Evidence that the Amino Acid Composition of the Particle Proteins of Plant Viruses is Characteristic of the Virus Group

II. Discriminant Analysis According to Structural Biological and Classification Properties of Plant Viruses

C. Fauquet, J. Dejardin, J.-C. Thouvenel
Phytovirology, ORSTOM, Abidjan, Ivory Coast

Summary. The amino acid composition (AAC) of the coat proteins (CPs) of 126 plant viruses or strains were analyzed by stepwise discriminant analysis. The criteria chosen for discrimination were: (i) the structure of virus particles (3 clusters); (ii) the mode of transmission of the viruses (6 clusters); and (iii) the grouping of viruses according to the classification of the International Committee on Taxonomy of Viruses (23 groups). Statistically significant correlations were obtained with different groups of discriminant amino acids. The results confirm that the AAC of the CPs contains all the information needed for a quantitative classification of plant viruses. These results and possible explanations of these clustering patterns are discussed.

Introduction

In a previous paper[1] we reported that the amino acid composition (AAC) of coat proteins (CPs) of particles of plant viruses is characteristic of the virus group concerned. The clusterings obtained by principal component analysis (PCA) correlated closely with the classification of the International

Address inquiries to: Dr. C. Fauquet, Phytovirology, ORSTOM, Abidjan (Ivory Coast)

Committee on Taxonomy of Viruses (ICTV) [2]. We also concluded that these virus particle proteins showed a general relationship to the variability of proteins in general, and that our groupings of them did not seem to correlate strictly with the shape of the virus particles, the serological relationships, or with the biological properties, but did correlate with a combination of these factors. Indeed, the AAC of CPs of plant viruses may be influenced by several factors that interact to produce the final arrangement. The resulting features may include: (i) features common to all plant viruses; (ii) features related to com-

patibility with the nucleic acid; (iii) features related to compatibility with the other proteins or protein subunits in the particle; (iv) features related to biological properties, such as the mode of transmission by vectors; and (v) features related to surface properties, such as serological relationships.

The first point was previously investigated by Tremaine and Argyle [3], who found in a multidimensional classification (MC) of all proteins based on their AACs, that the space occupied by plant virus proteins is no more than 5×10^{-4} of the total hyperspace occupied by proteins. This result indicates that there is an extremely strong similarity among all plant virus proteins. Whether this results from divergence from a common ancestor or is the result of a functional convergence is not known.

As for the features related to surface properties, Gibbs [4] found with tobamoviruses a high correlation (0.833) between the classification of viruses based on the AACs of their CPs and classification based on serological relationships. In contrast, Paul et al. [5] were unable to establish a significant correlation among the tymoviruses using the same criteria. The MC previously obtained [1] clearly shows that any such correlation applies at most within a group and not to plant viruses as a whole. A possible explanation may be that the part of CPs involved in serological relationships is very small and is not constant between different groups of viruses; for instance, if it is 5% for the tobamoviruses [6], it might be only 2% in the case of potyviruses and 1% for tombusviruses.

In the present paper we give the results obtained by testing hypotheses ii–iv above, relating the AACs of CPs of plant viruses to other features. We decided that hypotheses ii and iii could be tested together by searching

for a relation between the shape of the virus particles and the AACs of their CPs. Concerning the biological properties, we took into account only the mode of transmission of plant viruses. Finally, we intend to use the same procedure to establish the extent to which the AACs of CPs are characteristic of the virus groups in the ICTV classification.

Materials and Methods

We used the same data on AAC of CPs of plant virus particles as the previous paper [1] based on PCA. The AACs are expressed in numbers of amino acid (AA) residues per protein subunit.

The stepwise discriminant analysis (SDA) [7] used is the BMDP 7M program [8] from the BMDP library. The estimated number of AAs in each protein are the quantitative variables, and the discriminant criteria are the groups. Using the SDA, we attempted to explain a qualitative variable (structure, transmission, or classification groupings) by linear functions of quantitative variables (number of AAs in the AAC of CPs). These linear functions are called discriminant functions, and they are calculated with the mean of the quantitative variables to get the maximum ratio between the interpopulation variance and

Table I. Comparison of different SDA made with the AAC of CPs of particles of plant viruses according to different clusterings of plant viruses

Clustering criterion	Number of clusters	Percentage correctly classified	Number of discriminant AAs
Random Choice	3	0	0
Structure	3	93	9
Transmission	7	87	8
Classification	23	97	15

Table II. Discriminant AAs for each SDA by particle structure, mode of transmission, and ICTV classification clusterings with F value for each AA

Clustering		
Structure	Transmission	Groups
LYS 71	LYS 8	LYS 21
PRO 14	PRO 4	PRO 10
HIS 13	HIS 27	HIS 5
ASP 9	ASP 9	ASP 5
GLU 31		GLU 5
VAL 36		VAL 8
ALA 10		ALA 5
MET 4		MET 10
	PHE 7	PHE 8
	GLY 5	GLY 5
	TRP 5	TRP 6
THR 8	THR 8	
		ARG 12
		ILE 6
		TYR 6
		LEU 5

Table III. Classification matrix, indicating number and percentage of examples correctly classified into clusters I, R and F, by SDA based on the shape of plant virus particles

Clusters	Correctly clustered %	Number of viruses classified into clusters			Total
		I	R	F	
I	90	48	2	3	53
R	94	1	30	1	32
F	93	1	1	39	41
Total	93	50	33	43	126

the intrapopulation variance. A distance is then calculated between the clusters represented by their mean. The distance used here is one of the most often used, i.e., the D^2 distance of Mahalanobis.

It is obvious that all the AAs do not have the same power of discrimination. Some vary greatly in number or percentage from cluster to cluster; others are not different. The objective of the SDA was to determine for each clustering a selection of variables with a great ability to discriminate. Consequently, we first searched for the AA whose variability was greatest for all the clusters (using the Fisher test; F). Next, we searched for a second AA, which together with the first would give a linear combination of 2 AAs that discriminated best. This process was repeated until no AA could bring better discrimination to the analysis. These discriminant AAs thus formed a sufficient subset to discriminate between the clusters.

The ability of the discriminant function to distinguish between the clusters was assessed by the following method. For each individual with all the discriminant AAs, the D^2 distance of Mahalanobis was calculated for each cluster mean and a probability was calculated to reach them. The individual was finally classified in the cluster having the greatest probability. Then, this a posteriori classification was compared with the a priori classification, and a percentage of correctly classified individuals was calculated.

Finally, the program computed canonical discriminant variables and plotted the first two, to give an optimal two-dimensional diagram of the positions of the individual viruses, and hence the separation of the clusters. Furthermore, we can represent each mean point of the clusters with a 5% interval of confidence, which correlates with the number of individuals present in each cluster [9].

Initially, to test the value of the method, we grouped the individual viruses at random into 3 groups and searched for a discrimination of these clusters with the AAs. Then we made 3 different analyses of clusterings of increasing complexity, i.e., based on particle structure (3 clusters), mode of transmission (7 clusters), and the ICTV classification (23 groups).

In the first SDA, the 3 categories of virus particles were:

I. Viruses with isometric particles: 53 individuals including alfalfa mosaic virus.

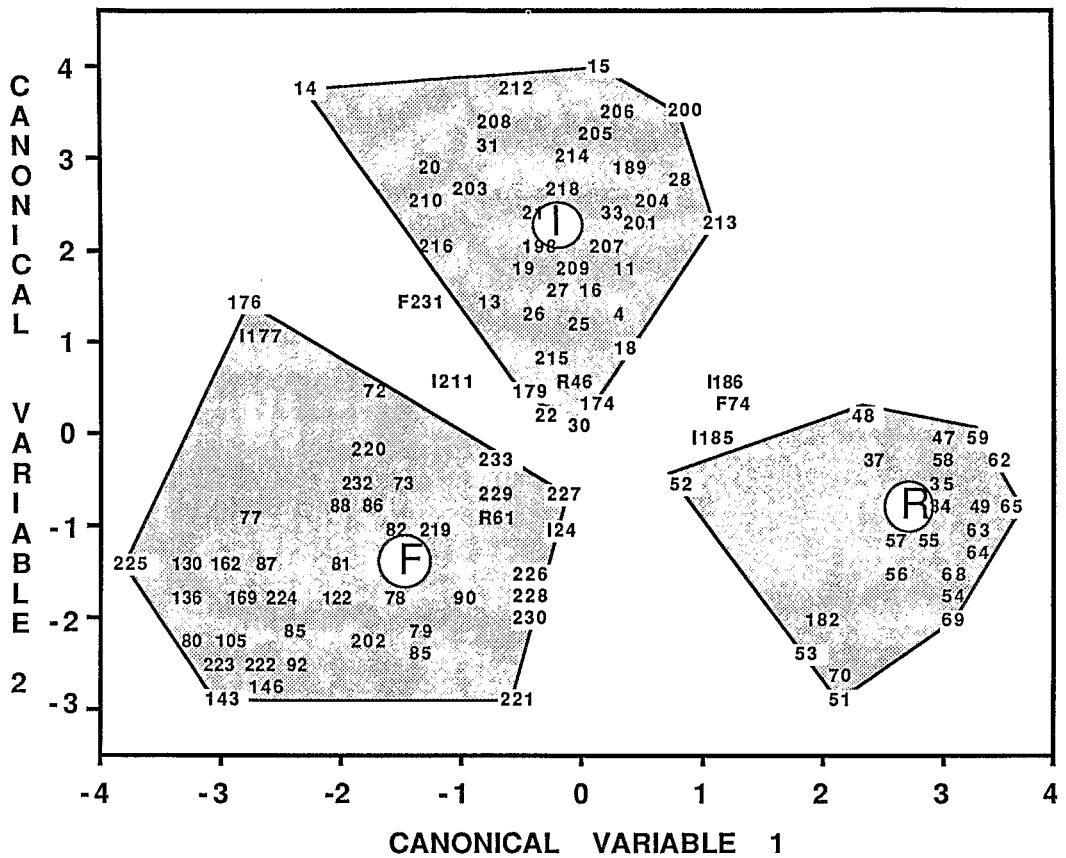R. Viruses with rod-shaped particles: 32 individuals.

Fig. 1. Two-dimensional diagram showing the first and second canonical variables of an SDA on the shape of plant virus particles. 126 data sets of CPs, compared by their AAC, were analyzed. The key for the code numbers is in table I of the MC [1]. The position of the mean for each cluster as well as the particle structure of the misclassified individuals are indicated by a letter: I for isometric, R for rod-shaped, and F for filamentous particles. The 5% interval of confidence of each cluster is visualized by a circle around the mean point. The limits of the clusters are indicated by a line joining all the outside members of each cluster.

F. Viruses with filamentous particles: 41 individuals.

In the second SDA, the 7 categories of mode of transmission, which excluded transmission by pollen, seed, or vegetative propagation were:

    A. Transmission by aphids: 39 viruses.
    C. Transmission by beetles: 30 viruses.
    W. Transmission by whiteflies: 3 viruses.
    N. Transmission by nematodes: 3 viruses.

F. Transmission by fungi: 12 viruses.
M. Mechanical transmission: 35 viruses.
U. Unknown mode of transmission: 4 viruses.
In the third SDA we used the ICTV plant virus classification, with 23 groups represented as follows:
    A. Bromovirus group: 3 viruses.
    B. Cocksfoot mild mosaic virus group: 1 virus.
    C. Comovirus group: 6 viruses.
    D. Cucumovirus group: 6 viruses.

Table IV. Classification matrix, indicating number and percentage of examples correctly classified into clusters A, C, W, N, F and M by SDA based on mode of transmission of plant viruses

Clusters	Correctly clustered %	Number of viruses classified into clusters						Total
		A	C	W	N	F	M	
A	85	33	1	1	2	2	0	39
C	87	2	26	0	0	2	0	30
W	100	0	0	3	0	0	0	3
N	100	0	0	0	3	0	0	3
F	75	0	0	0	2	9	1	12
M	91	1	1	0	1	0	32	35
Total	87	36	28	4	8	13	33	122

Table V. Classification matrix, indicating number and percentage of examples correctly classified into virus groups A to Y by SDA based on the ICTV classification

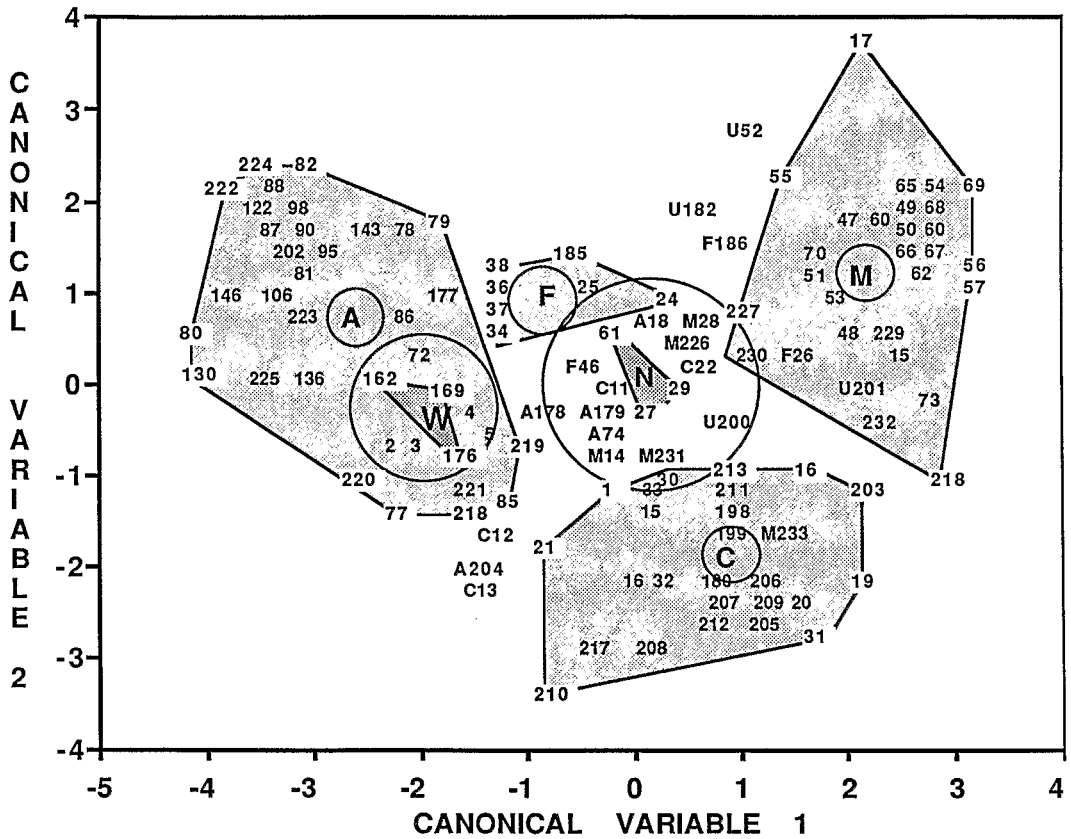| Clusters	Percentage in clusters	Number of viruses classified into clusters																							Total
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S	T	U	W	X	Y	
A	100	3																							3
B	100		1																						1
C	100			2																					2
D	100				6																				6
E	100					1																			1
F	100						2																		2
G	100							2																	2
H	100								1																1
I	100									5															5
J	100										2														2
K	80											4							1						5
L	100												17												17
M	100													3											3
N	100														3										3
P	100															1									1
Q	100																22								22
R	100																	1							1
S	75			1															6				1		8
T	100																			1					1
U	100																				3				3
W	100																					3			3
X	100																						9		9
Y	96																				1			24	25
Total	97	3	1	3	6	1	2	2	1	5	2	4	17	3	3	1	22	1	7	1	4	3	10	24	126

Fig. 2. Two-dimensional diagram showing the first and second canonical variables of a SDA on the mode of transmission of plant viruses. 126 data sets of CPs, compared by their AAC, were analyzed. The key for the code numbers is in table I of the MC [1]. The position of the mean for each cluster as well as the mode of transmission of the misclassified individuals are indicated by a letter: A for aphid transmission, C for beetle transmission, W for whitefly transmission, F for fungus transmission, N for nematode transmission, M for mechanical transmission, and U for unknown transmission. The 5% interval of confidence of each cluster is visualized by a circle around the mean point. The limits of the clusters are indicated by a line joining all the outside members of each cluster.

E. Dianthovirus group: 1 virus.

F. Ilarvirus group: 2 viruses.

G. Nepovirus group: 2 viruses.

H. Pea enation mosaic group: 1 virus.

I. Sobemovirus group: 5 viruses.

J. Tobacco necrosis virus group: 2 viruses.

K. Tombusvirus group: 5 viruses.

L. Tymovirus group: 17 viruses.

M. Tobacco necrosis satellite virus group: 4 viruses.

N. Alfalfa mosaic virus group: 3 viruses.

P. Hordeivirus group: 1 virus.

Q. Tobamovirus group: 22 viruses.

R. Tobravirus group: 1 virus.

S. Unclassified rod-shaped viruses: 8 viruses.

T. Carlavirus group: 1 virus.

U. 'Carlavirus' group (whitefly transmission): 3 viruses.

W. Closterovirus group: 3 viruses.

X. Potexvirus group: 12 viruses.

Y. Potyvirus group: 30 viruses.

Results

SDA of Viruses Placed in 3 Clusters at Random

When the 126 viruses were allocated to 3 clusters at random, the 18 AAs were unable to discriminate between the 3 clusters and consequently no viruses were classified a posteriori (table I). This proves that the 3 clusters considered do not represent 3 different entities according to the AAC of their CPs, and that the SDA method does not produce spurious clusters.

SDA According to Particle Structure

The SDA of 3 clusters (I, R and F) based on the shape of the virus particles gave the correct assignment to clusters in 93% of the examples (table I); the dispersion was 54% in the first canonical variable and 46% in the second. Figure 1 summarizes the information. Table II indicates the 9 AAs, which were discriminant for particle shape, with their respective F values. Figure 1 shows the position of the means of the 3 clusters, with a 5% confidence interval. The 3 clusters of individuals based on particle structure are completely separated, with only a few examples falling outside the boundaries of the clusters.

The percentage of examples correctly classified is consistently high for each particle shape (90, 94, and 95%) (table III). For the I cluster, only 5 viruses were misclassified. Two examples, tobacco necrosis virus (024) and alfalfa mosaic virus (177), were in the F cluster, and these examples also were wrongly placed in the MC [1]. In 3 other instances, tobacco necrosis satellite virus (185 and 186) and eggplant mosaic virus (211), the points were plotted in the intermediate zone between the clusters (fig. 1). Two viruses of the R cluster were misclassified, beet nec-

rotic yellow vein virus (046) in the I cluster and tobacco rattle virus (061) in the F cluster, perhaps because values for 1 or 2 discriminant AAs in their AAC are wrong. Only 2 viruses of the F cluster were misclassified, potato virus S (074) and white clover mosaic virus (231); both fall in the intermediate zone between the clusters (fig. 1).

SDA According to Mode of Transmission

This SDA with 6 clusters (A, C, W, N, F and M) classified 87% of the 122 viruses concerned (4 viruses with unknown mode of transmission were not used) (table I). The dispersion was 63% in the first canonical variable and 22% in the second. Thus, figure 3 represents 85% of the total information, and the 5 axes would be needed to get 100% of the information.

Table II lists the discriminant AAs and their respective F values. Figure 2 visualizes the 5% confidence interval, centered on the mean of each cluster. All the clusters were separated except the W and N clusters, because of the small numbers of individuals (3 for each). Nevertheless, considering the F-matrix of their means they were significantly distinct and the third axis would separate them. The two soil-transmitted clusters were in a central position, whereas the others were in an external situation. Figure 2 indicates the position of the 126 individuals together with the position of the clusters based on the mode of transmission using the two first axes.

Table IV lists the number of viruses in each cluster and the number misclassified. Between 75 and 100% of examples in each cluster were correctly classified. The 16 out of 122 viruses misclassified were from several of the clusters. Six examples were A viruses, namely pea enation mosaic virus (018), potato virus S (074), alfalfa mosaic virus
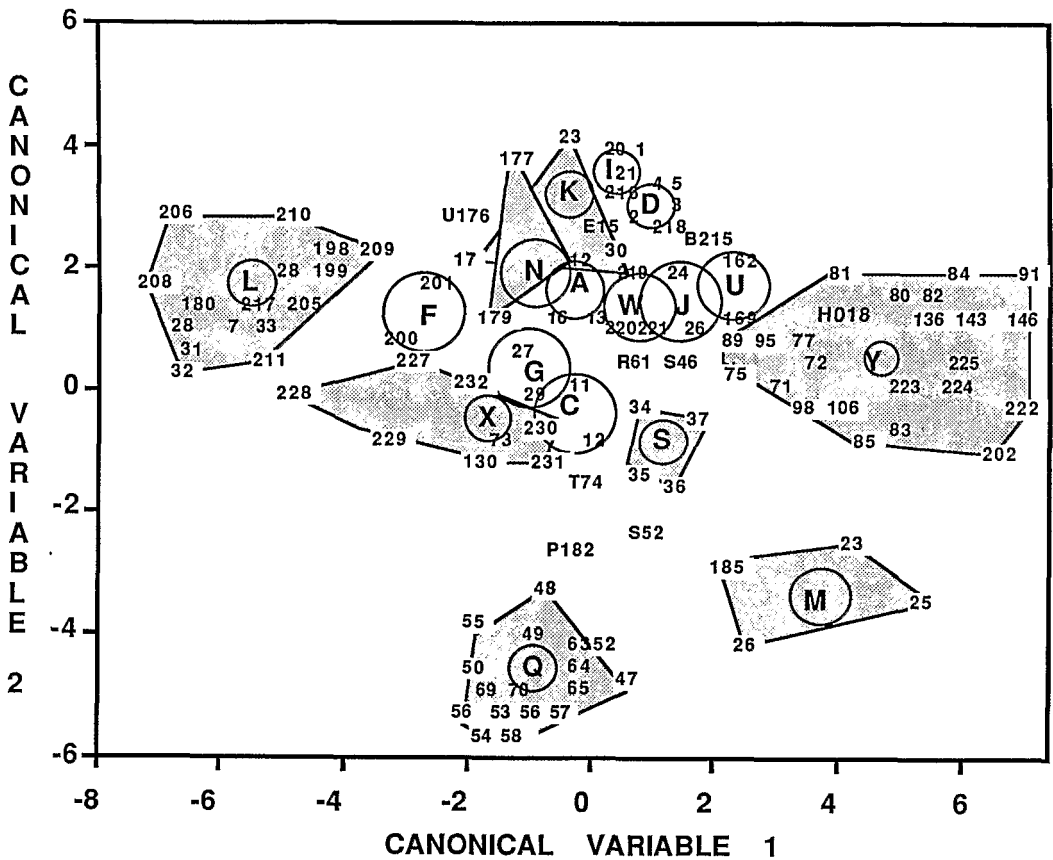
Fig. 3. Two-dimensional diagram showing the first and second canonical variables of a SDA on the grouping of plant viruses according to the ICTV classification. 126 data sets of CPs, compared by their AAC, were analyzed. The key for the code numbers is in table I of the MC [1]. The position of the mean for each cluster is indicated by a letter: A for bromovirus group, B for cocksfoot mild mosaic virus group, C for comovirus group, D for cucumovirus group, E for dianthovirus group, F for ilarvirus group, G for nepovirus group, H for pea enation mosaic virus group, I for sobemovirus group, J for tobacco necrosis virus group, K for tombusvirus group, L for tymovirus group, M for tobacco necrosis satellite virus group, N for alfalfa mosaic virus group, P for hordeivirus group, Q for tobamovirus group, R for tobravirus group, S for unclassified rod-shaped viruses, T for carlavirus group, U for whitefly transmitted carlavirus group, W for closterovirus group, X for potexvirus group, and Y for potyvirus group. The 5% interval of confidence of each cluster (except for clusters with only one individual) is visualized by a circle around the mean point. The limits of the clusters are indicated by a line joining all the outside members of each cluster.

(177–179), and peanut stunt virus (204). Four examples were C viruses, including two comoviruses, bean pod mottle virus (011) and squash mosaic virus (022), and two bromoviruses, broad bean mosaic virus (BBMV) (012) and brome mosaic virus (BMV) (013). However, for the comoviruses we do not know whether the AACs represent the major CP or both proteins that may not have equal relevance to transmission. This could explain the

Table VI. Repartition of the 23 virus groups of the ICTV classification in the clusterings of shape and mode of transmission of plant virus particles

	I	R	F
A	D, H, N		T, W, Y
C	A, B, C, I, K, L		
W			U
N	G	R	
F	J, M	S	
M	E	Q	X
U	F, K	P	

results obtained. The bromovirus group is also represented by cowpea chlorotic mottle virus (CCMV) (016), which was correctly classified. We considered this group to be transmitted by beetles even though this has not been proven for BMV [10]. CCMV, which is known to be transmitted by beetles [11], is really situated in the C cluster, and BBMV, which also is transmitted by beetles [12], is on the edge of the C cluster (fig. 2). The experimental transmission of BMV by nematodes and by mechanical inoculation [10] is not consistent with its position in figure 2.

Among the F viruses, 3 out of 12 were misclassified: tobacco necrosis satellite virus (186), tobacco necrosis virus (026), and beet necrotic yellow vein virus (046). Finally, 3 out of 35 M viruses were misclassified: cauliflower mosaic virus (014), white clover mosaic virus (231), and plantain virus X (233).

Four other viruses were not allocated to the above clusters because their mode of transmission is unknown (3 of them are pollen-borne, but this mode of transmission was not considered here). *Chara corallina* virus (052) was classified in the M cluster with a probability of 0.59; this hypothesis, if

confirmed by transmission tests, would consolidate the assignment of this virus to the tobamovirus group. Barley stripe mosaic virus (182) was classified in the F cluster with a probability of 0.61. This suggests it has a fungal vector and hence, it would be interesting to test whether it is transmitted by plasmodiophoraceous fungi. Prunus necrotic ringspot virus (200) and Tulare apple mosaic virus (201), both representing the ilarvirus group, are classified in the M and N clusters, respectively, confirming the assessment of the MC [1], i.e., that their proteins differ greatly and no conclusion can thus be reached about this group.

SDA According to ICTV Classification

The SDA according to the classification of plant viruses of the ICTV [2] within 23 groups (A to Y) correctly classifies 97% of the individuals (table I). The dispersion was 33% in the first canonical variable and 20% in the second; hence, figure 3 represents 53% of the information. A third axis would give 14% more, and it is necessary to reach the fifth axis to get 90%. Table II gives the 15 discriminant AAs, with their respective F values, which were necessary to discriminate the 23 groups of viruses.

Table V lists the 23 groups and gives the percentage of correctly classified viruses and the number of misclassified viruses for each group. This percentage was 100% for all groups, except tombusviruses (80%) and the unclassified rod-shaped viruses (75%). Six groups having only one virus are well classified automatically and, consequently, they have not been counted for a correct evaluation. The 4 misclassified viruses are: turnip crinkle virus (030) in the tombusvirus group; beet necrotic yellow vein virus (046) and the *Chara corallina* virus (052) for the unclassi-

fied rod-shaped virus group; and potato virus Y (077) in the potyvirus group.

The 5% intervals of confidence for the SDA of the 23 groups are indicated in figure 3, except for the 6 groups having only one individual, which are represented by letters. Most of the clusters were correctly isolated, but some were partially mixed in the center of the ordination. Considering the F-matrix of the 23 groups, all groups were significantly differentiated from the others, except 5 of the 6 groups composed of one individual. Consequently, the third axis will scatter the apparently mixed clusters in the center of the diagram. Figure 3 represents the position of the 126 individuals in the 2 first canonical variables. Some clusters are clearly separated from the others: potyviruses (Y), tymoviruses (L), tobamoviruses (Q), and the tobacco necrosis satellite virus group (S); some others are in the center of the ordination: potexviruses (X), tombusviruses (K), bromoviruses (A), comoviruses (C), etc.

Discussion

The failure of the SDA to find parameters to describe clusters of randomly selected viruses confirms that the method does not produce artifacts. Hence, the results of SDA according to particle shape, mode of transmission, and conventional classification support the ideas outlined in the 'Introduction'. The AAC of CPs of particles of plant viruses thus are adapted to the particle shape and the mode of transmission. Each CP is, therefore, suitable for particles of only one kind of shape and with only one mode of transmission. Common feature of CPs of several spherical viruses also were confirmed by high-resolution structural studies [13].

For each analysis there was good general correlation: 93, 87 and 97% of viruses were correctly classified in the three analyses, and almost all clusters (except groups represented by one member) were completely separated from other clusters within 3 axes. The good results obtained with the 23 groups are not a simple superposition of the results of the two first SDA because many groups have the same shape and different vectors, and vice versa (table VI). Nevertheless, they are differentiated by their AAC with a high degree of significance. In contrast, the results of the SDA on the shape and mode of transmission do not reflect those of the SDA on the 23 groups for the same reason as above, and also because the statistical probabilities of the discrimination F values are much greater in the first two SDA (table II). This might indicate a degree of importance of discrimination in the AAC: it is easier to discriminate first the shape (mean $F = 21.8$), then the mode of transmission (mean $F = 9.1$), and finally the groups (mean $F = 7.8$).

It is obvious that 14 out of 15 discriminant AAs used in the third SDA to discriminate the 23 groups of the ICTV classification correlated with the three axes of the MC [1]. Twelve AAs correlated with axis one, one correlated with axis two, and one correlated with axis three of the MC [1]. The 3 ionically charged AAs ASP, HIS, LYS, and PRO, which are involved in the structure of proteins, are present in the 3 different SDA. The AAs ALA, GLU, MET, and VAL are common to the SDA realized on the shape and on the groups. The AAs GLY, PHE, and TRP are common for the SDA on the mode of transmission and on the groups. THR is discriminant only for the first and the second SDA, whereas 4 more AAs are needed in the third SDA to discriminate all 23 groups.

In both SDA and MC, several data sets are always misclassified: 074, 177, 186, 231, ..., suggesting these AACs may be completely wrong. In other instances, data sets were correctly classified in the MC and in one SDA but not in the others; we consider that this is because SDA is more precise than MC. In contrast to the first axis of the MC, the molecular weight of the CPs is not important in the SDA; however, a change in one or two discriminant AAs can completely change the position of a virus.

The results of SDA, therefore, provide evidence that the AAC of plant virus CPs is adapted to formation of particles of a particular shape, to a particular mode of transmission, and probably to other unspecified properties. All such properties are taken into account in virus classification and it is therefore, not surprising to be able to discriminate by these criteria all 23 groups of the ICTV classification even though 15 AAs are needed to do this. The CPs of plant viruses, therefore, contain a large amount of information in a molecule of 150 to 450 AAs, representing 500 to 1,500 bases of nucleic acid.

References

1 Fauquet, C.; Dejardin, J.; Thouvenel, J.-C.: Evidence that the amino acid composition of particle proteins of plant viruses is characteristic of the virus group. I. Multidimensional classification of plant viruses. Intervirology 25:1–13 (1986).
2 Matthews, R.E.F.: Classification and nomenclature of viruses. Fourth Report of the International Committee on Taxonomy of Viruses. Intervirology 17:1–200 (1982).
3 Tremaine, J.H.; Argyle, E.: Cluster analysis of viral proteins. Phytopathology 60:654–659 (1969).
4 Gibbs, A.J.: Tobamovirus group. CMI/AAB Descriptions of plant viruses, No. 184 (1977).
5 Paul, H.L.; Gibbs, A.; Wittman-Liebold, B.: The relationships of certain tymoviruses assessed from the amino acid composition of their coat proteins. Intervirology 13:99–109 (1980).
6 Van Regenmortel, M.H.V.: Immunochemical studies; serology and immunochemistry of plant viruses, p. 302 (Academic Press, New York 1982).
7 Romeder, J.R.: Méthodes et programmes d'analyse discriminante (Dunod, Paris 1973).
8 Dixon, W.J.; Brown, M.B.: Stepwise discriminant analysis; in BMDP; 79th ed., Berkeley University (California Press, 1979).
9 Lefebvre, J.: Analyse discriminante; introduction aux analyses statistiques multidimensionnelles, p. 189 (Masson, Paris 1980).
10 Lane, L.C.: Brome mosaic virus. CMI/AAB Descriptions of plant viruses, No. 215 (1979).
11 Walters, H.J.; Dodds, N.L.: Identification and beetle transmission of an isolate of cowpea mottle virus from Desmodium. Phytopathology 59:1055 (1964).
12 Walters, H.J.; Surin, P.: Transmission and host-range studies of broad bean mottle virus. Pl. Dis. Reptr 57:833–836 (1973).
13 Harrison, S.C.: Virus structure: high resolution prospectives. Adv. Virus Res. 28:175–240 (1983).