

An Expert System to Classify Plant Viruses

J. Alonso, R. Lahoz-Beltra, A. Bailador¹, M. Levy and R. Diaz-Ruiz²

Departamento Matemática Aplicada, Facultad de Biología, Universidad Complutense, 28040 Madrid, Spain

¹Departamento Aplicaciones Informáticas, Ciemat, Avda. Complutense 22, 28040 Madrid, Spain

²UEI Fitopatología, CIB. Velázquez 144, 28006 Madrid, Spain

Received: 28 January 1992; revised 23 July 1992

We describe the building of a knowledge base for classification of a plant virus into a group or family. This knowledge base can be constructed using a simple expert system shell employing only the operator 'AND' and chained rulebases. We discuss some problems in the construction of the decision trees and present two different strategies in building them as a first approximation of the problem.

Introduction

Viruses that attack plants have been studied extensively during the last few years. These viruses have a great economic importance, attacking a wide range of plants. The techniques employed by pathologists provide enough information to identify a plant virus in a group or family (Smith, 1957).

The traditional methods used in biology to identify individuals were based on keys. In the beginning these were dichotomous, and then later tabular (Newell, 1970; 1972) and even computerized (Morse, 1975; Pankhurst, 1978), using classical programming techniques.

At present, techniques in artificial intelligence are being used with the aim of elaborating expert systems that carry out the function of these preceding keys to advantage (Alonso, 1990).

An expert system could be defined as "A kind of software with the aim of modelling expert human behaviour which has the task of solving problems for which there is no algorithm available" (Dussauchoy & Chatain, 1988).

All expert systems are principally composed of (Figure 1):

- (a) *A knowledge base* which contains all the knowledge on which the expert system is based. "....The power of an expert system comes from the knowledge it possesses and not from any particular formalism nor from the inference engine it uses..." (Feigenbaum, 1977).
- (b) *A fact base* containing all the intermediary results obtained during the deduction process carried out by the expert system.
- (c) *An inference engine* — a procedural program with the job of deducing an answer to the problem posed to the system from knowledge stored in the knowledge base.
- (d) *A user interface* allowing 'dialogue' between the system and the user is necessary to obtain the required input (knowledge, facts, data).

The aim of the present paper is to show the first steps in, and design problems of, the building of a knowledge base which allows identification and naming of plant viruses according to the rules given by the International

Committee on Taxonomy of Viruses (CITV) (Matthews, 1984).

The knowledge base was written using production rules which have allowed the construction of an expert system in which it is possible to use biochemical-type tests alone, or symptoms alone, as a means of classification.

The development of this expert system offers advantages to both its potential users and those who will have to maintain and update it:

- With an expert system, maintenance work is greatly reduced since it is possible to add or delete knowledge by simply modifying its knowledge base. Thus, it is kept independent from the program that drives it (the inference engine).
- There is a trace option available which can store a whole work session, and which allows later analysis of the reasoning process followed by the system. (For an explanation, see *Factbase* in the Results section).
- Extensive knowledge is not needed in either computers or virology to be able to use the expert system, due mainly to the help provided by user interface and the clear, simple language used in the questions posed throughout the session, which allow for an 'easy dialogue between the machine and the user'.

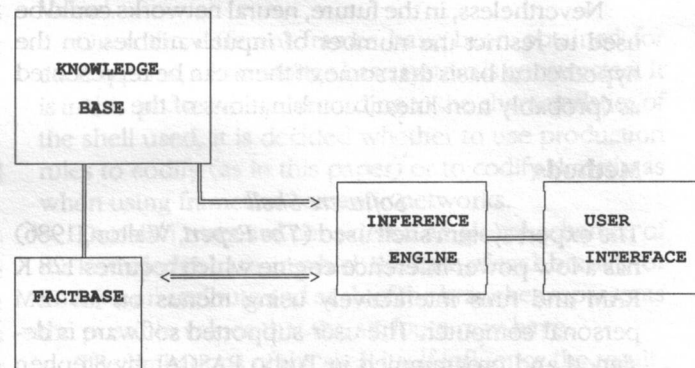


Figure 1 The basic architecture of an expert system.

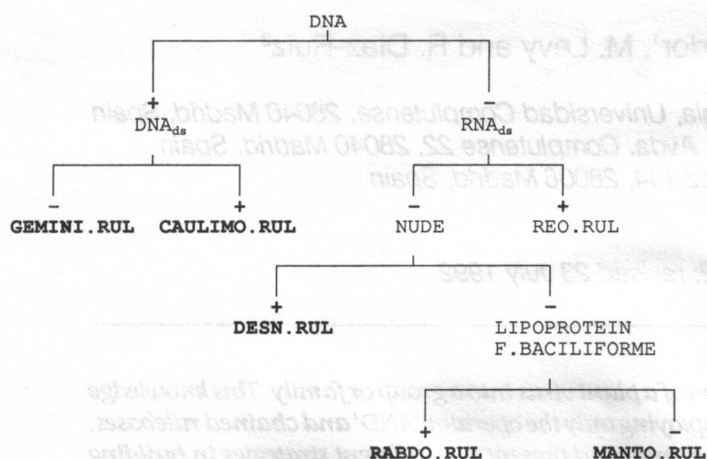


Figure 2 The origin of the decision tree with biochemical and morphological tests (analysis in amplitude).

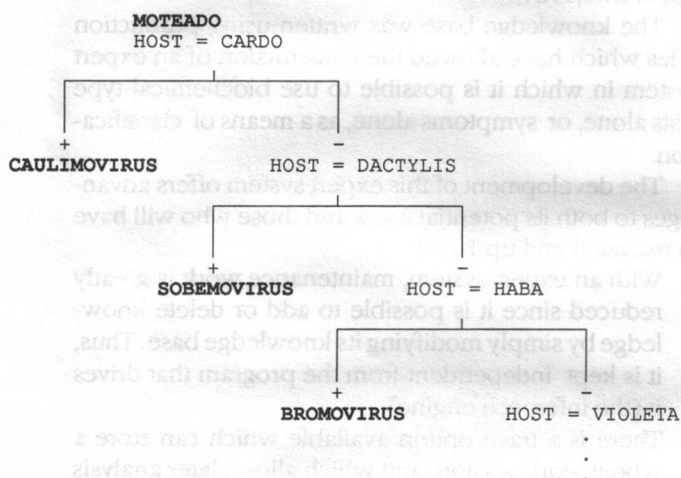


Figure 3 A fraction of the decision tree with symptoms which can be found in different hosts (analysis in depth).

There are other 'intelligent' classification systems such as artificial neural networks. In this particular case, where there are n dichotomous variables, the number of classifiable cases is finite and the relationship of the set of patterns with the output set is one-to-one, so that the generalizing power of the network is unimportant.

Nevertheless, in the future, neural networks could be used to restrict the number of input variables on the hypothetical basis that some of them can be represented as (probably non-linear) combinations of the rest.

Methods

Software Shell

The expert system shell used (*The Expert*, Walton, 1986) has a low power inference engine which requires 128 K RAM and runs interactively using menus on an IBM personal computer. The user-supported software is designed and programmed in Turbo PASCAL by Stephen Walton (496 Hudson St., Suite D15, New York, NY 10014, USA) who allows free distribution of the shell complete

with knowledge rulebases, by contacting the authors of this article.

Among the technical characteristics of this shell are backward chaining and forward chaining, which can be mixed by the user as search strategies, as well as being able to chain rulebases among themselves.

The inference engine of *The Expert* allows only the use of the logic operator 'AND', as is the case with other user-friendly expert system freeware such as MAE and ESIE.

Knowledge Rulebases

The knowledge base works with chained rulebases, presenting a set of characters or facts which allows classification of the virus into groups or families defined by the CITV between 1970 and 1981.

These groups are made up of 21 groups of plant viruses with international names (Caulimovirus, Geminivirus, Tobamovirus, Rhabdoviridae, etc.) and 17 groups without international names (tomato spotted wildvirus, maize chlorotic dwarfvirus, etc.).

The rulebases have been implemented as ASCII files with the extension .RUL, and have been organized as two independent decision trees with different facts and structure.

The first tree (Figure 2) presents a set of characters such as morphology, type of nucleic acid (DNA_{ds}, RNA_{ds}, ...) and other stable and useful properties.

This is a clear example of a search strategy in amplitude, as the inference engine acts by reviewing each level completely, before jumping to the next, reaching the identification of a virus family from the intersection of two or more traits.

The second tree (Figure 3) presents a set of symptoms such as necrosis, mottle, mosaic, ring spots, etc. and hosts (lettuce, tobacco, tomato, etc.). In this last tree the intersection between symptom and host allows us to find the virus group or family. In this case the inference engine applies a deep search strategy, reviewing each of the branches of the tree at all levels until it reaches the end. At this point, it returns to the starting point and begins the search in the next branch until it reaches the production rule which fulfills all, or the largest possible number of premises. It will then offer the conclusions of this rule as the answer.

Figure 4 shows these two different strategies: the first, an analysis in amplitude (tree 1) and, the second, an analysis in depth (tree 2).

Properties of the Rulebases

The knowledge rulebases have the following properties:

(a) Rules that have been built with only the operator 'AND'. Assigning true values to specific facts as assertions or denials of the facts in the particular case under consideration, we obtained the family of plant viruses (Z):

(i) In tree 1 the rules are:
IF A AND B AND C AND Y THEN Z,
where A, B, C Y are a set of biochemical and morphological properties.

- (ii) In tree 2 the rules are:
IF S AND H THEN Z
where S represents symptoms (S1, S2... Sn) and H,
the hosts (H1, H2, ... Hn).
- (b) The system is open. New facts can be incorporated into the roots and branches of the decision trees.
- (c) The system is made up of 1200 rules, working with chained rulebases which contribute to flexibility and debugging.

Results

The Rulebase

After carrying out the process of knowledge engineering, a rulebase comprising 112 files or modules that could be interlinked was built. The files take up about 200 Kbytes of space and are available to interested researchers.

This rulebase has been tested, checking the correct classification of all the virus types studied in each case.

The Factbase

The expert system shell used generates a factbase incorporating the results obtained in a work session.

A factbase summarizes the reasoning followed in a work session; it is the set of questions asked by the system and the answers given by the user, which allows the subsequent identification of a plant virus into a group or family. A factbase is saved as an ASCII file with the extension .FCT. Appendix 1 and Appendix 2 show the factbase generated in a work session on a tobacco mosaic virus (tobamovirus).

The factbase will list each question asked. If the question is answered positively the word 'given' will appear in the right-hand margin and will be related to the rule in question (e.g. Rule 12), also in the right-hand margin.

When a rulebase is finished and the facts are presented for solving, a chaining of rulebases will be produced (CHAIN 1 LINEAR.RUL) and the questions will begin again.

In this example, PRUEBA, DESN, LINEAR, ARBOL, ARBOLB and TOBA are the chained rulebases.

When a final goal is reached, the name of the virus identified will appear on screen. The phrase '-> END OF SESSION <-' will indicate that the consultation is over.

The system is optimized so as not to go on indefinitely unless there is a data error. The average time needed for the identification of a virus group or family is between 5 and 30 seconds where biochemical evidence is used, or between 15 and 70 seconds where symptoms are used.

Discussion

Expert systems based on production rules have proved to be very useful for species identification.

It should be noted that during the knowledge extraction stage, using knowledge engineering techniques, the best method for representing the knowledge in this type of expert system for identification is by the use of frames (Edwards *et al.*, 1987). This is as opposed to logic or semantic networks, since the use of frames allows the grouping in a single image of all the characteristics that

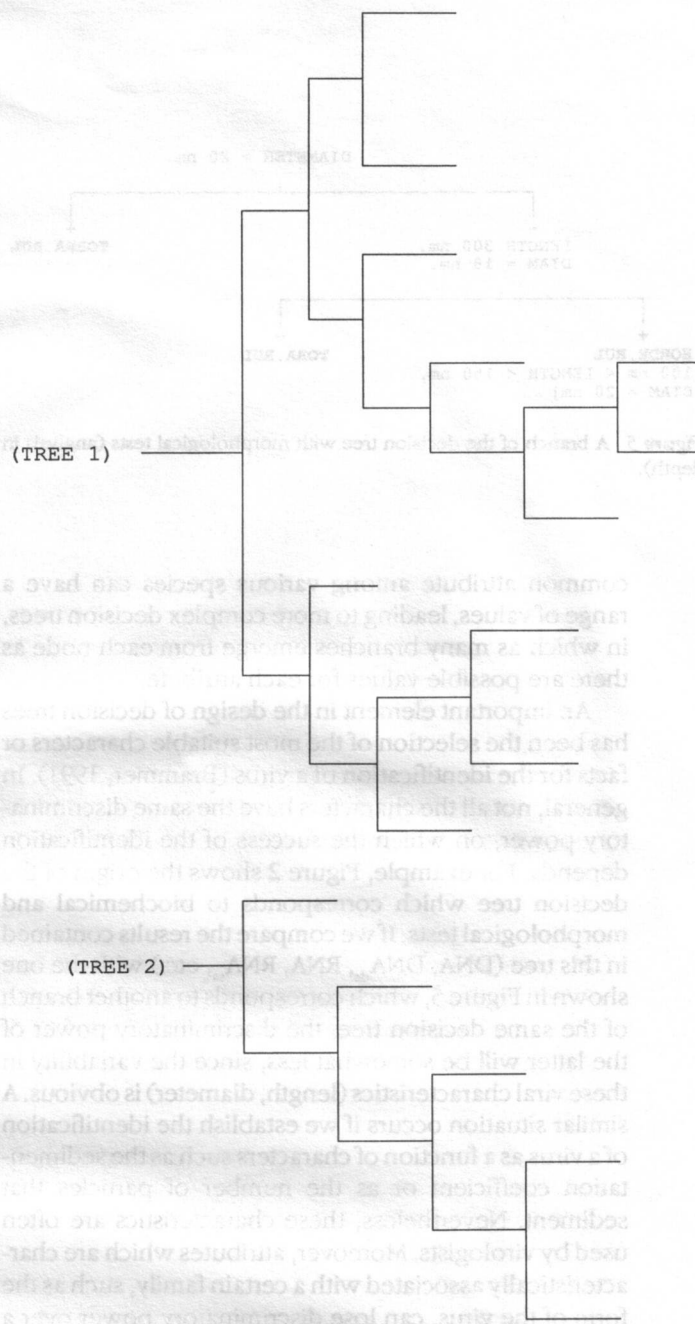


Figure 4 Two different strategies in the construction of the decision trees: analysis in amplitude (tree 1) and analysis in depth (tree 2).

define a species using the trinomies 'Object, Attribute, Value'.

Once the different frames have been obtained for each species, a knowledge base system is constructed. It is in this phase where, according to the characteristics of the shell used, it is decided whether to use production rules to codify (as in this paper) or to codify directly as when using frames or semantic networks.

Decision trees are considered as global schemes of the knowledge base, in which the base of each branch (or node) is an attribute and each of the branches represents the possible values that the attribute may have.

The limitations of the shell itself influence the resulting tree, so that this identification process is dichotomous. It is as well to remember, however, that one

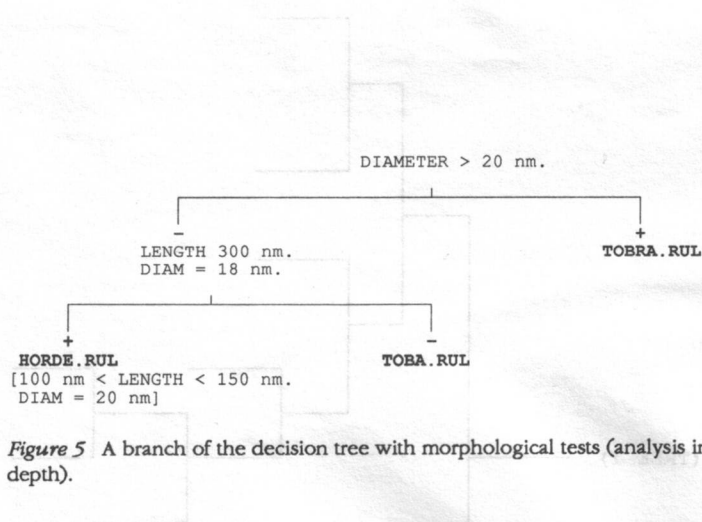


Figure 5 A branch of the decision tree with morphological tests (analysis in depth).

common attribute among various species can have a range of values, leading to more complex decision trees, in which as many branches emerge from each node as there are possible values for each attribute.

An important element in the design of decision trees has been the selection of the most suitable characters or facts for the identification of a virus (Brammer, 1991). In general, not all the characters have the same discriminatory power, on which the success of the identification depends. For example, Figure 2 shows the origin of the decision tree which corresponds to biochemical and morphological tests. If we compare the results contained in this tree (DNA, DNA_{ss}, RNA, RNA_{ss}, etc.) with the one shown in Figure 5, which corresponds to another branch of the same decision tree, the discriminatory power of the latter will be somewhat less, since the variability in these viral characteristics (length, diameter) is obvious. A similar situation occurs if we establish the identification of a virus as a function of characters such as the sedimentation coefficient or as the number of particles that sediment. Nevertheless, these characteristics are often used by virologists. Moreover, attributes which are characteristically associated with a certain family, such as the form of the virus, can lose discriminatory power over a period of time, when variants may appear.

An interesting criticism from our point of view is that when identification is established using symptoms, there may be more than one manifestation of the symptom in the host. However, in controlled laboratory conditions, this situation can normally be avoided.

Future developments include the incorporation of rules which correspond to genetic characteristics of each virus family not appearing in the current version of the expert system. The implementation of the system in a more powerful shell which incorporates a larger number of Boolean operators is also planned.

References

- ALONSO, J. (1990). Application of some expert systems to the Ascidiaceans identification. In *Boletín de la Real Sociedad Española de Historia Natural* 86, Núms. 1-4, 107-116.
- BRAMMER, R.J., BRYANT, T.N. & MAY, J.H.R. (1991). Investigation of an expert systems approach to bacterial identification. *CABIOS* 7, 461-469.
- DUSSAUCHOY, A. & CHATAIN, J.N. (1988). *Sistemas Expertos, Métodos y Herramientas*. Madrid: Paraninfo.

- EDWARDS, M., MORSE, D.R. & FIELDING, A.H. (1987). Expert systems: frames, rules or logic for species identification? *CABIOS* 3, 1-7.
- FEIGENBAUM, E. (1977). In DAVIS, R. & BUCHANAN, B.G. (1977). Production rules as a representation for a knowledge-based consultation system. *Artificial Intelligence* 8, 15-45.
- MATTHEWS, R.E.F. (1984). *Clasificación y Nomenclatura de los Virus*. Grupo de Virología. Sociedad Española de Microbiología.
- MORSE, L.E. (1975). Recent advances in the theory and practice of biological specimen identification. In *Biological Identification with Computers*, pp. 1-7. The Systematics Association Special Vol. 7. London: Academic Press.
- NEWELL, I.M. (1970). Construction and use of tabular keys. *Pacific Insects* 12, 25-37.
- NEWELL, I.M. (1972). Tabular keys - Further notes on their construction and use. *Transactions of the Connecticut Academy of Arts and Sciences* 44, 259-267.
- PANKHURST, R.J. (1978). *Biological Identification*. London: Edward Arnold.
- SMITH, K.M. (1957). *A Text Book of Plant Virus Diseases*. Boston: Little Brown & Co.
- WALTON, S. (1986). *Approvalware for the Implementation of User-specified Expert Systems*. Software and Manual. New York, NY 10014.

Appendix 1

- Do you want to make biochemical test?
- GIVEN ---->
- RULE 1 TEST.RUL
- RULE 1 TEST
- It is an RNA virus?
- GIVEN It is not the case that: It is a DNA virus?
- RULE 7 It is not the case that: It is bicatenary DNA?
- RULE 22 It is not the case that: It is monocatenary DNA?
- RULE 22 It is not the case that: Is the RNA bicatenary?
- GIVEN It is monocatenary RNA
- RULE 10 Is it a nude virus? (only considered in RNA viruses)
- GIVEN VIRUS
- RULE 12 It is not the case that: Lipoprotein virus (only considered in RNA virus)?
- RULE 12 CHAIN 1 DESN.RUL
- RULE 12 It is not the case that: It is a non-icosahedric virus, i.e. isometric?
- GIVEN Does your virus have linear form?
- GIVEN VIRUS
- RULE 26 CHAIN 1 LINEAR.RUL
- RULE 26 It is not the case that: It is a virus with an icosahedric form?
- RULE 25 It is not the case that: Is it a virus with an isometric form?
- RULE 25 Does it have a rigid form?
- GIVEN VIRUS
- RULE 1 Linear, nude, rigid-form RNA virus
- RULE 1 CHAIN 1 TREE.RUL
- RULE 1

Appendix 2

It is not the case that: Does your virus possess a $\phi > 20$ nm?
 GIVEN
 It is not the case that: Is the length of your virus approx. 300 nm
 and the $\phi = 18$ nm?
 GIVEN
 It is not the case that: Does the length of your virus vary from
 100 to 150 nm and the $\phi = 20$ nm?
 GIVEN
 VIRUS
 RULE 4
 CHAIN 1 TREE-B.RUL
 RULE 4
 Is the length of the viral particle greater than 250 nm?
 GIVEN
 VIRUS

RULE 1
 Linear, rigid RNA virus
 RULE 1
 CHAIN 1 TOBA.RUL
 RULE 1
 It is not the case that: Do you want to continue your consultation?
 GIVEN
 TOBAMOVIRUS
 RULE 2
 This is the virus identified by your data
 RULE 2
 At your request this consultation has ended. Thankyou.
 RULE 2
 -----> END OF SESSION <-----
 RULE 2