International Journal of Advanced Smart Convergence Vol.8 No.1 1-8 (2019)
http://dx.doi.org/10.7236/IJASC.2019.8.1.1

IJASC 19-1-1

HCoV-IMDB: Database for the Analysis of Interactions between HCoV and Host Immune Proteins

Mi-Ran Kim[1], Ji-Hae Lee[2], Hyeon Seok Son[3] , Hayeon Kim[4,*]

[1]*Special Chemistry, Laboratory Medicine, Asan Medical Center, Seoul, Korea*
[2]*Bioinforamtion Technology Division, Bioresources Industrialization Support Department, Nakdonggang National Institute of Biological Resources, Sangju, Gyeongsangbukdo, Korea*
[3]*Laboratory of Computational Biology & Bioinformatics, Institute of Public Health and Environment, Graduate School of Public Health, Seoul National University, Seoul, Korea*
[4]*Department of Biomedical Laboratory Science, Kyungdong University, Wonju, Gangwondo, Korea*
e-mail: {[1]grube, [2]genary08, [3]hss2003}@snu.ac.kr, [4,]hykim1984@kduniv.ac.kr*

Abstract

Coronaviruses are known respiratory pathogens. In the past, most human coronaviruses were thought to cause mild symptoms such as cold. However recently, as seen in the Severe Acute Respiratory Syndrome (SARS) and the Middle East Respiratory Syndrome (MERS), infectious diseases with severe pulmonary disease and respiratory symptoms are caused by coronaviruses, making research on coronaviruses become important. Considering previous studies, we constructed 'HCoV-IMDB (Human Corona Virus Immune Database)' to systematically provide genetic information on human coronavirus and host immune information, which can be used to analyze the interaction between human coronavirus and host immune proteins. The 'HCoV-IMDB' constructed in the study can be used to search for genetic information on human coronavirus and host immune protein and to download data. A BLAST search specific to the human coronavirus, one of the database functions, can be used to infer genetic information and evolutionary relationship about the query sequence.

Keywords: *Database, Coronavirus, Immune-protein, Bioinformatics*

1. Introduction

Viruses are the primary pathogens that cause acute respiratory diseases. Viral respiratory illnesses can be experienced in all ages and can cause more severe symptoms in children, elderly and immunocompromised patients [1]. The major viruses causing respiratory infections are respiratory syncytial virus (RSV), influenza viruses, human metapneumo virus (HMPV), rhinoviruses (RV), parainfluenza virus (PIV), adenovirus (AV),

Manuscript Received: Jan. 1, 2019 / Revised: Jan. 11, 2019 / Accepted: Jan. 17, 2019
Corresponding Author: hykim1984@kduniv.ac.kr
Tel: +82-33-738-1374, Fax: +82-33-738-0243
Department of Biomedical Laboratory Science, Kyungdong University, Korea

human coronavirus (hCoV), etc [2]. In general, these viruses infect airway epithelial cells, interact with host cell proteins to promote infection, activate innate immunity and acquired immunity of the host, and cause inflammation or diseases [3]. In the past, most human coronaviruses were considered responsible mild symptoms such as cold. However recently, as seen in the Severe Acute Respiratory Syndrome (SARS) and the Middle East Respiratory Syndrome (MERS), the mortality rate, accompanied by severe pulmonary disease and respiratory syndrome, is higher than other respiratory viruses. This calls for an increasing need for research in the coronavirus infection [4]. However, most studies on human coronavirus have been conducted epidemiologically, and there is a need to conduct studies using bioinformatics tools to detect the infectivity and immunological mechanism of human coronavirus.

Virus is a pathogen that can survive only in a living cell. If the host is destroyed, the virus cannot survive. Thus, viruses have evolved to interact with the host's immune system to maintain an appropriate balance [5]. When infected with a virus, the host displays an innate immune reaction and an acquired immune reaction to the virus [3]. The innate immune reaction is a nonspecific reaction that induces an antiviral state, leading to a primary defense by cell death, phagocytosis, and complement system. As the replication of the virus invading into the host gradually progresses, an acquired immune reaction occurs. At this time, the humoral immune response comes into action that produces viral-specific antibodies by B cells and the cell-mediated immunity that produces cytotoxic T cells to kill infected cells. As a result of this acquired immune reaction, the host produces memory cells, which allow the immune response to occur promptly, if re-infection by the same virus occurs [6]. However, viruses also have mechanisms against host defense reactions that cause acute infections and latent or persistent infections [6, 7]. Thus, we constructed ‘HCoV-IMDB (Human Corona Virus Immune Database)’ to systematically provide genetic information on human coronavirus and the immune information of host, which can be used to analyze the interaction between human coronavirus and host immune proteins.

2. Methods

To construct a database for human coronavirus research, we collected the data by reviewing published papers and searching other databases related to the virus and immune system. The data of human coronavirus protein sequence information (structural protein and nonstructural protein) was collected from 'GenBank (http://www.ncbi.nlm.nih.gov/genbank)’ and 'VIPR (http://www.viprbrc.org)’, host immune-related protein information was collected from ‘IMMPORT (http://www.immport.org)’, ‘Uniprot (http://www.uniprot.org)’, and ‘GenBank (http://www.ncbi.nlm.nih.gov/genbank)’. ‘HCoV-IMDB IMDB (Human Corona Virus Immune Database)’ was built using '8C AMD Opteron-6128 2.0Ghz, 8Gb RAM, 500Gb SATA 7200rpm 3Gbps HDD server’ based on 'HPC cluster systems' and ‘Linux v2.6.18’. We used ‘MySQL v5.5’ as 'Database Management System (DBMS)' to store collected data and 'Java Server Page (JSP v2.3)', 'Hyper Text Makeup Language (HTML)' and 'Java Script' to build a web-based database (Table 1). We also installed a Web BLAST package ‘wwwblast (ncbi-blast-2.2.26, for Linux)’ to build a BLAST (Basic Local Alignment Search Tool) server that can provide homology information of the query sequence. In the database, the categories of human coronavirus data were divided into ‘structural protein’ and ‘nonstructural protein’. The data stored in human coronavirus tables are all in character format and the gene sequence information is stored as text format(Table 2). Immune information is divided into three categories; ‘Cytokines & Receptor’, ‘Natural killer cell’ and ‘Antigen processing & presentation’. The Cytokine category includes ‘Interleukin & Receptor’, ‘Interferon & Receptor’, ‘TNF (tumor necrosis factor) - Family members & Receptor’ and ‘Chemokines & Receptor’. The data stored in immune information tables are all in character format and the gene sequence information is stored as text format (Table 2).

Table 1. System development environment

Category	System development environment
System	HPC cluster system
CPU	8C AMD Opteron-6128 2.0 GHz x 1 (8Core)
Memory	Master node(16GB),10 compute node(1 for 8GB)
HDD	500Gb SATA* 7200rpm 3Gbps x 1
Operating System	Linux
Web server	Apache
DBMS	MySQL
Programing language	JSP,HTML,JAVA

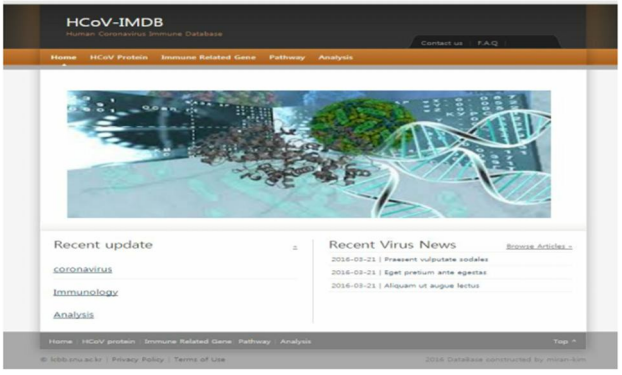
Table 2. Data type used for 'HCoV-IMDB'

Category	Data	Data type
Human Corona Virus ('Structural protein' and 'Nonstructural protein')	Gene symbol	Varchar(50) not null primary key
	ID	Varchar(10) null
	Protein name	Varchar(50) null
	Organism	Varchar(50) null
	Sequence	longtext null
Immune information ('Cytokines & Receptor', 'Natural killer cell' and 'Antigen processing & presentation'	Gene symbol	Varchar(50) not null primary key
	ID	Varchar(10) null
	Protein name	Varchar(50) null
	Synonymous	Varchar(50) null
	Sequence	longtext null
	Chromosome	Varchar(50) null

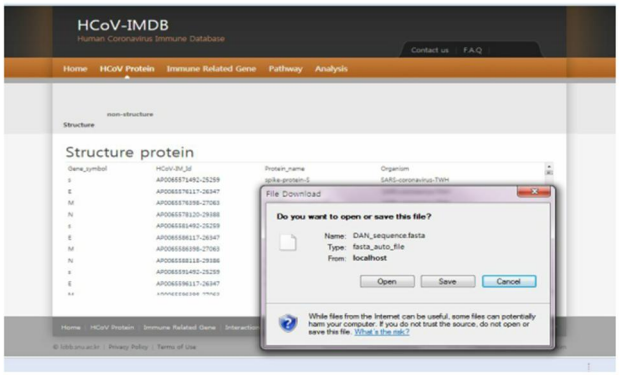
3. Results

3.1 HCoV-IMDB

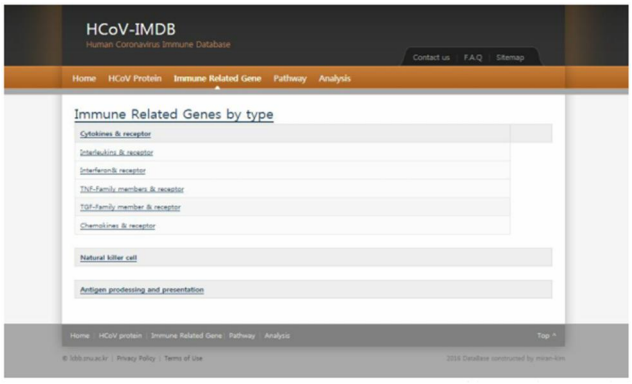
'HCoV-IMDB (Human Corona Virus Immune Database)' consists of pages for obtaining genetic information on human coronavirus and host immune protein, BLAST page and analysis result page (Figure 1). The number of structural proteins in human coronavirus are 1,478 (559 for S protein, 448 for N protein, 147 for M protein, 143 for E protein and 181 for HE protein) and the number of nonstructural proteins in human coronavirus are 4,109 (3660 for NSP1~16 proteins, 43 for NS5a protein, 80 for NS2a protein, 74 for RdRP protein, 168 for L protein and 84 for TM2 protein) in 'HCoV-IMDB' (Table 3). Immune-related proteins of 'Cytokines & receptor' category contains 89 proteins in 'Interleukins & receptor', 20 proteins in 'Interferons & receptor', 31 proteins in 'TNF-Family members & receptor', 46 proteins in 'TGF-Family members & receptor' and 155 proteins in 'Chemokines & receptor'. Immune-related proteins of 'Natural killer cell' category contain 112 proteins and 'Antigen processing & presentation' category contain 148 proteins, respectively (Table 3).



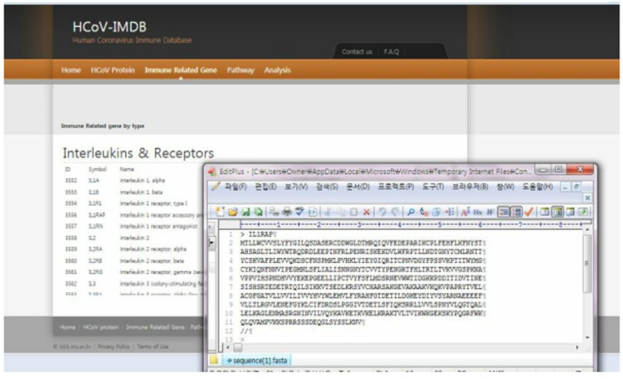
(a) Main page



(b) HCoV protein page



(c) Immune-related protein page



(d) Data search & download

Figure 1. HCoV-IMDB (Human Corona Virus Immune Database)

Users can search for genetic information on human coronavirus and host immune protein and download data in FASTA (.fasta) file format.

3.2 BLAST search based on ‘HCoV-IMDB’

BLAST page in ‘HCoV-IMDB’ was constructed by classifying human coronavirus proteins and immune related proteins. On this page, users can perform a homology search on their sequence data. To perform the BLAST search, users can click on 'Program' in the BLAST page and then select ‘blastn (nucleotide sequence)’ or ‘blastp (protein sequence)’ according to the query sequence type. Next, users can either copy the query sequence directly into 'Enter sequence in fasta format', or upload the sequence data file and click ‘submit sequence’. The result of homology search for a query sequence can be found in 'Sequence producing significant alignment'. In the results page, the sequence name and accession number of the file are displayed, and the contents of the file can be confirmed by selecting the file name. The ‘bit score’ of BLAST search result means the similarity of sequence. A higher ‘bit score’ indicates a high similarity to the query sequence. When the user clicks on the ‘bit score’ in the results page, they are connected to the database and can see the sequence alignment. An ‘E value’ means the expectation of finding that sequence by random chance when searching a database. A lower E-value indicates a better quality of the alignment BLAST search and suggests that sequences are homologous [8]. Using a BLAST search specific to the human coronavirus, users can infer genetic information and evolutionary relationship about the query sequence (Figure 2).

Table 3. Number of collected data in 'HCoV-IMDB'

Protein	No. of data	Protein	No. of data
HCoV structural protein		NSP12	307
E	143	NSP13	223
S	559	NSP14	134
M	147	NSP15	138
N	448	NSP16	137
HE	181	NS5a	43
HCoV non-structural protein		NS2a	80
NSP1	225	RdRP	74
NSP2	467	L	168
NSP3	241	TM2	84
NSP4	189	Cytokine & receptor	
NSP5	228	Interleukin & receptor	89
NSP6	223	Interferon & receptor	20
NSP7	222	TNF-Family member & receptor	31
NSP8	141	TGF-Family member & receptor	46
NSP9	398	Chemokine & receptor	155
NSP10	189	Natural killer cell	112
NSP11	198	Antigen processing	148

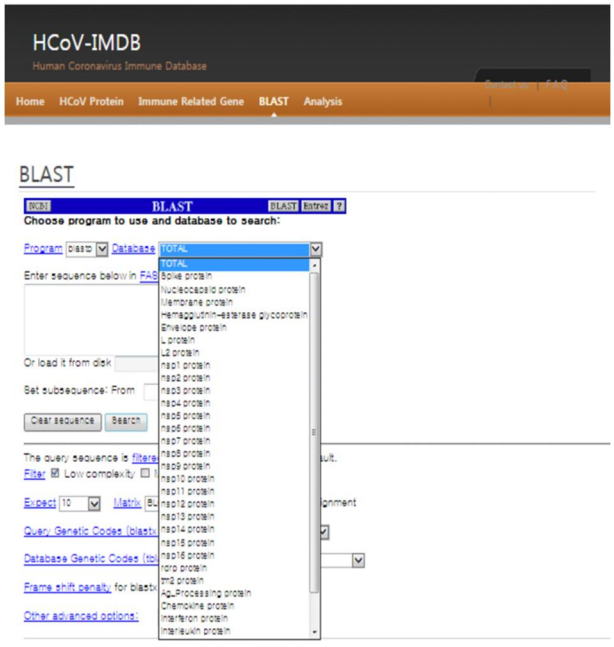
3.3 Phylogenetic analysis using data in 'HCoV-IMDB'

We performed phylogenetic analysis of viral proteins based on the 'HCoV-IMDB' database constructed in the study. S protein and N protein sequences of six coronaviruses (HCov-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1, MERS and SARS) were analyzed for evolutionary relationship. S (spike) protein binds to the receptor of the host cell membrane and mediates viral entry into host cells [9]. N (nucleocapsid) protein is known to be involved in RNA synthesis and regulation and is required for the encapsidation and packaging of viral RNA [10]. As a result of analysis of S protein, the evolutionary distance between 'HCov-229E' and 'HCoV-NL63' belonging to the 'Alpha corona virus genus' was found to be 0.407, and 'HCoV-OC43' and 'HCoV-HKU1' belonging to 'Beta corona virus genus (lineage A)' was found to be 0.374. They showed lower than the mean distance (1.063), which indicated a relatively close evolutionary relationship. The results of phylogenetic analysis of N protein appeared similar to S protein, but the distance of N protein between 'SARS' and 'MERS' was found to be 0.637, which was lower than the mean distance (1.067), indicating evolutionary similarity in sequences (Table 4).

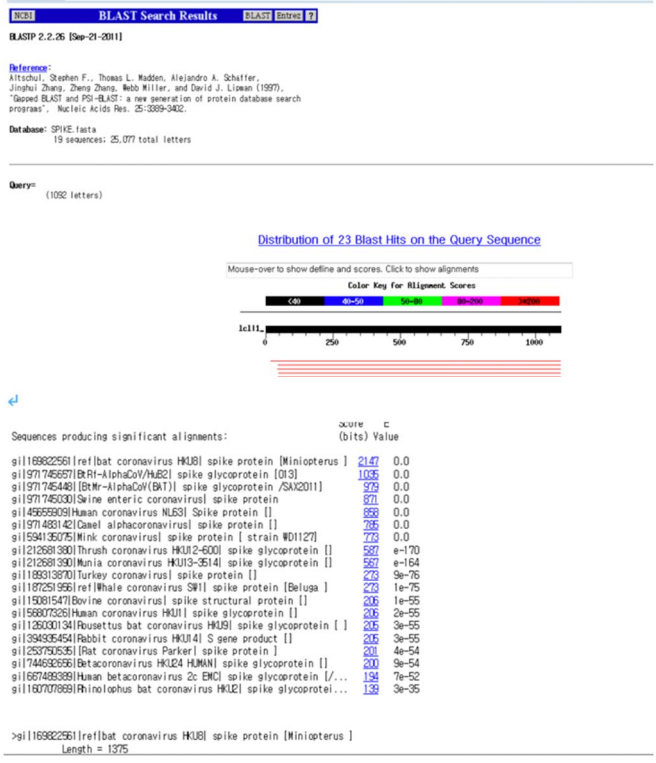
4. Discussion

We constructed HCoV-IMDB (Human Corona Virus Immune Database) to provide genetic information on coronaviruses and immune information that cause infections in humans. The 'HCoV-IMDB' enables convenient analysis by allowing the retrieval of genomic information of human coronavirus and immune-related protein together. The 'HCoV Protein' page of 'HCoV-IMDB' provides genetic information on the structural and nonstructural proteins of human coronaviruses, and the 'Immune Related Gene' page

provides genetic information on immune-related proteins in eight categories. The homology search page linked to the BLAST server is analyzed based on a database of human coronavirus proteins and immune-related proteins.



(a) Standalone BLAST web interface page



(b) Example of homology search by BLAST

Figure 2. Homology search with a query sequence in BLAST page of 'HCoV-IMDB'

BLAST is constructed on human coronavirus proteins and immune related proteins. Users can submit their sequence data, then a homology search is performed on the user's query sequence based on the sequence data

in the database.

Table 4. The result of pairwise distance

(a) Pairwise distance matrix of S (spike) protein

	HCoV-229E	HCoV-NL63	HCoV-OC43	HCoV-HKU1	MERS	SARS
HCoV-229E	-	-	-	-	-	-
HCoV-NL63	0.407	-	-	-	-	-
HCoV-OC43	1.236	1.239	-	-	-	-
HCoV-HKU1	1.239	1.243	0.374	-	-	-
MERS	1.236	1.251	0.946	1.015	-	-
SARS	1.285	1.350	1.074	1.087	1.018	-

(The overall average is 1.063).

(b) Standalone BLAST web interface page

	HCoV-229E	HCoV-NL63	HCoV-OC43	HCoV-HKU1	MERS	SARS
HCoV-229E	-	-	-	-	-	-
HCoV-NL63	0.706	-	-	-	-	-
HCoV-OC43	1.266	1.300	-	-	-	-
HCoV-HKU1	1.243	1.386	0.363	-	-	-
MERS	1.254	1.243	0.939	1.016	-	-
SARS	1.336	1.243	0.998	1.016	0.637	-

(The overall average is 1.067).

With the development of molecular biology technology, large amounts of biological data are accumulating, but traditional methods require considerable time and effort to understand complex immune systems. Bioinformatics research is needed to systematically and efficiently manage and analyze large-scale genetic information. Therefore, we built a database by collecting scattered information and classifying the data systematically, which will enable efficient analysis of the relationship between rapidly evolving viruses and immune proteins. The database has improved convenience by allowing the retrieval of genetic information of human coronavirus proteins and immune-related proteins and the downloading sequence data. If continuous data updates and improved bioinformatics methods are applied, 'HCoV-IMDB' could be a useful database for studies on coronavirus and immunological mechanisms.

5. Conclusion

Recently, the development of molecular biology techniques, such as Next Generation Sequencing (NGS), has resulted in rapid accumulation of large amounts of genetic data. Accordingly, databases are being actively constructed to efficiently utilize genetic information. In this study, we constructed a database for analyzing human coronaviruses, including MERS and SARS, among the infectious diseases viruses. This database was constructed by integrating genetic data of human coronavirus structural and nonstructural proteins and genetic data of immune related proteins, and can be implemented to search and analyze genetic information of stored proteins efficiently. Researchers can use ‘HCoV-IMDB’ to acquire the information needed for virus and host immune-specific studies. The data can be used for analysis via easy identification of the genetic characteristics and sequences of coronaviruses and of proteins involved in sequence and

antiviral immunity mechanisms. This can be useful for the study of unexpected infectious diseases caused by novel and variant viruses, and can be the basis for providing a fundamental measure to prevent epidemics by predicting host immune response to viral infection. In addition, information on the relationship between virus genetic information and the severity of the symptoms of an infected host can improve accurate prediction of the host immune response to coronavirus infection. 'HCoV-IMDB' provides immunological and molecular biological data useful for understanding the host immune response in relation to coronavirus infection, and will be able to ultimately assist in the development of vaccines and therapies by providing access to specific immune protein information.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT and MOE) (No. 2016R1C1B2015511 and No. 2017R1D1A1B03033413).

References

[1] T.R. Cate. Impact of influenza and other community-acquired viruses. Seminars in respiratory infections. Vol. 13, No. 1, pp. 17-23, Mar 1998.

[2] W.G. Nichols, A.J. Peck Campbell, M. Boeckh. Respiratory Viruses Other than Influenza Virus: Impact and Therapeutic Advances. Clinical microbiology reviews. Vol. 21, No. 2, pp. 274-290, Apr 2008.

[3] R.A. Tripp. Respiratory Syncytial Virus (RSV) Modulation at the Virus-Host Interface Affects Immune Outcome and Disease Pathogenesis. Immune Network. Vol. 13, No. 5, pp. 163-167, Oct 2013.

[4] A. Zumla, D.S. Hui, S. Perlman. Middle East respiratory syndrome. Lancet. Vol. 386, No. 9997, pp. 995-1007, Sep 2015.

[5] R.M. Zinkernagel. Immunology taught by viruses. Science. Vol. 271, No. 5246, pp. 173-178, Jan 1996.

[6] T.C. Merigan. Host defenses against viral disease. The New England journal of medicine..Vol. 290, No. 6, pp. 323-329, Feb 1974.

[7] E.S. Hwang, C.G. Park, C.Y. Cha. Immune Responses to Viral Infection. Immune Network. Vol. 4, No. 2, pp. 73-80, Jun 2004.

[8] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman. Basic local alignment search tool. Journal of molecular biology. Vol. 215, No. 3, pp. 403-410, Oct 1990.

[9] L. Enjuanes, C. Smerdou, J. Castilla, I.M. Antón, J.M. Torres, I. Sola, J. Golvano, J.M. Sánchez, B. Pintado. Development of protection against coronavirus induced diseases. A review. Advances in experimental medicine and biology. Vol. 380, pp. 197-211, 1995.

[10] S. Zúñiga, J.L. Cruz, I. Sola, P.A. Mateos-Gómez, L. Palacio, L. Enjuanes. Coronavirus nucleocapsid protein facilitates template switching and is required for efficient transcription. Journal of virology. Vol. 84, No. 4, pp. 2169-2175, Feb 2010.