S, Gao DZ, Luo YJ, Yang S, Musa SS, Wang MH, Cai YL, Wang WM, Yang L, He DH (2020) A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. Int J Infect Dis 93:211–216
2. Diethelm K (2010) The analysis of fractional differential equations. Springer, Heidelberg
3. Podlubny I (1999) Fractional differential equations. Academic Press, San Diego
4. Diethelm K (2013) A fractional calculus based model for the simulation of an outbreak of dengue fever. Nonlinear Dyn 71:613–619

5. Gonzalez-Parra G, Arenas AJ, Chen-Charpentier BM (2014) A fractional order epidemic model for the simulation of outbreaks of influenza A(H1N1). Math Methods Appl Sci 37:2218–2226
6. Almeida R, Brito da Cruz AMC, Martins N, Teresa M, Monteiro T (2019) An epidemiological MSEIRmodel described by the Caputo fractional derivative. Int J Dyn Control 7:776–784
7. Xu C, Yu Y, Chen Y, Lu Z (2020) Forecast analysis of the epidemics trend of COVID-19 in the USA by a generalized fractional-order SEIR model. Nonlinear Dyn 101:1621–1634
8. Knuth DE (1974) Postscript about NP-hard problems. ACM SIGACT News 6:15–16
9. Cong ND, Son DT, Siegmund S, Tuan HT (2017) An instability theorem for nonlinear fractional differential systems. Discrete Contin Dyn Syst Ser B 22:3079–3090
10. Li MY, Muldowney JS (1995) Global stability for the SEIR model in epidemiology. Math Biosci 125:155–164
11. Huang G, Takeuchi Y, Ma W, Wei D (2010) Global stability for delay SIR and SEIR epidemic models with nonlinear incidence rate. Bull Math Biol 72:1192–1207
12. Diethelm K (2013) A fractional calculus based model for the simulation of an outbreak of dengue fever. Nonlinear Dyn 71:613–619
13. Yang Y, Xu L (2020) Stability of a fractional order SEIR model with general incidence. Appl Math Lett 105:106303
14. Wu C, Liu X (2019) Lyapunov and external stability of Caputo fractional order switching systems. Nonlinear Anal-Hybri 34:131–146
15. Hanneken JW, Narahari Achar BN, Puzio R, Vaught DM (2009) Properties of the Mittag-Leffler function for negative alpha. Phys Scr T 136:014037
16. Khalil HK (2002) Nonlinear systems. Prentice-Hall, New Jersey
17. Ahmed E, El-Sayed AMA, El-Saka HAA (2007) Equilibrium points, stability and numerical solutions of fractional-order predator-prey and rabies models. J Math Anal Appl 325:542–553
18. Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Michigan
19. Diethelm K, Freed AD (1999) The Fracpece subroutine for the numerical solution of differential equations of fractional order. Forschung und wissenschaftliches Rechnen 1998:57–71

Chapter 8
Health-Based Geographic Information Systems for Mapping and Risk Modeling of Infectious Diseases and COVID-19 to Support Spatial Decision-Making

Xiao Huang, Renyi Zhang, Xiao Li, Bahar Dadashova, Lingli Zhu, Kai Zhang, Yu Li, and Bairong Shen

Abstract Infectious diseases remain an essential global challenge in public health. For instance, novel coronavirus (COVID-19) has resulted in significant negative impacts on public health, infecting more than 214 million people and causing 4.47 million deaths worldwide as of August 2021. Geographic Information Systems have played an essential role in managing, storing, analyzing, and mapping disease and related risk information. This article provides an overview of a broad topic on applications of GIS into infectious disease research. Our review follows the framework of human–environment interactions, focusing on the environmental and social factors that cause the disease outbreak and the role of humans in disease control, including public health policies and interventions such as social distancing/face covering practice and mobility modeling. The work identifies key spatial decision-making

X. Huang
Department of Geosciences, University of Arkansas, Fayetteville, AR, USA

R. Zhang
Department of Atmospheric Sciences, Texas A&M University, College Station, TX, USA

X. Li · B. Dadashova
Texas A&M Transportation Institute, Texas A&M University, College Station, TX, USA

L. Zhu
Department of Remote Sensing and Photogrammetry, National Land Survey of Finland, Helsinki, Finland

K. Zhang
Department of Environmental Health Sciences, University at Albany, State University of New York, Albany, NY, USA

Y. Li
School of Hydraulic Engineering, Dalian University of Technology, Dalian, China

B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_8

issues where GIS becomes valued in the agenda for infectious disease research and highlights the importance of adopting science-based policies to protect the public during the current and future pandemics.

Keywords Infectious disease · COVID-19 · Geographic information system (GIS) · Social distancing · Spatial decision-making · Public health policy

8.1 Introduction

Environmental pollution, disasters, urbanization, global warming, and rapid population growth have become the significant factors that cause infectious disease outbreaks [1–3]. Infectious diseases remain an essential global challenge in public health, causing over 13 million deaths each year. According to statistics, viral hepatitis, influenza, and tuberculosis stand among the leading causes of illness and death in the United States [4]. Since 2019, novel coronavirus (COVID-19) started to be detected from humans, which rapidly developed into a global pandemic, infecting more than 214 million people, and causing 4.47 million deaths worldwide as of August 2021 [5]. COVID-19 has changed human production and life behavior not only affected the water system, but also had a strong impact on a wider range of energy systems and food systems under the global background of high coupling of food, energy, water, and environment, and then affects the process of sustainable development of economy, society, and environment in the whole region. For the energy system, the reduction of power demand and the decline of fossil fuel use caused by the economic recession during COVID-19 have significantly reduced the carbon dioxide emissions of the global power sectors [6–8].

The development of computer-based geographic information systems (GIS) for integrating and analyzing spatially referenced data has provided new tools for medical geographic research on infectious disease control. Infectious diseases have revealed strong spatial patterns, where Geographic Information Systems (GISs) played a central role in managing, storing, analyzing, and mapping disease information. The Coronavirus Resource Center established by the Johns Hopkins University is one of the noteworthy examples of this practice (see https://coronavirus.jhu.edu/map.html). Disease cartography began with Koch's work, including the spatial mapping of pandemics such as the European plague and yellow fever [9]. Later, the GIS-based disease mapping tools also leveraged many other kinds of data such as demographic, social media, and environmental data to improve disease surveillance and decision-making [10–12].

Spatial decision-making and spatial decision support systems have been widely discussed in the GIS research for solving real-world problems such as disaster management, environmental and water resources management, agriculture risk management, and public health surveillance [13–22]. The existing literature describing GIS-based public health applications suggests that GIS diffusion into infectious diseases research and public health practice has moved beyond the early innovation phase [23]. Such publications can be identified in an extensive range of outlets, including multidisciplinary journals on public health, environmental science, social science studies, GIS conference proceedings, and government reports. For instance, numerous COVID-19 related research articles have been published since 2019 in the journals (or proceedings) of environmental science, geography, geosciences, infectious diseases, computer science, and multidisciplinary studies. Nevertheless, it is unclear to what extent and depth GIS has been utilized in infectious disease studies. For instance, which types of infectious diseases research have attracted most GIS applications? What kinds of GIS-based methodologies have been used in analyzing infectious diseases? Some infectious diseases such as COVID-19 are highly contagious, where public health policies (e.g., social distancing), human behavior, and mobility analysis have been extensively analyzed with the help of GIS-based data and methodologies in infectious disease studies.

This review article tends to systematically review and inductively summarize the influential literature on applications of GIS into infectious disease research. Figure 8.1 illustrates the workflow of the article. Our review follows the framework of human–environment interactions, where the term "environment" represents the environmental and social factors that contribute to disease outbreak and transmission. The term "human" represents the role of humans in disease control, including public health policies and interventions such as social distancing practice and mobility modeling. This reminder of this review paper is structured into the following sections. Section 8.2 systemically reviewed and summarized the typical applications of four types of GIS techniques in infectious disease-related research, including spatial clustering and statistics, spatial interpolation, WebGIS and spatial visualization, and spatial modeling. In Sect. 8.3, we conducted an in-depth review of COVID-19 related research works. We paid particular attention to an emerging

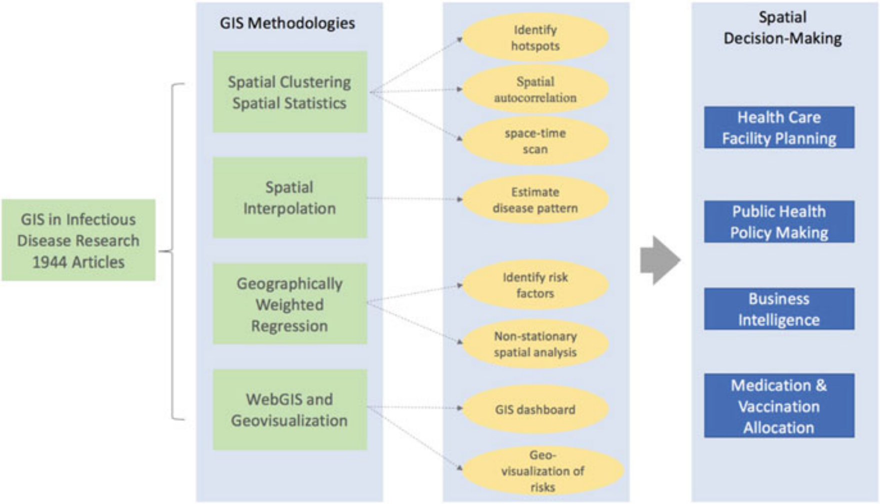
GIS in Infectious Disease Research 1944 Articles
GIS Methodologies
Spatial Clustering Spatial Statistics
Spatial Interpolation
Geographically Weighted Regression
WebGIS and Geovisualization
Identify hotspots
Spatial autocorrelation
space-time scan
Estimate disease pattern
Identify risk factors
Non-stationary spatial analysis
GIS dashboard
Geo-visualization of risks
Spatial Decision-Making
Health Care Facility Planning
Public Health Policy Making
Business Intelligence
Medication & Vaccination Allocation

Fig. 8.1 Use of GIS in infectious disease research for spatial decision-making

geographic data source—fine-grained mobility data, reviewed, and summarized the existing efforts about how to use mobility data to assess different COVID-19 protective measures (e.g., social distancing) and how to use mobility data to facilitate decision-making during different stages of the pandemic.

8.2 Environmental Distribution of Infectious Disease and GIS-Related Research

This article first developed a search strategy with terms relating to "GIS/Geographic Information Systems" and "Infectious Disease." This search was developed through an iterative process of incorporating new terms and refining those included based on results returned and identification of relevant citations. We conducted an electronic search on the Web of Science database with no restriction on the date or language of publication. We found 1944 peer-reviewed articles that focus on infectious disease and involved GIS or spatial analysis. Figure 8.2 illustrates the number of identified articles by different publishers, with Springer Nature publishing the most GIS-related infectious disease research, followed by Elsevier and Willey.

In the next step, we used the keyword "GIS" combined with different types of infectious disease keywords such as "HIV," "Influenza," and "COVID-19" to group the articles by different disease types. Table 8.1 illustrates the number of articles (with their corresponding citations) that applied GIS and spatial analysis for each type of infectious disease. According to Table 8.1, Malaria, COVID-19, and Human Immunodeficiency Virus (HIV) are the top three diseases that mostly utilized GIS and spatial analysis in their relevant research works.

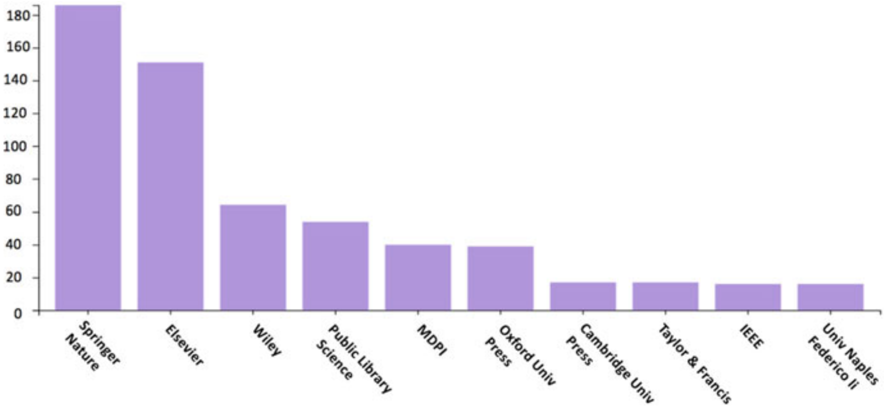

Fig. 8.2 Illustration of publishers for GIS-related infectious disease articles

Table 8.1 Illustration of several articles and citations that are related to applying GIS in infectious disease research

Infectious disease	Number of articles	Number of citations
Malaria	430	7233
COVID	226	1132
HIV/AIDS	212	2972
Escherichia Coli	193	3396
Tuberculosis	178	2299
Influenza	153	2213
West Nile virus	119	2153
Lyme	113	2554
Viral hepatitis	76	1202
Salmonella	57	1184
Severe acute respiratory syndrome	46	847
Pneumonia	36	716
Hand-foot-mouth disease	33	322
Measles	20	415
Meningitis	17	176
Whooping cough (pertussis)	11	139
Poliomyelitis	7	58
Diphtheria	6	561
Tetanus	5	75
Chickenpox	2	102
Giardiasis	1	12
Infectious mononucleosis	1	10
Mumps	2	11
Total	1944	39, 987

8.2.1 Use of Spatial Clustering and Spatial Statistics in Identifying Disease Hotspots

Spatial clustering and spatial statistics are two of the mostly used spatial analysis techniques for evaluating infectious diseases' geographic distribution (see Table 8.2). In this section, we searched for articles with keywords "infectious disease," "GIS," and "spatial clustering and statistics" in the Web of Science database. Results have returned with 49 articles. We removed duplicated and un-relevant articles and selected ten articles for analysis. Spatial clustering is used to partition spatial data (e.g., disease data) into a series of meaningful subclasses called spatial clusters, where spatial objects that are within the same cluster are similar to each other [36]. Spatial autocorrelation is often used in the GIS to identify how well objects correlate with other nearby objects across a spatial area [36]. As listed in Table 8.2, spatial autocorrelation was used in five articles for studying the spatial distribution of Hepatitis, Tuberculosis, HIV, Mumps, and SARS diseases. Spatial clustering methods such as Kulldorff's spatial scan and self-organizing maps were used in seven articles. In these articles, the spatial scan statistic identified statistically significant hotspots based on the number of disease cases by systematically scanning circular windows using varying sizes across the study area [26–28]. A space-time scan was used to test the statistically significant clusters of the disease cases across space and time [35]. Other spatial statistical models such as Local Moran's I are global clustering statistics that measure the tendency for points to occur closer together in space by chance across the entire study area [25, 31]. In contrast, the Kulldorff spatial scan statistic identifies local clusters in a particular region. Local clusters can exist in either the absence or presence of global clustering [26, 27].

8.2.2 Use of Spatial Interpolation in Estimating Disease Pattern

Another focused area of using GIS technology in infectious disease mapping is to create "heat maps'‘using data gathered in a limited number of locations to estimate values in unmeasured locations. Spatial interpolation is the process of using points with known values to estimate values at other points [36]. Traditional spatial interpolation methods include kriging interpolation, trend surface interpolation, and inverse distance weighted interpolation. As illustrated in Table 8.3, the kriging interpolation method has been used in six articles studying Burkholderia Pseudomallei, foot-and-mouth disease, norovirus, Tuberculosis, rotavirus, and influenza-like illness. Inverse distance weighted interpolation was used in four articles studying Malaria, Tuberculosis, Kala-azar disease, and Hepatitis. In these articles, spatial interpolation methods were often combined with spatial statistics to analyze spatial transmission patterns of infectious disease ([37, 40];). Spatial interpolation is often used to convert discrete data into continuous data for comparison with the spatial trend of infectious diseases [44, 45]. Others may consider spatial interpolation as a data processing method for spatial analysis [39, 40].

8.2.3 Spatial Visualization and Web-Based GIS Dashboard

With the advancement of web-based technologies (e.g., ArcGIS online), various web-based GIS platforms have been developed to visualize the infectious disease risks at different space-time scale. Some of the well-known dashboards include the WHO Coronavirus dashboard [46], John Hopkins University COVID-19 dashboard [47], the UK National Health Service (NHS) COVID-19 app [48], and CDC COVID-19 data tracker [5]. Spatial interpolation methods have often been combined with web-based geovisualization tools to predict the infectious disease spread patterns [37, 44, 45]. WebGIS and ESRI products such as ArcGIS dashboard are

Table 8.2 Selected studies using cluster detection and spatial statistics to characterize the spatial distribution of infectious disease

References	Study Country	Methods	Key findings
Stopka et al. [24]	United States	Hepatitis C virus/spatial autocorrelation Getis-Ord Gi* statistics	Largest clusters in Boston, New Bedford, Worcester, and Springfield HCV is positively associated with the race of the population
Rao et al. [25]	China	Tuberculosis Moran's I and spatial panel data model	The disease accidents are positively associated with temperature, precipitation, and wind speed
Gwitira et al. [26, 27]	Zimbabwe	HIV/AIDS and malaria Moran's I and space-time clustering Kulldorff's spatial scan	Identify risk areas based on clusters
Aturinde et al. [28]	Uganda	HIV-TB Moran's I and spatial scan statistics	Two clusters were identified in Lake Victoria and the presence of refugee camps
Yu et al. [29]	China	Mumps/spatial autocorrelation and Kulldorff space-time scan	Several clusters have been identified
Lai et al. [30] Lee and Wong [31]	Hongkong, China	SARS/spatial clustering, spatiotemporal clustering, global Moran's I	Origin-and-destination plots showed the directional bias and radius of the spread of superspreading events
Lantos et al. [32]	United States	Lyme/spatiotemporal cluster analysis	Northern Virginia experienced intensification and geographic expansion of Lyme disease cases.
Yang et al. [33]	Taiwan	HIV/spatial statistics	Spatial patterns of different HIV risk behaviors significantly differed in both local clustering patterns and global geographic distribution
Basara and Yuan [34]	United States	Infectious diseases/self-organizing maps	Identified positive relationship between environmental conditions and health outcomes in communities
Dong et al. [35]	China	Influenza (H7N9)/retrospective space-time permutation scan statistic	The epidemic moved from east to southeast coast, and hence to some central regions of China

Table 8.3 Use of spatial interpolation in infectious disease research

Reference	Country	Methods	Key findings
Saengnill et al. [37]	Thailand	Burkholderia Pseudomallei/Mann–Whitney U test, chi test, semivariogram model, and indicator kriging	Burkholderia Pseudomallei is not significantly associated with spatial soil factors. The lag distance between positive case is 90.51 m
Perez [38]	Pakistan	Foot-and-mouth disease/probability co-kriging	A higher risk of disease is associated with increased contact with infectious animal migration
Siya et al. [39]	Uganda	Malaria/inverse distance weighted interpolation (IDW) and Mann-Kendall trend test	Malaria is declining during the study period; rainfall plays an important role in malaria burdens. Altitudes can affect the key factors
Bhunia et al. [40]	India	Kala-azar disease/IDW, Moran index, Getis-Ord Gi*	Southeastern and northwestern part of the study area are with higher incidence rate; Kala-azar incidences are positively correlated for five consecutive years; the spatial trend of disease diffusion is shown
Liu et al. [41]	China	Hepatitis E/trend surface, IDW, spatial-temporal analysis	Higher incidences in northwestern counties of the study area; suggest the need for strengthened supervision and surveillance of sanitary water, sewage treatment, and food in high-risk areas
Inaida et al. [42]	Japan	Norovirus/kriging	Incidences increase in southern areas at first and extend to northern areas in Japan
Ding et al. [43]	China	Multidrug-resistant tuberculosis/kriging	The proportion of MDR-TB cases in all TB cases are higher during 2006–2009 and lower during 2010–2012
Török et al. [44]	USA	Rotavirus/kriging	Confirm the trends of rotavirus activity and identify the variability in the timing of peak disease activity
Sakai et al. [45]	Japan	Influenza-like illness/kriging	Two spreading patterns are observed

commonly used technology for geographical data sharing, visualizing [49]. WebGIS techniques were used in three articles for establishing visualization platforms [50, 51]. Google Maps were used in two articles for visualizing infectious disease information [50, 51]. As one of the most representative WebGIS platforms, ArcGIS Online provides various mapping and analysis functions, geographic data sources, and web-based applications, allowing users to effectively build up web applications

Table 8.4 Use of WebGIS techniques and geovisualization in infectious disease monitoring

Reference	Country	Methods	Key findings
Lu [52]	China	Infectious disease in general/WebGIS, J2EE based architecture is applied to construct a distrusted system infrastructure	A platform that contains georeferenced data can convert disease information into graphical and visual form
AI Manir et al. [53]	Global	Malaria/dashboard	Prototype of surveillance platform for accessing distributed disease data sources
Li et al. [50]	China	Infectious disease in general/WebGIS, Google maps	The platform can display infectious disease emergencies information and transfer information between workers in the field and decision-makers through the internet
Yang et al. [51]	China	Schistosomiasis/Google earth, WebGIS	A WebGIS platform that can operate search, evaluation, risk analysis and prediction. This platform can help identify early high-risk areas and provide detailed information
Patrick et al. [54]	USA	HIV/calculate the proportion of ever tested, tested positive and newly positive in the past year; chi-square test for trend	This dashboard can be used to complement the HIV care continuum
	USA	COVID-19/dashboard	An online dashboard that can display COVID-19 data for every county of 188 metropolitan areas in the USA
Cheng et al. [55]	China	Influenza/dashboard	An influenza surveillance dashboard with several data streams and indicators for monitoring disease activities
Ravinder et al. [56]	India	COVID-19/dashboard	A web-based dashboard that provides a 3-week prediction of COVID-19 incidences

without coding. Meanwhile, it also provides different GIS tools and APIs used by developers while it is not as functional as ArcGIS Desktop. The Google Maps API provides embedded Google Maps into web pages through JavaScript. The APIs provide many utilities to generate maps and customize the map content by adding additional information services. However, these APIs do not support complicated analysis functions. Table 8.4 illustrates selected articles that have used WebGIS techniques and geovisualization in infectious disease monitoring.

8.2.4 Exploring Environmental and Social Factors Using Spatial Regression Analysis

Several articles are focused on investigating the key factors that affect the occurrence and spread of infectious diseases. Geographically Weighted Regression (GWR) has a high utility in epidemiology, particularly for examining the relationship between the spread of infectious disease with different social, political, and environmental factors (e.g., built environment, health policies, and interventions). GWR is a local form of linear regression used to model spatially varying relationships [57]. Table 8.5 illustrates the key social and environmental factors that have been explored in infectious disease research. According to Table 8.5, environmental factors such as temperature, humidity, precipitation, wind speed, air pressure, altitude, and socioeconomic factors such as child population density and per capita Gross Domestic Product (GDP) are associated with Hand, Foot, and Mouth Disease (HFMD). Other environmental factors such as air pollution, brickfield density, land use, and public transportation facilities significantly impact on COVID-19 cases. Other sociodemographic factors such as gender, nationality, employment status, and occupation types are associated with malaria and tuberculosis.

8.3 Human-Centered Efforts to Address COVID-19 Challenges

Novel coronavirus (COVID-19) has significant negative impacts on public health, infecting more than 214 million people and causing 4.47 million deaths worldwide as of August 2021. The COVID-19 pandemic is much more pronounced than many of the previous outbreaks of infectious diseases, including the 2002/2003 SARS. The enormous scope and magnitude of the COVID-19 outbreak reflects a highly contagious nature and exceedingly efficient transmission for SARS-CoV-2. There exists two primary pathways for respiratory viruses to be transmitted from person to person (Fig. 8.3a). Virus-bearing particles are produced from breathing, talking, coughing, or sneezing by an infected person. Interhuman transmission occurs by the direct (deposited on persons) or indirect (deposited on objects) contact route via respiratory droplets (>5 μm) or the airborne route via respiratory aerosols (<5 μm). While large respiratory droplets readily settle out of air to cause person/object contamination, small virial-bearing respiratory aerosols are efficiently dispersed in air and inhaled by human to lead direct deposition along the respiratory tract and to cause infection [13, 14, 19]. Well-established public health measures to prevent interhuman transmission include face covering, social distancing, and testing/quarantine (Fig. 8.3b). There exists now compelling scientific evidence for the importance of airborne transmission in spreading the COVID-19 disease and face covering in preventing interhuman [11, 13, 14, 19]. Also, increasing ventilation in an enclosed community setting has been shown to effectively reduce viral

Table 8.5 Exploring environmental and social factors that cause infectious disease

References	Infectious disease/environmental factors	Methods	Key findings
Hassan et al. [2]	COVID-19/PM$_{2.5}$ pollution, population density, brickfield density, rainfall, wind speed, poverty level	GWR	Significant robust relationships between these factors were found in the middle and southern parts of the city, where the reported COVID-19 infection case was also higher
Wu and Zhang [58]	COVID-19/social, economic, and environmental factors such as population density, hospitalization, and age	GWR and principal component analysis (PCA)	In El Paso, Odessa, Midland, Randall, and Potter County areas in Texas, population, hospitalization, and age structures are presented as static, positive influences on COVID-19 cumulative cases
Ge et al. [1]	Hemorrhagic fever with renal syndrome (HFRS)/temperature, average humidity, average rainfall, area, rodent density, human population density, water area, and surface mean elevation	Seasonal difference-geographically and temporally weighted regression	Meteorological factors notably impacted the changing trends of HFRS outbreaks
Sun et al. [3] Hong et al. [59] Hu et al. [60] Dong et al. [61] Hu et al. [62]	Hand, foot, and mouth disease (HFMD)/temperature, humidity, precipitation, wind speed, air pressure, altitude, child population density, and per capita GDP	GWR, geographically and temporally weighted regression	The findings help to understand the seasonally and spatially relevant effects of natural environmental and socioeconomic factors on the HFMD
Yang et al. [63]	Malaria/occupation, annual average temperature, annual cumulative rainfall, rice yield per square kilometer and proportion of rural employees	GWR	Temperature, precipitation, rice cultivation, and proportion of rural employees were positively associated with malaria incidence

(continued)

Table 8.6 (continued)

References	Infectious disease/environmental factors	Methods	Key findings
Lak et al. [64]	COVID-19/12 quantitative place-based variables related to physical attributes, land use and public transportation facilities, and demographic status	GWR	Demographic composition and major neighborhood-level physical attributes are important factors explaining high rates of infection and mortality
Mohidem et al. [65]	Tuberculosis (TB)/gender, nationality, employment status, health care worker status, income status, residency, smoking status as well as; environmental factors such as AQI, CO, NO_2, SO_2, PM10, rainfall, relative humidity, temperature, wind speed, and atmospheric pressure	Moran’s I, Getis-Ord Gi* statistics Geographically weighted regression	Sociodemographic factors were associated with TB cases ($p < 0.05$). GWR model was the best model to determine the spatial distribution of TB cases

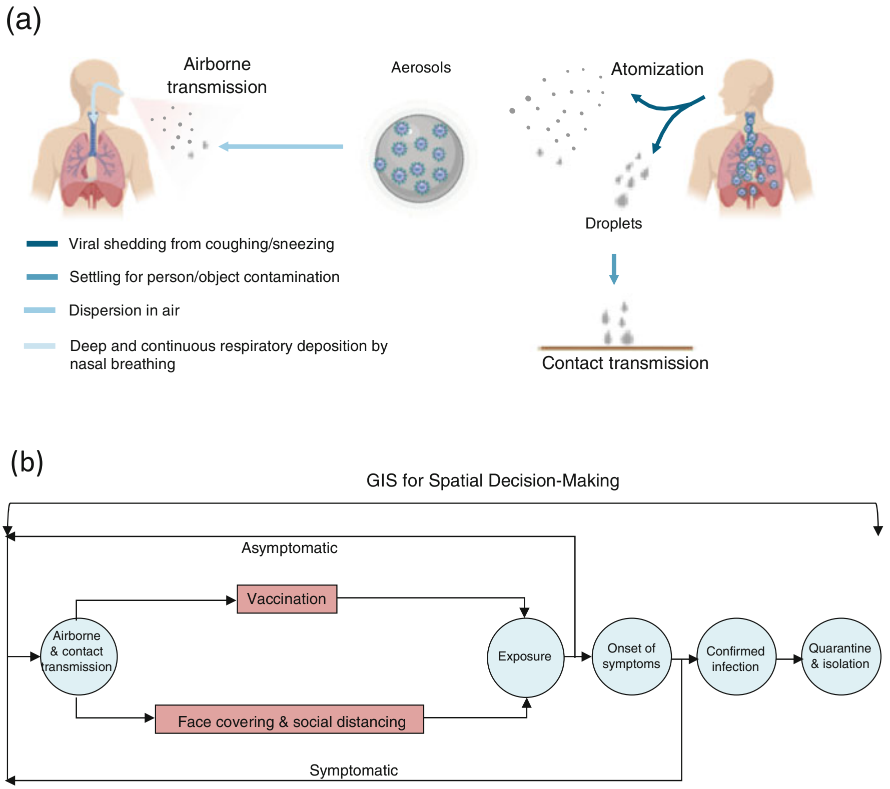
(a)
Airborne transmission
Aerosols
Atomization
Droplets
Viral shedding from coughing/sneezing
Settling for person/object contamination
Dispersion in air
Deep and continuous respiratory deposition by nasal breathing
Contact transmission
(b)
GIS for Spatial Decision-Making
Asymptomatic
Vaccination
Airborne & contact transmission
Exposure
Onset of symptoms
Confirmed infection
Quarantine & isolation
Face covering & social distancing
Symptomatic

Fig. 8.3 Transmission, science-based intervention, and application of GIS. (**a**) Illustration of viral transmission routes (adopted from Zhang et al. 2020). (**b**) Mitigation for preventing interhuman transmission and the application of GIS in decision-making. The boxes denote mitigation measures, and the circles depict the disease evolution

transmission [66]. Vaccination is commonly believed to mitigate viral transmission, albeit for the occurrence of break-through infections [67]. The effectiveness of vaccination has been clearly documented to significantly reduce hospitalization, severe syndromes, and mortality [68].

As the COVID-19 outbreak grew to an epidemic, and various GIS systems have been developed and implemented, leading the response to COVID-19 in many ways. For instance, Johns Hopkins University launched its COVID-19 dashboard using ESRI technology [47]. So far, social distancing plays an important role in controlling the spread of coronavirus. Governments issued different level of restrictions on traveling, institutions canceled gatherings, and citizens socially distanced themselves to limit the spread of the virus. Social distancing measures have significantly influenced the mobility patterns, which have been widely discussed in various COVID-19 related GIS applications. On the other hand, those literature are also tightly related to public health policy and social equity issues, which are worthy of future research. This article illustrates the key findings of using GIS in mobility and policy analysis during the COVID-19 pandemic. We structured our reviews by different stages of pandemic control, i.e., early stage, controlling stage, reopening stage, and post-pandemic recovering stage. We found 228 articles related to the topic. In the following four subsections, we discuss human-centered efforts that leverage mobility data in addressing COVID-19-related challenges.

8.3.1 Early in the Pandemic: Contact Tracing and Initial Control

At the early stage of the COVID-19 pandemic, location-based intelligence has been widely adopted to provide situational awareness for policy-makers and researchers. Human mobility records retrieved from cell phone users' location data (by way of GPS, cell phone towers, and/or Wi-Fi), electronic wristbands, credit card transactions, and closed-circuit television (CCTV) systems can assist in tracking disease spread and enforcing social isolation measures [69]. In China, Alipay and WeChat, two big providers of mobile payment systems, released apps that combine users' health, location, and financial data to generate a personal infection risk rating [70]. Other government-backed apps were also used in the early stage of the pandemic to collect users' essential information, and necessary user scanning was required at checkpoints to better gauge people's moving patterns. Besides efforts and guidelines by the officials, crowdsourcing efforts are also popular, as citizens themselves can contribute to contact tracing and surveillance by voluntarily sharing their whereabouts online. For instance, Private Kit (https://privatekit.mit.edu/), released by the Massachusetts Institute of Technology, is a crowdsourcing application that stores GPS location records from users every 5 min for up to 28 days. Users have the option to share their location data and notify health officials if they test positive for COVID-19. Numerous studies have proved that human mobility records with fine spatiotemporal granularity are essential for disease spread control, as reconstructed trajectories of individuals who have been tested positive can be used to alert those who may have been put at risk of infection [71, 72]. Zhang et al. [13, 14, 19] studied the relationship between human mobility and the cross-space infection in the early stage of the pandemic, based on which a variety of counterfactual analyses is developed to examine the necessity of lock-down and the other containment approaches.

8.3.2 During Control Measures: Compliance Monitoring

To contain the COVID-19 pandemic, one of the non-pharmacological epidemic control measures is to reduce the transmission rate of SARS-COV-2 in the population via social distancing or other similar quarantine measures [11, 73]. Besides the proof from epidemiologic simulations, many pieces of evidence have been found in numerous studies that the implementation of mobility-restricting measures is responsible for the declined transmission rates (e.g., [74, 75]). In certain cases, however, different countries, states/provinces, counties/towns, and other administrative units choose to handle COVID-19 in different ways, with great disparity in the implementation of policies and guidelines. Even in regions under the same level of restrictions, disparities in compliance tend to occur. Human mobility records, either at the individual level or aggregated to certain geographic units, can reflect how people adjust their travel patterns under the COVID-19 pandemic and whether policies are implemented in an effective manner. There are some notable efforts that Huang et al. [76] analyzed over 580 million tweets worldwide to investigate how people follow mobility-restricting measures at the global, country, and U.S. state levels. Their results revealed great discrepancies in responsiveness, evidenced by the contrasting mobility patterns in different epidemic phases at their investigated scales. Taking advantage of Google's COVID-19 mobility reports, Bargain and Aminjonov [77] investigated how policy compliance is linked with political trust at the regional level in Europe. Their findings indicate that high-trust regions decrease their mobility significantly more than low-trust regions, and the efficiency of policy stringency in terms of mobility reduction significantly increases with trust. Other efforts coupled mobility-related indices with sociodemographic factors, aiming to reveal the determinants that potentially lead to the disparity in policy compliance (e.g., [78]; Chiou and Tucker). The general findings point to the luxury nature of mobility-restricting measures (e.g., working from home and other virtual working conditions) with which socioeconomically disadvantaged groups cannot afford to comply. Zhu et al. [79] utilized network optimization to identify how the geographical centers of the pandemic moved spatially over time across the USA in the context of various intervention policies. The pandemic has also witnessed much mis- and dis-information. Network reconstruction methods can be employed to measure the interaction between the information diffusion and the outbreak of COVID-19 across space, and identify both positive and negative impact of information on the pandemic [12, 15]. The above evidence reveals the essential role of mobility data in policy compliance monitoring during the COVID-19 pandemic, which benefits further policymaking in terms of adjusting controlling measures and mitigating compliance disparity.

8.3.3 Reopening: When, How, and Where

After the implementation of mobility-restricting measures, federal and local government officials have been investigating reopening strategies, such as when and where to reopen borders and business, and how much activities are allowed in certain places. These reopening strategies, however, should be determined in a scientific manner with the assistance of epidemiological models that consider human mobility dynamics. Many studies have been conducted to assist in reopening decision-making taking advantage of fine-grained human mobility data. One notable effort is by Chang et al. [80], who built enormous mobility networks containing 5.4 billion hourly edges from mobile phone data that cover hourly movements of 98 million people from 56,945 U.S. census block groups to 552,758 points of interest (POIs). The results suggested that, coupled with detailed mobility records, their simulation can estimate the effects of specific reopening strategies in the USA. Using the same dataset, Andersen et al. [81] examined U.S. college reopenings' association with changes in human mobility within campuses and in COVID-19 incidence in the U.S. counties of the campuses over a 10-week period around college reopenings. They found that college reopenings were associated with increased campus mobility, responsible for the increased COVID-19 incidence by 2.7 cases per 100,000. Xiong et al. [82] investigated the partial reopening phases in the USA by leveraging anonymized mobile device location data from over 100 million monthly active users procured from multiple third-party data providers. The detailed mobility records coupled with their models revealed the high likelihood of a second spike in coronavirus in many early-opening regions. The above examples highlight the necessity of human mobility data in optimizing reopening decisions.

8.3.4 Post-Pandemic: Recovery and Transition Gauging

Human mobility data can be used to tell stories regarding how different regions recover after the lifting of strict mobility-restricting orders and the implementation of reopening policies by comparing the human moving patterns in post-pandemic situations to the ones in pre-pandemic situations. While some of the changes are temporary, such as the disruptive social, physical, and economic activities in urban and rural landscapes during the stay-at-home orders (most of which have largely recovered after the reopening), others seem to be permanent impacts that force multiperspective transitions in an irreversible manner. Human mobility data that cover multiple stages are expected to benefit the investigation of the dynamic, intertwining, long-term societal effects of the COVID-19 pandemic, filling the knowledge gaps in our understanding of how spatial and social interactions have shifted and transitioned in the post-pandemic world, and informing better adapting, responding, and recovering strategies that reduce inequalities and vulnerabilities. Despite the fact that it is difficult to decide when the post-pandemic era really starts, numerous efforts have been made to gauge recovery and transition when society functions resume. Kupfer et al. [83] investigated park visitation recovery by mapping and analyzing the spatiotemporal patterns of visitation for six national parks in the western USA, taking advantage of large mobility records sampled from mobile devices and released by SafeGraph as part of their Social Distancing Metric dataset. Huang et al. [78] leveraged multi-source mobility datasets from Google, Apple, Descartes Labs, and Twitter to investigate how people reduced their travels during the mobility-restricting period and how mobility recovered after the reopening at the county level in the USA. Their results revealed a great disparity in mobility dynamics in the recovery phase, as the poor countries tended to gain earlier and greater upward momentum than the wealthy counties. Such disparity in recovery has been noted by many studies that take advantage of mobility records (e.g., [76, 83]).

8.4 Conclusion and Discussion

Adopting science-based policies are paramount in protecting the public during the current and future pandemics. This article provides an overview and a summary on applying applications of GIS into infectious disease research, and application of GIS tools for analyzing and maintaining COVID-19. We paid special attention to COVID-19 related research in terms of human-environment interactions. The term "human" represents the role of humans in disease control, including public health policies such as social distancing practice and mobility modeling. A total of 1944 peer-reviewed GIS-based infectious disease research articles were identified, where Springer Nature published the most articles, followed by Elsevier and Willey. Spatial analysis methods such as spatial clustering, spatial statistics, and spatial interpolation (e.g., Kriging), and GWR analysis have been discussed in detail in those articles to demonstrate the important value of using GIS and spatial analysis in infectious disease monitoring. The article also provides the summary of web-based portals (e.g., GIS dashboards) in visualizing infectious disease risks.

The article also includes a review on human-centered methods for COVID-19 research, including the analysis of social distancing and mobility in COVID-19 disease control and policymaking. We structured this section by different pandemic stages, including early-pandemic, under strong control measures, reopening, and post-pandemic recovery. In the early stage, several articles discussed using human mobility records derived from emerging geo-data sources (e.g., cell phone location data, electronic wristbands, credit card transactions, and closed-circuit television (CCTV) to assist in tracking disease spread and enforcing social isolation measures. In the disease controlling stage, much evidence has been found that the implementation of mobility-restricting measures is responsible for the declined transmission rates. Later in the reopening and recovery stages, human mobility data has demonstrated effectiveness in determining how different regions recover after lifting social distancing orders by comparing the human moving patterns in post-pandemic situations to those in pre-pandemic situations.

According to the literature review performed in this study, GIS has been frequently used to prevent and control of infectious diseases to facilitate the appropriate spatial decision-making. By identifying spatial hot spots/patterns and potential risk factors of infectious diseases as well as vulnerable populations, the governmental and public health agencies, health care organizations, and other stakeholders, can put more efforts and resources into those regions and develop effective prevention strategies and mitigation actions. Furthermore, spatiotemporal disease modeling (e.g., Geographically and Temporally Weighted Regression) could also advance the understanding of spatiotemporal variation characteristics of the environmental and sociodemographic factors on the disease incidence and prevalence. Leveraging GIS techniques in COVID-19 research may produce broad impacts in spatial decision-making such as health care facility planning, public health policymaking, business intelligence, and health equity solutions.

Acknowledgments This work was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019), and the National Natural Science Foundation of China (32070671).

References

1. Ge L, Zhao Y, Sheng Z, Wang N, Zhou K, Mu X, Guo L, Wang T, Yang Z, Huo X (2016) Construction of a seasonal difference-geographically and temporally weighted regression (SD-GTWR) model and comparative analysis with GWR-based models for hemorrhagic fever with renal syndrome (HFRS) in Hubei Province (China). Int J Environ Res Public Health 13(11):1062
2. Hassan MS, Bhuiyan MAH, Tareq F, Bodrud-Doza M, Tanu SM, Rabbani KA (2021) Relationship between COVID-19 infection rates and air pollution, geo-meteorological, and social parameters. Environ Monit Assess 193(1):1–20
3. Sun J, Wu S, Yan Z, Li Y, Yan C, Zhang F, Liu R, Du Z (2021) Using geographically weighted regression to study the seasonal influence of potential risk factors on the incidence of HFMD on the Chinese mainland. ISPRS Int J Geo Inf 10(7):448
4. ODPHP (Office of Disease Prevention and Health Promotion) (2021) Immunization and infectious diseases. https://www.healthypeople.gov/node/3527/data-
5. CDC (2021) COVID data tracker. https://covid.cdc.gov/covid-data-tracker/#datatracker-home. Accessed 31 Aug 2021
6. Forster PM, Forster HI, Evans MJ, Gidden MJ, Jones CD, Keller CA, Lamboll RD, Quéré CL, Rogelj J, Rosen D, Schleussner C, Richardson TB, Smith CJ, Turnock ST (2020) Current and future global climate impacts resulting from COVID-19. Nat Clim Chang 10(10):913–919
7. Le Quéré C, Peters GP, Friedlingstein P, Andrew RM, Canadell JG, Davis SJ, Jackson RB, Jones MW (2021) Fossil CO2 emissions in the post-COVID-19 era. Nat Clim Chang 11(3):197–199
8. Shan Y, Ou J, Wang D, Zeng Z, Zhang S, Guan D, Hubacek K (2021) Impacts of COVID-19 and fiscal stimuli on global emissions and the Paris agreement. Nat Clim Change 11(3):200–206. https://doi.org/10.1038/s41558-020-00977-5

9. Koch T, Koch T (2005) Cartographies of disease: maps, mapping, and medicine. Esri Press, Redlands, CA, p 840
10. Gao S, Mioc D, Anton F, Yi X, Coleman DJ (2008) Online GIS services for mapping and sharing disease information. Int J Health Geogr 7(1):1–12
11. Li Y, Zhang R, Zhao J, Molina MJ (2020) Understanding transmission and intervention for the COVID-19 pandemic in the United States. Sci Total Environ 748:141560. https://doi.org/10.1016/j.scitotenv.2020.141560
12. Ye X, Du J, Gong X, Na S, Li W, Kudva S (2021) Geospatial and semantic mapping platform for massive COVID-19 scientific publication search. J Geovisualiz Spatial Anal 5(1):1–12
13. Zhang R, Li Y, Zhang AL, Wang Y, Molina MJ (2020) Identifying airborne transmission as the dominant route for the spread of COVID-19. Proc Natl Acad Sci U S A 117:14857–14863. https://doi.org/10.1073/pnas.2009637117
14. Zhang X, Ji Z, Zheng Y, Ye X, Li D (2020) Evaluating the effect of city lock-down on controlling COVID-19 propagation through deep learning and network science models. Cities 107:102869
15. Zhang X, Zhang ZK, Wang W, Hou D, Xu J, Ye X, Li S (2021) Multiplex network reconstruction for the coupled spatial diffusion of infodemic and pandemic of COVID-19. Int J Dig Earth 14(4):401–423
16. Zhang Z, Demšar U, Rantala J, Virrantaus K (2014) A fuzzy multiple-attribute decision-making modelling for vulnerability analysis on the basis of population information for disaster management. Int J Geogr Inf Sci 28(9):1922–1939
17. Zhang Z, Demšar U, Wang S, Virrantaus K (2018) A spatial fuzzy influence diagram for modelling spatial objects' dependencies: a case study on tree-related electric outages. Int J Geogr Inf Sci 32(2):349–366
18. Zhang Z, Hu H, Yin D, Kashem S, Li R, Cai H et al (2018) A cyberGIS-enabled multi-criteria spatial decision support system: a case study on flood emergency management. Int J Dig Earth
19. Zhang Z, Laakso T, Wang Z, Pulkkinen S, Ahopelto S, Virrantaus K et al (2020) Comparative study of AI-based methods—application of analyzing inflow and infiltration in sanitary sewer subcatchments. Sustainability 12(15):6254
20. Zhang Z, Zou L, Li W, Usery L, Albrecht J, Armstrong M (2021) Cyberinfrastructure and intelligent spatial decision support systems. Trans GIS 25(4):1651–1653
21. Zhang Z, Yin D, Virrantaus K, Ye X, Wang S (2021) Modeling human activity dynamics: an object-class oriented space–time composite model based on social media and urban infrastructure data. Computat Urban Sci 1(1):1–13
22. Zhao J, Zhang Z, Sullivan CJ (2019) Identifying anomalous nuclear radioactive sources using Poisson kriging and mobile sensor networks. PLoS One 14(5):e0216131
23. Cromley EK, McLafferty SL (2011) GIS and public health. Guilford Press, New York
24. Stopka TJ, Goulart MA, Meyers DJ, Hutcheson M, Barton K, Onofrey S, Church D, Donahue A, Chui KK (2017) Identifying and characterizing hepatitis C virus hotspots in Massachusetts: a spatial epidemiological approach. BMC Infect Dis 17(1):1–11
25. Rao HX, Zhang X, Zhao L, Yu J, Ren W, Zhang XL, Ma YC, Shi Y, Ma BZ, Wang X, Wei Z (2016) Spatial transmission and meteorological determinants of tuberculosis incidence in Qinghai Province, China: a spatial clustering panel analysis. Infect Dis Poverty 5(1):1–13
26. Gwitira I, Mukonoweshuro M, Mapako G, Shekede MD, Chirenda J, Mberikunashe J (2020) Spatial and spatio-temporal analysis of malaria cases in Zimbabwe. Infect Dis Poverty 9(1):1–14
27. Gwitira I, Murwira A, Mberikunashe J, Masocha M (2018) Spatial overlaps in the distribution of HIV/AIDS and malaria in Zimbabwe. BMC Infect Dis 18(1):1–10
28. Aturinde A, Farnaghi M, Pilesjö P, Mansourian A (2019) Spatial analysis of HIV-TB co-clustering in Uganda. BMC Infect Dis 19(1):1–10
29. Yu G, Yang R, Wei Y, Yu D, Zhai W, Cai J, Long B, Chen S, Tang J, Zhong G, Qin J (2018) Spatial, temporal, and spatiotemporal analysis of mumps in Guangxi Province, China, 2005–2016. BMC Infect Dis 18(1):1–13

30. Lai PC, Wong CM, Hedley AJ, Lo SV, Leung PY, Kong J, Leung GM (2004) Understanding the spatial clustering of severe acute respiratory syndrome (SARS) in Hong Kong. Environ Health Perspect 112(15):1550–1556
31. Lee SS, Wong NS (2011) The clustering and transmission dynamics of pandemic influenza A (H1N1) 2009 cases in Hong Kong. J Infect 63(4):274–280
32. Lantos PM, Nigrovic LE, Auwaerter PG, Fowler VG Jr, Ruffin F, Brinkerhoff RJ, Reber J, Williams C, Broyhill J, Pan WK, Gaines DN (2015) Geographic expansion of Lyme disease in the southeastern United States, 2000–2014. Open Forum Infect Dis 2(4):ofv143. Oxford University Press
33. Yang AC, Wen TH, Shih CC, Fang CT (2011) Differentiating geographic patterns of human immunodeficiency virus (HIV) infection with different risk factors in northern Taiwan: 1997–2008. Appl Geogr 31(2):519–524
34. Basara HG, Yuan M (2008) Community health assessment using self-organizing maps and geographic information systems. Int J Health Geogr 7(1):1–8
35. Dong W, Yang K, Xu Q, Liu L, Chen J (2017) Spatio-temporal pattern analysis for evaluation of the spread of human infections with avian influenza a (H7N9) virus in China, 2013–2014. BMC Infect Dis 17(1):1–13
36. Longley PA, Goodchild MF, Maguire DJ, Rhind DW (2005) Geographic information systems and science. John Wiley & Sons, London
37. Saengnill W, Charoenjit K, Hrimpeng K, Jittimanee J (2020) Mapping the probability of detecting Burkholderia pseudomallei in rural rice paddy soil based on indicator kriging and spatial soil factor analysis. Trans R Soc Trop Med Hyg 114(7):521–530
38. Perez AM (2004) Probability co-kriging estimation of foot and mouth disease spatial distribution in Pakistan. GISVET's 43
39. Siya A, Kalule BJ, Ssentongo B, Lukwa AT, Egeru A (2020) Malaria patterns across altitudinal zones of mount Elgon following intensified control and prevention programs in Uganda. BMC Infect Dis 20(1):1–16
40. Bhunia GS, Kesari S, Chatterjee N, Kumar V, Das P (2013) Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India. BMC Infect Dis 13(1):1–12
41. Liu K, Cai J, Wang S, Wu Z, Li L, Jiang T, Chen B, Cai G, Jiang Z, Chen Y, Wang Z (2016) Identification of distribution characteristics and epidemic trends of hepatitis E in Zhejiang Province, China from 2007 to 2012. Sci Rep 6(1):1–11
42. Inaida S, Shobugawa Y, Matsuno S, Saito R, Suzuki H (2013) The south to north variation of norovirus epidemics from 2006–07 to 2008–09 in Japan. PLoS One 8(8):e71696
43. Ding P, Li X, Jia Z, Lu Z (2017) Multidrug-resistant tuberculosis (MDR-TB) disease burden in China: a systematic review and spatio-temporal analysis. BMC Infect Dis 17(1):1–29
44. Török TJ, Kilgore PE, Clarke MJ, Holman RC, Bresee JS, Glass RI (1997) Visualizing geographic and temporal trends in rotavirus activity in the United States, 1991 to 1996. Pediatr Infect Dis J 16(10):941–946
45. Sakai T, Suzuki H, Sasaki A, Saito R, Tanabe N, Taniguchi K (2004) Geographic and temporal trends in influenzalike illness, Japan, 1992–1999. Emerg Infect Dis 10(10):1822
46. World Health Organization (2021) WHO coronavirus dashboard. https://covid19.who.int. Accessed 31 Aug 2021
47. Johns Hopkins University Center for Systems Science and Engineering (2021) Covid-19 dashboard. https://coronavirus.jhu.edu/map.html. Accessed 31 Aug 2021
48. NHS (2021) NHS COVID-19 app. https://covid19.nhs.uk. Accessed 1 Sept 2021
49. Alesheikh AA, Helali H, Behroz HA (2002) Web GIS: technologies and its applications. In Symposium on geospatial theory, processing and applications (Vol. 15)
50. Li YP, Fang LQ, Gao SQ, Wang Z, Gao HW, Liu P, Wang ZR, Li YL, Zhu XG, Li XL, Xu B (2013) Decision support system for the response to infectious disease emergencies based on WebGIS and mobile services in China. PLoS One 8(1):e54842
51. Yang K, Sun LP, Huang YX, Yang GJ, Wu F, Hang DR, Li W, Zhang JF, Liang YS, Zhou XN (2012) A real-time platform for monitoring schistosomiasis transmission supported by Google

earth and a web-based geographical information system. Geospat Health 6(2):195–203

52. Lu X (2009) Web GIS based information visualization for infectious disease prevention. In 2009 Third International Symposium on Intelligent Information Technology Application (Vol. 1, pp. 148–151). IEEE
53. Al Manir MS, Brenas JH, Baker CJ, Shaban-Nejad A (2018) A surveillance infrastructure for malaria analytics: provisioning data access and preservation of interoperability. JMIR Public Health Surveill 4(2):e10218
54. Patrick R, Greenberg A, Magnus M, Opoku J, Kharfen M, Kuo I (2017) Development of an HIV testing dashboard to complement the HIV care continuum among MSM, PWID, and heterosexuals in Washington, DC, 2007–2015. J Acquir Immune Defic Syndr 75(Suppl 3):S397
55. Cheng CK, Ip DK, Cowling BJ, Ho LM, Leung GM, Lau EH (2011) Digital dashboard design using multiple data streams for disease surveillance with influenza surveillance as an example. J Med Internet Res 13(4):e85
56. Ravinder R, Singh S, Bishnoi S, Jan A, Sharma A, Kodamana H, Krishnan NA (2020) An adaptive, interacting, cluster-based model for predicting the transmission dynamics of COVID-19. Heliyon 6(12):e05722
57. Jiang H, Hu H, Li B, Zhang Z, Wang S, Lin T (2021) Understanding the non-stationary relationships between corn yields and meteorology via a spatiotemporally varying coefficient model. Agric For Meteorol 301:108340
58. Wu X, Zhang J (2021) Exploration of spatial-temporal varying impacts on COVID-19 cumulative case in Texas using geographically weighted regression (GWR). Environ Sci Pollut Res:1–15
59. Hong Z, Mei C, Wang H, Du W (2021) Spatiotemporal effects of climate factors on childhood hand, foot, and mouth disease: a case study using mixed geographically and temporally weighted regression models. Int J Geogr Inf Sci:1–23
60. Hu B, Qiu W, Xu C, Wang J (2020) Integration of a Kalman filter in the geographically weighted regression for modeling the transmission of hand, foot and mouth disease. BMC Public Health 20(1):1–15
61. Dong W, Yang P, Liao H, Wang X, Wang Q (2016) The effects of weather factors on hand, foot and mouth disease in Beijing. Sci Rep 6(1):1–9
62. Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, Guo Y, Zhao D, Yang W (2012) Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. PLoS One 7(6):e38978
63. Yang D, Xu C, Wang J, Zhao Y (2017) Spatiotemporal epidemic characteristics and risk factor analysis of malaria in Yunnan Province, China. BMC Public Health 17(1):1–10
64. Lak A, Sharifi A, Badr S, Zali A, Maher A, Mostafavi E, Khalili D (2021) Spatio-temporal patterns of the COVID-19 pandemic, and place-based influential factors at the neighborhood scale in Tehran. Sustain Cities Soc:103034
65. Mohidem NA, Osman M, Hashim Z, Muharam FM, Mohd Elias S, Shaharudin R (2021) Association of sociodemographic and environmental factors with spatial distribution of tuberculosis cases in Gombak, Selangor, Malaysia. PLoS One 16(6):e0252146
66. Bazant MZ, Bush WM (2021) A guideline to limit indoor airborne transmission of COVID-19. Proc Natl Acad Sci U S A 118(17):e2018995118. https://doi.org/10.1073/pnas.2018995118
67. Fowlkes A, Gaglani M, Groover K, Thiese MS, Tyner H, Ellingson K (2021) Effectiveness of COVID-19 vaccines in preventing SARS-CoV-2 infection among frontline workers before and during B.1.617.2 (Delta) variant predominance — eight U.S. Locations, December 2020–August 2021. MMWR Morb Mortal Wkly Rep 70:1167–1169. https://doi.org/10.15585/mmwr.mm7034e4
68. Tregoning JS, Flight KE, Higham SL, Wang Z, Pierce BF (2021) Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. Nat Rev Immunol. https://doi.org/10.1038/s41577-021-00592-1
69. Rosenkrantz L, Schuurman N, Bell N, Amram O (2021) The need for GIScience in mapping COVID-19. Health Place 67:102389

70. Calvo RA, Deterding S, Ryan RM (2020) Health surveillance during covid-19 pandemic. BMJ:369
71. Colizza V, Grill E, Mikolajczyk R, Cattuto C, Kucharski A, Riley S, Kendall M, Lythgoe K, Bonsall D, Wymant C, Abeler-Dörner L (2021) Time to evaluate COVID-19 contact-tracing apps. Nat Med 27(3):361–362
72. Li J, Guo X (2020) COVID-19 contact-tracing apps: a survey on the global deployment and challenges. arXiv preprint arXiv:2005.03599
73. Huang X, Li Z, Lu J, Wang S, Wei H, Chen B (2020) Time-series clustering for home dwell time during COVID-19: what can we learn from it? ISPRS Int J Geo Inf 9(11):675
74. Hadjidemetriou GM, Sasidharan M, Kouyialis G, Parlikad AK (2020) The impact of government measures and human mobility trend on COVID-19 related deaths in the UK. Transport Res Interdisc Perspect 6:100167
75. Kraemer MU, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, Du Plessis L, Faria NR, Li R, Hanage WP, Brownstein JS (2020) The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 368(6490):493–497
76. Huang X, Li Z, Jiang Y, Li X, Porter D (2020) Twitter reveals human mobility dynamics during the COVID-19 pandemic. PLoS One 15(11):e0241957
77. Bargain O, Aminjonov U (2020) Trust and compliance to public health policies in times of COVID-19. J Public Econ 192:104316
78. Huang X, Li Z, Jiang Y, Ye X, Deng C, Zhang J, Li X (2021) The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the US during the COVID-19 pandemic. Int J Dig Earth 14(4):424–442
79. Zhu D, Ye X, Manson S (2021) Revealing the spatial shifting pattern of COVID-19 pandemic in the United States. Sci Rep 11(1):1–9
80. Chang S, Pierson E, Koh PW, Gerardin J, Redbird B, Grusky D, Leskovec J (2021) Mobility network models of COVID-19 explain inequities and inform reopening. Nature 589(7840):82–87
81. Andersen MS, Bento AI, Basu A, Marsicano C, Simon K (2021) College openings, mobility, and the incidence of COVID-19 cases. medRxiv:2020–2009
82. Xiong C, Hu S, Yang M, Luo W, Zhang L (2020) Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. Proc Natl Acad Sci U S A 117(44):27087–27089
83. Kupfer JA, Li Z, Ning H, Huang X (2021) Using Mobile device data to track the effects of the COVID-19 pandemic on spatiotemporal patterns of National Park Visitation. Sustainability 13(16):9366

Chapter 9
5G, Big Data, and AI for Smart City and Prevention of Virus Infection

Shumin Ren and Bairong Shen

Abstract With the development of urbanization, artificial intelligence, communication technology, and the Internet of Things, cities have evolved a new ecology from traditional city structures, that is, smart city. Combining 5G and big data, the applications of smart cities have been extended to every aspect of residents' lives. Based on the popularization of communication equipment and sensors, the great improvement in data transmission and processing technology, the production efficiency in medical field, industrial field, and security field has been improved. This chapter introduces the current research related to smart cities, including its architecture, technologies, and equipment involved. Then it discussed the challenges and opportunities of explainable artificial intelligence (XAI), which is the next important development direction of AI, especially in the medical field, where patients and medical personnel have non-negligible needs for the interpretability of AI models. Then, taking COVID-19 as an example, it discussed how smart cities play a role during virus infection and introduced the specific applications designed so far. Finally, it discussed the shortcomings of the current situation and the aspects that can be improved in the future.

Keywords Smart city · 5G · Big data · Artificial intelligence · XAI · Virus infection

S. Ren
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China

Department of Computer Science and Information Technology, University of A Coruña, A Coruña, Spain
e-mail: renshumin@wchscu.cn

B. Shen (✉)
Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu, Sichuan, China
e-mail: bairong.shen@scu.edu.cn

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7_9

9.1 Introduction

The settlements of humans in cities have been promoting the development of urbanization, which has become an unstoppable trend in human society. In the process of urbanization development, due to denser housing and more intensive use of infrastructure, socio-economic development is accelerating, and data from various fields such as transportation, public safety, urban management, education, and medical care are gradually increasing. Cities need to coordinate the deployment, networking, and data processing of various infrastructures to achieve good management of the cities. During this period, a large amount of data is generated, forming big data, with typical 3v, 4v, or even 5v characteristics, including volume, variety, velocity, veracity, and value [1]. In the following content, we will explore how big data and emerging communication technology, 5G, are combined with smart cities, and related applications they brought.

The popularization of machine learning (ML) has led to a lot of artificial intelligence (AI) applications, and AI is an indispensable part of smart city. Through the construction of big data databases and the increase in computing ability, key technologies have been invented and provided for the fields of science and industry. A key function of an artificial intelligence system is the ability to explain the process and reason of the decisions or predictions it makes. However, unfortunately, many of the advanced ML models are "black boxes" because they cannot explain the reasons for their decision-making due to their nested non-linear structure. However, in smart cities, interpretability is vital for some special fields, which is related to making professional decisions. Therefore, we need to explore the future and challenges of XAI.

While large numbers of people gather in cities, cities have also become the source of pollution and diseases. In particular, infectious diseases can have permanent negative effects [2, 3]. Due to the high population density in the city, it will directly affect the prevalence of the virus [4] thus affecting virus control. For example, the spread of COVID-19 has not been fully controlled up to now. The epidemic has severely endangered the lives of citizens and society's economy. Under this situation, it is urgent to accurately predict the epidemic trend of viruses, make prevention, and control strategies. Smart cities have the advantages of big data, powerful algorithms, and infrastructure, which can play an important role in local optimization and overall scheduling in disaster emergencies. In the end, we will discuss how smart cities can play a specific role in the prevention and control of virus infection.

9.2 5G and Big Data Promote the AI and Smart City

9.2.1 5G and Big Data

The development of society leads to changes in the use of mobile and wireless communication systems. Basic services such as e-health, e-government, e-transportation, e-commerce, and e-education will keep developing quickly, and the number of communications equipment in the future will increase rapidly. These developments will lead to an avalanche of mobile and wireless communications. It is predicted that there will be 50 billion connected devices worldwide by 2020 [5], making it impossible for existing network technologies to support this huge growth. Under the social development conditions with such big data and high-circulation, 5G has become the latest generation of mobile communication technology. 5G is supposed to achieve high data rates, high throughput, ultra-reliable and low-latency communications, energy-saving, costs reduction, system capacity increase, and massive machine type communications by introducing high carrier frequencies, a large number of antennas, and new features such as sensors, embedded systems, cyber physical systems (CPS), device-to-device communication (D2D), and fog computing [6], etc. On the basis of cloud computing paradigm, fog computing places resources close to the edge of the network to cope with the growth of connected devices [7]. Meanwhile, with the development of Internet of Things (IoT), the integration of 5G wireless systems will change people's lifestyles, provide solutions for business models, and be applied to advanced fields such as robots, drives, and drones [8].

9.2.2 Artificial Intelligence

Based on the development of big data and the latest mobile communication technologies, the development of AI is also in full swing. Artificial intelligence is described as imitating the cognitive functions of the normal brain, such as problem solving and learning, with the help of data [9]. Generally speaking, artificial intelligence applications can be divided into two modules. The first and most important module is the model learning module, which is mainly responsible for effective data collection, data training, and modeling. On the other hand, the other module is the prediction module, which is responsible for reacting to the current situation. Artificial intelligence algorithms are used to make decisions, usually using real-time data from different sources, such as sensors, digital data, or remote input. The technology has been applied to many sectors, such as finance, national security, healthcare, and transportation. Well known examples include Google Maps, Facebook, and autopilots [10]. In these examples, the use of AI helps to analyze large amounts of data to make reasonable predictions and decisions [11]. The applications of artificial intelligence usually include rule-based systems, classic machine learning (ML), and deep learning (DL). Rule-based system is used to interpret information in a useful way by storing and manipulating knowledge. Generally, rule-based system applies to systems that are formed based on manually formulated rules. ML is a branch of artificial intelligence. It belongs to a rule-based system built using automatic rule inference. It focuses on using filtering data and algorithms to imitate the way humans learn and solve problems, and aims to make accurate predictions or decisions without limited instructions from users [12]. It's able to analyze large amounts of data at high speed. According to the types of input data, ML can be divided into two categories: supervised ML and unsupervised ML. The supervised ML model is constructed based on labeled data, that is, training data, to make predictions on new data. In contrast, in unsupervised ML, the provided data is not labeled and classified before training. Commonly used machine learning algorithms include support vector machines, random forest classifiers, k-means, hierarchical clustering, and artificial neural networks (ANN), while deep learning includes advanced ML algorithms based on artificial neural networks, in which multi-layer processing units are used to infer higher-level features from data [13].

9.2.3 The Development of Smart City

The city is a complex system. The economic and social activities in the city are diversified and interdependent, resulting in economic specialization and the division of labor [14]. Various types of data are distributed in different levels and departments, while their infrastructure and economic components are closely related [15], so urban development decision-makers must adopt a more dynamic and unified vision, emphasize the cooperation across disciplines and professions, promoting the development of intelligence of cities in a scientific, predictable, and quantitative manner.

9.2.3.1 ICT, the Internet of Things, and Smart City

Information and communications technology (ICT) has played a key role in urban development. ICT generally refers to all communication equipment and related services, including radios, televisions, mobile phones, computer and network hardware, satellite systems, etc. [16]. The concept of telecity and smart city occurred due to the use of ICT in the provision of municipal services. Senbab defines telecity as a city that uses information technology for transportation, economics, and other public services, which consists of a series of houses that can achieve simple and a wide range of functions [17]. Telecity can also be regarded as the predecessor of smart city. Going further than telecity, smart city pays more attention to the combination of AI, IoT, and ICT.

The origin of the IoT started in the 1980s, and its goal is to embed technology in daily life [18]. On one hand, the IoT plays a key role in improving people's living standards through e-medicine, smart home, and e-education. On the other hand, the IoT can be used in automation, intelligent logistics, remote monitoring. Driven by various technologies such as sensor nodes (SNs), the IoT is undergoing revolutionary changes to the existing framework by integrating wireless sensor networks (WSNs) into the base layer of IoT framework. As the basic components of the IoT, sensors and actuators can be controlled remotely to perceive and transmit a large amount of data. The data is then processed and analyzed according to user queries. These data will be distributed on various applications and platforms in smart cities. This trend prompted researchers to use the IoT as the basic framework for conceiving smart cities [19]. The WSNs and their various technologies are completely integrated into the infrastructure of cities, which forms a digital skin [20]. The massive amount of information generated by ubiquitous embedded devices will be shared among different platforms and applications, transforming cities into smart cities [21].

Smart city is a concept that integrates all other concepts related to smart applications into the same space [22, 23]. As the future direction of urban development, the smart city can provide people with services and in many fields such as traffic management, medical care, shopping, security, and government decision-making. It aims at optimizing resource use. Under smart cities, citizens will have more opportunities to participate in various services and decision-making, thus improving service efficiency of infrastructure and the quality of life of citizens. Combined with AI, various types of data collected by the urban internet can be used for intelligent decision-making, increasing transparency, thereby increasing productivity and efficiency in various industries. Smart city initially only focused on smart home, which is usually the basic unit of smart city. In recent years, as the IoT gradually replaces the traditional network, the products and components of smart homes cooperate to work, gradually expanding to areas other than home.

9.2.3.2 Smart City and Its Components

Most of the current projects on smart cities are limited to some conventional urban management parts, such as parking management, waste treatment and recycling, infrastructure management, etc. The effectiveness of urban service provision is negatively affected by the lack of the whole city network. Smart cities should be more complete, flexible, and efficient. Today, IoT consists of a large number of smart devices, and IoT related applications are supporting and accelerating the development of smart cities [24]. Through integrating its advanced computing, embedded devices (such as actuators, smart phones, etc.), sensor networks, Internet/WSN protocols, and applications, IoT is incorporating heterogeneous objects into smart life. Smart city is composed of various scenarios, including smart industry, smart security, smart energy, smart transportation, smart healthcare, smart shopping, smart governance, etc. Specifically, for example, smart industry can improve production efficiency and reduce waste of human resources; smart security can monitor the dangerous elements in the city (natural disasters, military terror, etc.) in real time to protect the safety of cities and residents; smart energy can optimize energy supply and reduce energy waster; smart transportation can plan the best travel route, save user time and reduce energy consumption. Smart health can prevent, predict, and treat diseases according to the patient's situation. Smart governance can facilitate residents to understand real-time policies and promote the interaction between the government and the public; smart shopping can provide self-service shopping, and help users find suitable products based on the recommendation system. In the above smart city application scenarios, the data from heterogeneous devices are processed and analyzed to help users make more rational decisions and provide more efficient services. The smart city function distribution is shown in Fig. 9.1.

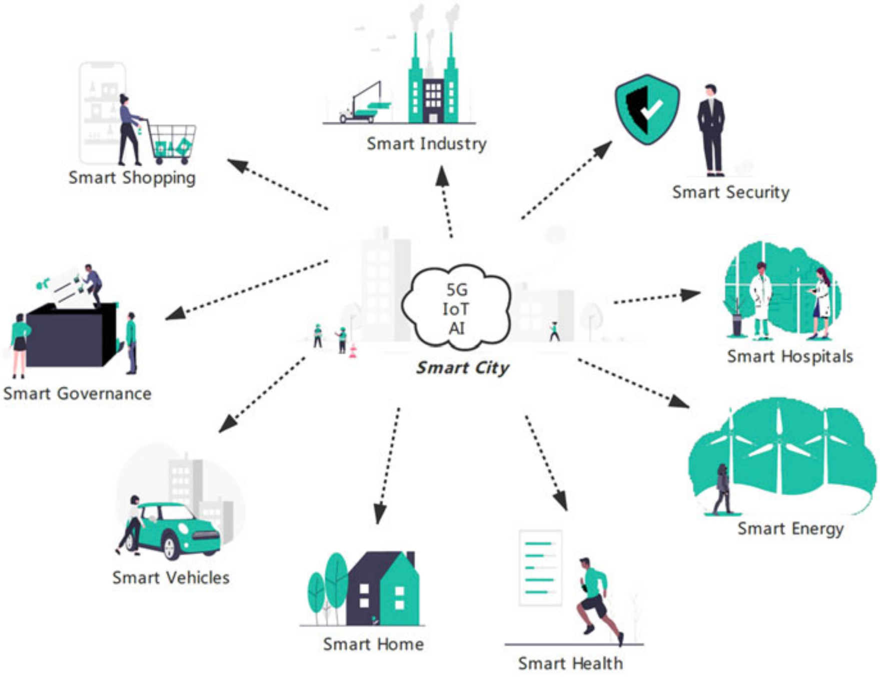

Fig. 9.1 The function of smart city

Traditional artificial intelligence applications are usually developed for specific scenarios, which always require deep customization [25]. Most of the customized solutions used in smart products can only be understood by service providers [26]. This situation usually results in unnecessary fragmentation of information and only presents a partial view of the whole urban network. Therefore, traditional artificial intelligence applications may be difficult to apply to common scenarios composed of heterogeneous objects. However, many different types of devices or objects coexist in smart city, while traditional AI cannot integrate these devices or objects as a whole. Generally speaking, the Internet of Things is defined as a collection of heterogeneous objects, which are uniquely addressable and can collect and share information through symbolic human–computer interaction [20, 27–29]. In order to promote correct and effective communication between different intelligent objects in a smart city, it is increasingly necessary to continuously collect, analyze, process, and transmit data from all devices with sensor function. Therefore, it is necessary to extract knowledge about the "things" in the IoT and assign their service functions so as to better promote the cooperation among research and development (ICT), service industry (provider/consumer), and government (policy developers) [30]. Since the expertise and requirements of stakeholders are different, smart products must provide users with an intelligent interface to browse and use the stored data and knowledge. Smart services require strong data processing capabilities where

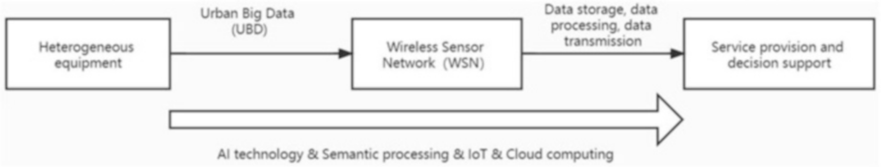
Heterogeneous equipment
Urban Big Data (UBD)
Wireless Sensor Network (WSN)
Data storage, data processing, data transmission
Service provision and decision support
AI technology & Semantic processing & IoT & Cloud computing

Fig. 9.2 The workflow of smart city

data will be calculated and analyzed in the context awareness. In this way, the system becomes intelligent by being aware of its environment and can change its service with the different environments. Therefore, an autonomous and compatible model should be developed for smart cities, which should make different fields interact with each other, thereby improving urban life. The objects in smart city may follow different standards and patterns when accessing the Internet through different devices. As a result, in order to ensure that artificial intelligence can effectively solve the heterogeneity problem of service, it is essential to build a platform that maps the different behaviors and patterns of objects to a unified model which can support the fusion of heterogeneous objects and systems, ultimately benefiting the system's intelligence and interoperability. At the same time, in order to achieve the opening and fusion of urban data, the open protocols for IoT devices should be supported to achieve seamless and continuous communication between them, while data integrity and security should be protected during transmission and processing [31].

In conclusion, a large number of heterogeneous devices are deployed in smart cities, which results in urban big data (UBD). Our goal is to standardize the data. The heterogeneous devices and data will be connected to a wireless sensor network. By the current advanced artificial intelligence technology, the data of the urban network can be stored, processed, transmitted, and finally be analyzed to obtain service provision and decision support. The operation flowchart of the entire smart city is shown in Fig. 9.2.

9.2.3.3 Smart City Related Technologies and Applications

The current improvement ideas and related technology about smart city are discussed as follows. In the development of smart cities driven by the Internet of Things, the connection between heterogeneous devices and the support of smart urban big data databases should be taken into account to promote data based and context-aware computing. Wenge et al. [32] proposed a hierarchical smart city architecture containing urban network data stream, which consists of event-driven application, domain service, support service, data storage and vitalization, data transmitting, data acquisition. Nandury et al. [33] proposed a smart city architecture concept consisting of a data collection layer, a data processing layer, and a query processing layer to maintain communication between heterogeneous devices continuous. Rathore et al. [34] proposed a four-level smart city architecture, which consists of multiple heterogeneous devices and relay nodes. Kun Guo [35] proposed an artificial intelligence-based Semantic Internet of Things (AI-SIoT) hybrid service architecture to integrate heterogeneous devices in IoT to provide related smart services. This architecture enables flexible connections between heterogeneous devices through AI-based semantics technology. In addition, the service delivery efficiency of smart city is closely related to the performance of data collection, data management, data processing, and decision-making. The expansion of WSNs has increased the flow of urban data, resulting in urban big data (UBD). With the rapid data growth, traditional data processing mechanisms can no longer meet real-time data processing needs. And in smart city applications, because resources may be requested by multiple devices at different locations at the same time, there is a great demand for low-latency data transmission. Real-time big data processing is important in designing and managing cities, thereby promote the transformation of smart cities. Analysis of big data enables the government to make key decisions in a scientific way. Therefore, it is necessary to introduce advanced technology to collect big data, manage big data, and complete the analysis of big data. Alaa Alsaig [36] proposed a data-centric IoT pattern which conceptualized things from a service-oriented perspective, and discussed effective methods for identifying, integrating, and managing big data. The data-centric approach is designed to benefit the management of the IoT big data with complexity features. Jara et al. [37] proposed a smart city architecture embedded in big data analysis to handle massive data analysis needs. The smart city with big data analysis is designed to provide services from two perspectives [38]. First, the application of urban big data plays a key role in smart city design and management. Second, smart city requires big data analysis to achieve and manage a large number of urban planning and services. In addition, many researches paid attention to the development of cloud-based smart cities [39–41]. For example, SmartCityWare is a cloud-based service middleware platform that aims to promote the development of smart city and improve municipal service quality [42]. SmartCityWare provides a virtual environment for deploying smart city applications, where the various components and services could be integrated and cooperate seamlessly and effectively. In order to achieve this function, services are distributed in various clouds components. José Santos [43] proposed a fog computing framework to support autonomous management in 5 g smart cities. Following the guidelines of the European Telecommunications Standards Institute (ETSI) NFV MANO architecture, the framework is extended with additional software components, achieving a significant reduction in network latency. In Tang et. al's research [44], another four-layer structure based on fog calculation was proposed. The architecture continuously monitors and identifies the public infrastructure and its changes. The framework selects low-cost and non-invasive sensors to collect big data from the bottom layer. Shalli Rani [21] proposed a new method to improve the energy-saving of micro sensor nodes and reduce the delay of big data collection under IoT framework. In order to realize smart city scenarios, an efficient and energy-saving wireless sensor network providing well service is required. Therefore, in this research, a new protocol namely QoS-IoT (quality of service enabled IoT) is also proposed. In addition, the city data analysis platform (CiDAP) developed on Hadoop was proposed to improve the design and use of smart city services [45]. Through testing, the platform improves the throughput of the city network. Similarly, SCOPE [46] and FIWARE [47] are two commercial products that aim to maximize the benefits of big data analysis. In particular, SCOPE is an open ecosystem platform based on cloud service with smart city. FIWARE is a framework that promotes the future development of intelligent Internet applications. Furthermore, IBM [48] and AGT [49] have also developed some other data platforms for smart cities.

9.3 From AI for XAI: Challenges and Opportunities

The popularization of machine learning (ML) has led to more and more artificial intelligence (AI) applications. Through the construction of big data databases and the increase in computing ability, key technologies have been invented and provided for the fields of science and industry. A key function of an artificial intelligence system is the ability to explain the process and reason of the decisions or predictions it makes. However, unfortunately, many of the advanced ML models are "black boxes" because they cannot explain the reasons for their decision-making due to their nested non-linear structure. These systems cannot provide enough reasons for their autonomous decisions to human users as their pattern is to process and analyze the data without knowing the domain. The complexity of black box models makes them seem opaque to human reasoning. For some artificial intelligence applications, explanation may not be a necessary step. However, recently, interpretability has become extremely important for the applications of artificial intelligence models especially in certain fields, such as defense, healthcare, finance, and law, etc., where interpretation is essential for users to understand and manage these artificial intelligence tools [50]. Therefore, experts and practitioners in some fields tend not to accept the black box AI model in practice, as they need sufficient evidence to reason and verify the output of the model before making a final decision. Exploring how the input of an artificial intelligence model affects the output will help us better understand the behavior and logic of the model and contribute to building a more robust and interpretable artificial intelligence model [51]. For example, in the clinical scenario, patients and doctors expect to acquire the explanation for the actions, especially high-risk decision-making, including medical diagnosis and therapeutic regimen [52]. In this situation, the output of the model should promote professional decision-making and improve expert clinical experience.

The recent success and popularization of artificial intelligence are largely due to the new ML technology, including support vector machines (SVM), random forests, and deep learning (DL) neural networks, etc. Although these models may perform well in certain tasks, they are opaque from the perspective of interpretability. Actually, there may be inherent conflicts between ML performance (such as prediction accuracy) and interpretability. Generally, the best-performing method (e.g., deep learning) is the least easy to interpret and does not apply to mathematical analysis, while the method with the easiest logic (e.g., decision tree) is the least accurate.

Transparency is generally considered to be the key to the effective deployment and applications of intelligent systems in the practice. Therefore, explained AI (XAI) has gradually attracted attention. XAI refers to the AI models where the output can be understood by humans, which is the opposite of the pattern of black box. The XAI model provides an in-depth understanding of the reasons for model output. In addition, interpretability also can help identify unknown vulnerabilities and flaws in AI systems, thereby enabling developers to correct errors timely [51]. The purpose of an explainable artificial intelligence (XAI) system is to make it easier for humans to understand its mechanisms and behavior by providing explanations. The current two main research branches of XAI domain are interpretable models, and prediction interpretation and justification. For the former, its goal is to make the operation of the system understandable by humans, while for the latter, it can explain why a decision is best, but it cannot guarantee to explain how the decision was made [53]. In order to enable AI system end users to understand and cooperate with the system, artificial intelligence researchers have produced many user-centric XAI toolkits with visualized interfaces (e.g. Pytorch Captum and TensorFlow tf-explain), to meet the demands of users in different disciplines with even insufficient artificial intelligence literacy.

In some circumstances, explanation is important for the acceptance and satisfaction of users. It has proved that explanations have significantly increased the trust of users for systems [54]. In a study, doctors regarded the ability to explain decisions as to the most ideal and needed feature of a decision-making assistance system [55, 56]. At the same time, the application of interpretability evaluation criteria (such as reliability, causality, and usability) helps to track the use of algorithms, so as to figure out how to improve the algorithms and provide guidance for further development [57–59]. Some methods have been proposed to measure the effectiveness of interpretation by now. Some of views are from the perspective of users' satisfaction and other subjective feeling, while other measurement may focus more on task performance, improvement of decision-making, and other objective criterion [60]. Each explanation is produced in a context that is related to the task, ability, and AI system user expectations. Therefore, the definition of interpretability largely depends on specific domain [60]. Meanwhile, an effective explanation will take into account the demands of different target user groups. For example, for expert users and lay users, they may have total different level of AI literacy, professional knowledge, and concerns for their tasks [61]. Finally, the workflow of XAI is shown in Fig. 9.3.

At present, there have been some researches focusing on XAI in the urban field for clinical diagnosis and treatment prediction [62], drug development, urban building safety management [63], aerial image rendering, autonomous driving [64], etc. In order to create an effective and more understandable AI system, the development of XAI should follow the following principles: The XAI system needs to explain its capabilities and understanding of the task; the XAI system needs to

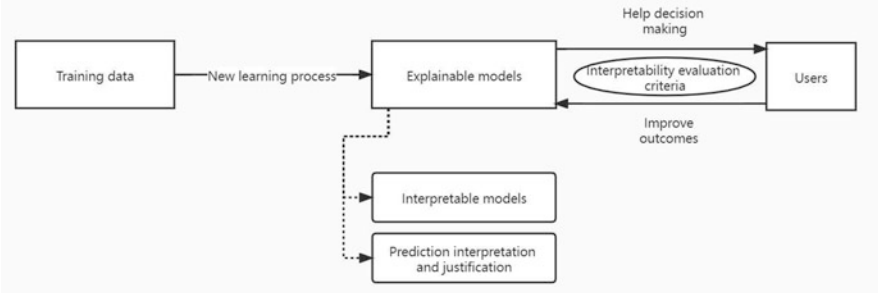

Fig. 9.3 XAI workflow

explain the procedures it has completed, the procedures it is going through, and the subsequent procedures; the XAI system needs to provide a basis for making certain decisions or predictions [65].

There are still many challenges in the XAI field. For instance, (1) how to formulate effective explanations according to the cognitive level of different users. Because users of the XAI system have different knowledge backgrounds and task requirements, as well as different cognitive abilities and operation skills of the AI system, XAI should focus on the process of human–computer interaction, so that users can obtain the most suitable explanations in the exchange and improve their ability to learn XAI. (2) The current technology on interpretability still has limitations. Because certain contradictions exist between interpretability and model accuracy, how to achieve a balance between interpretability and model accuracy is an important issue. (3) How to establish an evaluation model for the interpretability of the XAI system. Since interpretability is still a vague concept, it is necessary to measure the level of interpretability to evaluate user satisfaction, improve decision-making quality, and improve system efficiency. (4) In terms of time cost, since each interpretation needs to be generated separately, it will be a very time-consuming task to generate a systematic interpretation database.

Due to the wider range of application areas for current models, they are being more complex and difficult to interpret than ever before. They are used for tasks in different disciplines and are more common in daily life and they are increasingly allowed to make more autonomous decisions. Therefore, the importance of explaining these models is critical. In the future, XAIs will not only include independent interpretation, but also can coordinate with the AI systems in other disciplines to fuse and integrate knowledge across interdisciplinary backgrounds, so as to promote the development of cross-disciplines. There are still many challenges and opportunities in the future of XAI. Once we achieve a breakthrough in the development of XAI, the entire society and industry will undergo tremendous changes, and smart cities will also make great progress on this basis.

9.4 Smart City and Prevention of Virus Infection

In order to prevent epidemics and its adverse effect, smart city is playing an important role. In this process, the contribution of artificial intelligence (AI), big data, the Internet of Things, and other emerging technology support such as 5G is essential. As the data generated in cities around the world is encouraged to be shared among laboratories, the technology tools and treatments can be designed and invented by working together. Subsequently, smart cities can take advantage of these tools and data, promote the cooperation in various fields, thereby providing better platform for health works with latest information on infection tracking, patient diagnosis, and information consultation [66]. For example, a wealth of technical products were developed based on smart cities [67, 68], which can help early detection of epidemics. The data sets from various technical products and devices can enrich the health database, provide more accurate and real-time information about the epidemic, thereby helping to provide more effective risk management decision-making for cities [69]. City managers and decision-makers need to control the epidemic and take effective actions without harming the society and economy.

9.4.1 Virus Spread Process and Corresponding Responses

In general, during the period of the emergence and spread of a virus, several processes would be gone through as follows: (1) The first is the dissemination stage of the information. At this time, people do not have enough knowledge and vigilance, which is also a large-scale ambush period for the virus. During this time, the misinformation could cause people to make wrong responses to the virus. Based on this, the government should release the exact message as soon as possible. The task of this stage is to achieve early detection and early reporting. (2) The establishment and implementation of epidemic prevention policies is an important aspect during the process of virus control, including medical resources allocation, isolation policies, treatment policies, traffic restrictions, and commercial business plans, etc. This stage is mainly for early diagnosis, early isolation, and early treatment. (3) The third stage focus on tracking, consolidation, and prevention, including tracking of personal conditions (e.g. health codes), vaccine production and arrangements, and education for public awareness. In this process, with the cooperation of public health personnel, epidemiologists, scientists, clinicians, and other professionals, combined with the rich technical network under the smart city, the best epidemic prevention effect could be achieved. In conclusion, the response steps to virus can be summarized as in Fig. 9.4.

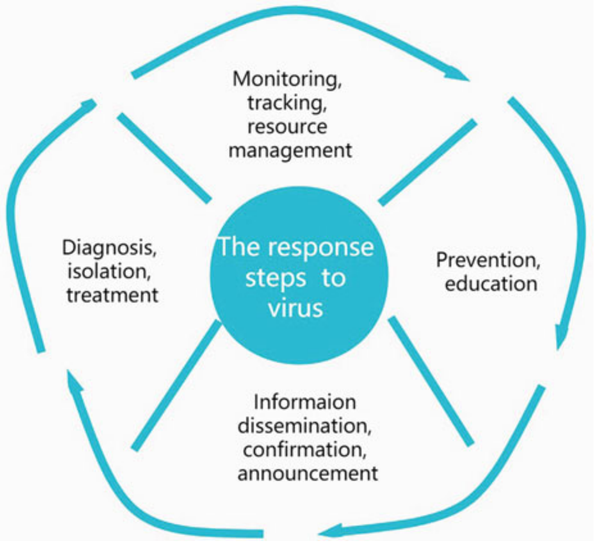
Monitoring, tracking, resource management
Prevention, education
Informaion dissemination, confirmation, announcement
Diagnosis, isolation, treatment
The response steps to virus

Fig. 9.4 The response steps to virus

9.4.2 *Smart Applications for Virus Prevention*

The development of AI computing methods has led to a paradigm shift in research methods related to infectious diseases [70] [71, 72]. Due to the development of high-performance algorithms (deep learning algorithms, machine learning, and neural networks) and cloud computing, the application of intelligent data analysis has become more and more widespread, which enables researchers to collect and process large amounts of data and acquire analysis results. In order to control the impact of the virus pandemic, a large number of technical methods and products were applied. Among them, the Internet of Things, artificial intelligence as emerging technology, and the cutting-edge media transmission networks such as 5G are at the forefront [73, 74].

Complex and large amounts of data could be collected and processed using ML-based technology. These techniques have been widely applied to predict epidemic patterns. For example, Bullock et al. [75] divided the applications of artificial intelligence into three levels. Artificial intelligence models can be applied to disease diagnosis and treatment by analyzing proteins (molecular level), analyze patient data such as diagnosis images, personal health records to improve patient care quality (clinical level), as well as analyze current cases and online information (such as social media) to predict disease development (social level). In Harrus and Wyndham's research, artificial intelligence applications were divided into five categories, including prediction, diagnosis, containment and monitoring, drug development and treatment, and social and medical management [76]. In addition, the intelligent layout of smart cities with efficient information processing capacity and logistics chain decreases the speed of virus spreading, make the allocation of resources faster, such as delivering medicines and vaccines by point-to-point, thereby reducing the burden of preventing and controlling the epidemic. Recent development in the field of artificial intelligence has greatly contributed to the screening, diagnosis, and prediction of virus. These models and applications can be used on a large scale, provide the feedback in time, and perform much better than humans in certain tasks. (Sipior, 2020; Beck et al., 2020; Pant et al., 2020) [77–79]. There is already a lot of research on how machine learning plays a role in the epidemic such as COVID-19.

Under the structure of smart city, the process of dealing with the spread of the virus can be summarized into the following paradigm, as shown in Fig. 9.5.

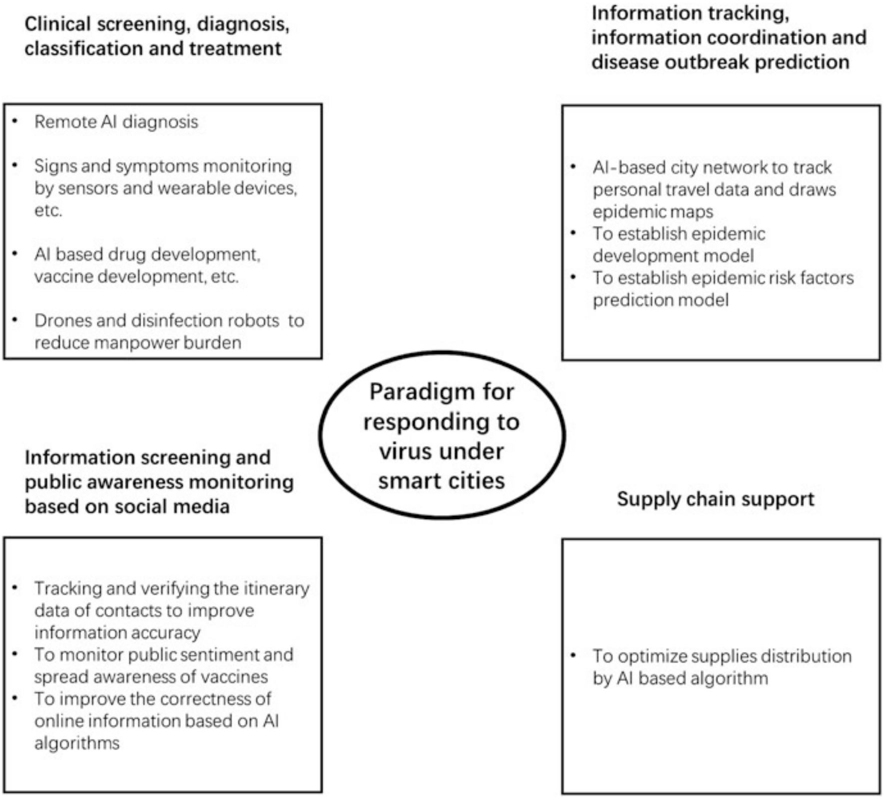
Clinical screening, diagnosis, classification and treatment
• Remote AI diagnosis
• Signs and symptoms monitoring by sensors and wearable devices, etc.
• AI based drug development, vaccine development, etc.
• Drones and disinfection robots to reduce manpower burden
Information tracking, information coordination and disease outbreak prediction
• AI-based city network to track personal travel data and draws epidemic maps
• To establish epidemic development model
• To establish epidemic risk factors prediction model
Paradigm for responding to virus under smart cities
Information screening and public awareness monitoring based on social media
• Tracking and verifying the itinerary data of contacts to improve information accuracy
• To monitor public sentiment and spread awareness of vaccines
• To improve the correctness of online information based on AI algorithms
Supply chain support
• To optimize supplies distribution by AI based algorithm

Fig. 9.5 Paradigm for responding to virus under smart cities

9.4.2.1 Smart Applications for Clinical Screening, Diagnosis, Classification, and Treatment

In terms of clinical screening, diagnosis, classification, and treatment, taking COVID-19 as an example, many countries have applied artificial intelligence technologies to CT scans, X-ray images, and even cough sounds after infection [80–82], thereby helping diagnosis decision. At the same time, due to the need to limit physical contact and quickly track COVID-19 positive patients, rapid, point-of-care (POC) testing for COVID-19 is found to be increasingly attractive, such as kit for detecting antibodies or antigens. Some COVID test kits can detect antibodies faster with specialized portable test equipment (such as Abbott ID NOW) [83]. These kits are able to identify contacts who may be infected, and in order to prevent them from infecting the public, appropriate isolation measures should be taken in time. At the same time, a series of innovative technologies have been invented due to the progress of mobile wireless networks, wireless sensors, and the Internet of Medical Things (IoMT). Under this trend, knowledge can be shared in real time and patient information can be kept consistent in medical system, making telehealth has become an important communication and treatment method during the COVID-19 pandemic. Through telehealth, doctors can evaluate, analyze, treat, and communicate with patients without face to face contact [74], thus improving the efficiency of clinical management and preventing the spread of the virus in the hospitals. Meanwhile, smartphone applications for COVID-19 detection have gradually emerged and been applied. For example, SANOFI, a medical firm in French, has designed a home test for COVID-19 [84]. It uses a nanoparticle that can glow in the dark, which can be detected by a smartphone's camera, and then the smartphone will process the signal by AI algorithms and automatically send results to telehealth platform within 30 min. This kind of over-the-counter (OTC) method for COVID-19 detection is not only easy to use, but also reduce the risk of infection [84]. Similarly, there are other AI test methods, such as MDBio COVID-19 test kit [85], AI Tool-Chest X-ray [86]. As for the treatment and management of patients, mobile devices can be used to collect clinical health data of patients and then provide these data to clinicians to obtain real-time monitoring of patient vital signs. At the same time, the combination of artificial intelligence and drug research can help to develop drugs against COVID-19. Some studies have used omics data to find drug candidates for the treatment of COVID-19 [87, 88]. In addition, vaccine development can also cooperate with AI technology. For instance, through combining data analysis tools from Oracle cloud computing, a vaccine against the COVID-19 pandemic was developed [89].

9.4.2.2 Smart Applications for Information Tracking, Information Coordination and Disease Outbreak Prediction

In terms of information tracking, information coordination, and disease outbreak prediction, there have been quite a few applications combined with artificial intelligence. Integrated modeling that combines different types of individual data (such as travel data, GPS tracking, individual health data, and behavior pattern data) is the key to building a successful epidemic surveillance system [90]. This comprehensive surveillance system can help detect threats from viruses and be used for epidemic surveillance in early stages. This pattern also takes the integration of mathematical models into account to estimate the spread of large-scale infectious diseases and simulate the effects of health interventions from the communities [91], cities, and national level, thereby enabling the authorities to make the effective decisions. The interoperability and information sharing pattern is the key value of these systems and tools, so the central and local health systems can communicate and synchronize information in a timely manner, thereby improving the efficiency of epidemic prevention and control [83]. Taking COVID-19 as an example, one method widely used in China for information tracking is the health code. Health code is an application implanted in Wechat (the most popular social media app in China) based on smartphone quick response (QR) code generation. With real-name information system, it collects self-reported and networked recorded health data, travel history, and contact history to assess the user's infection risk. The program uses different colors to divide individuals' risk profiles, including green (individuals are allowed to travel and work normally), amber (7-day home quarantine is required), and red (14-day medical quarantine is required [92]). Meanwhile, the information is integrated and delivered to the data platform of the local government to map the population flow and detect the contacts, so as to make a timely isolation decision with contacts and transportation [93]. For example, since January 31, 2020, the Shenzhen Municipal Government website [94] had been releasing related information on each confirmed and suspected case of COVID-19, including gender, age, travel route, date of diagnosis, and number of close contacts. After that, local joint IT companies quickly updated this information into online maps. Residents can check maps to confirm whether there are cases in their neighbor to prevent themselves from being exposed to the virus [95]. Regarding the study of epidemic outbreak prediction, for example, some researchers in China used comprehensive modeling methods to predict the infectious disease vulnerability index (IDVI) during COVID-19 with multiple indicators such as travel information, national socio-economic status, infrastructure, etc. [96–99]. Ye et al. developed an artificial intelligence program called α-Satellite based on data from social media, demographics, travel data during COVID-19. α-Satellite uses a advanced heterogeneous graphical auto-encoder (GAE) to integrate and process data from nearby communities so as to assess the risk [100]. Other prediction methods also include multi-layer perceptron (MLP), adaptive network-based fuzzy inference system (ANFIS), DL, and other ML algorithms, which are all beneficial to predict the outbreak and spread of COVID-19 in the future [101–103]. Chen et al. [104] developed a time-dependent mathematical model for predicting the total number of confirmed cases, and Hu et al. [105] developed prediction model for the spread of COVID-19 spreading period using a modified stacked auto-encoder. The results show that AI plays a key role in predicting the virus outbreak and spread. In addition, the travel history and physical signs and symptoms recorded online can be took advantage of to establish a model for predicting risk factors of virus spread based on artificial intelligence [106]. Pirouz et al. [107] used artificial intelligence algorithms to explore the correlation between environmental parameters and COVID-19, which shows a significant correlation between the city and climate parameters and the number of confirmed COVID-19 cases. Haghshenas [108] also analyzed the effects of certain environmental parameters on virus spread based on artificial intelligence, including daily average temperature, humidity, wind speed, etc., and found that urban parameters and relative humidity are the most priority variables for predicting confirmed cases of COVID-19. Zaheer Allam [109] analyzed the virus outbreak from the perspective of cities, and proposed the establishment of standardization protocols of smart city networks in order to promote data sharing and global cooperation during outbreaks of virus.

9.4.2.3 Smart Applications for Information Screening and Public Awareness Monitoring Based on Social Media

In addition to tracking the journey information of contacts and predicting disease outbreaks, artificial intelligence is also used to screen COVID-19 releted information and assess public perception for epidemic situation based on mobile networks and social media [110]. During epidemic, tracking contacts and updating the related information plays an important role in minimizing the spread of infection. Using smartphone-based GPS and social media data is one way for contact tracing and risk assessment [111]. Although this method may lead to high false positives, a solution has been proposed that using data from six different smartphone sensors to track contacts simultaneously can reduce information errors [112]. At the same time, the development of digital communications and social medias provides vast amount of real-time data for the content circulating in online communities. As a result, social network data and public sentiment analysis are important tools to manage the COVID-19 pandemic [113]. However, social media platforms are also regarded as a medium for dissemination of disinformation, especially for popular social media platforms with a mass of data exposed, such as Twitter, YouTube [114, 115], Wechat, Weibo, etc. Under this trend, the phenomenon called infodemic appeared [116]. Infodemic refers to a large amount of information is provided during public health crisis, including false or misleading information, which provokes confusion and wrong behavior of public, and brings harm to authorities of public health department [117]. It may have an inverse effect on the interventions for pandemics. As a result, in order to help the authority understand the public's concerns and emotions about the epidemic, and to track misinformation spreading and information gap, a lot of initiatives and research has been conducted. Information monitoring, online social media listening, content pretesting, and other computational methods for social science are considered to be effective methods for detecting and analyzing misinformation [118] and information voids for virus. The WHO Information Network for Epidemics (EPI-WIN), in cooperation with digital research institutions, has developed a digital media data analyzing method, which detects infodemic signals to analyze and summarize the main concerns and information gaps detected in online communications [119], so as to make more effective responses and strategies to virus spread. Tina D Purnat [116] developed a taxonomy that divides online conversations and contents in English and French about COVID-19 into 5 topics 35 subtopics. Each subtopic will be analyzed in terms of quantity, speed, and emerging issues to detect misinformation or information void. In addition, artificial intelligence tools can help local governments assess the public's awareness of vaccination and contribute to spreading vaccination awareness to the public.

9.4.2.4 Smart Applications for Supply Chain Support

As for necessary material manufacture and distribution, such as face masks and vaccines, artificial intelligence algorithms can have a positive effect on manufacturing, storage, and logistics. E-commerce companies and logistics companies can cooperate with artificial intelligence technology companies to develop personal protective equipment (PPE) intelligent supply chain management systems by using big data and advanced algorithms, which can achieve the automatic supply match and distribution. Through Software as a Service (SaaS) platform, suppliers of personal protective equipment, suppliers of necessary materials, parts and manufacturing equipment will be connected together to promote the efficiency of the production of personal protective equipment [92].

9.5 Summary and Perspectives

With the development of urbanization, artificial intelligence, Internet of Things, and communication technologies, the concept and applications of smart cities have emerged. A smart city includes various aspects, from people's daily shopping, transportation, power supply, to government decision-making, mobile health, medical prevention and control, etc. From the perspective of the overall architecture of smart cities, it is necessary to strengthen the popularization of basic equipment, upgrade the technical architecture of smart cities, and strengthen the importance of the Internet of Things in various fields. At Smart city managers should expand the use of smartphones, sensors, wearable devices, drones, robots, etc., and expand the scope of data collection. The successful realization of a smart city depends on the efficient transmission and management of the urban big data. Therefore, in the process of generating data and transmitting the data to the data server or base station, it is necessary to combine new communication technologies and computing methods (such as 5G, cloud computing, fog computing, etc.) to further meet the stringent requirements in IoT scenarios, such as low latency, high energy efficiency, and high mobility [43]. At the same time, it is also important to improve the smart city application framework and model. Combined with wireless sensor network, the data needs to be effectively aggregated, transmitted, analyzed, classified, and managed, thereby providing the basis for related services. In addition, the future smart city network still needs improvement in transmission speed, throughput, service quality, etc. Furthermore, since artificial intelligence in smart cities involves different scenarios applied with different data sets, the context model with narrow definition may not work at the general level [120]. As a result, it is necessary to improve the generalizability of artificial intelligence algorithms, that is, the ability to execute and adjust algorithms efficiently in different contexts.

For data collection and processing, due to the integration and interoperability of all smart components and technical tools in smart city, there is mass of multi-dimensional data. At the same time, smart cities generate urban big data at a very high speed and big scale, which provides real-time and high-quality information and services for a variety of smart applications, contributing to the convenience for city people. Therefore, it is necessary to encourage data sharing and collaboration between different regions and industries. The generated data should be first collected and managed uniformly, and be processed and analyzed in real time to update future services. Meanwhile, another key point in the development of smart cities is how to deal with concurrent services with complex data and how to provide the service orderly. However, one of the main challenges for this problem is the lack of standard data sets.

A large amount of data is needed for the development of epidemic forecasting tools, however, the data could be complicated and varied. At present, various models of artificial intelligence have been proposed. However, most of them have used heterogeneous data sets. Due to the use of different samples, which model is best for detecting viruses is not clear. The data structure and data collection standards should be set to improve the reliability and accuracy of epidemic prediction. On the one hand, efficient database systems could be established to standardize data structure. On the other hand, semantic technology such as natural language processing can better process information with complexity characteristics and transform them into structured multi-dimensional data. Therefore, databases based on standardized structures and artificial intelligence analysis technologies (such as natural language semantic models) will be the future keys to smart cities. At the same time, the lack of standardization among smart city technology and application suppliers may cause inefficient communication between regions and data platforms. In the case of a virus outbreak, this may also lead to reduced production efficiency in industries, as it has an inverse effect on products' early detection and management. Therefore, there is an urgent need to achieve the standardization of smart city communication protocols and cooperation among technology suppliers, improving the data's fairness and transparency between stakeholders [109].

At the same time, most of the current artificial intelligence and machine learning algorithms are still "black boxes." As AI models are increasingly used in fields that require high interpretability, such as medical care and government decision-making. Researchers in different fields such as philosophy, psychology, cognitive science, human–computer interaction, etc. need to cooperate with artificial intelligence companies to develop more interpretable AI frameworks connected with human's interpretability demand.

As for the response to infectious diseases, smart cities based on artificial intelligence have great potential in preventing the rapid spread of infectious diseases (such as the recent COVID-19). It is necessary to integrate artificial intelligence, optimize data sets and algorithms, and develop more infectious disease diagnosis and prediction models, prevention, control and treatment paradigms. At the same time, advanced technologies such as drones, disinfection robots, and video temperature detectors can be used to monitor and report the epidemic. In addition, the basic application of smart cities in the supply chain also contributes to the medical resources allocation and virus epidemics prediction. By integrating supply chain and demand information in various regions, relevant medical supplies (such as vaccines, masks, etc.) can be accurately transported to spots in need, saving human resources to the greatest extent and improving virus defense efficiency. At the social media level, artificial intelligence semantic technologies such as natural language processing can be used to manage misleading information on the Internet, thereby improving the accuracy of Internet epidemic information and the efficiency of epidemic prevention and control.

At last, a noteworthy problem about smart cities is data privacy. In the process of using digital tools to deliver smart services, because artificial intelligence is embedded in different applications, it is important to think about how to design and implement these technologies in a reliable and fair way. At the same time, due to seamless wireless network, people need to share personal information in the cloud, which will also increase the threat to personal data privacy. Especially in medical institutions, since artificial intelligence involves multiple stakeholders, the lack of transparency in the clinical models, the privacy of patient data, and related ethical issues are the main data regulatory challenges that AI faces [121]. Meanwhile, as mentioned above, interpretability is also very important for AI models in smart cities to ensure the fairness, transparency, and accountability of AI technologies [122]. In particular, when smart city provides services, the service suppliers will analyze user demands based on users' personal data. For example, during COVID-19 pandemic, related prediction models need to use data including X-ray images, CT scans, travel history, personal health records, GPS location, and other personal information. However, if there is no formal privacy law or rules, few people will allow their data to be shared to the database or online. Therefore, related departments should create formal procedures to collect the privacy data according to the type, accessibility, and utilization of data [123]. In conclusion, how to protect personal information security in the background of information sharing is a huge challenge in the future.

Acknowledgments This work was supported by the COVID-19 Research Projects of West China Hospital Sichuan University (Grant no. HX-2019-nCoV-057), the Regional Innovation Cooperation between Sichuan and Guangxi Provinces (2020YFQ0019) and the National Natural Science Foundation of China (32070671).

References

1. Jain A, Jain A (2016) The 5 V's of big data-Watson Health perspectives
2. Pirouz B, Golmohammadi A, Masouleh HS, Delazzari C, Violini G, Pirouz B (2020) Relationship between average daily temperature and average cumulative daily rate of confirmed cases of COVID-19
3. Pirouz B, Shaffiee Haghshenas S, Pirouz B, Shaffiee Haghshenas S, Piro P (2020) Development of an assessment method for investigating the impact of climate and urban parameters in confirmed cases of COVID-19: a new challenge in sustainable development. Int J Environ Res Public Health 17(8):2801
4. Palermo SA, Zischg J, Sitzenfrei R, Rauch W, Piro P (2018) Parameter sensitivity of a microscale hydrodynamic model. International Conference on Urban Drainage Modelling; Springer
5. Ericsson L (2011) More than 50 billion connected devices. White Paper 14(1):124
6. Gupta A, Jha RK (2015) A survey of 5G network: architecture and emerging technologies. IEEE Access 3:1206–1232
7. Vaquero LM, Rodero-Merino L (2014) Finding your way in the fog: towards a comprehensive definition of fog computing. ACM SIGCOMM Comput Commun Rev 44(5):27–32
8. Niyato D, Maso M, Kim DI, Xhafa A, Zorzi M, Dutta A (2017) Practical perspectives on IoT in 5G networks: from theory to industrial challenges and business opportunities. IEEE Communic Magaz 55(2):68–69
9. Dias R, Torkamani A (2019) Artificial intelligence in clinical and genomic diagnostics. Gen Med 11(1):1–12
10. Matheny M, Israni ST, Ahmed M, Whicher D (2019) Artificial intelligence in health care: the hope, the hype, the promise, the peril. National Academy of Medicine, Washington, DC, p 154
11. Shah P, Kendall F, Khozin S, Goosen R, Hu J, Laramie J et al (2019) Artificial intelligence and machine learning in clinical development: a translational perspective. NPJ Dig Med 2(1):1–5
12. Deo R (2015) Machine learning in medicine. Circulation 132(20):1920–1930
13. Cao C, Liu F, Tan H, Song D, Shu W, Li W et al (2018) Deep learning and its applications in biomedicine. Genomics Proteomics Bioinformatics 16(1):17–32
14. Bettencourt L, West G (2010) A unified theory of urban living. Nature 467(7318):912–913
15. Jacobs J (1961) The Death and Life of Great American Cities. Randoms House, New York
16. Kondra I (2020) Use of IT in higher education. UGC Care J India 40:280
17. Siembab W (1996) Telecity development strategy for sustainable, livable communities. The blue line televillage in Compton, California. Proceedings from Urban Design, Telecommuting and Travel Forecasting Conference; September 8
18. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54(15):2787–2805
19. Brock DL (2001) The electronic product code (EPC)-A naming scheme for physical objects. White Paper
20. Jin J, Gubbi J, Marusic S, Palaniswami M (2014) An information framework for creating a smart city through internet of things. IEEE Intern Things J 1(2):112–121
21. Rani S, Chauhdary SH (2018) A novel framework and enhanced QoS big data protocol for smart city applications. Sensors 18(11):3980
22. Silva BN, Khan M, Han K (2018) Towards sustainable smart cities: a review of trends, architectures, components, and open challenges in smart cities. Sustain Cit Soc 38:697–713
23. Alvi AN, Bouk SH, Ahmed SH, Yaqub MA, Sarkar M, Song H (2016) BEST-MAC: bitmap-assisted efficient and scalable TDMA-based WSN MAC protocol for smart cities. IEEE Access 4:312–322
24. Gil D, Ferrández A, Mora-Mora H, Peral J (2016) Internet of things: a review of surveys based on context aware intelligent services. Sensors 16(7):1069

25. Cosma G, Brown D, Archer M, Khan M, Pockley AG (2017) A survey on computational intelligence approaches for predictive modeling in prostate cancer. Exp Syst Applic 70:1–19
26. Vermesan O, Friess P (2014) Internet of things-from research and innovation to market deployment. River Publishers, Aalborg
27. Silva BN, Khan M, Han K (2018) Internet of things: a comprehensive review of enabling technologies, architecture, and challenges. IETE Techn Rev 35(2):205–220
28. Rani S, Talwar R, Malhotra J, Ahmed SH, Sarkar M, Song H (2015) A novel scheme for an energy efficient Internet of Things based on wireless sensor networks. Sensors 15(11):28603–28626
29. Jung C, Kim K, Seo J, Silva BN, Han K (2017) Topology configuration and multihop routing protocol for bluetooth low energy networks. IEEE Access 5:9587–9598
30. Washburn D, Sindhu U, Balaouras S, Dines RA, Hayes N, Nelson LEJG (2009) Helping CIOs understand "smart city". Initiatives 17(2):1–17
31. Weber M, Podnar Žarko I (2019) A regulatory view on smart city services. Sensors 19(2):415
32. Wenge R, Zhang X, Dave C, Chao L, Hao S (2014) Smart city architecture: a technology guide for implementation and design challenges. China Communic 11(3):56–69
33. Nandury SV, Begum BA (2015) Smart WSN-based ubiquitous architecture for smart cities. 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI); IEEE
34. Rathore MM, Ahmad A, Paul A, Rho S (2016) Urban planning and building smart cities based on the internet of things using big data analytics. Comp Netw 101:63–80
35. Guo K, Lu Y, Gao H, Cao R (2018) Artificial intelligence-based semantic internet of things in a user-centric smart city. Sensors 18(5):1341
36. Alsaig A, Alagar V, Chammaa Z, Shiri N (2019) Characterization and efficient management of big data in iot-driven smart city development. Sensors 19(11):2430
37. Jara AJ, Genoud D, Bocchi Y (2014). Big data in smart cities: from poisson to human dynamics. 2014 28th International Conference on Advanced Information Networking and Applications Workshops; IEEE
38. Mora-Mora H, Gilart-Iglesias V, Gil D, Sirvent-Llamas AJS (2015) A computational architecture based on RFID sensors for traceability in smart cities 15(6):13591–13626
39. Suciu G, Vulpe A, Halunga S, Fratu O, Todoran G, Suciu V (2013) Smart cities built on resilient cloud computing and secure internet of things. 2013 19th international conference on control systems and computer science; IEEE
40. Talari S, Shafie-Khah M, Siano P, Loia V, Tommasetti A, Catalão J (2017) A review of smart cities based on the internet of things concept. Energies 10(4):421
41. Ng ST, Xu FJ, Yang Y, Lu M (2017) A master data management solution to unlock the value of big infrastructure data for smart, sustainable and resilient city planning. Procedia Eng 196:939–947
42. Mohamed N, Al-Jaroodi J, Jawhar I, Lazarova-Molnar S, Mahmoud S (2017) SmartCityWare: a service-oriented middleware for cloud and fog enabled smart city services. IEEE Access 5:17576–17588
43. Santos J, Wauters T, Volckaert B, De Turck FJE (2018) Fog computing: Enabling the management and orchestration of smart city applications in 5G networks. Entropy 20(1):4
44. Tang B, Chen Z, Hefferman G, Wei T, He H, Yang Q (2015) A hierarchical distributed fog computing architecture for big data analysis in smart cities. Proc ASE BigData SocialInform 2015:1–6
45. Cheng B, Longo S, Cirillo F, Bauer M, Kovacs E (2015) Building a big data platform for smart cities: Experience and lessons from santander. 2015 IEEE International Congress on Big Data; IEEE
46. Bestavros A, Hutyra L, Terzi E. SCOPE: Smart-city cloud based open platform and ecosystem. Boston University: Boston, MA 2016

47. FIWARE. FIWARE Consolidates as Open Source IoT-enabled Smart Services Platform of Reference With Launch of FIWARE Foundation. https://www.fiware.org/news/fiware-consolidates-as-open-source-iot-enabled-smart-services- platform-of-reference-with-launch-of-fiware-foundation/FIWARE2016; https://www.fiware.org/news/fiware-consolidates-as-open-source-iot-enabled-smart-services-platform-of-reference-with-launch-of-fiware-foundation/
48. Naccarati F, Hobson S (2011) IBM Smarter City Solutions on Cloud. IBM, Somers Ny
49. Strohbach M, Ziekow H, Gazis V, Akiva N (2015) Towards a big data analytics framework for IoT and smart city applications. Modeling and processing for next-generation big-data technologies. Springer, Cham, pp 257–282
50. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R (2019) Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature
51. Adadi A, Berrada MJ (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160
52. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv JL Tech 31:841
53. Biran O, Cotton C (2017) Explanation and justification in machine learning: a survey. IJCAI-17 workshop on explainable AI (XAI)
54. Symeonidis P, Nanopoulos A, Manolopoulos Y (2009) MoviExplain: a recommender system with explanations. Proceedings of the third ACM conference on Recommender systems
55. Teach RL, Shortliffe EH (1981) An analysis of physician attitudes regarding computer-based clinical consultation systems. Comput Biomed Res 14(6):542–558
56. Ye LR, Johnson PE (1995) The impact of explanation facilities on user acceptance of expert systems advice. Mis Quart:157–172
57. Doshi-Velez F, Kim BJ (2017) Towards a rigorous science of interpretable machine learning
58. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A (2019) What clinicians want: contextualizing explainable machine learning for clinical end use. Machine learning for healthcare conference; PMLR
59. Herlocker JL, Konstan JA, Riedl J, (2000). Explaining collaborative filtering recommendations. Proceedings of the 2000 ACM conference on Computer supported cooperative work
60. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z (2019) XAI—Explainable artificial intelligence. Sci Robot 4(37)
61. Kulesza T, Burnett M, Wong W-K, Stumpf S (2015) Principles of explanatory debugging to personalize interactive machine learning. Proceedings of the 20th international conference on intelligent user interfaces
62. Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Transactions on Neural Networks and Learning Systems.
63. Matin SS, Pradhan B (2021) Earthquake-induced building-damage mapping using Explainable AI (XAI). Sensors 21(13):4489
64. Abdollahi A, Pradhan BS (2021) Urban vegetation mapping from aerial imagery using Explainable AI (XAI). Sensors 21(14):4738
65. Bellotti V, Edwards K (2001) Intelligibility and accountability: human considerations in context-aware systems. Hum Comp Interac 16(2–4):193–212
66. Siriwardhana Y, Gür G, Ylianttila M, Liyanage M (2020) The role of 5G for digital healthcare against COVID-19 pandemic: opportunities and challenges. ICT Express
67. Allam Z (2020) On culture, technology and global cities, Cities and the Digital Revolution. Springer, pp 107–124
68. Allam Z (2020) Data as the new driving gears of urbanization, Cities and the Digital Revolution. Springer, pp 1–29
69. Boulos MNK, Peng G, VoPham T (2019) An overview of GeoAI applications in health and healthcare. Int J Health Geogr 18(1):1–9
70. Arora G, Misra R, Sajid A (2017) Model systems for pulmonary infectious diseases: paradigms of anthrax and tuberculosis. Curr Top Med Chem 17(18):2077–2099

71. Dash S, Shakyawar SK, Sharma M, Kaushik S (2019) Big data in healthcare: management, analysis and future prospects. J Big Data 6(1):1–25
72. Emmert-Streib FJML, Extraction K (2021) From the digital data revolution toward a digital society: pervasiveness of artificial intelligence. Mach Learn Knowle Extract 3(1):284–298
73. Ting DSW, Carin L, Dzau V, Wong TY (2020) Digital technology and COVID-19. Nat Med 26(4):459–461
74. Rodríguez-Rodríguez I, Zamora-Izquierdo M-Á, Rodríguez J-V (2018) Towards an ICT-based platform for type 1 diabetes mellitus management. Appl Sci 8(4):511
75. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the landscape of artificial intelligence applications against COVID-19. J Artif Intellig Res 69:807–845
76. Harrus I, Wyndham J (2021) Artificial intelligence and COVID-19: applications and impact assessment. In AAAS AI Report. https://www.aaas.org/sites/default/files/2021-05/AIandCOVID19_2021_FINAL.pdf
77. Sipior JC (2020) Considerations for development and use of AI in response to COVID-19. Int J Inform Manag 55:102170
78. Beck BR, Shin B, Choi Y, Park S, Kang K (2020) Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Computat Struct Biotechnol J 18:784–790
79. Pant S, Singh M, Ravichandiran V, Murty U, Srivastava HKJ (2020) Peptide-like and small-molecule inhibitors against Covid-19. J Biomol Struct Dyn
80. Abbasi WA, Abbas SA, Andleeb S (2020) COVIDX: Computer-aided diagnosis of Covid-19 and its severity prediction with raw digital chest X-ray images
81. Garg T, Garg M, Mahela OP, Garg AR (2020) Convolutional neural networks with transfer learning for recognition of COVID-19: a comparative study of different approaches. AI 1(4):586–606
82. Jain R, Gupta M, Taneja S, Hemanth DJ (2021) Deep learning based detection and analysis of COVID-19 on chest X-ray images. Appl Intellig 51(3):1690–1700
83. Dabla PK, Gruson D, Gouget B, Bernardini S, Homsak E (2021) Lessons learned from the COVID-19 pandemic: emphasizing the emerging role and perspectives from artificial intelligence, mobile health, and digital laboratory medicine. Ejifcc 32(2):224
84. Whelan K (2021) Covid-19: Smartphone-Based tests to do at home. http://emag.medicalexpo.com/covid-19-smart-phone-based-tests-to-do-at-home/. Accessed 28 Sept 2021
85. Mobile Detect-Bio BCC19 Coronavirus Test, COVID-19 Smartphone Testing Kit (2020). https://www.fda.gov/media/141791/download. Assessed 28 Sept 2021
86. HandMed Handheld X-ray Camera (2020). Assessed 28 Sept 2021
87. Barh D, Tiwari S, Weener ME, Azevedo V, Góes-Neto A, Gromiha MM et al (2020) Multi-omics-based identification of SARS-CoV-2 infection biology and candidate drugs against COVID-19. Comput Biol Med 126:104051
88. Ciliberto G, Cardone L (2020) Boosting the arsenal against COVID-19 through computational drug repurposing. Drug Discov Today 25(6):946
89. Khan M, Mehran MT, Haq ZU, Ullah Z, Naqvi SR (2021) Applications of artificial intelligence in COVID-19 pandemic: a comprehensive review. Exp Syst Appl 185:115695
90. Zeng D, Cao Z, Neill DB (2021) Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control. Artif Intellig Med:437–453
91. Siettos CI, Russo LJV (2013) Mathematical modeling of infectious disease dynamics. Virulence 4(4):295–306
92. Chen M, Xu S, Husain L, Galea G (2021) Digital health interventions for COVID-19 in China: a retrospective analysis. Intelligent Med
93. Jao N, Jao D, Udemans C (2020) How China is using QR code apps to contain Covid-19. https://technode.com/2020/02/25/how-china-is-using-qr-code-apps-to-contain-covid-19/
94. Zhu D, Ye X, Manson S (2021) Revealing the spatial shifting pattern of COVID-19 pandemic in the United States. Sci Rep 11(1):8396

95. Zou H, Shu Y, Feng T (2020) How Shenzhen, China avoided widespread community transmission: a potential model for successful prevention and control of COVID-19. Infect Dis Poverty 9(1):1–4
96. Malik YS, Sircar S, Bhat S, Ansari MI, Pande T, Kumar P et al (2021) How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future. Rev Med Virol 31(5):1–11
97. Bogoch II, Watts A, Thomas-Bachli A, Huber C, Kraemer MU, Khan K (2020) Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel. J Travel Med 27(2):taaa008
98. Gilbert M, Pullano G, Pinotti F, Valdano E, Poletto C, Boëlle P-Y et al (2020) Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. Lancet 395(10227):871–877
99. Cartaxo ANS, Barbosa FIC, de Souza Bermejo PH, Moreira MF, Prata DN (2021) The exposure risk to COVID-19 in most affected countries: A vulnerability assessment model. PLoS One 16(3):e0248075
100. Ye Y, Hou S, Fan Y, Qian Y, Zhang Y, Sun S, et al. (2020) $\alpha $-Satellite: an AI-driven System and Benchmark Datasets for Hierarchical Community-level Risk Assessment to Help Combat COVID-19
101. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U et al (2020) Covid-19 outbreak prediction with machine learning. Algorithms 13(10):249
102. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R (2020) COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics 8(6):890
103. Punn NS, Sonbhadra SK, Agarwal S (2020) COVID-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv
104. Chen Y, Lu P, Chang C (2020) A time-dependent SIR model for COVID-19
105. Hu Z, Ge Q, Li S, Jin L, Xiong M (2020) Artificial intelligence forecasting of covid-19 in China
106. Rao ASS, Vazquez JA (2020) Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine. Infect Control Hosp Epidemiol 41(7):826–830
107. Pirouz B, Shaffiee Haghshenas S, Shaffiee Haghshenas S, Piro P (2020) Investigating a serious challenge in the sustainable development process: analysis of confirmed cases of COVID-19 (new type of coronavirus) through a binary classification using artificial intelligence and regression analysis. Sustainability 12(6):2427
108. Shaffiee Haghshenas S, Pirouz B, Shaffiee Haghshenas S, Pirouz B, Piro P, Na K-S et al (2020) Prioritizing and analyzing the role of climate and urban parameters in the confirmed cases of COVID-19 based on artificial intelligence applications. Int J Environ Res Public Health 17(10):3730
109. Allam Z, Jones DS (2020) On the coronavirus (COVID-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (AI) to benefit urban health monitoring and management. Healthcare; Multidisciplinary Digital Publishing Institute
110. Cresswell K, Tahir A, Sheikh Z, Hussain Z, Hernández AD, Harrison E et al (2021) Understanding public perceptions of COVID-19 contact tracing apps: artificial intelligence–enabled social media. Analysis 23(5):e26618
111. Wang S, Ding S, Xiong L (2020) A new system for surveillance and digital contact tracing for COVID-19: spatiotemporal reporting over network and GPS. JMIR mHealth uHealth 8(6):e19457
112. Nguyen KA, Luo Z, Watkins C (2020) Epidemic contact tracing with smartphone sensors. J Locat Based Serv 14(2):92–128
113. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S et al (2020) Social network analysis of COVID-19 sentiments: application of artificial intelligence. J Med Intern Res 22(8):e22590

114. Medford RJ, Saleh SN, Sumarsono A, Perl TM, Lehmann CU (eds) (2020) An “infodemic”: leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. Open Forum Infectious Diseases. Oxford University Press US
115. Tsao S-F, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA (2021) What social media told us in the time of COVID-19: a scoping review. Lancet Dig Health
116. Purnat TD, Vacca P, Czerniak C, Ball S, Burzo S, Zecchin T et al (2021) Infodemic signal detection during the COVID-19 pandemic: development of a methodology for identifying potential information voids in online conversations. JMIR Infodemiol 1(1):e30971
117. Tangcharoensathien V, Calleja N, Nguyen T, Purnat T, D’Agostino M, Garcia-Saiso S et al (2020) Framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. J Med Internet Res 22(6):e19659
118. Purnat T, Wilhelm EJL (2020) Building systems for respond to infodemics and build resilience to misinformation
119. Coronavirus disease 2019 (COVID-19) Situation Report 100. 29 April 2020. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
120. Wang F, Preininger AJ (2019) AI in health: state of the art, challenges, and future directions. Yearbook Med Inform 28(01):016–026
121. Kuziemski M, Misuraca GJ (2020) AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. Telecommunic Policy 44(6):101976
122. Goodman K, Zandi D, Reis A, Vayena EJ (2020) Balancing risks and benefits of artificial intelligence in the health sector. Bull World Health Organ 98(4):230
123. Berman G, Carter K, Herranz MG, Sekara V (2020) Digital contact tracing and surveillance during COVID-19. General and child-specific ethical issues. https://www.unicef-irc.org/publications/pdf/. Accessed 28 Sept 2021

Index

B. Shen (ed.), *Translational Informatics*, Advances in Experimental Medicine and Biology 1368, https://doi.org/10.1007/978-981-16-8969-7