Arch Virol (1998) 143: 1513–1521

Archives of Virology

Computer-aided virus identification on the World Wide Web

A. S. Kolaskar and **P. S. Naik**

Bioinformatics Centre, University of Pune, Pune, India

Accepted March 26, 1998

Summary. An attempt has been made to devise computer software that will aid virologists to identify unknown virus isolates using the World Wide Web. Computerized information from the Animal Virus Information System was used to obtain data on various characters of a virus species. Sequence data banks are used to obtain the molecular data. A probabilistic method of virus identification based on Willcox's implementation of Bayes' theorem is implemented. The program provides hints to the users to carry out additional tests required to obtain higher confidence in identification of virus species. Signature peptides of the virus can also be used to confirm identification. The software is implemented on a UNIX machine and is written in C, UNIX shell scripts and HTML to run on the World Wide Web. This is the first species identification software that allows the user to carry out identification online through Internet.
[see `http://bioinfo.ernet.in/www/prob.html`]

Introduction

The number of users who use Internet and the World Wide Web (WWW) has increased many fold in the last few years. The availability of scientific databases on the Internet in the area of biology, is also on the increase. Many Internet-accessible databases in the area of biology deal with molecular data such as nucleotide sequences, though some databases have information at the cellular and organism or species level. The species 2000 project (http: `//www.species2000.org`), is aimed at increasing the number of species databases on the Internet. It has not gone unnoticed that at the time when the biological species are becoming extinct, the numbers of experts in the area of taxonomy of biological species are decreasing at a rapid rate. This makes it essential to develop user-friendly computer software that will aid biologists in the classification and identification of organisms. If made available on the WWW, such software has the capability to improve the quality of identification of species with little efforts. With this aim, an attempt has been made to develop software with an expert system approach to classify

and identify viruses on the WWW. An expert system is computer-based software that performs functions similar to those normally performed by a human expert [12].

The two principal approaches used to identify organisms are deterministic and probabilistic. In the deterministic method, one can use monothetic or polythetic approaches. Numerical taxonomy and identification programs used to classify and identify higher organisms normally use deterministic methods. The DELTA system by Dallwitz and Paine [5] and PANKEY package by Pankhurst [11] are examples of a deterministic approach of identification of organisms. On the other hand, bacteriologists, and microbiologists in general, seem to more commonly use methods based on a probabilistic approach. The probabilistic approach developed by Willcox et al. [14,15] used for bacterial identification has been critically reviewed by Bryant [2]. Both schools have shown the usefulness of computer-aided identification and classification.

The hybrid method that makes use of deterministic and probabilistic approach for classification and identification is rare. The International Committee on Taxonomy of Viruses (ICTV) keys follow the hierarchical rule based system, and the order of presentation of families and groups follows three criteria: (i) nature of viral genome, (ii) the strandedness of the viral genome and (iii) the replication strategy of the virus [10]. The species identification thus makes use of hierarchical as well as rule-based approach. In the present study we have developed computer software based on a hybrid virus identification system that uses deterministic as well as probabilistic approach. In the software discussed below, a deterministic monothetic approach has been used to assign the family to an animal virus species. The probabilistic approach uses character weight matrices of an assigned virus family to identify the virus species. The computer program has been written to extract data for character weight matrices from the Animal Virus Information System (AVIS) [7]. AVIS is a computerized database on animal viruses and can be accessed through the Internet from the URL `http://bioinfo.ernet.in/avis/avis.html`. AVIS has been organized under 16 different categories, which include: virus status and distribution, the original source of the virus, method of isolation and validity, physicochemical properties of the virus, stability of infectivity and virulence, virion morphology, morphogenesis, hemagglutination, antigenic relationship, susceptibility of cell systems, natural host range, experimental viremia, histopathology, human disease, links with other data banks and references.

Materials and methods

The software developed for virus identification uses the diagram given in Fig. 1. As can be seen from this figure, the process of virus identification has been divided into two parts. Thus the software consists of two main modules, input module and compute module.

Input module

The input module has three components: (i) input for identification of virus family (ii) input for user character values to create identification matrix and (iii) direct input of users identification

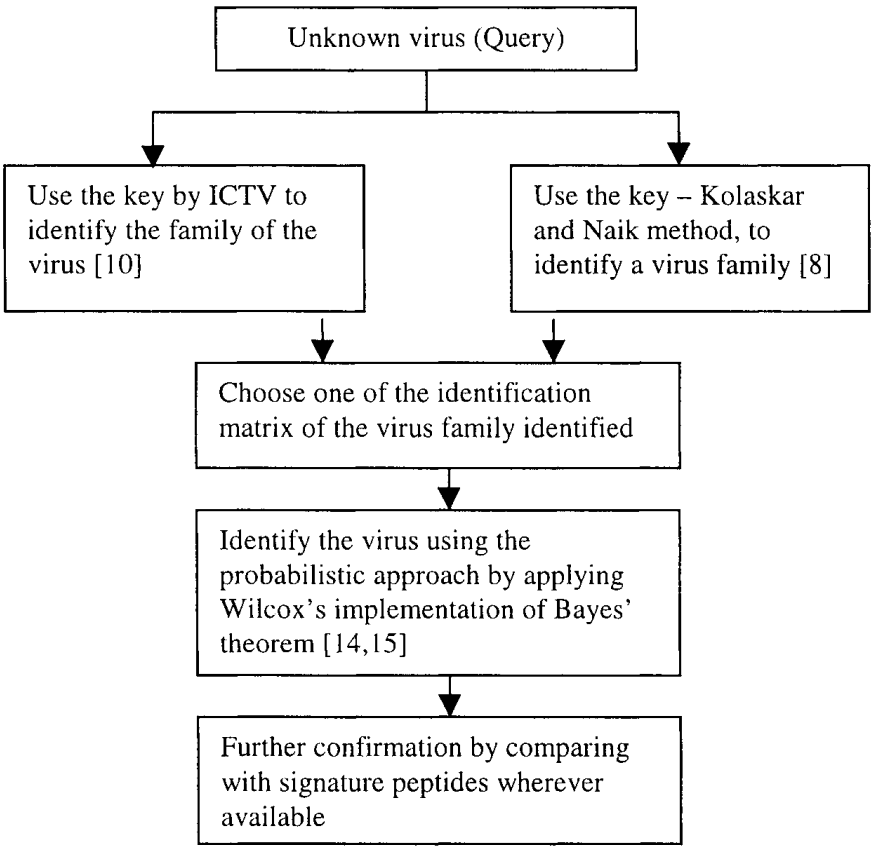

Fig. 1. Diagram illustrating the computer aided identification of viruses on the World Wide Web

matrix. In each of these input modules, a pre-designed screen allows user to feed in data, by clicking one of the possible alternatives. These pop-up screens are user-friendly and allow input of precise information. These programs for the input module are written in HTML and CGI scripts. At present the user can feed identification matrix of a maximum order of [193 × 58] (This limitation on the size of the identification matrix is mainly because of our present hardware SUN spark 5).

Compute module

The compute module is divided into three parts: (i) assignment of a family (ii) creation of the identification matrix and (iii) identification of virus species and display of results. This compute module is written in C and thus can be transported easily across computer platforms. The family assignment to the unknown virus as mentioned in the diagram in Fig. 1 is carried out deterministically by using the method described earlier [8] or using the keys developed by ICTV [10]. From the display the user can choose either of the approaches. The compute module then assigns the family using input character data fed by the user.

The identification matrices are created using the character value information, either extracted from AVIS or supplied by the user. The procedure followed to create the matrix is explained in Fig. 2 with the example of 5 Operational Taxonomic Units (OTUs) from the family *Flaviviridae*. In Fig. 2 character values are extracted from AVIS. Jaccard's coefficient [6] approach is used to calculate euclidean distances (dP_k dP_l) between OTUs P_k and P_l. The higher values of euclidean distance between two OTUs suggest large difference in the

OTUs (Viruses)	Characters	*Ae. albopictus*	LLC-MK2	Vero	BHK-21	Human epith.	Duck embryo	PS	HeLa	C6/36
P1	Kunjin	1	0	1	1	0	1	1	0	0
P2	Japanese encephalitis	0	1	1	0	1	0	1	1	1
P3	Powassan	1	1	1	0	0	0	1	0	0
P4	Dengue type 4	1	1	1	1	0	0	1	0	0
P5	West Nile	1	0	1	0	0	0	0	0	0

The data is represented as follows: The cytopathic effects observed on these cell lines by the five chosen viruses are represented as: 1 for yes, 0 for no and unknown. Number of OTUs = n = 5

The Euclidean distances dP_k dP_1 between the OTUs P_k and P_1 is obtained using the Jaccard's coefficient [6]:

dP_1	0.00				
dP_2	0.89	0.00			
dP_3	0.71	0.76	0.00		
dP_4	0.58	0.80	0.45	0.00	
dP_5	0.78	0.93	0.71	0.78	0.00
	dP_1	dP_2	dP_3	dP_4	dP_5

The average distance for the P_kth row with respect to other OTUs:

$$W_k = \left[\frac{\sum_{\substack{l=1 \\ k \neq l}}^{n} dP_k dP_1}{n-1} \right] \cdot 100$$

Identification matrix, W_k values for above OTUs:

P_1	74	01	74	74	01	74	74	01	01
P_2	01	84	84	01	84	01	84	84	84
P_3	65	65	65	01	01	01	65	01	01
P_4	65	65	65	65	01	01	65	01	01
P_5	80	01	80	01	01	01	01	01	01

Fig. 2. Procedure used to develop the Virus Identification Matrix to identify the virus probabilistically. An example of a few viruses of the *Flaviviridae* family

character values of these OTUs. The weighing scheme to create the identification matrices uses these euclidean distances among OTUs to calculate the weight values W_k. The example of such weight values calculated for five members of the *Flaviviridae* family using information on cytopathic effects on cell lines is given in Fig. 2. The user can also input identification matrices derived by using different criteria or weighing schemes [4] in an interactive fashion as mentioned above. In most cases the user is expected to use identification matrices made available by the authors to identify the virus.

Probabilistic virus identification is carried out using Willcox's implementation of Bayes' theorem [14, 15]. The Bayes' theorem is represented below:

$$P(Ti \mid a) = \frac{P(a|Ti)\,P(Ti)}{P(a)}. \tag{1}$$

Where $P(Ti|a)$ is the probability of OTU with character value a, belonging to the taxon Ti
$P(a|Ti)$ is the probability of observing character state a in taxon Ti
$P(Ti)$ is the prior probability of finding a specimen belonging to taxon Ti
$P(a)$ is the probability of observing character state a, that can be written as in Eq. (2).

$$P(a) = \sum_{i=1}^{n} P(a \mid Ti)\,P(Ti) \tag{2}$$

$P(Ti)=1$, as in most cases these values are unknown or constant [6]. Also in Bacterial identification $P(Ti)=1$ is assumed [3].

A case study is presented in Fig. 3. This example uses identification matrix given in Fig. 2. Further, the results of the tests carried out using unknown virus to find out the cytopathic effects on the cell lines, *Aedes albopictus,* LLC-MK2 and Vero are used in identification as shown in Fig. 3. The calculated likelihood values and identification scores (L/S) are given in Fig. 3 as an example.

Test carried out and the results for the unknown virus

	Tests carried out		
	CPE observed on *Aedes albopictus* cell line	CPE observed on LLC-MK2 cell line	CPE observed on Vero cell line
Results for the unknown virus	Yes	Unknown	Yes

The likelihoods and the identification score are calculated as follows:

Taxa	Likelihoods (L)	Identification score (L/S)	Final identification score (L/S).100
P_1	$(0.74)\cdot(0.74)=0.5476$	0.2682	26.82
P_2	$(0.01)\cdot(0.84)=0.0084$	0.0041	00.41
P_3	$(0.65)\cdot(0.65)=0.4225$	0.2070	20.70
P_4	$(0.65)\cdot(0.65)=0.4225$	0.2070	20.70
P_5	$(0.80)\cdot(0.80)=0.6400$	0.3135	31.35
	Sum(S) = 2.0410	Total = 0.998	

Note: The final identification score is low because only few tests are used

Fig. 3. An example to calculate the Identification score. Identification matrix used is from Fig. 2

The output consists of names of the viruses and their identification score in the decreasing order of the score value. On the second screen of the output, additional tests required to improve the identification score and thus the confidence are also listed.

Results and discussion

As mentioned in materials and methods, the software assigns the family to the unknown virus in a deterministic fashion by using two key based approaches: (i) Kolaskar and Naik method [8] and (ii) ICTV keys [10]. If one clicks on the first option, and inputs the data for the unknown virus as follows: (i) genome is RNA, (ii) genome is single stranded, (iii) genome is positive sense, (iv) virion is enveloped, (v) virion is spherical, (vi) the hosts are insects and vertebrates and (vii) the carbohydrate content is $> 8\%$, the software assigns the family as *Flaviviridae* using the above data. On the other hand, if one clicks on the ICTV keys and inputs data in an interactive fashion (i) genome is RNA (ii) genome does not encode reverse transcriptase; virus genome does not integrate, (iii) RNA single stranded, (iv) RNA negative sense or ambisense, (v) RNA linear, (vi) genome multipartite, (vii) genome in < 5 segments, (viii) virions not filamentous and (ix) genome tripartite; virion does not contain host ribosomes, (x) all RNA segments negative sense (xi) S RNA < 1 kb; S RNA encodes NSS protein + N protein. Then the software assigns the family as *Bunyaviridae*, and also suggests that the unknown virus belongs to the genus *Bunyavirus*. Once the family of the unknown virus is assigned, then, the software module for probabilistic virus species identification is used.

A probabilistic identification is carried out using the character matrices of the corresponding family. If one uses amino acid composition identification matrix of the Envelope glycoprotein (Egp), then the pop-up screen demands inputs of amino acid composition in percentage of Egp of the unknown virus. Let us say the unknown virus Egp contains values of Ala, 9.1%; Val, 7.5%; Arg, 3.5%; Leu, 8.1%; Ile, 4.3%; Gln, 2.3%; Glu, 4.5%; Asp, 4.1%; Lys, 5.1%; Asn, 3.7%; Pro, 2.9%; His, 2.9%; Thr, 7.9%; Ser, 8.5%; Tyr, 2.1%; Trp, 1.7%; Phe, 4.5%; Met, 2.3%; Cys, 2.3%; and Gly, 10.7%. The software identifies the virus as Japanese encephalitis with 99% confidence limit. Note one has used *Flaviviridae* amino acid composition of the Egp matrix and identification is upto species level. Similar amino acid composition matrix for *Bunyaviridae* family, which contains approximately 309 virus species, could not be prepared as protein composition data is available for about 35 virus species of this family. On the other hand, use of a matrix on cytopathic effects on cell lines for *Bunyaviridae* family identifies the unknown virus with 98% accuracy. Characters studied for most virus species of the family that have a high discrimination power should be used in the creation of a character matrix. A negative example can be mentioned of an identification matrix for the *Flaviviridae* family that uses the data from the tests on cross-neutralization of the polyclonal hyperimmune antisera. The software could not identify the virus in an acceptable range when the matrix was used to identify an unknown virus. The output of the program listed, antigenically closely related virus species to the unknown virus. A single large matrix with different characters, viz. symp-

Table 1. Use of identification matrix of the family *Flaviviridae* with 72 viruses and 112 characters

Step 1: Identification of viruses using above-mentioned [72 × 112] matrix

Stages	Data input	Identification score	Virus identified
Stage 1	Virus causes following symptoms: fever, headache, stiff-neck, CNS-signs, encephalitis, CNS-pleocytosis	0.205 0.205 0.205 0.180 0.180	Hypr virus Japanese encephalitis virus St. Louis encephalitis virus Murray valley encephalitis virus Powassan virus
Stage 2	Polyclonal hyperimmune antisera cross-reactivity was positive for the following viruses : Japanese encephalitis virus, St. Louis encephalitis virus, Murray valley encephalitis virus, West Nile virus, Kunjin virus, Kokobera virus, Uusutu virus, Stratford virus, Alfuy virus, Koutango virus.	0.369 0.369 0.260 0.000	Japanese encephalitis virus St. Louis encephalitis virus Murray valley encephalitis virus West Nile virus
Stage 3	Cytopathic effects observed on the following cell lines: Vero, LLC-MK2, PS, HeLa, C6/36	0.999 0.000 0.000 0.000	Japanese encephalitis virus St. Louis encephalitis virus Murray valley encephalitis virus West Nile virus

Step 2: Identification of the virus using [71 × 112] matrix with deletion of Japanese encephalitis virus character values row

Stages	Data input	Identification score	Virus identified
Stage 1	Virus causes following symptoms: fever, headache, stiff-neck, CNS-signs, encephalitis, CNS-pleocytosis	0.205 0.205 0.205	Hypr virus St. Louis encephalitis virus Murray valley encephalitis virus
Stage 2	Polyclonal hyperimmune antisera cross-reactivity was positive for the following viruses: Japanese encephalitis virus, St. Louis encephalitis virus, Murray valley encephalitis virus, West Nile virus, Kunjin virus, Kokobera virus, Usutu virus, Straford virus, Alfuy virus, Koutango virus	0.586 0.413 0.000	St. Louis encephalitis virus Murray valley encephalitis virus West Nile virus
Stage 3	Cytopathic effects observed on the following cell lines: Vero, LLC -MK2, PS, HeLa, C6/36	0.601 0.398 0.000	St. Louis encephalitis virus Murray valley encephalitis virus West Nile virus

The matrix obtained by extracting information from AVIS on Symptoms for the human disease, Cross-neutralization tests data, Cytopathic effects on cell lines data and the percentage amino acid composition of the envelope glycoproteins. Note: The increase in identification score and resolution with increase of data

toms of human disease, cross-neutralization test results, cytopathic effects on cell lines, and amino acid composition, when used was quite useful to identify the virus species with high confidence. The matrix identifies the unknown virus as Japanese encephalitis virus (see Table 1). These results are in accordance with the findings of bacteriologists where multiple weight matrices/large matrices give better identification score [3].

Character based identification matrices have been prepared using AVIS to identify viruses belonging to *Togaviridae* (natural host range, experimental viremia), *Rhabdoviridae* (natural host range), *Flaviviridae* (natural host range, experimental viremia, amino acid composition of Egp, cytopathic effects on cell lines, symptoms of human disease), *Reoviridae* (natural host range) and *Bunyaviridae* (natural host range - mosquitoes, cytopathic effects on cell lines). Additional matrices for various other families will soon be added to make the web based identification more useful for Virologists. The users are requested to visit the web site `http://bioinfo.ernet.in/www/prob.html` at regular intervals to find out the new matrices.

Sequence information of viruses is increasing at a rapid rate. Though for many viruses the protein sequence data is not fully available, the existing protein sequence data [1] can be analyzed carefully [13]. Such analysis helps to identify peptides that are unique to a particular virus protein and thus can be used as PCR probes to identify viruses as is described in earlier studies [9]. Such unique sequences in the envelope glycoprotein of *Flaviviruses* are given in Table 2. The last column in Table 2 indicates that even if there are mismatches upto three residues, the sequence is unique for the virus. Thus allowing the experimentalist to identify the virus correctly using probes generated using these peptides.

Thus, in short a novice in either computer application or Virology can use the computer-aided identification method available at the URL `http://bioinfo.ernet.in/www/prob.html`. The increase in virus data will help to generate more matrices with higher discrimination power and thus a better utility of the identification tool on the Internet.

Table 2. Signature peptides for members of the family *Flaviviridae*. The peptides are from the envelope glycoproteins from the respective viruses

Virus	Signature peptide	Unique upto no. of mismatches
St. Louis encephalitis virus	VNPFISTGGAN	3
Murray valley encephalitis virus	VTANPYVASSTA	3
Japanese encephalitis virus	LDVRMINIEA[S/V]Q	3
West Nile virus	TTKATGWIIQK	3
Kunjin virus	STKATGRTILKE	3
Langat virus	DGAEAWNEAGR	3
Yellow fever virus	MRVTKDTN[D/G][N/S]NL	3
Powassan virus	KDNQDWNSVE	3

Acknowledgement

We acknowledge the help of the Department of Biotechnology, Government of India, New Delhi for financial support.

References

1. Barker W, George D, Hunt L (1990) Protein sequence database. In: Doolittle RF (ed) Molecular evolution: computer analysis of protein and nucleic acid sequences. Academic Press, San Diego, pp 31–49 (Methods in Enzymology, vol 183)
2. Bryant TN (1991a) Software for the identification and evaluation of probabilistic identification matrices. Comput Appl Biosci 7: 189–193
3. Bryant TN (1991b) Bacterial identifier – A utility for probabilistic identification of bacteria. Blackwell, London
4. Bryant TN (1994) Freak, a program for analysing microbial binary characteristics. Binary 6: 97–100
5. Dallwitz MJ, Paine TA (1986) User's guide to the DELTA system. A general system for processing taxonomic descriptions 3rd ed. CSIRO Division of Entomology Publication, Canberra
6. Dunn G, Everitt BS (1982) An introduction to mathematical taxonomy. Cambridge University Press, Cambridge
7. Kolaskar AS, Naik PS (1992) Computerization of virus data and its usefulness in virus classification. Intervirology 34: 133–141
8. Kolaskar AS, Naik PS (1996) Concerted use of multiple database for taxonomic insights. In: Dubois JE, Gershon N (eds) The information revolution: impact of science and technology, Springer, Berlin Heidelberg New York Tokyo, pp 236–270
9. Leary TP, Muerhoff AS, Simons JN, Pilot-Matias TJ, Erker JC, Chalmers ML, Schlauder GG, Dawson GJ, Desai SM, Mushahwar IK (1996) Consensus oligonucleotide primers for the detection of GB virus C in human cryptogenic hepatitis. J Virol Methods 56: 119–121
10. Murphy FA, Fauquet CM, Bishop DHL, Ghabrial SA, Jarvis AW, Martelli GP, Mayo MA, Summers MD (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)
11. Pankhurst RJ (1986) A package of computer programs for handling taxonomic databases. Comput Appl Biosci 2: 33–39
12. Pankhurst RJ (1991) Practical taxonomic computing. Cambridge University Press, Cambridge
13. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW – improving the sensitivity of progressive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4 673–4 680
14. Willcox WR, Lapage SP, Bascomb S, Curtis MA (1973) Identification of bacteria by computer: theory and programming. J Gen Microbiol 77: 317–330
15. Willcox WR, Lapage SP, Holmes B (1980) A review of numerical methods in bacterial identification. Antonie van Leeuwenhoek 46: 233–299

Authors' address: Dr. A. S. Kolaskar, Bioinformatics Centre, University of Pune, Pune 411007, India.

Received January 6, 1998