Towards a universal virus database – progress in the ICTVdB

C. Buechen-Osmond and **M. Dallwitz**

Research School of Biological Sciences, The Australian National University, Canberra, Australia

Origins and goals

In 1991, the Executive Committee of the ICTV decided to develop a universal virus database (ICTVdB), in response to, among other things, a petition brought forward by the American Type Culture Collection (ATCC). Drafted during a workshop in March 1990, the petition from leading virologists and database experts agreed on the need for a database for all viruses, which would be in conformity with the ICTV Reports and prepared under the auspices of the Committee. This paper concentrates on progress since that reported at the CODATA conference, Chambery, September 1994 (Buechen-Osmond et al. [1]).

The goal of the ICTVdB is to describe all viruses of animals (vertebrates, invertebrates, protozoa), plants (higher plants and algae), bacteria, fungi, and archaea from the family level down to strains and isolates. The lower levels of classification have important applications in medicine and agriculture, but also give insight into evolutionary trends. The database will thus benefit research and applications at all levels of expertise.

The language

The ICTV decided that the database should use the DELTA system, a DEscription Language for TAxonomy developed by Dallwitz in 1980 [4, 5]. The DELTA system is an integrated set of programs based on the DELTA format. The DELTA format is a flexible and powerful method of recording taxonomic descriptions for computer processing. Now adopted as a standard for data exchange by the International Taxonomic Databases Working Group (TDWG), it is in use worldwide for such diverse organisms as corals, crustaceans, insects, fish, fungi and plants. The TDWD is a group of plant taxonomists with members from major botanical gardens, herbariums and similar institutions (a previous Chairman was Frank Bisby, who is the coordinator of „Species 2000“, the initiative to collate all living species names that was started during a special CODATA meeting during the conference in Chambery, 1994).

The DELTA programs are continually refined and enhanced in response to feedback from users, through http://muse.bio.cornell.edu/delta/. The facilities available include the generation and typesetting of descriptions and conventional keys, conversion of DELTA data for use by classification programs, and the construction of Intkey packages for interactive identification and information retrieval.

Intkey is available for MS-DOS and MS-Windows, and offers better and more comprehensive features than any similar program. These features include entry and deletion of attributes in any order during an identification, calculation of the 'best' characters for use in identification; the ability to allow for errors (whether made by the user or in the data, or lack of data), and the ability to express variability or uncertainty in attributes. A comprehensive list of these features is available at http://muse.bio.cornell.edu/delta/www/delta.htm.

DELTA data can be converted to the forms required by programs for phylogenetic analysis, e.g. PAUP, HENNIG86, and MacClade. The characters and taxa required for these analyses can be selected from the full data set. Numeric characters, which cannot be handled by these programs, are converted to multistate characters. Printed descriptions can be generated to facilitate checking of the data, and Intkey can be used for further data checking, and for finding differences, similarities, and correlations among the taxa. Images illustrating particular features of the virus morphology, for example, are also an integral part of Intkey, thus simplifying the identification process.

The DELTA system is capable of producing high-quality printed descriptions as well as descriptions in HyperText Markup Language (HTML) for display on the Web. DELTA data can include any amount of text to qualify or amplify the coded information, and this text can be carried through into the descriptions. Common features can be omitted from the data and the descriptions, while remaining available for identification and analysis. These attributes are exemplified in books such as Viruses of Plants in Australia (Buechen-Osmond et al. [1]) and The Grass Genera of the World (Watson and Dallwitz [7]), which were generated automatically from DELTA databases.

The DELTA system is particularly useful in the international context envisaged by ICTV. For example, Intkey packages can be prepared in different spoken languages by translating the character list. Intkey is particularly easy to translate into other languages, as all of the program text (menus, commands, prompts, diagnostic messages, and help) are in simple text files (ASCII) separate from the program files. English, French, German, Malay, Portuguese, and Spanish versions are currently available.

Decimal code – identifiers in virus classification

From the early beginnings of ICTVdB activities the difficulties were realized to uniquely identify a virus. Virus nomenclature at the levels of the family, genus and species can be very similar, and vernacular names usually do not indicate the family or genus to which a virus belongs. Thus, for the purposes of the database it was found convenient to introduce a decimal numbering system similar to that used for enzyme nomenclature (Buechen-Osmond et al. [2]). The families have been sorted in alphabetical order and each has been assigned a number which represents a particular family, or genus, if the genus is not yet assigned to a family (Table 1). The system can carry more levels to accommodate strains and isolates, and in common with present practice in genomic and protein databases, new families, revisions etc, are added as they appear without alphabetical consideration.

This type of numbering system enables the user to follow the path linking the various features of one genus or family, and permits the presentation of similarities between different groups on more than one level. The numbering system also gives an internal

Table 1. Decimal classification illustrated from family to species in the *Parvoviridae*

50. Family	*Parvoviridae*
50.1. Subfamily	*Parvovirinae*
50.1.1. Genus	*Parvovirus*
50.1.1.0. Subgenus	(level not used in *Parvoviridae*, thus the number is set to 0)
50.1.1.0.001 Type Species	mice minute virus
50.1.1.0.002 Species	Aleutian mink disease virus

structure to the database that indicates the characters needed for completing the coding of a specific virus or data set. In the case of revisions, the original numbering system is retained as a pointer to the new assignation.

With this numbering system the database is structured such that we do not need to repeat the same information between levels. When calling up a description of a strain we can supply the pointers to the next higher level where the full description of the species is stored as illustrated in Table 2.

The numbers assigned to each virus also serve as locators numbers within the database and as an unchanging reference. The locator number is easily transformed into a file accession number by removing the decimal stops and filling each number up to 8 digits as demonstrated here on our example of the family *Parvoviridae* where the family level becomes 50000000, the genus level, that is the genus *Parvovirus*, 50110000, and for the type species mice minute virus 50110001. These identifiers are used throughout the ICTVdB as file names to access the computer generated virus descriptions and provide the basis of many of the links between components of the project.

Table 2. Natural language excerpt from ICTVdB illustrating the economic, non-redundant accumulation of information using the decimalised virus nomenclature

50. *Parvoviridae*

Taxonomic level of description: family. Taxa included in this taxon: 50.1 *Parvovirinae*; 50.2 *Densovirinae*. Taxon infects invertebrates and vertebrates.

50.1 *Parvovirinae*

Taxonomic level of description: subfamily. Taxon belongs to family 50. *Parvoviridae*. Taxa included in this taxon: 50.1.1 *Parvovirus*; 50.1.2 *Erythrovirus*; 50.1.3 *Dependovirus*. Taxon infects vertebrates.

50.1.1 *Parvovirus*

Taxonomic level of description: genus. Taxon belongs to subfamily: 50.1 *Parvovirinae*; family: 50. *Parvoviridae*. Taxa included in this taxon: Type species: 50.1.1.0.001 mice minute virus. Other species: 50.1.1.0.002 Aleutian mink disease virus; 50.1.1.0.003 bovine parvovirus; 50.1.1.0.004 canine minute virus; 50.1.1.0.005 chicken parvovirus; 50.1.1.0.006 feline panleukopenia virus; etc.

50.1.1.0.001 *Mice minute virus*

Acronym(s): MMV. Taxonomic level of description: species. Taxon belongs to genus *Parvovirus*; subfamily: 50.1 *Parvovirinae*; family: 50. *Parvoviridae*.

The character lists

The 1990 ATCC sponsored workshop focused on the need for standardized virus descriptions to be used in the database. The heart of the universal virus database is a standardised character list capable to describe unambiguously all viruses of humans, animals, plants, invertebrates, protozoa, bacteria and fungi. The currently accepted guidelines from the most recent ICTV Report (Murphy et al. [6]) provide the basis for the section layout of the character list and the framework for the virus descriptions in the database. The ICTV descriptions of the virus families and genera are also the source from which the single property statements are drawn to prepare a character list. However, the standardized character list for the ICTVdB must be extendable to allow the incorporation of detailed information that is required to differentiate between strains and isolates.

Although many virus characters have been in use for many years, it has proven difficult to produce a terminology acceptable to all cultures in virology. By the same token, advances in molecular biology brought so many new data for the concise description of a virus, that it has become difficult to select the most basic characteristics without loosing specificity important to the identification process at one level or another. Therefore, it has been necessary to rethink the concept of a list of standardised characters and the basic layout of the virus database.

Different terminologies for similar or identical features prevail in the different branches of virology, and these make the preparation of a universally acceptable character list very difficult. In each case, and at each level, traditional virus descriptions have to be reduced to single property statements, removing redundancies to evolve an agreed standardised nomenclature for similar features and characteristics. Thus progress in the development of a standardised character list, and in particular the reviewing of the list by the Data Subcommittee and Study Groups of the ICTV, has been much slower than anticipated.

The present state of the database

Standardized characters for morphology and genome properties have been used to code all virus families and genera including their type species. The description of particle morphology has been tackled first because the shape of a virus is the most important characteristic for identification to the level of genus. At this stage of development, the descriptions also contain genome accession numbers and reference lists. Thus far, only a few descriptions accommodate images (mainly electron micrographs of the virus particle), but it is planned to include many more, such as gene maps, distribution maps, illustrations of hosts and vectors, of symptoms and histopathology, as the database grows. Primary lists of characters exist for most other particle and biological properties and are in the process of culling, amalgamation, sorting and reviewing.

The ICTVdB on the World Wide Web

Recognizing that all aspects of the generation and maintenance of the ICTVdB will be greatly facilitated by the World Wide Web (WWW), over the last year efforts have been directed to making the database available on the WWW. The natural language translations of the descriptions of all families and genera, including type species treated thus far in

ICTVdB, are available on http://life.anu.edu.au/viruses/ICTVdB/index.html. They are presented with a few examples of electron micrographs of virus morphology. In cases where genome sequence data are available links to EMBL/GenBank have been established. Links to other databases on the Web relevant to the ICTVdB are listed and as the project progresses many more links to other databases will be established. In the process of its construction, ICTVdB on the WWW will be a fully federated database (Wertheim [8]).

Layout and WWW access of the ICTVdB

WWW browser

The ICTVdB Web[3] site has been set-up using Netscape, but other viewing software can be used, although the positioning of images and the layout might not be optimal. *Home Page – Introduction to the virus databases on-line:* The user can best access the ICTVdB Web site through the Home Page (http://life.anu.edu.au/viruses/welcome.html) which also introduces other virus databases on-line that have been developed by researchers in the Molecular Evolution and Systematics Group, Research School of Biological Sciences, Australian National University. The information belonging to the ICTVdB project is marked with the ICTV logo, and contains about 800 files. Other databases on the Web relevant to the ICTVdB and which have been used to establish links, for example, to electron micrographs or genomic sequence data, are also listed on the Home Page.

Indexes

The ability to select records is a quintessential feature of a functional database. As shown in Fig. 1, in the ICTVdB Web site facilitates this by making data accessible both alphabetically, in the Index Virum (the complete lists of virus families, genera and species from the VIth ICTV Report) and decinumerically in the ICTVdB Index. For example, the Index Virum may be searched from a species name or synonym in alphabetical order, or from a family name in alphabetical order, by host range, or from nucleic acid composition. As pointed out above, classification beyond family in the ICTVdB is numerically based following alphabetic listing. It is clear in Fig. 1 that a variety of paths all lead to the current natural language translation of the ICTVdB descriptions.

The Index Virum is the main entry point to the Classification and Nomenclature section of the ICTVdB. It provides easy access to the records describing an individual virus, as well as to information on its relatives. In addition to the Index Virum there is an index to the virus descriptions that have been generated from DELTA records. This index is automatically regenerated and updated every time new information is added to the database and translated into hypertext. Thus Fig. 1 summarises the present functional basis of ICTVdB, which will be greatly expanded by future data input, and improved interoperability with other related databases. Already, in cases where the genomic sequence accession number for a virus species is listed, links to either GenBank (NCBI) or EMBL (EBI) have been established. It is planned to have mirror sites in Europe and North America, so that the access time can be reduced for the user and the mirror sites will have links to appropriate genome databank. Figure 1 further indicates that it is planned to add a search engine to improve the speed and flexibility by which a particular virus can be retrieved, even if the user does not know the correct name.

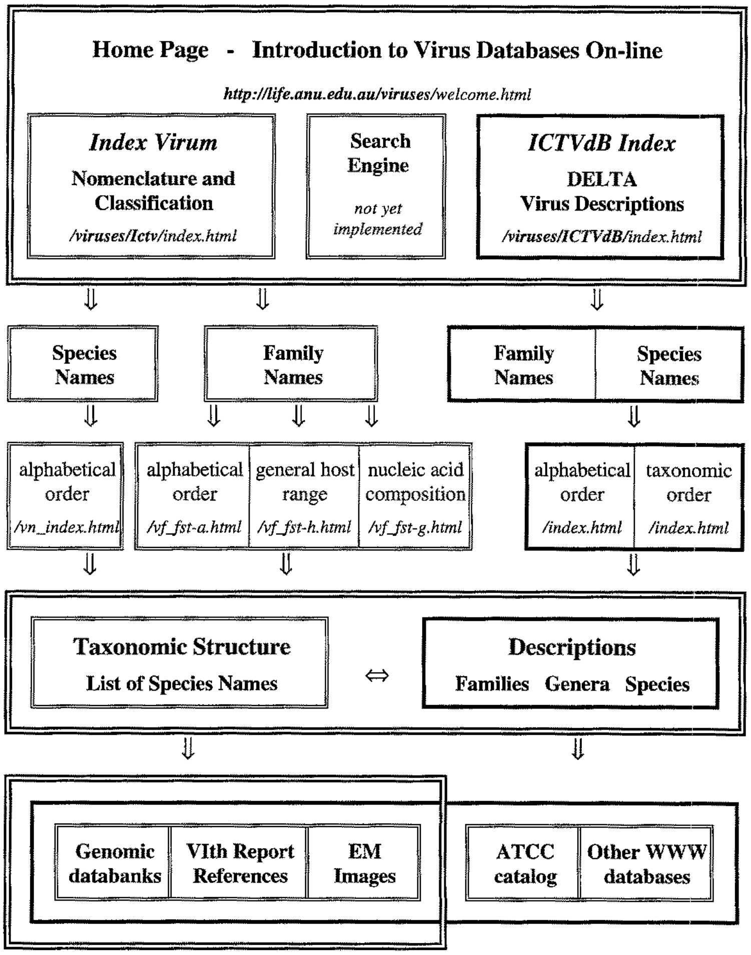

Fig. 1. ICTVdB on the Web. This diagram displays the various index files that are guiding the browser through the different parts of the ICTVdB project to the virus descriptions

Intkey

In future, Intkey will be available to interrogate the database on the Web. Using Intkey, a virus can be identified by comparing its attributes with stored descriptions of taxa. At present, the required data files and images must be down loaded to the users PC (via ftp, gopher, WWW), but a future version of Intkey will be able to access these files directly from the WWW. Images which are part of the database are used by Intkey. All images in the database can be accessed via hyperlinks from other databases. By the same token, images from other databases can be linked to the ICTVdB and will thus become accessible through the database, without becoming physically a part of it.

The tasks ahead

The WWW format should now greatly facilitate the formerly cumbersome process of data acquisition by printed questionnaires, shipped to the potential suppliers of data. This method, used to construct the VIDE plant virus database (eg. Buechen-Osmond et al. [1])

often did not attract enthusiastic cooperation of colleagues through the sheer complexity of the questionnaires, and the repetitive handling of data.

The first task will be to devise a new way of data acquisition. To this end, an electronic questionnaire based on standardised characters, will be devised as the primary data sheet. It will contain a highly structured index of keywords and headings to the different sections within the character list, that are collapsible or expandable. From previous experience we know that it helps greatly if the questionnaire contains already available data, the expert recipient being invited to fill in the gaps, and review existing data. It also helps if expert opinion is restricted to characters appropriate to the particular virus family. The data will be transformed into DELTA format and will go through the same reviewing process engaged for any description prepared for the ICTV Reports. Only after the reviewing committee is satisfied with the new submission, the coded description will be placed permanently into the database. The Web accessibility of the virus descriptions at all stages of preparation will also facilitate the reviewing process by the Study Groups of the ICTV.

The second task is to maintain and coordinate data acquisition and entry. Data entries must be checked and regularly updated, to take account of developments in the field. However, the coordinator will not be able to keep up with all the latest movements in virus research. New findings must be provided to the ICTVdB from the virological community on a regular basis. This is of utmost importance if the database is to provide the community with be a reliable up-to-date source of information.

A third task is to use ICTVdB to generate future ICTV Reports on the Nomenclature and Classification of Viruses, now laboriously compiled the ICTV Study Groups. It is envisaged that Reports will be generated from the DELTA database. In the future, most of the descriptions in the database will be of species and strains. DELTA programs can summarise the characteristics of all species of one genus, for example, thus generating an accurate summary that will reflect much more objectively the feature of a genus. The description of genera and families in the future reports can be based on these summaries. Judging by recent comments on other microbial databases (Wertheim [8]), the ICTVdB may be already one of the most advanced, interoperable databases in biology, in structural terms at least. A major effort is now required to fill in the database, drawing on the expertise of the virological community as a whole. The future success of ICTVdB thus depends heavily on the help and goodwill of all virologists, and the continued enthusiasm of the experts in the ICTV Study Groups.

Acknowledgements

Progress thus far has depended heavily on the sponsorship of this project by Lois Blaine and the ATCC, and Adrian Gibbs at the Australia National University. The preparation of a standardised character list has been supported by an NSF Grant (DIR-91- 07464).

References

1. Buechen-Osmond C, Crabtree K, Gibbs A, Maclean G (1988) Viruses of plants in Australia. Australian National University, Canberra
2. Buechen-Osmond C, Blaine L, Horzinek MC (1996) ICTVdB: The universal virus database. In: Data and Knowledge in a Changing World: The Quest for a Healthier Environment; Chambery, 94 CODATA Conference, Ed PS Glaeser. CODATA, Paris (in press)

3. Buechen-Osmond C (1995) http://life.anu.edu.au/viruses/welcome.html
4. Dallwitz MJ (1980) A general system for coding taxonomic descriptions. Taxon 29: 41–46
5. Dallwitz MJ, Paine TA, Zurcher EJ (1993) DELTA User's Guide: a general system for processing taxonomic descriptions, 4th ed. CSIRO, Canberra
6. Murphy FA, Fauquet CM, Mayo MA, Jarvis AW, Ghabrial SA, Summers MD, Martelli GP, Bishop DHL (1995) Virus Taxonomy. Classification and Nomenclature of Viruses. Sixth Report of the International Committee on Taxonomy of Viruses. Springer, Wien New York (Arch Virol [Suppl] 10)
7. Watson L, Dallwitz MJ (1992). Grass genera of the world. CABI, Wallingford
8. Wertheim M (1995) Call to desegregate microbial databases. Science 269: 1516

Authors' address: Dr. Cornelia Büchen-Osmond[1] and Dr. Michael J. Dallwitz[2]
[1] Research School of Biological Sciences, Institute of Advanced Studies, Australian National University, GPO Box 475, Canberra ACT 2601, Australia.
[2] CSIRO, Division of Entomology, GPO Box 1700, Canberra ACT 2601, Australia.