Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Introduction

In 1955, Niu and Frankel-Conrat published the C-terminal amino acid sequence of tobacco mosaic virus capsid protein. The complete 158-amino-acid sequence of this protein was published in 1960. The first completely sequenced viral genome published was that of bacteriophage MS2 in 1976 (GenBank accession number V00642). Sanger used DNA from bacteriophage phiX174 (J02482) in developing the dideoxy sequencing method, while the first animal viral genome, SV40 (J02400), was sequenced using the Maxam and Gilbert method and published in 1978. Viruses therefore played a pivotal role in the development of modern-day sequencing methods, and viral sequence information (both protein and nucleotide) formed a substantial subset of the earliest available biological databases. In 1965, Margaret O. Dayhoff published the first publicly available database of biological sequence information. This Atlas of Protein Sequence and Structure was available only in printed form and contained the sequences of approximately 50 proteins. Establishment of a database of nucleic acid sequences began in 1979 through the efforts of Walter Goad at the US Department of Energy’s Los Alamos National Laboratory (LANL) and separately at the European Molecular Biology Laboratories (EMBL) in the early 1980s. In 1982, the LANL database received funding from the National Institutes of Health (NIH) and was christened GenBank. In December of 1981, the Los Alamos Sequence Library contained 263 sequences of which 50 were from eukaryotic viruses and 12 were from bacteriophages. By its tenth release in 1983, GenBank contained 1865 sequences (1 827 214 nucleotides) of which 449 (457 721 nucleotides) were viral. In August of 2006, GenBank (release 154) contained approximately 59 000 000 records, including 367 000 viral sequences.

The number of available sequences has increased exponentially as sequencing technology has improved. In addition, other high-throughput technologies have been developed in recent years, such as those for gene expression and proteomic studies. All of these technologies generate enormous new data sets at ever-increasing rates. The challenge, therefore, has been to provide computational systems that support the storage, retrieval, analysis, and display of this information so that the research scientist can take advantage of this wealth of
resources to ask and answer questions relevant to his or her work. Every article in this encyclopedia contains knowledge that has been derived in part from the analysis of large data sets. The ability to effectively and efficiently utilize these data sets is directly dependent on the databases that have been developed to support storage of this information. Fortunately, the continual development and improvement of data-intensive biological technologies has been matched by the development and improvement of computational technologies. This work, which includes both the development and utilization of databases as well as tools for storage and analysis of biological information, forms a very important part of the bioinformatics field. This article provides an overview of database structure and how that structure supports the storage of biological information. The different types of data associated with the analysis of viruses are discussed, followed by a review of some of the various online databases that store general biological information as well as virus-specific information.

**Databases**

**Definition**

A database is simply a collection of information, including the means to store, manipulate, retrieve, and share that information. For many of us, lab notebook fulfilled our initial need for a ‘database’. However, this information storage vehicle did not prove to be an ideal place to archive our data. Backups were difficult, and retrieval more so. The advent of computers – especially the desktop computer – provided a new solution to the problem of data storage. Though initially this innovation took the form of spreadsheets and electronic notebooks, the subsequent development of both personal and large-scale database systems provided a much more robust solution to the problems of data storage, retrieval, and manipulation. The computer program supplying this functionality is called a ‘database management system’ (DBMS). Such systems provide at least four things: (1) the necessary computer code to guide a user through the process of database design; (2) a computer language that can be used to insert, manipulate, and query the data; (3) tools that allow the data to be exported in a variety of formats for sharing and distribution; and (4) the administrative functions necessary to ensure data integrity, security, and backup. However, regardless of the sophistication and diverse functions available in a typical modern DBMS, it is still up to the user to provide the proper context for data storage. The database must be properly designed to ensure that it supports the structure of the data being stored and also supports the types of queries and manipulations necessary to fully understand and efficiently analyze the properties of the data.

**Data**

The development of a database begins with a description of the data to be stored, all of the parameters associated with the data, and frequently a diagram of the format that will be used. The format used to store the data is called the database schema. The schema provides a detailed picture of the internal format of the database that includes specific containers to store each individual piece of data. While databases can store data in any number of different formats, the design of the particular schema used for a project is dependent on the data and the needs and expertise of the individuals creating, maintaining, and using the database. As an example, we will explore some of the possible formats for storing viral sequence data and provide examples of the database schema that could be used for such a project.

*Figure 1(a)* provides an example of a GenBank sequence record that is familiar to most biologists. These records are provided in a ‘flat file’ format in which all of the information associated with this particular sequence is provided in a human-readable form and in which all of the information is connected in some manner to the original sequence. In this format, the relationships between each piece of information and every other piece of information are only implicitly defined, that is, each line starts with a label that describes the information in the rest of the line, but it is up to the investigator reading the record to make all of the proper connections between each of the data fields (lines). The proper connections are not explicitly defined in this record. As trained scientists, we are able to read the record in *Figure 1(a)* and discern that this particular amino acid sequence is derived from a strain of Ebola virus that was studied by a group in Germany, and that this sequence codes for a protein that functions as the virus RNA polymerase. The format of this record was carefully designed to allow us, or a computer, to pull out each individual type of information. However as trained scientists, we already understand the proper connections between the different information fields in this file. The computer does not. Therefore, to analyze the data using a computer, a custom software program must be written to provide access to the data.

Extensible markup language (XML) is another widely used format for storing database information. *Figure 1(b)* shows an example of part of the XML record for the Ebola virus polymerase protein. In this format, each data field can be many lines long; the start and end of a data record contained within a particular field are indicated by tags made of a label between two brackets (’<label>... </label>’). Unlike the lines in the GenBank record in *Figure 1(a)*, a field in an XML record can be placed inside of another, defining a structure and a relationship between them. For example, the *TSeq_orgname* is placed inside of the *TSeq* record to show that this
GenBank record

LOCUS       NP_066251               2212 aa            linear   VRL 30-MAR-2006
DEFINITION   polymerase [Zaire ebolavirus].
ACCESSION   NP_066251.1
VERSION     NP_066251.1
GI:10313999
ORIGIN      Zaire ebolavirus
ORGANISM    Zaire ebolavirus Virus; ssRNA negative-strand viruses; Mononegavirales;
            Filoviridae; Ebola-like viruses.
REFERENCE 1 (residues 1 to 2212)
AUTHORS    Volchkov,V.E., Volchkova,V.A., Chepurnov,A.A., Blinov,V.M.,
            Dolnik,O., Netesov,S.V. and Feldmann,H.
TITLE      Characterization of the L gene and 5' trailer region of Ebola virus
JOURNAL    J. Gen. Virol. 80 (Pt 2), 355-362 (1999)
PUBMED      10073695
...
REFERENCE 8 (residues 1 to 2212)
AUTHORS    Volchkov,V.E.
TITLE      Direct Submission
JOURNAL    Submitted (20-AUG-1998) Institute of Virology, Philipps-University
            Marburg, Robert-Koch-Str. 17, Marburg 35037, Germany
COMMENT    PROVISIONAL: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AAD14589.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
source          1..2212
/organism="Zaire ebolavirus"
/strain="Mayinga"
/db_xref="taxon:186538"
Protein        1..2212
/product="polymerase"
/function="synthesis of viral RNAs; transcriptional RNA editing"
/note="L"
/calculated_mol_wt=252595
CDS            1..2212
/gene="L"
/locus_tag="ZEBOVgp7"
/coded_by="NC_002549.1:11581..18219"
/citation=[1]
/db_xref="GeneID:911824"

ORIGIN
1 matqhtqpypd qcdtvsrag lyssyslnpq lrncikpahi yrlkydtvt
2 61 kfdsvdpvlp lkasgmgfz pvepccqgfz dkskystdul algkybzyln
... 2101 ndynqqrqsr tgyhdfirta kgrittvlnvd ylfkfluvsq lkhngtwqas fkklpelisv
2161 cserfjhirrc nowerfiqvq jiyhmgdske vklieirg gtsq lpsdqylrd fd
//

(a)

XML record

<?xml version="1.0"?>
<!DOCTYPE TSeqSet PUBLIC "NCBI_Tseq.dtd">
<TSeqSet>
<TSeq>
<TSeq_seqtype value="protein"/>
<TSeq_gi>10313999</TSeq_gi>
<TSeq_accver>NP_066251.1</TSeq_accver>
<TSeq_taxid>186538</TSeq_taxid>
<TSeq_orgname>Zaire ebolavirus</TSeq_orgname>
<TSeq_defline>polymerase [Zaire ebolavirus]</TSeq_defline>
<TSeq_length>2212</TSeq_length>
<TSeq_sequence>
MATQHTQYPDARLSSPIVLDQCDTVACGYSSYSLNSQNLNCKLRRHLYKVDVTVPFLEVFVAT
LDIYVFLVYKIKAXSNGCTCFVPEFCQQLSEDIEHTYQGVNLKLYLYKVLWNGVSCDVDEMHPQKRELLSI
...
QLQIQSPYWPULQ7QDCEMLSYRGLFSPKLYKLYKLYIRLWSSLKRIPLQVISITQKLHNLAPREFELT
NDYQQQQRGTQRTFIRANGRTKLYKLYKLYKLYQALKHGNTQAEPEKFLPELISVFCHFMYHRC
NCEBFVQTLYLRLQSBHVKLLERLCLGSLIFQDVYRFD
</TSeq_sequence>
</TSeq>
</TSeqSet>

(b)

Figure 1  Data formats. Examples of two different formats for organizing sequence data are shown. (a) An example of a GenBank flat file sequence record. (b) Part of the same record using a hierarchical XML format.
a user (with an appropriate search program) needs to traverse the whole file in order to pull out the particular items of data that are of interest. Therefore, while an XML file may be an excellent format for defining and exchanging data, it is often not the best vehicle for efficiently storing and querying that data. That is still the realm of the relational database.

‘Relational database management systems’ (RDBMSs) are designed to do two things extremely well: (1) store and update structured data with high integrity, and (2) provide powerful tools to search, summarize, and analyze the data. The format used for storing the data is to divide it into several tables, each of which is equivalent to a single spreadsheet. The relationships between the data in the tables are then defined, and the RDBMS ensures that all data follow the rules laid out by this design. This set of tables and relationships is called the schema. An example diagram of a relational database schema is provided in Figure 2. This Viral Genome Database (VGD) schema is an idealized version of a database used to store viral genome sequences, their associated gene sequences, and associated descriptive and analytical information. Each box in Figure 2 represents a single object or concept, such as a genome, gene, or virus, about which we want to store data and is contained in a single table in the RDBMS. The names listed in the box are the columns of that table, which hold the various types of data about the object. The ‘gene’ table therefore contains columns holding data such as the name of the gene, its coding strand, and a description of its function. The RDBMS is

![Figure 2: VGD Database schema](image-url)

Underlying structure of a Viral Genome Sequence Database (VGD). Data tables are grouped according to the type of information contained in each set of tables. Each table contains a set of fields that hold a particular type of data. Lines and arrows display the relationships between fields as defined by the foreign key (FK) and primary key (PK) that connect two tables. (Each arrow points to the table containing the primary key.) Tables are color-coded according to the source of the information they contain: yellow, data obtained from the original GenBank sequence record and the ICTV Eighth Report; pink, data obtained from automated annotation or manual curation; blue, controlled vocabularies to ensure data consistency; green, administrative data.
able to enforce a series of rules for tables that are linked by defining relationships that ensure data integrity and accuracy. These relationships are defined by a foreign key in one table that links to corresponding data in another table defined by a primary key. In this example, the RDMS can check that every gene in the ‘gene’ table refers to an existing genome in the ‘genome’ table, by ensuring that each of these tables contains a matching ‘genome_id’. Since any one genome can code for many genes, many genes may contain the same ‘genome_id’. This defines what is called a one-to-many relationship between the ‘genome’ and ‘gene’ tables. All of these relationships are identified in Figure 2 by arrows connecting the tables.

Because viruses have evolved a variety of alternative coding strategies such as splicing and RNA editing; it is necessary to design the database so that these processes can be formally described. The ‘gene_segment’ table specifies the genomic location of the nucleotides that code for each gene. If a gene is coded in the traditional manner — one ORF, one protein — then that gene would have one record in the ‘gene_segment’ table. However, as described above, if a gene is translated from a spliced transcript, it would be represented in the ‘gene_segment’ table by two or more records, each of which specifies the location of a single exon. If an RNA transcript is edited by stuttering of the polymerase at a particular run of nucleotides, resulting in the addition of one or more non-templated nucleotides, then that gene will also have at least two records in the ‘gene_segment’ table. In this case, the second ‘gene_segment’ record may overlap the last base of the first record for that gene. In this manner, an extra, non-templated base becomes part of the final gene transcript. Other more complex coding schemes can also be identified using this, or similar, database structures.

The tables in Figure 2 are grouped according to the type of information they contain. Though the database itself does not formally group tables in this manner, database schema diagrams are created to benefit database designers and users by enhancing their ability to understand the structure of the database. These diagrams make it easier to both populate the database with data and query the database for information. The core tables hold basic biological information about each viral strain and its genomic sequence (or sequences if the virus contains segmented genomes) as well as the genes coded for by each genome. The taxonomy tables provide the taxonomic classification of each virus. Taxonomic designations are taken directly from the Eighth Report of the International Committee on Taxonomy of Viruses (ICTV). The ‘gene properties’ tables provide information related to the properties of each gene in the database. Gene properties may be generated from computational analyses such as calculations of molecular weight and isoelectric point (pl) that are derived from the amino acid sequence. Gene properties may also be derived from a manual curation process in which an investigator might identify, for example, functional attributes of a sequence based on evidence provided from a literature search. Assignment of ‘gene ontology’ terms (see below) is another example of information provided during manual curation. The BLAST tables store the results of similarity searches of every gene and genome in the VGD searched against a variety of sequence databases using the National Center for Biotechnology Information (NCBI) BLAST program. Examples of search databases might include the complete GenBank nonredundant protein database and/or a database comprised of all the protein sequences in the VGD itself. While most of us store our BLAST search results as files on our desktop computers, it is useful to store this information within the database to provide rapid access to similarity results for comparative purposes; to use these results to assign genes to orthologous families of related sequences; and to use these results in applications that analyze data in the database and, for example, display the results of an analysis between two or more types of viruses showing shared sets of common genes. Finally, the ‘admin’ tables provide information on each new data release, an archive of old data records that have been subsequently updated, and a log detailing updates to the database schema itself.

It is useful for database designers, managers, and data submitters to understand the types of information that each table contains and the source of that information. Therefore, the database schema provided in Figure 2 is color-coded according to the type and source of information each table provides. Yellow tables contain basic biological data obtained either directly from the GenBank record or from other sources such as the ICTV. Pink tables contain data obtained as the result of either computational analyses (BLAST searches, calculations of molecular weight, functional motif similarities, etc.) or from manual curation. Blue tables provide a controlled vocabulary that is used to populate fields in other tables. This ensures that a descriptive term used to describe some property of a virus has been approved for use by a human curator, is spelled correctly, and when multiple terms or aliases exist for the same descriptor, the same one is always chosen.

While the use of a controlled vocabulary may appear trivial, in fact, misuse of terms, or even misspellings, can result in severe problems in computer-based databases. The computer does not know that the terms ‘negative-sense RNA virus’ and ‘negative-strand RNA virus’ may both be referring to the same type of virus. The provision and use of a controlled vocabulary increases the likelihood that these terms will be used properly, and ensures that the fields containing these terms will be easily comparable. For example, the ‘genome_molecule’ table contains the following permissible values for ‘molecule_type’: ‘ambisense ssRNA’, ‘dsRNA’, ‘negative-sense ssRNA’, ‘positive-sense ssRNA’, ‘ssDNA’, and ‘dsDNA’. A particular viral genome must then have one of these values entered into the ‘molecule_type’ field
of the 'genome' table, since this field is a foreign key to the 'molecule_type' primary key of the 'genome_molecule' table. Entering 'double-stranded DNA' would not be permissible.

**Annotation**

Raw data obtained directly from high-throughput analytical techniques such as automated sequencing, protein interaction, or microarray experiments contain little-to-no information as to the content or meaning. The process of adding value to the raw data to increase the knowledge content is known as annotation and curation. As an example, the results of a microarray experiment may provide an indication that individual genes are up- or downregulated under certain experimental conditions. By annotating the properties of those genes, we are able to see that certain sets of genes showing coordinated regulation are a part of common biological pathways. An important pattern then emerges that was not discernable solely by inspection of the original data. The annotation process consists of a semiautomated analysis of the information content of the data and provides a variety of descriptive features that aid the process of assigning meaning to the data. The investigator is then able to use this analytical information to more closely inspect the data during a manual curation process that might support the reconstruction of gene expression or protein interaction pathways, or allow for the inference of functional attributes of each identified gene. All of this curated information can then be stored back in the database and associated with each particular gene.

For each piece of information associated with a gene (or other biological entity) during the process of annotation and curation, it is always important to provide the evidence used to support each assignment. This evidence may be described in a Standard Operating Procedure (SOP) document which, much like an experimental protocol, details the annotation process and includes a description of the computer algorithms, programs, and analysis pipelines that were used to compile that information. Each piece of information annotated by the use of this pipeline might then be coded, for example, 'IEA: Inferred from Electronic Annotation'. For information obtained from the literature during manual curation, the literature reference from which the information was obtained should always be provided along with a code that describes the source of the information. Some of the possible evidence codes include 'IDA: Inferred from Direct Assay', 'IGI: Inferred from Genetic Interaction', 'IMP: Inferred from Mutant Phenotype', or 'ISS: Inferred from Sequence or Structural Similarity'. These evidence codes are taken from a list provided by the Gene Ontology (GO) Consortium (see below) and as such represent a controlled vocabulary that any data curator can use and that will be understood by anyone familiar with the GO database. This controlled evidence vocabulary is stored in the 'evidence' table, and each record in every one of the gene properties tables is assigned an evidence code noting the source of the annotation/curation data.

As indicated above, the use of controlled vocabularies (ontologies) to describe the attributes of biological data is extremely important. It is only through the use of these controlled vocabularies that a consistent, documented approach can be taken during the annotation/curation process. And while there may be instances where creating your own ontology may be necessary, the use of already available, community-developed ontologies ensures that the ontological descriptions assigned to your database will be understood by anyone familiar with the public ontology. Use of these public ontologies also ensures that they support comparative analyses with other available databases that also make use of the same ontological descriptions. The GO Consortium provides one of the most extensive and widely used controlled vocabularies available for biological systems. GO describes biological systems in terms of their biological processes, cellular components, and molecular functions. The GO effort is community-driven, and any scientist can participate in the development and refinement of the GO vocabulary. Currently, GO contains a number of terms specific to viral processes, but these tend to be oriented toward particular viral families, and may not necessarily be the same terms used by investigators in other areas of virology. Therefore it is important that work continues in the virus community to expand the availability and use of GO terms relevant to all viruses. GO is not intended to cover all things biological. Therefore, other ontologies exist and are actively being developed to support the description of many other biological processes and entities. For example, GO does not describe disease-related processes or mutants; it does not cover protein structure or protein interactions; and it does not cover evolutionary processes. A complementary effort is under way to better organize existing ontologies, and to provide tools and mechanisms to develop and catalog new ontologies. This work is being undertaken by the National Center for Biomedical Ontologies, located at Stanford University, with participants worldwide.

**Access (Searching for Information)**

The most comprehensive, well-designed database is useless if no method has been provided to access that database, or if access is difficult due to a poorly designed application. Therefore, providing a search interface that meets the needs of intended users is critical to fully realizing the potential of any effort at developing a comprehensive database. Access can be provided using a number of different methods ranging from direct query of the database using the relatively standardized 'structured query language' (SQL), to customized applications designed to provide the ability to ask sophisticated questions regarding the data.
contained in the database and mine the data for meaningful patterns. Web pages may be designed to provide simple-to-use forms to access and query data stored in an RDBMS.

Using the VGD schema as a data source, one example of an SQL query might be to find the gene_id and name of all the proteins in the database that have a molecular weight between 20,000 and 30,000, and also have at least one transmembrane region.

Many database providers also provide users with the ability to download copies of the database so that these users may analyze the data using their own set of analytical tools.

**Output (Utilizing Information)**

When a user queries a database using any of the available access methods, the results of that query are generally provided in the form of a table where columns represent fields in the database and the rows represent the data from individual database records. Tabular output can be easily imported into spreadsheet applications, sorted, manipulated, and reformatted for use in other applications. But while extremely flexible, tabular output is not always the best format to use to fully understand the underlying data and the biological implications. Therefore, many applications that connect to databases provide a variety of visualization tools that display the data graphically, showing patterns in the data that may be difficult to discern using text-based output. An example of one such visual display is provided in **Figure 3** and shows conservation of synteny between the genes of two different poxvirus species. The information used to generate this figure comes directly from the data provided in the VGD. Every gene in the two viruses (in this case crocodilepox virus and molluscum contagiosum virus) has been compared to every other gene using the BLAST search program. The results of this search are stored in the BLAST tables of the VGD. In addition, the location of each gene within its respective genomic sequence is stored in the ‘gene_segment’ table. This information, once extracted from the database server, is initially text but it is then submitted to a program running on the server that reformats the data and creates a graph. In this manner, it is much easier to visualize the series of points formed along a diagonal when there are a series of similar genes with similar genomic locations present in each of the two viruses. These data sets may contain gene synteny patterns that display deletion, insertion, or recombination events during the course of viral evolution. These patterns can be difficult to detect with text-based tables, but are easy to discern using visual displays of the data.

Information provided to a user as the result of a database query may contain data derived from a combination of sources, and displayed using both visual and textual methods.

**Figure 3** Gene synteny plot. A comparison of gene sequence and genomic position conservation between crocodilepox virus (horizontal axis) and molluscum contagiosum virus (vertical axis). All predicted proteins encoded by each virus were compared to each other using BLASTP. Each pair of proteins showing some measure of similarity, as determined by a BLAST expect (E) value \(<0.00001\), was plotted according to the location of each gene on each respective genome. The color of the points reflects the identity of the coding strand of each gene. Black points along either of the two axes represent proteins unique to that genome.
feedback. Figure 4 shows the web-based output of a query designed to display information related to a particular virus gene. The top of this web page displays the location of the gene on the genome visually showing surrounding genes on a partial map of the viral genome. Basic gene information such as genome coordinates, gene name, and the nucleotide and amino acid sequence are also provided. This information was originally obtained from the original GenBank record and then stored in the VGD database. Data added as the result of an automated annotation pipeline are also displayed. This includes calculated values for molecular weight and pI; amino acid composition; functional motifs; BLAST similarity searches; and predicted protein structural properties such as transmembrane domains, coiled-coil regions, and signal sequences. Finally, information obtained from a manual curation of the gene through an extensive literature search is also displayed. Curated information includes a mini review of gene function; experimentally determined gene properties such as molecular weight, pI, and protein structure; alternative names and aliases used in the literature; assignment of ontological terms describing gene function; the availability of reagents such as antibodies and clones; and also, as available, information on the functional effects of mutations. All of the information to construct the web page for this gene is directly provided as the result of a single database query. (The tables storing the manually curated gene information are not shown in Figure 2.) Obviously, compiling the data and entering it into the database required a substantial amount of effort, both computationally and manually; however, the information is now much more available and useful to the research scientist.

Errors

No discussion of databases would be complete without considering errors. As in any other scientific endeavor, the data we generate, the knowledge we derive from the data, and the inferences we make as a result of the analysis of the data are all subject to error. These errors can be introduced at many points in the analytical chain. The original data may be faulty: using sequence data as one example, nucleotides in a DNA sequence may have been misread or miscalled, or someone may even have mistyped the sequence. The database may have been poorly designed; a field in a table designed to hold sequence information may have been set to hold only 2000 characters, whereas the sequences imported into that field may be longer than 2000 nucleotides. The sequences would have then been automatically truncated to 2000 characters, resulting in the loss of data. The curator may have mistyped an Enzyme Commission (EC) number for an RNA polymerase, or may have incorrectly assigned a genomic sequence to the wrong taxonomic classification. Or even more insidious, the curator may have been using annotations provided by other groups that had justified their own annotations on the basis of matches to annotations provided by yet another group. Such chains of evidence may extend far back, and the chance of propagating an early error increases with time. Such error propagation can be widespread indeed, affecting the work of multiple sequencing centers and database creators and providers. This is especially true given the dependencies of genomic sequence annotations on previously published annotations. The possible sources of errors are numerous, and it is the responsibility of both the database provider and the user to be aware of, and on the lookout for, errors. The database provider can, with careful database and
application design, apply error-checking routines to many aspects of the data storage and analysis pipeline. The code can check for truncated sequences, interrupted open reading frames, and nonsense data, as well as data annotations that do not match a provided controlled vocabulary. But the user should always approach any database or the output of any application with a little healthy skepticism. The user is the final arbiter of the accuracy of the information, and it is their responsibility to look out for inconsistent or erroneous results that may indicate either a random or systemic error at some point in the process of data collection and analysis.

**Virus Databases**

It is not feasible to provide a comprehensive and current list of all available databases that contain virus-related information or information of use to virus researchers. New databases appear on a regular basis; existing databases either disappear or become stagnant and outdated; or databases may change focus and domains of interest. Any resource published in book format attempting to provide an up-to-date list would be out-of-date on the day of publication. Even web-based lists of database resources quickly become out-of-date due to the rapidity with which available resources change, and the difficulty and extensive effort required to keep an online list current and inclusive. Therefore, our approach in this article is to provide an overview of the types of data that are obtainable from available biological databases, and to list some of the more important database resources that have been available for extended periods of time and, importantly, remain current through a process of continual updating and refinement. We should also emphasize that the use of web-based search tools such as Google, various web logs (Blogs), and news groups, can provide some of the best means of locating existing and newly available web-based information sources. Information contained in databases can be used to address a wide variety of problems. A sampling of the areas of research facilitated by virus databases includes

- taxonomy and classification;
- host range, distribution, and ecology;
- evolutionary biology;
- pathogenesis;
- host–pathogen interaction;
- epidemiology;
- disease surveillance;
- detection;
- prevention;
- prophylaxis;
- diagnosis; and
- treatment.

Addressing these problems involves mining the data in an appropriate database in order to detect patterns that allow certain associations, generalizations, cause–effect relationships, or structure–function relationships to be discerned. Table 1 provides a list of some of the more useful and stable database resources of possible interest to virus researchers. Below, we expand on some of this information and provide a brief discussion concerning the sources and intended uses of these data sets.

**Data**

**Major repositories of biological information**

The two major, overarching collections of biological databases are at the NCBI, supported by the National Library of Medicine at the NIH, and the EMBL, part of the European Bioinformatics Institute. These large data repositories try to be all-inclusive, acting as the primary source of publicly available molecular biological data for the scientific community. In fact, most journals require that, prior to publication, investigators submit their original sequence data to one of these repositories. In addition to sequence data, NCBI and EMBL (along with many other data repositories) include a large variety of other data types, such as that obtained from gene-expression experiments and studies investigating biological structures. Journals are also extending the requirement for data deposition to some of these other data types. Note that while much of the data available from these repositories is raw data obtained directly as the result of experimental investigation in the laboratory, a variety of ‘value-added’ secondary databases are also available that take primary data records and manipulate or annotate them in some fashion in order to derive additional useful information.

When an investigator is unsure about the existence or source of some biological data, the NCBI and EMBL websites should serve as the starting point for locating such information. The NCBI Entrez Search Engine provides a powerful interface to access all information contained in the various NCBI databases, including all available sequence records. A search engine such as Google might also be used if NCBI and EMBL fail to locate the desired information. Of course PubMed, the repository of literature citations maintained at NCBI, also represents a major reference site for locating biological information. Finally, the journal *Nucleic Acids Research* (NAR) publishes an annual ‘database’ issue and an annual ‘web server’ issue that are excellent references for finding new biological databases and websites. And while the most recent NAR database or web server issue may contain articles on a variety of new and interesting databases and websites, be sure to also look at issues from previous years. Older issues contain articles on many existing sites that may not necessarily be represented in the latest
| Information resource name | Sponsor | URL | Description |
|---------------------------|---------|-----|-------------|
| **Indices of databases**  |         |     |             |
| All the Virology on the WWW | David M. Sander; Tulane University | http://www.virology.net/ | Collection of virology information on the Internet |
| Nucleic Acids Research 2006 Database Issue | Nucleic Acids Research Journal | http://nar.oxfordjournals.org/content/vol34/suppl_1/index.dtl | Issue describing molecular biology databases |
| Nucleic Acids Research 2006 Web Server Issue | Nucleic Acids Research Journal | http://nar.oxfordjournals.org/content/vol34/suppl_2/index.dtl | Issue describing molecular biology websites |
| **General sites**         |         |     |             |
| NCBI                      | National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health (NIH) | http://www.ncbi.nlm.nih.gov/ | Resource for molecular biology information |
| PubMed                    | National Library of Medicine, NIH, NCBI | http://www.ncbi.nlm.nih.gov/Genbank/ | Biomedical literature database |
| GenBank Sequence Database | NCBI | http://www.ncbi.nlm.nih.gov/Genbank/ | Annotated collection of publicly available nucleotide sequences |
| NCBI Trace Archive        | NCBI | http://www.ncbi.nlm.nih.gov/Traces/ | NCBI repository of sequence trace files |
| NCBI Assembly Archive     | NCBI | http://www.ncbi.nlm.nih.gov/Traces/assembly/ | NCBI repository of sequence assembly files |
| EMBL Sequence Database    | European Molecular Biology Laboratory (EMBL) | http://www.ebi.ac.uk/embl/ | Annotated collection of publicly available nucleotide sequences |
| DDBJ Sequence Database    | DNA Data Bank of Japan | http://www.ddbj.nig.ac.jp/ | Annotated collection of publicly available nucleotide sequences |
| UniProt (Universal Protein Resource) | EMBL; Protein Information Resource, Georgetown University; Swiss Institute of Bioinformatics | http://www.ebi.ac.uk/uniprot/ | Protein sequence repository |
| **Taxonomy**              |         |     |             |
| ICTVnet                   | International Committee on Taxonomy of Viruses (ICTV) | http://www.danforthcenter.org/ltab/ictvnet/ | ICTV official website |
| ICTVdb                    | ICTV | http://www.ncbi.nlm.nih.gov/ICTVdb | ICTV database of virus taxonomy |
| NCBI Taxonomy Browser     | NCBI | http://www.ncbi.nlm.nih.gov/Taxonomy/ | Taxonomy-based retrieval of sequence data |
| Virus-oriented            |         |     |             |
| NCBI Viral Genomes Resource | NCBI | http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html | NCBI resources for the study of viruses |
| ASV                      | American Society for Virology (ASV) | http://www.asv.org/ | Society to promote the exchange of information and stimulate discussion and collaboration among scientists active in all aspects of virology |
| IUUMS-Virology            | International Union of Microbiological Societies | http://www.iums.org/divisions/divisions-virology.html | Society to promote the study of microbiological sciences internationally |

Continued
| Information resource name | Sponsor | URL | Description |
|---------------------------|---------|-----|-------------|
| Viruses: From Structure to Biology | ASV; Washington University School of Medicine | [http://medicine.wustl.edu/~virology/](http://medicine.wustl.edu/~virology/) | Historical overview of virus research |
| BRCs (Bioinformatics Resource Centers for Biodefense and Emerging or Re-Emerging Infectious Diseases) | National Institute of Allergy and Infectious Diseases (NIAID), NIH | [http://www.brc-central.org/](http://www.brc-central.org/) | NIAID/NIH-funded Bioinformatics Resource Centers providing web-based genomics resources to the scientific community on viruses in category A, B, and C priority pathogens |
| VBRC (Viral Bioinformatics Resource Center) | University of Alabama at Birmingham; NIAID | [http://www.vbrc.org/](http://www.vbrc.org/) | Bioinformatics resources directed at Arenaviridae, Bunyaviridae, Filoviridae, Paramyxoviridae, Poxviridae, and Togaviridae |
| PATRIC (PathoSystems Resource Integration Center) | University of Victoria, Canada Virginia Bioinformatics Institute; NIAID | [http://www.virology.ca/](http://www.virology.ca/) | Bioinformatics resources directed at Caliciviridae, Coronaviridae, hepatitis A virus, hepatitis E virus, Lyssaviridae |
| VirOligo | Oklahoma State University | [http://viroligo.okstate.edu/](http://viroligo.okstate.edu/) | Virus-specific oligonucleotides for PCR and hybridization |
| VirGen | Bioinformatics Centre, University of Pune, India | [http://bioinfo.ernet.in/virgen/](http://bioinfo.ernet.in/virgen/) | Annotated and curated database for complete viral genome sequences |
| VIDE (Virus Identification Data Exchange) | Australian Centre for International Agricultural Research; University of Idaho | [http://image.fs.uidaho.edu/vide/](http://image.fs.uidaho.edu/vide/) | Online descriptions and lists of plant viruses |
| RNAs and Proteins of dsRNA Viruses | Institute for Animal Health, UK Biotechnology and Biological Sciences Research Council | [http://www.iah.bbsrc.ac.uk/dsRNA_virus_proteins/](http://www.iah.bbsrc.ac.uk/dsRNA_virus_proteins/) | Databases for the study of Reoviridae, Cystoviridae, Birnaviridae, Totiviridae |
| euHCVdb (European HCV Database) | Institute for the Biology and Chemistry of Proteins, CNRS and Lyon University | [http://euhcvdb.ibcp.fr/](http://euhcvdb.ibcp.fr/) | European hepatitis C virus database |
| HCV Database Project | Los Alamos National Laboratories; NIAID | [http://hcv.lanl.gov/](http://hcv.lanl.gov/) | HCV sequence and immunology database |
| Hepatitis Virus Database | National Institute of Genetics (Japan) | [http://is2as02.genes.nig.ac.jp/](http://is2as02.genes.nig.ac.jp/) | Database for the study of hepatitis B, C, and E |
| HERVd (Human Endogenous Retrovirus Database) | Institute of Molecular Genetics, Academy of Sciences of The Czech Republic | [http://herv.imag.cas.cz/](http://herv.imag.cas.cz/) | Human endogenous retrovirus database |
| HIV Drug Resistance Database | Stanford University | [http://hivdb.stanford.edu/](http://hivdb.stanford.edu/) | HIV drug resistance database |
| HIV Sequence Database | Los Alamos National Laboratories; NIAID | [http://www.hiv.lanl.gov/](http://www.hiv.lanl.gov/) | HIV sequence, resistance, immunology, and vaccine trials databases |
| HIV-1, Human Protein Interaction Database | Southern Research Institute, Birmingham, Alabama; NCBI | [http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/](http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/) | Interactions of HIV-1 proteins with those of the host cell |
| Database Name                                      | Website                                                                 | Description                                                                 |
|---------------------------------------------------|-------------------------------------------------------------------------|------------------------------------------------------------------------------|
| BioHealthBase (Influenza)                         | http://www.biohealthbase.org/                                           | Bioinformatics resources directed at influenza                               |
| Influenza Sequence Database                       | http://www.flu.lanl.gov/                                                | Influenza sequence database                                                  |
| Influenza Virus Resource Database                 | http://www.ncbi.nlm.nih.gov/genomes/FLU/                                | Influenza sequence database                                                  |
| IVDB (Influenza Virus Database)                   | http://influenza.genomics.org.cn/                                       | Integrated information resource and analysis platform for influenza virus     |
| Influenza Genome Sequencing Project               | http://www.niaid.nih.gov/dmid.genomes/mscs/influenza.htm               | NIAID influenza sequencing project                                            |
| Poxvirus Bioinformatics Resource Center          | http://www.poxvirus.org/                                                | Poxvirus genomic sequences, gene annotations, and analysis                  |
| SARS-CoV RNA SSS Database                         | http://www.liuweibo.com/sarsdb/                                         | SARS Secondary Structural Sequence database                                  |
| ACLAME (A Classification of Genetic Mobile Elements)| http://aclame.uib.ac.be/                                                | Database of mobile genetic elements                                           |
| Influenza Genome Conformation and Bioinformatics  | University of Ottawa                                                   | Viroids and viroid-like RNAs                                                |
| Subviral RNA Database Genomes of the T4-like phages| http://subviral.med.uottawa.ca/                                         | Sequences of T4-like bacteriophages                                          |
| DPhWeb (Descriptions of plant viruses)            | http://phage.bioc.tulane.edu/                                           | Plant virus database                                                        |
| ILTAB (International Laboratory for Tropical Agricultural Biotechnology) | http://www.danforthcenter.org/iltab/                                   | Caulimoviridae, Geminiviridae, and Potyviridae database                     |
| Functional motifs                                 | European Bioinformatics Institute (EBI)                                 | Database of protein families, domains, and functional sites                  |
| CDD (Conserved Domain Database)                   | http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml                    | Conserved Domain Database and search service                                 |
| PFAM (Protein families)                           | http://pfam.wustl.edu/                                                 | Database of protein families and Hidden Markov models                        |
| PRINTS (Protein fingerprints)                     | http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/                          | Conserved motifs used to characterize a protein family                      |
| PROSITE                                            | http://ca.expasy.org/prosite/                                          | Database of protein families and domains                                      |
| VIDA (Virus database of homologous protein families) | http://www.biochem.ucl.ac.uk/bsm/virus_database/                      | Homologous protein families from complete and partial virus genomes          |
| VOCI (Viral Orthologous Clusters)                 | http://athena.bioc.uvic.ca/workbench.php?tool=vocs                   | Viral orthologous clusters database                                          |
| Information resource name               | Sponsor                                                                 | URL                                                                 | Description                                                                 |
|----------------------------------------|-------------------------------------------------------------------------|----------------------------------------------------------------------|----------------------------------------------------------------------------|
| VOGs (Viral Clusters of Orthologous Groups) | NCBI                                                                    | http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/vog.html                | Viral clusters of orthologous groups database                               |
| Structural                              |                                                                         |                                                                      |                                                                            |
| PDB (Protein Data Bank)                | Research Collaboratory for Structural Bioinformatics                    | http://www.rcsb.org/                                                 | Three-dimensional structure database                                       |
| Macromolecular Structure Database       | EBI                                                                     | http://www.ebi.ac.uk/msd/                                           | Macromolecular structure database                                          |
| Dali                                    | Institute of Biotechnology, University of Helsinki, Finland             | http://ekhidna.biocenter.helsinki.fi/dali/                          | Database of 3D structure comparisons                                       |
| The Big Picture Book of Viruses        | David M. Sander; Tulane University                                      | http://www.virology.net/Big_Virology/                               | Catalog of virus pictures on the Internet                                   |
| VIPERdb (Virus particle explorer)      | Scripps Research Institute Structural Biology                           | http://viperdb.scripps.edu/                                         | Icosahedral virus capsid structures                                         |
| Virus World                             | Institute for Molecular Virology, University of Wisconsin-Madison       | http://virology.wisc.edu/virusworld/                                 | Images of virus capsids from X-ray crystallography and cryo-EM data         |
| Pathway                                 |                                                                         |                                                                      |                                                                            |
| Reactome                                | Cold Spring Harbor Laboratory, EBI, Gene Ontology Consortium           | http://www.reactome.org/                                             | Curated knowledgebase of biological pathways                               |
| BioCyc                                 | SRI International, Menlo Park, California                              | http://www.biocyc.org/                                              | Collection of databases describing genome and metabolic pathways            |
| KEGG (Kyoto Encyclopedia of Genes and Genomes) | Kyoto University and the Human Genome Center at University of Tokyo | http://www.genome.jp/kegg/                                          | Database of biological pathways and genes                                   |
| Other                                   |                                                                         |                                                                      |                                                                            |
| GO (Gene Ontology Database)             | Gene Ontology Consortium                                               | http://www.geneontology.org/                                       | Development of controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components, and molecular functions |
| National Center for Biomedical Ontology | Stanford University                                                    | http://bioontology.org/                                             | Development of innovative technologies and methods that allow scientists to create, disseminate, and manage biomedical ontologies |
| GEO (Gene Expression Omnibus)           | NCBI                                                                    | http://www.ncbi.nlm.nih.gov/geo/                                    | A gene expression/molecular abundance data repository                       |
| Virus Databases |
|----------------|
| IEDB (Immune Epitope Database and Analysis Resource) | La Jolla Institute for Allergy and Immunology | http://www.immuneepitope.org/ | Dissemination of immune epitope information |
| IntAct (Protein interactions) | EBI | http://www.ebi.ac.uk/intact/ | Database and analysis system for protein interactions |
| Analytical tools | | | |
| ExPASy | Swiss Institute of Bioinformatics (SIB) | http://ca.expasy.org/ | Tools for the analysis of protein sequences, structures, etc. |
| PASC (Pairwise Sequence Comparison) | NCBI | http://www.ncbi.nlm.nih.gov/sutils/pasc/ | Analysis of pairwise identity distribution within viral families |
| NCBI Viral Genotyping Tool | NCBI | http://www.ncbi.nlm.nih.gov/projects/genotyping/ | Identification of viral sequence genotypes |
| Medically oriented | | | |
| AIDSinfo | US Department of Health and Human Services | http://aidsinfo.nih.gov/ | Information on HIV/AIDS treatment, prevention, and research |
| AMA Infectious diseases | American Medical Association | http://www.ama-assn.org/ama/pub/category/1797.html | Online information resources |
| CDC NCID (National Center for Infectious Diseases) | US Centers for Disease Control and Prevention (CDC) | http://www.cdc.gov/ncidod/diseases/ | Centers for Disease Control and Prevention (USA) disease index |
| CI PHI Fact Sheets | Canadian Institute of Public Health Inspectors | http://action.web.ca/home/ciphiot/readingroom.shtml | Information pages for public health, including viral diseases |
| CSEI (Center for the Study of Emerging Infections) | Institute for BioSecurity, St Louis University, School of Public Health | http://www.emerginginfections.slu.edu/ | Information on emerging infections |
| Diseases & Conditions | New York State Department of Health | http://www.health.state.ny.us/diseases/ | General public information |
| ENIVD (European Network for Diagnostics of "Imported" Viral Diseases) | European Commission | http://www.enivd.org/ | European Network for Diagnostics of "Imported" Viral Diseases |
| EPR (Epidemic and Pandemic Alert and Response) | World Health Organization (WHO) | http://www.who.int/csr/disease/en/ | WHO integrated global alert and response system for epidemics and other public health emergencies |
| European Centre for Disease Prevention and Control | The European Union | http://www.ecdc.eu.int/ | European Centre for Disease Prevention and Control |
| MedScape | WebMD | http://www.medscape.com/ | Information and educational tools for specialists |
| Merck Manual of Diagnosis and Therapy, Infectious Diseases | Merck & Co., Inc. | http://www.merck.com/mrkshared/rmanual/section13/sec13.jsp | Online reference of infectious diseases |
| MMWR (Morbidity and Mortality Weekly Report) | CDC | http://www.cdc.gov/mmwr/ | CDC Morbidity and Mortality Weekly Report |
| NIH Vaccine Research Center | Dale and Betty Bumpers' Vaccine Research Center, NIAID, NIH | http://www.vrc.nih.gov/ | NIH vaccine research |
Taxonomy and classification

One of the strengths of storing information within a database is that information derived from different sources or different data sets can be compared so that important common and distinguishing features can be recognized. Such comparative analyses are greatly aided by having a rigorous classification scheme for the information being studied. The International Union of Microbiological Societies has designated the International Committee on Taxonomy of Viruses (ICTV) as the official body that determines taxonomic classifications for viruses. Through a series of subcommittees and associated study groups, scientists with expertise on each viral species participate in the establishment of new taxonomic groups, assignment of new isolates to existing or newly established taxonomic groups, and reassessment of existing assignments as additional research data become available. The ICTV uses more than 2600 individual characteristics for classification, though sequence homology has gained increasing importance over the years as one of the major classifiers of taxonomic position. Currently, as described in its Eighth Report, the ICTV recognizes 3 orders, 73 families, 287 genera, and 1950 species of viruses. The ICTV officially classifies viral isolates only to the species level. Divisions within species, such as clades, subgroups, strains, isolates, types, etc., are left to others. The ICTV classifications are available in book form as well as from an online database. This database, the ICTVdb, contains the complete taxonomic hierarchy, and assigns each known viral isolate to its appropriate place in that hierarchy. Descriptive information on each viral species is also available. The NCBI also provides a web-based taxonomy browser for access to taxonomically specified sets of sequence records. NCBI’s viral taxonomy is not completely congruent with that of ICTV, but efforts have been under way to ensure congruency with the official ICTV classification.

Nucleotide sequence data

The primary repositories of existing sequence information come from the three organizations that comprise the International Nucleotide Sequence Database Collaboration. These three sites are GenBank (maintained at NCBI), EMBL, and the DNA Data Bank of Japan (DDBJ). Because all sequence information submitted to any one of these sites is shared with the others, a researcher need query only one of these sites to get the most up-to-date set of available sequences. GenBank stores all publicly available nucleotide sequences for all organisms, as well as viruses. This includes whole-genome sequences as well as partial-genome and individual coding sequences. Sequences are also available from large-scale sequencing projects, such as those from shotgun sequencing of environmental samples (including viruses), and high-throughput low- and high-coverage genomic sequencing projects. NCBI provides separate database divisions for access to these sequence datasets. The sequence provided in each GenBank record is the distillation of the raw data generated by (in most cases these days) automated sequencing machines. The trace files and base calls provided by the sequencers are then assembled into a collection of contiguous sequences (contigs) until the final sequence has been assembled. In recognition of the fact that there is useful information contained in these trace files and sequence assemblies (especially if one would like to look for possible sequencing errors or polymorphisms), NCBI now provides separate Trace File and Assembly Archives for GenBank sequences when the laboratory responsible for generating the sequence submits these files. Currently, the only viruses represented in these archives are influenza A, chlorella, and a few bacteriophages.

An important caveat in using data obtained from GenBank or other sources is that no sequence data can be considered to be 100% accurate. Furthermore, the annotation associated with the sequence, as provided in the GenBank record, may also contain inaccuracies or be out-of-date. GenBank records are provided and maintained by the group originally submitting the sequence to GenBank. GenBank may review these records for obvious errors and formatting mistakes (such as the lack of an open reading frame where one is indicated), but given the large numbers of sequences being submitted, it is impossible to verify all of the information in these records. In addition, the submitter of a sequence essentially ‘owns’ that sequence record and is thus responsible for all updates and corrections. NCBI generally will not change any of the information in the GenBank record unless the sequence submitter provides the changes. In some cases, sequence annotations will be updated and expanded, but many, if not most, records never change following their initial submission. (These facts emphasize the responsibility that submitters of sequence data have to ensure the accuracy of their
original submission and to update their sequence data and annotations as necessary.) Therefore, the user of the information has the responsibility to ensure, to the extent possible, its accuracy is sufficient to support any conclusions derived from that information. In recognition of these problems, NCBI established the Reference Sequence (RefSeq) database project, which attempts to provide reference sequences for genomes, genes, mRNAs, proteins, and RNA sequences that can be used, in NCBI's words, as "a stable reference for gene characterization, mutation analysis, expression studies, and polymorphism discovery". RefSeq records are manually curated by NCBI staff, and therefore should provide more current (and hopefully more accurate) sequence annotations to support the needs of the research community. For viruses, RefSeq provides a complete genomic sequence and annotation for one representative isolate of each viral species. NCBI solicits members of the research community to participate as advisors for each viral family represented in RefSeq, in an effort to ensure the accuracy of the RefSeq effort.

**Protein sequence data**

In addition to the nucleotide sequence databases mentioned above, UniProt provides a general, all-inclusive protein sequence database that adds value through annotation and analysis of all the available protein sequences. UniProt represents a collaborative effort of three groups that previously maintained separate protein databases (PIR, SwissProt, and TrEMBL). These groups, the National Biomedical Research Foundation at Georgetown University, the Swiss Institute of Bioinformatics, and the European Bioinformatics Institute, formed a consortium in 2002 to merge each of their individual databases into one comprehensive database, UniProt. UniProt data can be queried by searching for similarity to a query sequence, or by identifying useful records based on the text annotations. Sequences are also grouped into clusters based on sequence similarity. Similarity of a query sequence to a particular cluster may be useful in assigning functional characteristics to sequences of unknown function. NCBI also provides a protein sequence database (with corresponding RefSeq records) consisting of all protein-coding sequences that have been annotated within all GenBank nucleotide sequence records.

**Virus-specific sequence databases**

The above-mentioned sequence databases are not limited to viral data, but rather store sequence information for all biological organisms. In many cases, access to nonviral sequences is necessary for comparative purposes, or to study virus–host interactions. But it is frequently easier to use virus-specific databases when they exist, to provide a more focused view of the data that may simplify many of the analyses of interest. Table 1 lists many of these virus-specific sites. Sites of note include the NIH-supported Bioinformatics Resource Centers for Biodefense and Emerging and Reemerging Infectious Diseases (BRCs). The BRCs concentrate on providing databases, annotations, and analytical resources on NIH priority pathogens, a list that includes many viruses. In addition, the LANL has developed a variety of viral databases and analytical resources including databases focusing on HIV and influenza. For plant virologists, the Descriptions of Plant Viruses (DPV) website contains a comprehensive database of sequence and other information on plant viruses.

**Structural information**

The three-dimensional structures for quite a few viral proteins and virion particles have been determined. These structures are available in the primary database for experimentally determined structures, the Protein Data Bank (PDB). The PDB currently contains the structures for more than 650 viral proteins and viral protein complexes out of 38,000 total structures. Several virus-specific structure databases also exist. These include the VIPERdb database of icosahedral viral capsid structures, which provides analytical and visualization tools for the study of viral capsid structures; Virus World at the Institute for Molecular Virology at the University of Wisconsin, which contains a variety of structural images of viruses; and the Big Picture Book of Viruses, which provides a catalog of images of viruses, along with descriptive information.

**Functional motifs and orthologous clusters**

Ultimately, the biology of viruses is determined by genomic sequence (with a little help from the host and the environment). Nucleotide sequences may be structural, functional, regulatory, or protein coding. Protein sequences may be structural, functional, and/or regulatory, as well. Patterns specified in nucleotide or amino acid sequences can be identified and associated with many of these biological roles. Both general and virus-specific databases exist that map these roles to specific sequence motifs. Most also provide tools that allow investigators to search their own sequences for the presence of particular patterns or motifs characteristic of function. General databases include the NCBI Conserved Domain Database; the Pfam (protein family) database of multiple sequence alignments and hidden Markov models; and the PROSITE database of protein families and domains. Each of these databases and associated search algorithms differ in how they detect a particular search motif or define a particular protein family. It can therefore be useful to employ multiple databases and search methods when analyzing a new sequence (though in many cases they will each detect a similar set of putative functional motifs). InterPro is a database of protein families, domains, and functional sites that combines many other existing motif databases.
provides a search tool, InterProScan, which is able to utilize several different search algorithms dependent on the database to be searched. It allows users to choose which of the available databases and search tools to use when analyzing their own sequences of interest. A comprehensive report is provided that not only summarizes the results of the search, but also provides a comprehensive annotation derived from similarities to known functional domains. All of the above databases define functional attributes based on similarities in amino acid sequence. These amino acid similarities can be used to classify proteins into functional families. Placing proteins into common functional families is also frequently performed by grouping the proteins into orthologous families based on the overall similarity of their amino acid sequence as determined by pairwise BLAST comparisons. Two virus-specific databases of orthologous gene families are the Viral Clusters of Orthologous Groups database (VOGs) at NCBI, and the Viral Orthologous Clusters database (VOCs) at the Viral Bioinformatics Resource Center and Viral Bioinformatics, Canada.

Other information
Many other types of useful information, both general and virus-specific, have been collected into databases that are available to researchers. These include databases of gene-expression experiments (NCBI Gene Expression Omnibus – GEO); protein–protein interaction databases, such as the NCBI HIV Protein–Interaction Database; The Immune Epitope Database and Analysis Resource (IEDB) at the La Jolla Institute for Allergy and Immunology; and databases and resources for defining and visualizing biological pathways, such as metabolic, regulatory, and signaling pathways. These pathway databases include Reactome at the Cold Spring Harbor Laboratory, New York; BioCyc at SRI International, Menlo Park, California; and the Kyoto Encyclopedia of Genes and Genomes (KEGG) at Kyoto University in Japan.

Analytical Tools
As indicated above, the information contained in a database is useless unless there is some way to retrieve that information from the database. In addition, having access to all of the information in every existing database would be meaningless unless tools are available that allow one to process and understand the data contained within those databases. Therefore, a discussion of virus databases would not be complete without at least a passing reference to the tools that are available for analysis. To populate a database such as the VGD with sequence and analytical information, and to utilize this information for subsequent analyses, requires a variety of analytical tools including programs for

- sequence record reformatting,
- database import and export,
- sequence similarity comparison,
- gene prediction and identification,
- detection of functional motifs,
- comparative analysis,
- multiple sequence alignment,
- phylogenetic inference,
- structural prediction, and
- visualization.

Sources for some of these tools have already been mentioned, and many other tools are available from the same websites that provide many of the databases listed in Table 1. The goal of all of these sites that make available data and analytical tools is to provide – or enable the discovery of – knowledge, rather than simply providing access to data. Only in this manner can the ultimate goal of biological understanding be fully realized.

See also: Evolution of Viruses; Phylogeny of Viruses; Taxonomy, Classification and Nomenclature of Viruses; Virus Classification by Pairwise Sequence Comparison (PASC).

Further Reading
Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25(1): 25–29.
Bao Y, Federhen S, Leipe D, et al. (2004) National Center for Biotechnology Information Viral Genomes Project. Journal of Virology 78(14): 7291–7296.
Fauquet CM, Mayo MA, Maniloff J, Desselberger U, and Ball LA (2005) Virus Taxonomy: Classification and Nomenclature of Viruses. Eighth Report of the International Committee on Taxonomy of Viruses. San Diego, CA: Elsevier Academic Press.
Galperin MY (2006) The Molecular Biology Database Collection: 2006 update. Nucleic Acids Research 34(database issue): D3–D5.
Joshi-Tope G, Gillespie M, Vastrik I, et al. (2005) Reactome: A knowledgebase of biological pathways. Nucleic Acids Research 33(database issue): D429–D432.
Kellam P and Alba MM (2002) Virus bioinformatics: Databases and recent applications. Applied Bioinformatics 1(1): 37–42.
Korber B, LaButé M, and Yusim K (2006) Immunoinformatics comes of age. PLoS Computational Biology 2(6): e71.
Kuiken C, Korber B, and Shafer RW (2003) HIV sequence databases. AIDS Reviews 5(1): 52–61.
Kuiken C, Mizokami M, Deleage G, et al. (2006) Hepatitis C databases, principles and utility to researchers. Hepatology 43(5): 1157–1165.
Leffkowitz EJ, Upton C, Changayil SS, et al. (2005) Poxvirus bioinformatics resource center: A comprehensive Poxviridae informational and analytical resource. Nucleic Acids Research 33(database issue): D311–D336.
Lesik AM (2005) Database Annotation in Molecular Biology. Chichester, UK: Wiley.
Lindler LE, Lebeda FJ, and Korch G (2005) Biological Weapons Defense: Infectious Diseases and Counterterrorism. Totowa, NJ: Humana Press.
Natarajan P, Lander GC, Shepherd CM, et al. (2005) Exploring icosahedral virus structures with VIPER. Nature Reviews Microbiology 3(10): 809–817.
Rubin DL, Lewis SE, Mungall CJ, et al. (2006) National Center for Biomedical Ontology: Advancing biomedicine through structured organization of scientific knowledge. Omics 10(2): 185–198.
Yusim K, Richardson R, Tao N, et al. (2005) Los Alamos Hepatitis C Immunology Database. Applied Bioinformatics 4(4): 217–225.