Systems Biology

SARS2020: An integrated platform for identification of novel coronavirus by a consensus sequence-function model

Dachuan Zhang[1, $], Tong Zhang[1, $], Sheng Liu[1], Dandan Sun[1], Shaozhen Ding[1], Xingxiang Cheng[1], Pengli Cai[1, 2], Ailin Ren[2], Mengying Han[1], Dongliang Liu[1], Cancan Jia[1], Linlin Gong[1], Rui Zhang[1], Huadong Xing[1], Weizhong Tu[3], Junni Chen[3] & Qian-Nan Hu[1]*

[1] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. [2] Tianjin Institute of Industrial Biotechnology, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Tianjin, China. [3] Wuhan LifeSynther Science and Technology Co. Limited, Wuhan, China.

[$]These authors contributed equally to the work.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The 2019 novel coronavirus outbreak has significantly affected global health and society. Thus, predicting biological function from pathogen sequence is crucial and urgently needed. However, little work has been performed to identify viruses by the enzymes that they encode, and which are key to pathogen propagation.

Results: We built a comprehensive scientific resource, SARS2020, that integrates coronavirus-related research, genomic sequences, and results of anti-viral drug trials. In addition, we built a consensus sequence-catalytic function model from which we identified the novel coronavirus as encoding the same proteinase as the Severe Acute Respiratory Syndrome virus. This data-driven sequence-based strategy will enable rapid identification of agents responsible for future epidemics.

Availability: SARS2020 is available at http://design.rxnfinder.org/sars2020/.

Contact: qnhu@sibs.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The 2019 novel coronavirus (2019-nCoV) outbreak is an ongoing pandemic. As of 20 May 2020, 4,900,647 cases were confirmed, and 320,107 deaths were attributed to the virus. On 30 Jan. 2020, the World Health Organization declared the outbreak a Public Health Emergency of International Concern.

Identification of the virus is crucial for public health authorities to contain the spread of the disease and for researchers to find methods to cure the disease (Wang, et al., 2020). The genome sequence of the 2019-nCoV became available on 10 January 2020 (Wu, et al., 2020). However, sequence alone is insufficient for accurate identification because pathogens are not defined by "taxonomy". To circumvent this limitation, we built an integrated 2019-nCoV scientific resource platform and a consensus sequence-catalytic function model with which we developed novel methodology to analyze pathogen sequences for catalytic functions. This model predicted that the 2019-nCoV has an enzymatic activity unique to SARS viruses.

2 System and methods

Platform construction and data curation: We systematically collected reports of coronavirus-related research, genomic sequences, biochemical reactions, government policies, media public opinion, and anti-viral drugs in clinical trial (Table S1, Hu, et al., 2011; Khan, et al., 2020; Shu and McCauley, 2017). This information was used to build SARS2020, an integrated scientific resource about 2019-nCoV, to provide foundation data for researchers in various fields. For data quality, we imposed strict evaluation and validation criteria. All 2019-nCoV related data were checked one-by-one to ensure authenticity. In addition, we integrated a consensus sequence-function model (Zhang, et al., 2020), a genome browser (Ham, et al., 2012), and a catalytic function annotation tool (Dawson, et al., 2017) into the platform to assist in the research of novel viruses.

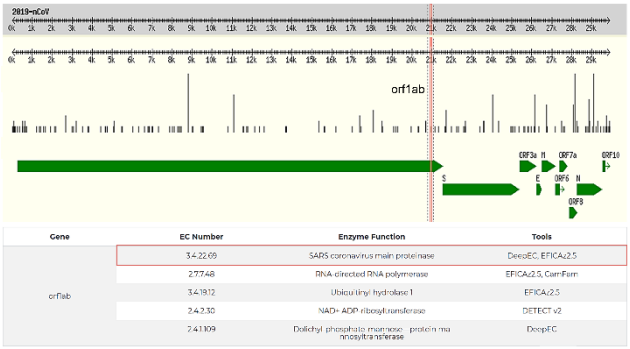
Sequence-function model: We adopted a consensus strategy to annotate enzymatic functions of biological sequences. For sequence function annotation, the family classification method captures common properties from the samples and extracts their feature vectors using machine learning algorithms, then merges the sequences into clusters or families. This consensus strategy enables efficient integration of these computational resources to maximize the accuracy and comprehensiveness of enzyme function prediction.

Web server: SARS2020 runs on a Linux server under a Nginx environment. The backend program and algorithm were written in Python using the Django framework in combination with MySQL to manage the data. Bootstrap, CSS, and JavaScript were used to implement the front-end data presentation and interactions.

Identification of 2019-nCoV: We obtained the coding sequences of 2019-nCoV from NCBI (NC_045512) and constructed a gene model from sequence based on an interpolated Markov model. We used the long-orfs tool from Glimmer3 (Delcher, et al., 2007) to identify the coding regions of bacterial, archaeal, and viral genomes. Protein translation of coding regions was performed with Biopython (Cock, et al., 2009). Then we used a consensus sequence-catalytic function model provided by SARS2020 to analyze the pathogen sequence for likely catalytic functions.

3 Results

The SARS2020 system is an integrated scientific resource platform about 2019-nCoV. At present, the system includes ~60,000 units of information. It provides powerful assistance for scientists to grasp the progress of 2019-nCoV research and to share data. SARS2020 is also a platform to assist in the identification of new viruses. We analysed the 2019-nCoV genome by the method described above. All predicted catalytic functions were derived from orf1ab (GeneID: 43740578), which seems to encode multiple proteins (Fig 1). The most likely predicted catalytic activity was SARS coronavirus main proteinase, which Enzyme Commission (EC) number is 3.4.22.69. This prediction suggested that 2019-nCoV was most likely a SARS virus, and this result was consistent with the conclusion of the International Committee on Taxonomy of Viruses. At the same time, we also predicted other possible catalytic functions in the 2019-nCoV genome, including RNA-directed RNA polymerase (EC: 2.7.7.48), dolichyl-phosphate-mannose—protein mannosyltransferase (EC: 2.4.1.109), NAD+ ADP-ribosyltransferase (EC: 2.4.2.30), and Ubiquitinyl hydrolase 1(EC: 3.4.19.12). These predicted functions will provide valuable reference for further study of biological activity and pathogenesis of the 2019-nCoV.



Gene	EC Number	Enzyme Function	Tools
orf1ab	3.4.22.69	SARS coronavirus main proteinase	DeepEC, EFICAz2.5
	2.7.7.48	RNA-directed RNA polymerase	EFICAz2.5, CamFam
	3.4.19.12	Ubiquitinyl hydrolase 1	EFICAz2.5
	2.4.2.30	NAD+ ADP-ribosyltransferase	DETECT v2
	2.4.1.109	Dolichyl-phosphate-mannose—protein mannosyltransferase	DeepEC

Fig. 1. Function analysis that identified 2019-nCoV as a SARS/SARS-like coronavirus.

4 Summary

We built an integrated platform to assist 2019-nCoV research, and we proposed a novel consensus sequence-function model for using genome sequence data to identify unknown species. Our data-driven sequence-based strategy will enable rapid identification of constantly emerging pathogens.

Funding

This work was supported by the National Key Research and Development Program of China [grant number 2019YFA0904300, 2018YFA0900700], National Natural Science Foundation of China [grant number 31700081, 31570092], Scientific Research Conditions and Technical Support System Program [grant number KFJ-BRP-009], the CAS STS program [grant number QYZDB-SSW-SMC012], International Partnership Program of Chinese Academy of Sciences of China [grant number 153D31KYSB20170121], and the Natural Science Foundation of Tianjin [15JCYBJC54300]. *Conflict of Interest:* none declared.

References

Cock, P.J., *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25(11):1422-1423.

Dawson, N.L., *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res* 2017;45(D1):D289-D295.

Delcher, A.L., *et al.* Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007;23(6):673-679.

Ham, T.S., *et al.* Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Res* 2012;40(18):e141.

Hu, Q.N., *et al.* RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. *Bioinformatics* 2011;27(17):2465-2467.

Khan, A., *et al.* Phylogenetic Analysis and Structural Perspectives of RNA-Dependent RNA-Polymerase Inhibition from SARs-CoV-2 with Natural Products. *Interdiscip Sci* 2020.

Shu, Y. and McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 2017;22(13).

Wang, C., *et al.* Human Intestinal Defensin 5 Inhibits SARS-CoV-2 Invasion by Cloaking ACE2. *Gastroenterology* 2020.

Wu, F., *et al.* Author Correction: A new coronavirus associated with human respiratory disease in China. *Nature* 2020;580(7803):E7.

Zhang, T., *et al.* Bio2Rxn: sequence-based enzymatic reaction predictions by a consensus strategy. *Bioinformatics* 2020.