G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(4):e22
https://doi.org/10.5808/GI.2018.16.4.e22

ORIGINAL ARTICLE

Genome-Based Virus Taxonomy with the ICTV Database Extension

Shinduck Kang, Young-Chang Kim*

Department of Microbiology, Chungbuk National University, Cheongju 28644, Korea

In 1966, the International Classification of Viruses (ICNV) was established to standardize the naming of viruses. In 1975, the organization was renamed "International Committee on Taxonomy of Viruses (ICTV)," by which it is still known today. The primary virus classification provided by ICTV in 1971 was for viruses infecting vertebrates, which includes 19 genera, 2 families, and 24 unclassified groups. Presently, the 10th virus taxonomy has been published. However, the early classification of viruses was based on clinical results "*in vivo*" and "*in vitro*," as well as on the shape of the Phenotype virus. Due to the development of next-generation sequencing and the accompanying bioinformatics analysis pipelines, a reconstruction of the classification system has been proposed. At a meeting held in Boston, USA between June 9–11, 2016, there was even an in-depth discussion regarding the classification of viruses using metagenomic data. One suggested activity that arose from the meeting was that viral taxonomy should be reconstructed, based on genotype and bioinformatics analysis "*in silico*." This article describes our efforts to achieve this goal by construction of a web-based system and the extension of an associated database, based on ICTV taxonomy. This virus taxonomy web system was designed specifically to extend the virus taxonomy up to strain and isolation, which was then connected with the NCBI database to facilitate searches for specific viral genes; there are also links to journals provided by the EMBL RESTful API that improves accessibility for academic groups.

Keywords: ICTV DB extension, virus history, virus taxonomy searching web

Introduction

Presently, there are 3,279 virus reference genomes registered in NCBI. More than 1.8 million sequences are included in GenBank (https://www.ncbi.nlm.nih.gov/genbank/) [1]. The number of whole-genome sequences in GenBank is rapidly increasing, as shown in Fig. 1. Currently, only about 1,800 genome sequences have been assigned to species in International Committee on Taxonomy of Viruses (ICTV); the remaining 1,400 sequences have not been classified as species. Although ICTV is responsible for viral classification, it does not have the capacity to immediately formulate the naming conventions and taxonomy for the large number of viral sequences that is submitted to the organization.

However, with the advent of next-generation sequencing and the enhancement of NCBI GenBank data, the classical ICTV method of viral classification based on phenotypic parameters has been converting to a classification based on genotypic classification due to improvements in the speed and accuracy associated with virus taxonomy. Recently, a metagenomic method based on genotype was proposed as an approach to aid virus taxonomy [2]. However, this creates a requirement for an appropriate data handling and analysis pipeline to cope with such needs.

Since the development of web servers in 1993, bioinformatics data have been provided through browser-based systems. Many analysis tools, such as MSA, BLAST, and Genome Browser, have been developed for end users. In the case of ICTV taxonomy and naming, the initial ICTVdB was developed with flat data (DELTA: DEscription Language for TAxonomy), which was not connected to other databases. ICTVdB did not contain any sequence information but was used for phylogenetic analysis [3]. Presently, the 10th ICTV virus taxonomy has been published and is available on the ICTV website (http://ictv.global/report/). However, there is no easy approach to NCBI GenBank data based on ICTV

Received May 17, 2018; Revised December 23, 2018; Accepted December 23, 2018; Published online December 28, 2018

***Corresponding author:** Tel: +82-43-261-2302, Fax: +82-50-4477-6563, E-mail: youngkim@chungbuk.ac.kr

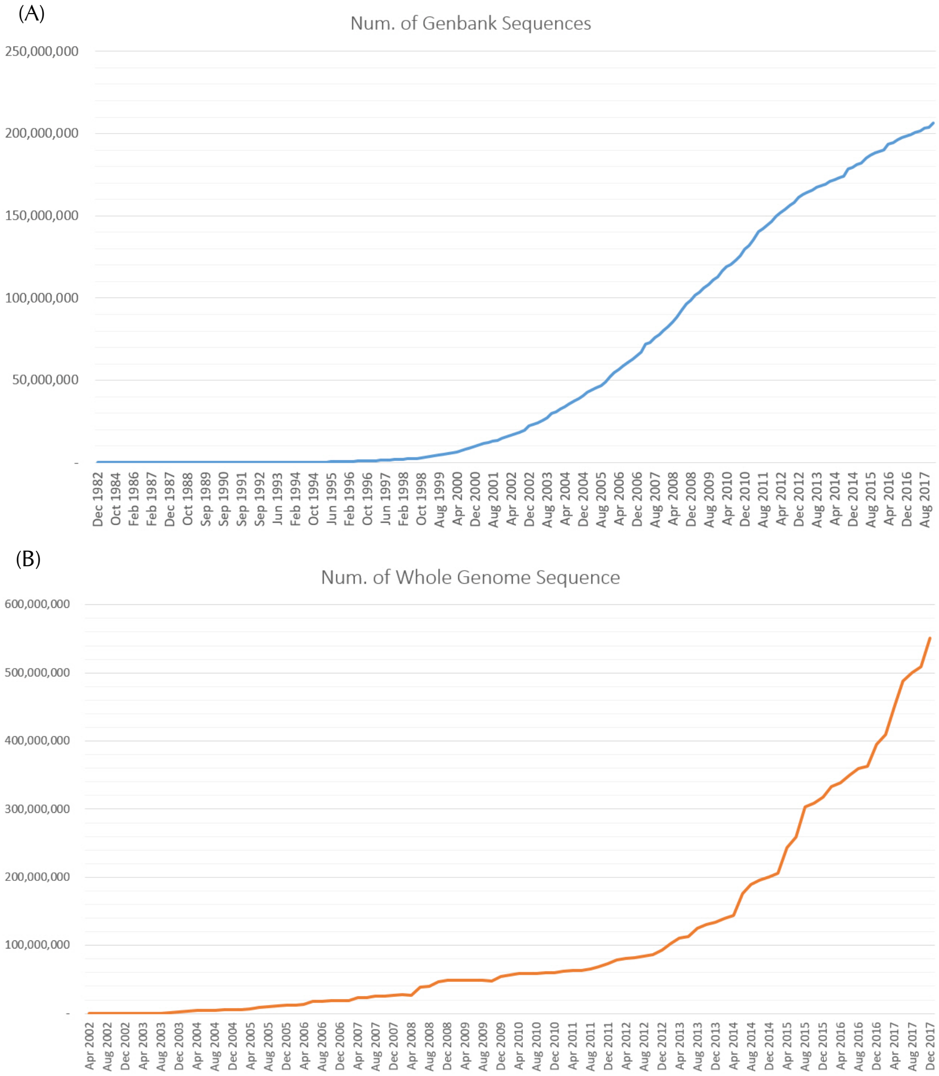


Fig. 1. GenBank sequences (A) and whole-genome sequences (B) published over time.

taxonomy, strain, or isolation information for selected viral species, because ICTV taxonomy has only been providing up to the species level. Also, web-based ICTV taxonomy does not provide direct PubMed access, which facilitates academic searches. As a result, our virus taxonomy website reinforces this problem and extends the related tables in the ICTV database.

Methods

ICTV taxonomy and virus history

The gene sequences submitted to NCBI are recorded in GenBank format with a unique key that is generated by the combination of the accession number and version number. The accession number consists of 1 letter and 5 numerals or 2 letters and 6 numerals for nucleotides and 3 letters and 5 numerals for proteins. The GenBank format is structurally divided into meta information, feature information, and sequence information. In this system, due to the types of viral targets, only "gbvrl" data among NCBI GenBank data are collected and used (ftp://ftp.ncbi.nlm.nih.gov/genbank). Currently, it is available from "gbvrl1.seq.gz" to "gbvrl51.seq.gz" (2017/12/20). However, GenBank data are highly redundant due to frequently overlapping submissions. This means that computing or parsing after collecting or manipulating GenBank data is an extremely inefficient process. Therefore, NCBI has provided RefSeq data to minimize redundancy, and there are presently 9,557 complete viral genomic RefSeq sequences. Meanwhile, in International Classification of Viruses (ICNV; the name before being revised to ICTV), the first virus taxonomy of 1,971 included 19 genera and 2 families (*Papovaviridae* and *Picornaviridae*), while 24 groups were unassigned until the appropriate classification levels were determined [4]. In the current 10th virus taxonomy on the ICTV website, based on the final version (“ICTVMasterSpeciesList2016v1.3”), there are 4,404 species, whereas there are 9,556 complete genome sequences of viral species in GenBank RefSeq (Table 1).

The goal in our web-based system is to extend the basic information in the ICTV taxonomy database in order to include strain and isolate group and to provide raw data of genomic sequences, as well as history and PubMed information for user-chosen viruses. As a prerequisite, the 10th ICTV taxonomy, which is the most recent, must be parsed. However, ICTV does not provide taxonomy history through OpenAPI. Thus, we collected the data on the taxonomy history and the linked node information via web scraping.

Virus taxonomy database

We collected the ICTV taxonomy from the “ICTV Master Species List,” which was officially announced in ICTV in 2016 (Table 2); the taxonomy history was obtained by web scraping. Furthermore, in order to extend the resource including strain and isolation information and to connect to the viral GenBank information, we downloaded “gbvrl1.seq.gz”~“gbvrl51.seq.gz” (the GenBank virus file) using an FTP protocol and classified the data according to ICTV taxonomy criteria. Currently, the classification table, which is designed in the current ICTV database, includes classification name, classification level, release number and year, classification ID (composed of 8 digits), the most recent classification change ID (composed of 8 digits), parent classification name, change status, and proposal documents [5].

However, in our web-based system, the current ICTV database was redesigned and divided as the tables in our database (Table 3). Specifically, to enhance the database and to make useful linkages for NCBI accession, the NCBI Taxonomy items described in Table 3 and the items parsed by web scraping were built as an “ICTV history” table and “ICTV Taxonomy” table in Table 3, respectively. The “2016 ICTV Species” table consists of the data parsed by the ICTV Master Species List (2016, v1.3).

Table 1. Number of viruses classified or sequenced in the 10th ICTV Taxonomy and GenBank

Virus taxonomy	ICTV	RefSeq
Order	8	-
Family	122	-
Subfamily	35	-
Genus	735	-
Species	4,404	9,556

ICTV, International Committee on Taxonomy of Viruses.

Table 2. Current ICTV DB

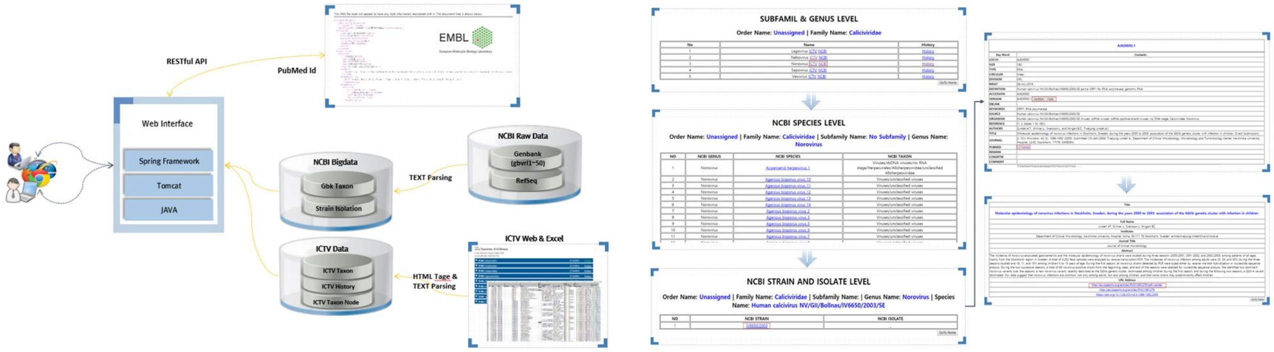
ICTV items	Example 1	Example 2
Taxon name	Measles virus	Measles morbillivirus
Taxon level	Species	Species
Release number	30	31
Release year	2015	2016
Taxon ID (stable)	19750163	19750163
Node ID (new with each release)	20151044	20161044
Parent taxon	Morbillivirus	Morbillivirus
Last change	Move	Rename
Proposal	2015.Pneumoviridae.pdf	2016.Paramyxoviridaespren.pdf

ICTV DB, International Committee on Taxonomy of Viruses database.

Table 3. Modified virus taxonomy tables (PK: primary key)

ICTV history	2016 ICTV species	ICTV taxonomy	NCBI taxonomy
ICTV taxon node	ICTV order	Taxon node (PK)	NCBI accession (PK)
ICTV name	ICTV family	Taxon type	NCBI genus
ICTV new taxon	ICTV subfamily	Taxon name	NCBI species
Modification year	ICTV genus		NCBI taxon
Modification status	ICTV species		NCBI strain
ICTV old taxon	ICTV main species		NCBI isolate
ICTV proposal	NCBI accession		
	Isolation name		
	Gene type		
	ICTV status		
	ICTV proposal		

Refer to “Supplementary Material” for a detailed table description of the International Committee on Taxonomy of Viruses (ICTV) Extension database.
Accessible URL address: http://synb.chungbuk.ac.kr:8080/ICTV.


RESTful API
PubMed Id
EMBL
Web Interface
Spring Framework
Tomcat
JAVA
NCBI Bigdata
Gbk Taxon
Strain Isolation
TEXT Parsing
NCBI Raw Data
Genbank (gbvrl1~50)
RefSeq
ICTV Data
ICTV Taxon
ICTV History
ICTV Taxon Node
HTML Tage & TEXT Parsing
ICTV Web & Excel
SUBFAMIL & GENUS LEVEL
Order Name: Unassigned | Family Name: Caliciviridae
NCBI SPECIES LEVEL
Order Name: Unassigned | Family Name: Caliciviridae | Subfamily Name: No Subfamily | Genus Name: Norovirus
NCBI STRAIN AND ISOLATE LEVEL
Order Name: Unassigned | Family Name: Caliciviridae | Subfamily Name: | Genus Name: Norovirus | Species Name: Human calicivirus NV/GII/Bollnas/IV6650/2003/SE

(A) Flow diagram of web-based virus taxonomy with the extended database (using Cubird DBMS)

(B) Web-based search interfaces based on ICTV taxonomy

Fig. 2. (A, B) Search for strain and isolate from NCBI GenBank files and connection to PubMed via the EMBL RESTful API. ICTV, International Committee on Taxonomy of Viruses.

Web construction

Our web-based system includes the traditional taxonomy (order, family, subfamily, genus, and species), as well as the information regarding strain and isolate. Furthermore, users can easily access journals containing information related to the published virus GenBank via the EMBL RESTful protocol [6] and directly download and reuse NCBI FASTA and “gbk” file via the Entrez openAPI [7]. However, the information of the chosen PubMed and strain and isolate are based on NCBI accession in (Table 3). The journal search is connected by the parameters of the HTML *get* method, which is indicated by the PubMed ID. Web scraping methods were used to build the “ICTV history” table in our database after extracting meaningful information from the NCBI raw data. The tables in our database form the foundation of the web system. In the Spring framework, the web system consists of Java, which is independent of the operating system. According to the user commands, the internal parsing process is executed by pipelines that are implemented by the BioPython module. The internal parsing process extracts the information of the virus taxonomy, history, and reference articles from XML data, which are produced by the EMBL RESTful API, and text files of the NCBI virus GenBank. The overall process map for the web system is described in Fig. 2.

Results and Discussion

The aim of this study was to evaluate and develop a computerized system that is fused with bioinformatics. Specifically, we focused on implementing an environment that extends the capabilities of the ICTV web system and connects to PubMed in order to enhance searches performed by academic groups. We extended and rebuilt the database and extracted meaningful data using a pipeline that parses XML, text, and web contents. Henceforth, this computerized system will be continually extended and used as a web tool that can detect new viral types and classify them rapidly and accurately. Recently, a new virus classification system based on metagenomics has been proposed. Thus, web-based virus taxonomy could augment the quality by adding virus classification, which is derived by viral metagenomics analysis [8]. We suggest that the web system, analytical pipelines, and extended database we describe herein could be used to add these metagenomics data to ICTV taxonomy data.

ORCID: Shinduck Kang: https://orcid.org/0000-0002-0624-9451; Young-Chang Kim: https://orcid.org/0000-0001-8285-0882

Authors’ contribution

Conceptualization: YCK
Data curation: SK
Formal analysis: SK
Methodology: SK, YCK
Writing – original draft: SK
Writing – review & editing: SK, YCK

Conflicts of Interest

No potential conflicts of interest relevant to this article was reported.

Acknowledgments

This work was supported by an intramural research grant of Chungbuk National University in 2015.

Supplementary material

Supplementary data can be found with this article online at https://doi.org/10.5808/GI.2018.16.4.e22.

References

1. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, *et al*. GenBank. *Nucleic Acids Res* 2013;41:D36-D42.
2. Simmonds P. Methods for virus classification and the challenge of incorporating metagenomic sequence data. *J Gen Virol* 2015;96:1193-1206.
3. Cientific S, Atabases D. The universal virus database ictvdb. *Syst Zool* 1974;23:50-57.
4. Adams MJ, Lefkowitz EJ, King AM, Harrach B, Harrison RL, Knowles NJ, *et al*. 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. *Arch Virol* 2017;162:1441-1446.
5. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* 2018;46:D708-D717.
6. Lopez R, Cowley A, Li W, McWilliam H. Using EMBL-EBI services via web interface and programmatically via web services. *Curr Protoc Bioinformatics* 2014;48:3.12.1-3.12.50.
7. McEntyre J. Linking up with Entrez. *Trends Genet* 1998; 14:39-40.
8. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, *et al*. Consensus statement: virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* 2017;15:161-168.