Intervirology
Editor-in-Chief: J.L. Melnick, Houston, Tex.

Taxonomy

Reprint
Publisher: S. Karger AG, Basel
Printed in Switzerland

Intervirology *25:* 1–13 (1986)

0300–5526/86/0251–0001$2.75/0

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

I. Multidimensional Classification of Plant Viruses

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

Key Words. Coat protein · Amino acid composition · Plant virus · Classification · Principal components analysis

Summary. The amino acid (AA) contents of the coat proteins of 134 plant viruses and strains were classified by principal components analysis. The virus groupings that were obtained correlated well with the classification of *Matthews*. The relationships of each virus were dependent on the number of AA residues (axis 1) and on the percentage composition of each AA in the proteins (axes 2–4). The classification indicated which data were anomalous and needed confirmation. There seemed to be more anomalies in estimates of protein size than of protein composition.

Tremaine and Goldsack [1] attempted, without success, to determine if there was a relationship between the amino acid composition (AAC) of the coat proteins (CPs) of the particles of plant viruses and the shapes of those particles. *Tremaine and Argyle* [2], using an agglomerative method of sorting strategy and the Euclidean distance metric, could not correlate the AAC of the CPs of plant viruses with groupings based on other classifications [3–5]. *Gibbs* [6] chose the same criterion in an attempt to classify 66 plant viruses by using the nonmetric coefficient of *Lance and Williams* [7] and the principal coordinates method [8]. The analysis distinguished only tobamoviruses and tymoviruses. Nevertheless, a hierarchical agglomerative classification of those viruses not separated by the ordination, using a nonmetric coefficient and flexible sorting [9], showed a general clustering of viruses belonging to the same group, e.g., bromoviruses and sobemoviruses.

Similarly, *Gibbs and Harrison* [10] studied tobamoviruses and found a close correlation between a classification based on the AAC of the CPs and the groupings proposed by *Tsugita* [11] and *Van Regenmortel* [12]. They also demonstrated a close correlation (0.832) be-

Address inquiries to: Dr. C. Fauquet, Phytovirologie, ORSTOM, BP V-51, Abidjan (Côte d'Ivoire)

Received: April 18, 1984
Revised: April 9, 1985

tween a computer classification based on the amino acid (AA) sequences of the CPs of 6 tobamoviruses and one based on their AAC. *Gibbs* [13] and *Paul* et al. [14] showed that for tobamoviruses there is a linear correlation (0.833) between similarities based on the AAC of the CPs and those based on serological relationships. Consequently, for tobamoviruses it seems clear that groupings based on biochemical criteria are correlated with serological relationships; such groupings are related to those based on sequences of the AAs of these CPs. By contrast, in a study of the tymoviruses, *Paul* et al. [14] concluded that, although there is a general similarity between the classification based on the AAC of the CPs and that obtained from serological relationships, the coefficient of correlation (0.369) is poor. *Moghal and Francki* [15], working with potyviruses, concluded that: 'the AAC of antigenically closely related viruses were very similar, but similarities of those distantly related were no greater than those of the apparently unrelated viruses'.

We have reexamined the potential uses of these methods, using new data on the AAC of the CPs and improved methods of statistical analysis. Several classification methods were tested, and the results obtained by the method that gave the best correlation with classifications obtained by other methods using different data are presented here. Our aim was to compare all known data on the AAC of the CPs to discover how well the classification obtained correlated with that of *Matthews* [16], which is now widely used.

Materials and Methods

We collected all published data on the AAC, as well as some new AACs obtained for viruses isolated in the Ivory Coast [17]. If the amount of a particular AA was unknown (e.g., cys or trp), we replaced it either by the average amount in the CPs of the other strains of the same virus or by the average amount in the CPs of the other viruses belonging to the same group. When it was not possible to estimate values in this way, it was assumed that one residue of the AA was present. The AAC data used, expressed in numbers of AA residues per molecule and grouped according to the usual accepted classification, are given in table I.

The classification method used was a principal components analysis by the ANCOMP program from the ADDAD library.[1] The estimated numbers of AA residues in each protein were the quantitative variables, and the principal components analysis was done with a Euclidean metric of the data after standardizing them to zero mean and unit variance, i.e., the Eigenstructur was searched in the correlation matrix [18].

The objective of a principal components analysis is to find a small number of linearly independent combinations (principal components) that keep the maximum information of the original variables. The results can be expressed graphically by representing the cluster of individuals as 3-dimensional diagrams that have a minimum of anomalies. The total variation is expressed by a few components without any great loss of information: the first principal component is that which accounts for most of the information (variability) and corresponds to the longest axis of the total cluster of individuals; the second component is orthogonal to it (uncorrelated) and takes a maximum of the residual variability; etc.

Results

The first four axes obtained with the principal components analysis accounted, respectively, for 39.6, 14.8, 7.6 and 6.1% of the total information available in the AAC of the CPs of the viruses. In other words, axes 1, 2

[1] ADDAD (Association pour le Développement et la Diffusion de l'Analyse des Données) library is available at the CIRCE (Centre Interrégional de Calcul Electronique), CNRS-Orsay (Centre National de la Recherche Scientifique).

and 3 (fig. 1) account for 62% of the variability and result in clusters that correlate well with the currently accepted groups; only the bromoviruses seem to be intermingled with the potexviruses. Axis 1 correlates with 12 of the 18 AAs, and the relative contribution of any one AA does not exceed 10% (table II). This axis mostly represents the molecular weights (MWs) of the CPs, which range from 17,500 (17.5K) for tobamoviruses to 45K for tombusviruses. Axes 2, 3 and 4 correlate to particular AAs. For example, axis 3 is statistically correlated only to tryptophan (trp) content, which represents 42% of the variability in this dimension (table II). Axes 2, 3 and 4 (fig. 2) represent only 28.5% of the total information, but clearly differentiate the viruses into groups, although they are less well separated. Whether considering figure 1 or 2, most of the virus groups are clearly separated from one another; however, the potexviruses are very close to the bromoviruses, comoviruses, and nepoviruses.

Viruses with Rod-Shaped Particles

The tobamoviruses (23 data sets) were all situated in a restricted part of the ordination and showed great homogeneity. The exception was CCV (No. 052), a tentative member of the tobamovirus group [13].

Tobraviruses (1 data set; No. 061) were classified very close to *Chara corallina* tobamovirus (No. 052).

The hordeiviruses, represented only by barley stripe mosaic virus (No. 182), and the furoviruses [19], represented by beet necrotic yellow vein virus (No. 046) and peanut clump virus (No. 034–039), were distinct from the tobamoviruses and the single tobravirus (No. 061). Peanut clump virus seemed to be the most clearly differentiated from all the rod-shaped viruses.

All the rod-shaped virus groups were relatively close together in the ordination, showing that the AAC of their CPs is homogeneous.

Viruses with Filamentous Particles

The carlaviruses were represented by 4 data sets: potato virus S (No. 074) and 3 viruses related to cowpea mild mottle virus (No. 162, 169, and 176). These 3 viruses have properties similar to those of carlaviruses, but they are transmitted by whiteflies instead of aphids and their intracellular inclusions are different [20]. Except for No. 074, all are clustered and are close to the potyvirus group.

Potexviruses were represented by 12 data sets (4 of potato virus X and 2 of white clover mosaic virus). This group was the most scattered, perhaps because of the difficulty in determining the MWs of their CPs; estimates range from 103 AAs for data set No. 184 to 463 AAs for No. 183. *Gibbs and McIntyre* [21] suggested that the AA number for potexviruses is around 210–215, which agrees with that published by *Miki and Knight* (No. 073) and by *Short* (No. 226–232). Because the differences are very large and would unnecessarily complicate the figures, we avoided representation of the value 133 AAs for white clover mosaic virus (No. 076) and of the values 103 and 463 AAs for potato virus X (No. 184 and 183). Nevertheless, the AAC of these data are not wrong, and their position in figure 2 is accurate. Except for the last 3 examples, the cluster of potexviruses is clearly delimited in space and close to several virus groups with isometric particles (fig. 1, 2).

The potyvirus group (29 data sets) was the best represented group. Just as for potexviruses, there is uncertainty in the MWs of their

Table I. List of the AAC of the CPs of 134 plant viruses with isometric, bacilliform, rod-shaped, and filamentous *particles*[1]

DATA N°	VIRUS NAME	ASP	THR	SER	GLU	PRO	GLY	ALA	CYS	VAL	MET	ILE	LEU	TYR	PHE	HIS	LYS	ARG	TRP	TOTAL	REF
FUROVIRUS GROUP																					
034	PCV	27	8	13	25	12	26	21	1	20	0	11	18	5	7	5	7	19	3	228	#6
035	PCV	27	8	13	25	13	27	21	1	19	0	11	18	6	8	5	7	19	3	231	#6
036	PCV MJ	27	7	12	25	12	26	21	1	20	0	11	17	6	8	5	7	19	3	227	#6
037	PCV MJ	26	8	13	25	13	26	21	1	19	0	11	18	6	8	5	8	19	3	230	#6
038	PCV S	28	6	13	24	12	30	22	1	18	0	11	18	6	8	4	7	19	3	230	#6
039	PCV S	27	8	14	25	13	26	21	1	18	0	11	18	6	8	5	7	19	3	230	#6
046	BNYVV	24	15	18	14	10	14	19	2	14	7	5	19	4	6	2	12	10	4	198	#36
HORDEIVIRUS GROUP																					
182	BSMV	25	9	9	19	12	8	20	0	10	0	6	21	8	7	4	7	17	5	187	#38
TOBAMOVIRUS GROUP																					
047	FMV	17	13	14	16	4	9	14	1	13	0	11	13	5	7	1	4	11	5	158	#3
048	SOV	12	11	9	12	6	5	12	1	9	1	7	12	5	6	0	4	8	2	122	#3
049	TMV	18	16	16	16	8	6	14	1	14	0	9	12	4	8	0	2	11	3	158	#3
050	ToMV	18	16	15	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3	157	#3
051	HRV	17	13	13	22	8	4	18	1	10	3	8	11	7	6	1	2	10	2	156	#3
053	HRV	16	13	16	21	9	3	17	1	10	4	7	12	7	5	1	2	11	3	158	#3
054	T2MV	22	19	10	16	10	4	18	1	12	2	8	11	6	8	0	1	8	2	158	#3
055	CGMMV	20	10	24	10	6	9	21	0	7	0	7	18	4	9	1	4	8	2	160	#3
056	ORSV1	20	21	12	15	9	7	11	1	10	3	8	14	6	7	0	1	10	3	158	#3
057	ORSV2	20	21	12	15	9	7	11	1	9	3	9	14	5	7	0	1	10	3	157	#3
058	CV41	18	11	24	10	9	6	19	0	12	0	7	14	4	11	0	4	9	1	159	#3
059	CV42	20	12	23	10	8	5	20	0	13	0	6	13	4	11	0	4	10	1	160	#3
060	SHMV	18	19	18	16	8	4	12	0	12	0	10	15	8	6	1	1	12	1	161	#3
062	TMV J1481	17	16	17	15	8	6	14	1	14	0	9	12	4	8	0	3	11	3	158	#2
063	TMV YA	19	17	14	16	8	6	14	1	14	0	8	12	4	8	0	2	12	3	158	#2
064	TMV GA	19	17	15	16	8	5	14	1	14	0	8	12	4	8	0	2	12	3	158	#2
065	TMV OM	19	15	16	16	8	6	14	1	15	0	8	12	4	8	0	2	12	3	159	#39
066	DAHLE	17	17	16	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3	158	#2
067	Y TAMV	18	17	15	19	8	6	11	1	15	1	7	13	5	8	0	2	9	3	158	#2
068	G TAMV	22	19	10	16	10	4	18	1	12	2	8	11	6	8	0	1	8	2	158	#2
069	U2	22	19	10	16	10	5	17	1	12	2	8	11	6	8	0	1	8	2	158	#40
070	HR	17	13	13	22	8	4	18	1	10	3	8	11	7	6	1	2	10	2	156	#41
052	CCV	25	14	15	15	9	12	14	0	8	3	12	10	4	14	1	10	8	0	174	#3
TOBRAVIRUS GROUP																					
061	TRV	20	10	21	16	13	7	21	1	8	3	3	14	5	11	1	15	10	1	180	#37
CARLAVIRUS GROUP																					
074	PVS	16	8	10	17	11	9	13	1	10	6	10	9	3	4	3	4	11	1	146	#1
"CARLAVIRUS" GROUP																					
162	VoMV	36	20	22	32	16	20	27	3	13	7	14	24	9	11	7	19	15	2	297	#6
169	GCV	32	18	22	34	15	26	27	5	14	8	13	24	9	12	7	18	14	3	301	#6
176	PNMV	30	22	22	26	19	27	34	5	17	3	18	26	7	13	7	19	10	3	308	#6
CLOSTEROVIRUS GROUP																					
219	BYV	16	15	16	22	8	17	17	6	7	1	7	26	4	11	5	14	12	0	204	#52
220	BYV	22	17	21	18	8	21	17	3	6	1	9	31	3	12	6	17	12	0	224	#53
221	CNFV	23	20	15	24	9	15	18	3	9	0	5	30	9	13	1	14	12	0	220	#53
POTEXVIRUS GROUP																					
073	PVX	19	24	14	15	14	11	38	2	11	6	10	8	2	10	2	10	8	6	210	#44
076	WCMV	12	11	10	9	8	7	19	2	7	2	9	10	3	6	2	8	6	2	133	#46
183	PVX	43	41	23	31	29	25	74	1	22	12	19	18	4	22	3	16	15	5	403	#49
184	PVX	9	13	7	9	8	5	17	1	6	3	5	4	1	5	1	5	4	2	105	#1
226	FMV	25	16	9	20	14	7	27	2	11	3	7	11	7	8	1	13	9	2	192	#51
227	NMV	24	13	15	18	20	13	32	2	13	3	8	21	6	9	2	10	10	3	222	#51
228	VMV	14	25	17	18	17	14	27	2	8	1	7	24	3	10	1	5	9	2	204	#51
229	PMV	18	17	23	21	18	8	27	2	11	4	11	13	4	12	1	10	5	2	207	#51
230	CYMV	20	17	19	20	10	10	21	2	7	1	8	15	6	9	3	10	8	3	189	#51
231	WCMV	16	17	15	14	13	11	27	3	10	2	13	14	4	9	4	12	8	3	195	#51
232	PVX	19	24	14	16	15	11	38	3	11	6	10	8	2	10	2	11	9	4	213	#51
233	PlCV	19	20	20	28	22	17	27	2	18	5	11	17	6	6	1	10	13	2	244	#51
POTYVIRUS GROUP																					
071	TEV	25	13	9	23	8	13	19	1	12	10	5	13	7	5	6	10	13	2	194	#42
072	TUMV	29	16	10	23	9	15	17	1	12	10	11	20	8	9	8	13	17	2	230	#43
075	PVY	22	13	10	23	11	13	16	1	13	8	12	10	6	5	4	13	11	2	193	#45
077	PVY	33	24	18	34	18	18	26	1	16	7	15	18	10	6	6	18	16	3	287	#5
078	BYMV	42	20	15	33	11	21	21	1	16	9	14	22	11	9	6	20	17	5	293	#5
079	PMV	40	20	15	33	11	22	22	1	16	7	15	22	13	9	4	19	17	4	290	#5
080	LMV	44	19	12	32	11	23	26	1	11	12	12	20	14	6	9	18	16	4	290	#5
081	SPMV	42	20	13	33	10	22	22	1	16	8	15	21	11	8	5	22	18	4	291	#5
082	SCMV	47	19	22	25	11	19	26	1	16	9	11	16	11	9	6	22	16	4	290	#5
083	PVY	22	13	10	22	11	13	16	1	13	8	12	10	6	5	4	13	11	2	192	#45
084	TEV	25	13	9	23	8	13	19	1	12	10	5	13	7	5	6	10	13	2	194	#42
085	MDMV B	27	25	20	29	10	34	23	1	12	11	8	13	9	7	5	12	14	4	264	#48
087	PWV	46	18	14	31	10	21	27	1	19	18	7	21	10	8	5	22	14	3	295	#5
088	BCMV	47	16	16	28	15	19	22	1	18	18	7	20	10	8	6	20	17	3	291	#5
089	PVY	17	16	9	22	10	10	15	1	13	11	9	12	7	5	5	10	13	2	187	#47
090	PVMV	39	15	17	38	10	19	21	2	20	12	13	25	10	9	8	14	18	2	292	#6
095	PRSV	40	18	16	36	12	21	23	3	20	12	9	21	10	9	6	17	19	2	294	#6
098	GGMV A	41	14	17	35	14	21	25	3	12	12	14	17	11	11	8	16	19	2	292	#6
105	GGMV C	46	19	18	33	23	26	26	2	11	11	11	13	10	9	7	18	17	2	302	#6
122	GGMV B	42	13	16	32	13	17	23	4	15	12	12	19	12	10	8	19	19	3	289	#6
130	YMV	35	15	19	39	14	23	26	3	16	15	15	24	13	11	9	16	11	1	305	#6
136	CuMV	37	19	13	34	10	20	27	4	14	14	14	24	11	8	8	18	15	3	293	#6
143	CanMV	42	17	15	35	13	19	25	4	12	13	12	22	10	9	6	18	16	3	291	#6
146	GESV	39	16	19	39	10	19	23	3	14	11	13	23	10	8	8	17	16	2	290	#6
202	PSBMV	39	17	18	41	12	22	27	1	20	16	15	18	10	9	8	12	21	1	307	#50
222	GGMV A	44	15	18	37	14	22	26	3	13	13	14	18	12	11	9	17	21	2	309	#6
223	GGMV C	47	20	19	34	24	27	26	2	11	11	12	13	10	9	7	18	17	2	309	#6
224	GGMV B	45	14	17	35	15	19	25	5	16	13	13	21	13	11	8	20	21	3	314	#6
225	YMV	35	15	19	40	14	23	26	4	16	15	15	25	13	11	9	18	11	2	311	#6

[1] See footnote on p. 6

DATA N°	VIRUS NAME	ASP	THR	SER	GLU	PRO	GLY	ALA	CYS	VAL	MET	ILE	LEU	TYR	PHE	HIS	LYS	ARG	TRP	TOTAL	REF
BROMOVIRUS GROUP																					
012	BBMV	14	10	18	17	9	10	23	2	23	2	7	19	4	7	2	15	12	0	194	#8
013	BMV	10	11	13	18	7	10	33	1	18	3	8	15	5	5	4	13	13	2	189	#9
016	CCMV	11	17	16	16	7	10	25	2	19	1	7	16	5	4	2	12	9	4	183	#11
CMMV GROUP																					
215	CMMV	20	18	19	20	16	21	21	3	16	8	9	18	10	6	3	8	17	3	236	#4
COMOVIRUS GROUP																					
011	BPMV	21	14	18	17	13	21	14	1	15	7	12	18	2	11	3	9	6	1	203	#7
022	SqMV	21	17	16	14	10	15	19	1	9	4	14	19	3	10	3	8	7	1	191	#15
CUCUMOVIRUS GROUP																					
002	CMV	28	13	31	18	17	19	20	2	21	4	10	22	9	7	4	15	20	1	261	#6
003	CMV	28	14	31	17	17	18	20	2	21	3	10	22	9	7	4	15	20	1	259	#6
004	CMV	29	14	31	18	16	20	24	2	19	4	10	22	10	6	3	13	19	1	261	#6
005	CMV	26	14	26	17	16	14	23	2	20	4	10	23	10	6	3	14	24	1	253	#6
204	PSV	16	15	19	14	13	10	13	2	17	0	6	15	5	5	5	12	12	1	180	#32
218	CMV	22	13	24	15	14	12	13	0	16	6	12	20	8	4	3	14	18	1	215	#35
DIANTHOVIRUS GROUP																					
015	CaRSV	34	37	37	23	20	20	24	3	36	7	16	26	16	12	2	14	16	4	347	#10
ILARVIRUS GROUP																					
200	PNRSV	25	16	11	18	25	13	11	5	22	4	8	16	4	8	6	9	17	5	223	#30
201	TuAMV	16	10	16	11	19	13	16	1	14	3	4	5	6	9	2	11	7	7	170	#30
NEPOVIRUS GROUP																					
027	ToRSV	17	13	14	14	11	15	15	5	11	3	11	14	6	10	6	9	8	5	187	#19
029	TomRSV	17	15	16	18	11	18	15	5	10	3	13	24	7	14	5	10	11	5	217	#21
PEMV GROUP																					
018	PEMV	21	13	16	14	11	21	17	3	13	3	7	10	5	7	4	11	21	2	199	#12
SOBEMOVIRUS GROUP																					
001	RYMV	24	19	30	13	19	18	28	6	20	9	8	19	9	5	3	10	17	3	260	#6
019	SBMV	18	32	26	18	14	19	24	3	21	7	12	28	10	4	2	7	20	5	270	#13
020	SBMV	21	30	17	19	18	16	28	4	23	9	14	23	9	4	2	12	16	5	270	#13
021	SoMV	16	13	14	12	12	16	15	2	13	5	9	12	7	4	3	12	8	3	176	#14
216	CFMV	19	22	24	17	17	22	21	2	16	7	7	18	7	9	4	12	19	8	251	#4
TNV GROUP																					
024	TNV	34	19	21	24	23	25	41	5	18	5	20	21	15	10	1	10	20	1	313	#17
026	TNV	18	16	14	20	15	8	13	2	14	6	11	10	11	12	1	12	14	1	198	#18
TOMBUSVIRUS GROUP																					
014	CaMV	36	36	28	31	24	29	31	5	35	9	19	28	10	14	1	25	19	2	382	#1
017	CuNV	46	31	32	24	23	32	41	0	33	1	20	33	12	20	3	16	17	7	391	#1
028	ToBSV	44	45	35	21	16	38	37	3	40	3	13	43	10	14	5	13	20	2	402	#20
030	TUCV	14	14	12	16	9	15	17	1	12	2	5	11	4	6	1	12	9	4	164	#22
203	SaCV	28	36	36	25	25	23	37	8	31	3	17	24	12	12	2	15	23	4	361	#32
TYMOVIRUS GROUP																					
031	TYMV	11	26	16	15	20	8	15	4	14	4	15	18	3	5	3	7	3	2	189	#22
032	TYMV C	17	20	20	15	20	7	13	5	14	4	13	19	3	3	5	4	6	2	190	#22
033	WCuMV	15	13	26	11	19	9	16	2	12	1	13	24	3	8	3	9	5	1	190	#22
180	BeMV	11	16	24	17	16	13	16	2	15	2	18	18	4	5	0	9	5	1	192	#26
198	KYMV	14	24	23	12	17	9	20	1	11	2	15	22	6	4	4	5	0	1	190	#28
199	EMV	16	20	20	13	19	7	25	1	18	3	13	20	4	6	3	7	4	1	200	#29
205	ScrMV	14	21	27	16	22	12	13	0	16	4	14	17	3	5	3	7	8	2	204	#33
206	APLV	12	17	28	11	22	11	19	2	16	4	11	24	5	6	2	7	3	2	202	#4
207	BMV	12	17	25	19	17	13	18	2	16	2	17	18	5	5	0	10	5	1	202	#4
208	CYVV	13	22	26	10	20	8	22	4	11	0	20	22	8	4	5	7	1	1	204	#4
209	DMV	10	19	20	20	15	13	20	1	19	3	13	21	5	6	0	10	6	2	203	#4
210	DYMV	15	24	18	15	19	8	21	2	12	2	12	25	9	3	6	10	1	1	203	#4
211	EMV	17	22	19	15	18	8	27	3	8	3	13	20	4	7	3	8	4	1	200	#4
212	DMV	15	26	19	10	20	11	22	4	15	2	18	18	6	6	5	6	4	1	208	#4
213	OYMV	15	12	32	17	21	9	10	2	15	3	14	22	4	6	2	6	8	2	200	#4
214	SCMV	15	20	25	16	21	12	13	3	15	4	15	17	3	6	3	8	8	2	206	#4
217	EryLV	11	20	28	21	21	17	14	0	16	3	8	22	3	7	7	4	5	1	208	#34
STNV GROUP																					
023	STNV	57	34	24	32	9	28	28	3	25	9	25	29	8	14	8	15	25	1	374	#16
025	STNV	27	25	12	18	4	8	9	2	13	4	13	20	6	11	6	11	24	1	214	#18
185	STNV	31	19	14	17	4	18	16	1	14	5	14	16	4	7	4	8	14	2	208	#27
186	STNV	25	26	17	18	3	12	18	2	19	3	13	9	4	11	3	8	15	1	207	#27
AMV GROUP																					
177	AMV	30	15	20	25	23	21	30	5	16	4	7	29	5	20	8	21	16	2	297	#23
178	AMV	17	8	9	17	13	19	17	2	11	2	5	16	4	11	4	11	7	1	174	#24
179	AMV	16	10	12	16	14	14	17	2	9	2	9	2	4	13	5	11	8	1	165	#25

CPs, with estimates ranging from 21K to 37K. The sizes of these proteins have not been tested by the Fitmol method [21], but it is recognized that the MW of the CPs of potyviruses is about 34K [22]. As for the potexviruses, we avoided the representation of the values corresponding to low MWs of CPs (No. 071, 075, 083, 084, and 089). Nevertheless, the AAC of these data are possibly quite accurate, because in figure 2 (which does not take into account the MW factor) they would integrate well into the potyvirus cluster. This suggests that the main error may be in MW rather than in percentage AAC. The potyvirus group was always strictly differentiated from other viruses. (fig. 1, 2) and filled a volume of ⅛th of the ordination.

Closteroviruses were represented only by 3 analyses; nevertheless, they were similar and were separated from the other viruses.

When only the filamentous viruses are considered, they separate into subclusters that do not overlap.

Viruses with Isometric Particles

The bromoviruses were represented only by 3 viruses, but the results were closely similar and the cluster was compact.

[1] Data sources: Reference #1: Tremaine, J.H.; Goldsack, D.E.: Virology *35:*102–107 (1968). #2: Tsugita, A.: J. molec. Biol. *5:*293–300 (1962) [cf. ref. 11]. #3: Gibbs, A.J.: CMI/AAB No. 184 (1977) [13]. #4: Paul, H.L. et al.: Intervirology *13:*99–109 (1980) [14]. #5: Moghal, S.M.; Francki, R.I.B.: Virology *73:*350–362 (1976) [15]. #6: Fauquet, C.; Thouvenel, J.-C.: Init. Doc. Tech., vol. 46 (ORSTOM, Paris 1980) [17]. #7: Semancik, K.S.: Virology *30:*698–704 (1966). #8: Yamazaki, H.; Kaesberg, P.: J. molec. Biol. *6:*455–473 (1963). #9: Stubbs, J.D.; Kaesberg, P.: J. molec. Biol. *8:*314–323 (1964). #10: Kalmakoff, J.; Tremaine, J.H.: Virology *33:*10–16 (1976). #11: Bancroft, J.B. et al.: Virology *34:*224–229 (1968). #12: Shepherd, R.J. et al.: Virology *35:*255–267 (1968). #13: Tremaine, J.H.: Virology *30:*348–354 (1966). #14: Kado, C.I.: Virology *31:*217–229 (1967). #15: Mazzone, H.M. et al.: Biochim. biophys. Acta *55:*164–175 (1962). #16: Reichmann, R.E.: Proc. natn. Acad. Sci. USA *52:*1009–1017 (1964). #17: Lesnaw, J.A.; Reichmann, R.E.: Virology *39:*729–737 (1969). #18: Uyemoto, J.K.; Grogan, R.G.: Virology *39:*79–89 (1969). #19: Stace-Smith, R. et al.: Virology *25:*487–494 (1965). #20: De Fremery, D.; Knight, C.A.: J. biol. Chem. *214:* 559–566 (1955). #21: Tremaine, J.H.; Stace-Smith, R.: Virology *35:*102–107 (1968). #22: Symons, R.H. et al.: J. molec. Biol. *6:*1–15 (1963). #23: Hull, R. et al.: Virology *37:*404–415 (1969). #24: Kelly, J.J.; Kaesberg, P.: Biochim. biophys. Acta *61:*865–871 (1962). #25: Tremaine, J.H.; Stace-Smith, R.: Phytopathology *59:*521–522 (1969). #26: Jankulova, M. et al.: Phytopathologische Z. *63:* 177–185 (1968). #27: Rees, M.W. et al.: Virology *40:* 448–461 (1970). #28: Gibbs, A.J.: CMI/AAB No. 194 (1978). #29: Gibbs, A.J.; Harrison, B.D.: CMI/AAB No. 124 (1973). #30: Barnett, O.W.; Fulton, R.W.: Virology *39:*556–561 (1969). #31: Nelson, M.R.; Tremaine, J.H.: Virology *65:*309–319 (1975). #32: Mink, G.I.: CMI/AAB No.92 (1972). #33: Bercks, R.: CMI/AAB No. 113 (1973). #34: Shukla, D.D. et al.: Phytopathology *70:*382–384 (1980). #35: Van Regenmortel, M.H.V. et al.: Virology *49:* 647–653 (1972). #36: Putz, C.: J. gen. Virol. *35:* 317–401 (1977). #37: Semancik, J.S.: Phytopathology *56:* 1190–1193 (1966). #38: Gumpf, D.J.; Hamilton, R.I.: Virology *35:*87–93 (1968). #39: Nozu, Y.; Okada, Y: J. molec. Biol. *35:*643–646 (1968). #40: Rentschler, L.: Mol. gen. Genet. *100:*84–95 (1967). #41: Funatsu, G.; Funatsu, M.: Phytopathol. Soc. Japan, 1–9 (1968). #42: Damirdagh, I.S.; Shepherd, R.J.: Virology *40:* 84–89 (1970). #43: Hill, J. H.; Shepherd, R.J.: Virology *47:*807–816 (1972). #44: Miki, T.; Knight, C.A.: Virology *36:*168–173 (1968). #45: Stace-Smith, R.; Tremaine, J.H.: Phytopathology *60:*1785–1789 (1970). #46: Miki, T.; Knight, C.A.: Virology *31:*55–63 (1967). #47: Miki, T.; Oshima, N.: J. gen. Virol. *15:*179–182 (1972). #48: Hill, J.H. et al.: J. gen. Virol. *20:*327–339 (1973). #49: Shaw, J. G.; Larson, R.H.: Phytopathology *52:* 170–171 (1962). #50: Knesek, J.E. et al.: Phytopathology *64:*1076–1081 (1974). #51: Short, M.M.: personal commun. (1981). #52: Carpenter, J.M. et al.: Virology *77:*101–109 (1977). #53: Short, M.N. et al.: Virology *77:*408–412 (1977).

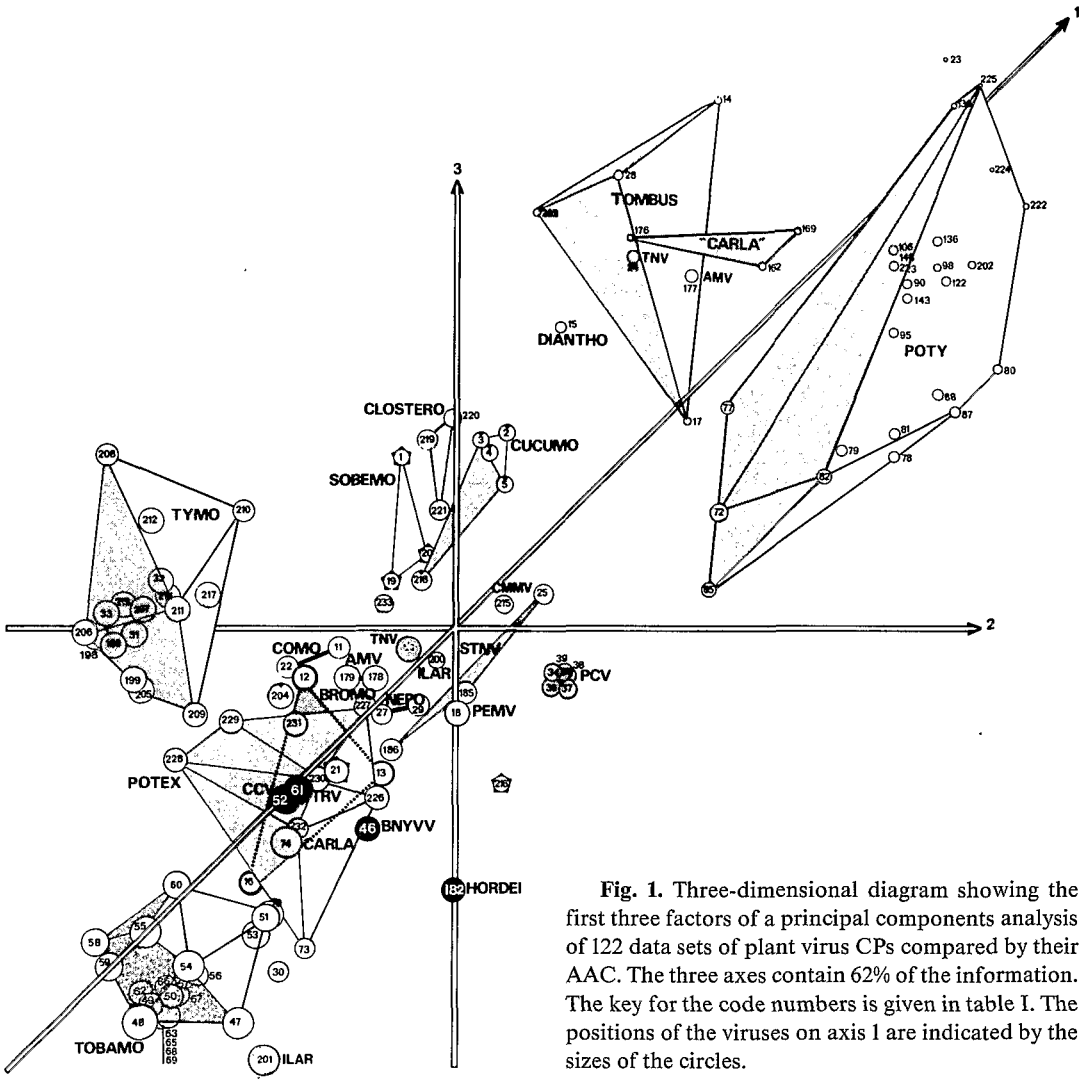

Fig. 1. Three-dimensional diagram showing the first three factors of a principal components analysis of 122 data sets of plant virus CPs compared by their AAC. The three axes contain 62% of the information. The key for the code numbers is given in table I. The positions of the viruses on axis 1 are indicated by the sizes of the circles.

The comoviruses were illustrated by bean pod mottle virus (No. 011) and squash mosaic virus (No. 022). They have 2 capsid proteins (22K and 42K), but the correspondence of the AAC used here is unknown and the results are tentative. The 2 AACs are always very close and are near the centroid of the general cluster.

The cucumoviruses were represented by cucumber mosaic virus (5 data sets: No. 002–005 and 218), and by peanut stunt virus (No. 204). They cover a large space along axis 1, again revealing possible inaccuracies concerning the MW (185 AAs for peanut stunt virus and 261 AAs for some estimates for cucumber mosaic virus). In the other axes

Table II. Correlation coefficients (COR) between the axes of the ordination and the AA contents of the viral CPs, and percentage of contribution (%C) of the AA considered in the total variance of the axis

Axis 1			Axis 2			Axis 3			Axis 4		
AA	COR	%C	AA	COR	%C	AA	COR	%C	AA	COR	%C
Asp	0.84	10	Ser	0.70	19	Trp	0.76	42	Ala	0.44	17
Gly	0.83	10	Pro	0.58	13				Val	0.43	17
Lys	0.79	9	Thr	0.56	12				Arg	0.43	16
Glu	0.77	9	Met	0.47	8						
Tyr	0.75	8	Glu	0.46	8						
Leu	0.66	6	His	0.44	7						
Arg	0.65	6	Val	0.42	6						
Ala	0.65	6									
Met	0.64	6									
His	0.64	6									
Val	0.62	5									
Ile	0.59	5									

(fig. 2) the AACs are homogeneous. As the MW of cucumber mosaic virus has been revised to about 287 AAs [23] and 235 AAs [24], the real position of the group is probably much closer to the cluster of data sets 002–005. However, the cucumoviruses are well separated from the other groups of spherical viruses.

The sobemoviruses were represented by 5 data sets: 2 strains of southern bean mosaic virus (No. 019 and 020), rice yellow mottle virus (No. 001), sowbane mosaic virus (No. 021), and cocksfoot mosaic virus (No. 216). Only data sets 001, 019, and 020 were always related, and hence probably indicate the position of the group. Data set 216 is remote from the others, and No. 021 is probably not a sobemovirus, insofar as the AAC of the CPs is concerned, which seems to be correct [21].

Four of the five tombusviruses are situated within one subcluster and clearly indicate the position of the group. The tombusvirus group is the most distant from the center of the ordination. The cluster is determined by factors other than the MW of their particle protein, because it is also quite distinct in figure 2. In fact, the tombusvirus group is represented by one definitive member, tomato bushy stunt virus (No. 028), and by 4 tentative members (No. 014, 017, 030, and 203) [25, 26]. Three of those tentative members (No. 014, 017, and 203) reveal apparent affinities with tomato bushy stunt virus, providing a supplementary element for their classification in the tombusvirus group.

The tymovirus group (17 viruses) is the best represented group of viruses with isometric particles and shows the greatest homogeneity along each axis. It is well separated from the other groups, and its body forms a reference mark for the others. It is noteworthy that erysimum latent virus, which is a tentative member of the group, is contained in the tymovirus cluster.

Only 2 nepoviruses represented this group: tomato ringspot virus (No. 027) and

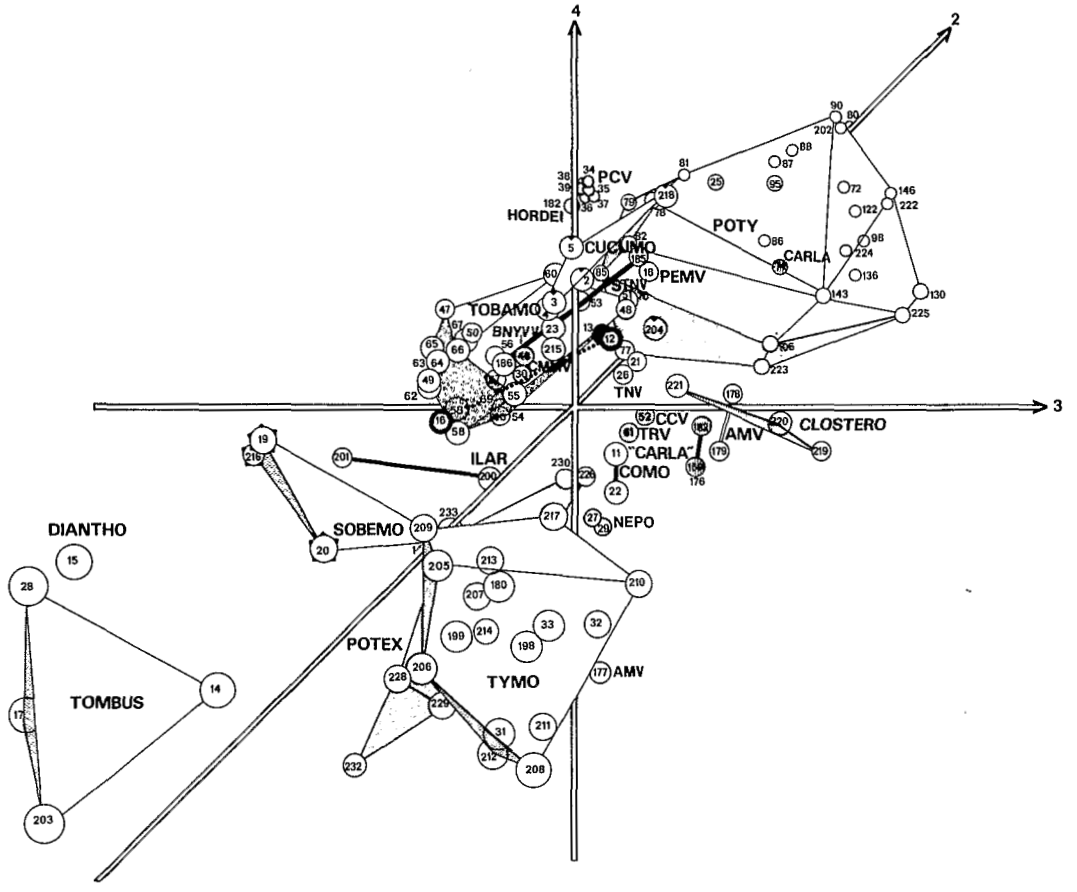

Fig. 2. Three-dimensional diagram illustrating factors 2, 3 and 4 of a principal components analysis of 122 data sets of plant virus CPs compared by their AAC. The three axes contain 28.5% of the information. The key for the code numbers is given in table I. The positions of the viruses on axis 2 are indicated by the sizes of the circles.

tobacco ringspot virus (No. 029). These were always associated and placed near the center of the ordination. The MW of their CPs had been determined to be 53–60K [27, 28] and was then revised to 13–19K [29]. We used the AAC corresponding to about 20K; consequently, if the value of 53–60K is verified, the position of this group will have to be revised.

Other groups were represented by only 1 or 2 individuals, and consequently their positions in the diagrams are uncertain, i.e., tobacco necrosis virus group (No. 024 and 026), pea enation mosaic virus group (No. 018), dianthoviruses (No. 015), and ilarviruses (No. 200 and 201). Most of these groups are found near the center of the ordination, as is the satellite virus of tobacco necrosis virus (4 data sets; No. 023, 025, 185, and 186). Three of these data sets are clustered (No. 025, 185, and 186).

When spherical viruses only are considered, they are spread through a large proportion of the ordination, and some of them, e.g., tombusviruses and tymoviruses, occupy relatively large volumes in the diagrams.

Viruses with Bacilliform Particles

There were 3 data sets for alfalfa mosaic virus (No. 177–179), but its MW is uncertain. No. 177 has 297 AAs and correlates well with the Fitmol analysis [21]; in contrast, No. 178 and 179 were assessed to have 172 and 177 AAs, respectively. The primary structure of the coat protein [30] has been shown to have 217 AAs, and consequently we must imagine a migration of the group (No. 178, 179) in the positive direction of axis 1 to get the correct position of this virus in figure 1.

Discussion

The analysis of principal components used in this work is a reliable method for representing the relationships of individuals and clusters of individuals, when there is no evidence to indicate that they are phylogenetically related, for which a more realistic classification is a hierarchical one. This method enables a multivariate analysis to be represented in multidimensional space, thus giving a precise picture of the relationships of the viruses [2, 6].

The hyperspace filled by plant viruses in an ordination of all proteins represents only 5×10^{-4} of the total hyperspace [2]. The CPs of plant viruses therefore constitute a very dense subcluster of all known proteins. This cluster is not organized at random, and the most important conclusion of our study is that subclusters within it correlate well with currently accepted virus groups [16] that are formed on biochemical, structural, biological, and serological criteria. Thus, the product of one gene of each of these viruses provides classificatory information which is closely similar to that provided by all genes of the viruses.

It is noteworthy that, despite the great range of sources of information and of analyses used in our study, the classification obtained is close to the currently accepted classification [16]. There are some exceptions, and it is not known whether these are real or a result of experimental error. Our study showed that axis 1 correlates most closely with the MW of the CPs and consequently must be determined precisely. Nevertheless, the MW is not the sole discriminatory element; figure 2, which represents 28.5% of the information and excludes the MW axis, provides the same clustering pattern. Obviously, more data sets of the AAC of these and additional viruses would bring a greater precision to the ordination and would increase the density of the clusters.

Only 28.5% of the total information included in the AACs is needed to provide a meaningful classification, and there is a similarity of CPs of plant viruses within the protein hyperspace. These apparent similarities may reflect a common origin in evolution, with only small, but real, differences. Our classification does not correlate only with the shape of virus particles; within one part of the diagram, viruses can be found whose particles are filamentous, rod-shaped, or isometric. Serologically related viruses are grouped in clusters, but the distances between the clusters do not reflect distances in serological relationship. The AAC of the CPs of plant viruses seems to contain information derived from several sources that may be diverse and may interfere with the AAC of

the CPs. Nevertheless, there is a basic similarity of all plant virus CPs; this is presumably because the CPs protect the nucleic acid genomes and form large soluble macromolecules. Plant virus CPs also have a structure that is related to biological factors (e.g., transmission mode), and they have a specific basis reflected and measured by serological relationships [31].

The principal components method of classification, like hierarchical methods, shows close relationships clearly. Unlike the latter, it also gives a measure of the relationship between subclusters. Therefore, as the close groupings within our classification correlate well with currently accepted groupings of viruses, it is worth examining the correspondence between the higher-order relationships (inter-cluster) shown by our classification and the recently discovered 'intergroup' or 'inter-genus' relationships indicated by nucleotide sequence analysis. Distant relationships of this sort have been found between viruses with RNA and DNA genomes [32], between plant and animal viruses [33], among those with rod-shaped, isometric or bacilliform particles, and between those whose particles have a lipid envelope and those that do not [34].

Such sequence homologies indicate, for example, that at least some of the genes of alfalfa mosaic virus, brome mosaic bromovirus, cucumber mosaic cucumovirus, tobacco streak ilarvirus and Sindbis alphavirus have homologous sequences [34–37] and hence probably have a common ancestor. Thus, it is of interest that all these viruses (except Sindbis alphavirus, which was not included in the classification) are close to one another in the central region of the ordination (fig. 2). A similar distant relationship has been found among cowpea mosaic comovirus and polio- and encephalomyocarditis picornaviruses [33]; each of these viruses has a divided RNA genome and a 5′-linked protein (VPg) [38–40]. Other viruses of this type are the nepoviruses [41] and pea enation mosaic virus [42]. It is noteworthy that the single comovirus and nepovirus in our classification group close to pea enation mosaic virus (fig. 2). However, other viruses that have a 5′-linked VPg but an undivided genome, e.g., potyviruses [43], sobemoviruses [44], and luteoviruses [45], are widely dispersed in our classification. Thus at least some of the relationships between subclusters that are illustrated in figures 1 and 2 may correlate with more distant, possibly more ancient, relationships between the currently accepted groups.

Acknowledgments

The authors thank Dr. *A.J. Gibbs* and Dr. *B.D. Harrison* for help and advice in the preparation of this paper.

References

1 Tremaine, J.H.; Goldsack, D.E.: The structure of regular viruses in relation to their subunit amino-acid composition. Virology *35:*227–237 (1968).
2 Tremaine, J.H.; Argyle, E.: Cluster analysis of viral proteins. Phytopathology *60:*654–659 (1969).
3 Brandes, J.; Bercks, R.: Gross morphology and serology as a basis for classification of elongated plant viruses. Adv. Virus Res. *11:*1–24 (1965).
4 Haselkorn, R.: Physical and chemical properties of plant viruses. A. Rev. Plant Physiol. *17:*137–154 (1966).
5 Kaper, J.M.: The small RNA viruses of plants, animals and bacteria. A. Physical properties; in Fraenkel-Conrat, Molecular basis of virology, pp. 1–133 (Reinhold, New York 1968).
6 Gibbs, A.J.: Plant virus classification. Adv. Virus Res. *14:*263–328 (1969).

7 Lance, G.N.; Williams, W.T.: Computer programs for hierarchical polythetic classification. Computer J. *9:* 60–66 (1966).
8 Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika *53:* 325–338 (1966).
9 Lance, G.N.; Williams, W.T.: A general theory of classification sorting strategies. 1. Hierarchical systems. Computer J. *9:* 373–380 (1967).
10 Gibbs, A.J.; Harrison, B.D.: Plant virology, the principles (Arnold, London 1976).
11 Tsugita, A.: The proteins of mutants of TMV. Classification of spontaneous and chemically evoked strains. J. molec. Biol. *5:* 293–300 (1962).
12 Van Regenmortel, M.H.V.: Serological studies on naturally occurring strains and chemically induced mutants of TMV. Virology *31:* 467–480 (1967).
13 Gibbs, A.J.: Tobamovirus group. CMI/AAB Descriptions of plant viruses, No. 184 (1977).
14 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).
15 Moghal, S.M.; Francki, R.I.B.: Towards a system for the identification and classification of potyviruses. I. Serology and amino-acid composition of six distinct viruses. Virology *73:* 350–362 (1976).
16 Matthews, R.E.F.: Classification and nomenclature of viruses. Fourth Report of the International Committee on Taxonomy of Viruses. Intervirology *17:* 1–200 (1982).
17 Fauquet, C.; Thouvenel, J.-C.: Viral diseases of crop plants in the Ivory Coast. Init. Doc. Techn., vol. *46* (ORSTOM, Paris 1980).
18 Benzecri, J.P. et al.: L'analyse des données. I. La taxinomie. II. L'analyse des corrrespondances (Dunod, Paris 1973).
19 Shirako, Y.; Brakke, M.K.: Two purified RNAs of soil-borne wheat mosaic virus are needed for infection. J. gen. Virol. *65:* 119–127 (1984).
20 Brunt, A.A.; Atkey, P.T.; Woods, R.D.: Intracellular occurrence of cowpea mild mottle virus in two unrelated plant species. Intervirology *20:* 137–142 (1983).
21 Gibbs, A.J.; McIntyre, G.A.: A method for assessing the size of a protein from its composition: its use in evaluating data on the size of the protein subunits of plant virus particles. J. gen. Virol. *9:* 51–67 (1970).
22 Hollings, M.; Brunt, A.A.: Potyviruses; in Kurstak, Handbook of plant virus infections and comparative diagnosis, pp. 731–807 (Elsevier, Amsterdam 1981).
23 Van Regenmortel, M.H.V.: Biochemical and biophysical properties of cucumber mosaic virus. Virology *31:* 391–396 (1967).
24 Kaper, J.M.; Waterworth, H.E.: Cucumoviruses; in Kurstak, Handbook of plant virus infections and comparative diagnosis, pp. 257–332 (Elsevier, Amsterdam 1981).
25 Martelli, G.P.; Russo, M.; Quacquarelli, A.: Tombusvirus group; in Maramorosch, The atlas of insect and plant viruses, pp. 257–279 (Academic Press, New York 1977).
26 Martelli, G.P.: Tombusviruses; in Kurstak, Handbook of plant virus infections and comparative diagnosis, pp. 61–90 (Elsevier, Amsterdam 1981).
27 Murant, A.F.: Nepoviruses; in Kurstak, Handbook of plant virus infections and comparative diagnosis, pp. 197–238 (Elsevier, Amsterdam 1981).
28 Mayo, M.A.; Murant, A.F.; Harrison, B.D.: New evidence on the structure of nepoviruses. J. gen. Virol. *12:* 175–178 (1971).
29 Chu, P.W.G.; Francki, R.I.B.: The chemical subunit of tobacco ringspot virus coat protein. Virology *93:* 398–412 (1979).
30 Collot, D.; Peter, R.; Das, B.; Wolff, B.; Duranton, H.: Primary structure of alfalfa mosaic virus coat protein. Virology *74:* 236–238 (1976).
31 Van Regenmortel, M.H.V.: Tobamoviruses; in Kurstak, Handbook of plant virus infections and comparative diagnosis, pp. 541–564 (Elsevier, Amsterdam 1981).
32 Toh, H.; Hidenori, H.; Yakahashi, M.: Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature, Lond. *305:* 827–829 (1983).
33 Franssen, H.; Leunissen, J.; Goldbach, R.; Lomonosoff, G.; Zimmern, D.: Homologous sequences in the non-structural proteins from cowpea mosaic virus and picornaviruses. EMBO J. *3:* 855–861 (1984).
34 Haseloff, J.; Goelet, P.; Zimmern, D.; Ahlquist, P.; Dasgupta, R.; Kaesberg, P.: Striking similarities in amino acid sequence among nonstructural proteins encoded by RNA viruses that have dissimilar genomic organization. Proc. natn. Acad. Sci. USA *81:* 4358–4362 (1984).

35 Cornelissen, B.J.C.; Bol, J.F.: Homology between the proteins encoded by tobacco mosaic virus and two tricornaviruses. Plant molec. Biol. *3:*379–384 (1984).
36 Cornelissen, B.J.C.; Janssen, H.; Zuidema, D.; Bol, J.F.: Complete nucleotide sequence of tobacco streak virus RNA3. Nucl. Acids Res. *12:* 1253–1255 (1984).
37 Rezaian, M.A.; Williams, R.H.V.; Gordon, K.H.J.; Symons, R.H.: The primary structure of the four RNAs of cucumber mosaic virus. Abstr. 6th Int. Congr. Virology, P42–6 (1984).
38 Daubert, S.D.; Bruening, G.; Najarian, R.C.: Protein bound to the genome RNAs of cowpea mosaic virus. Eur. J. Biochem. *92:*45–51 (1978).
39 Hruby, D.E.; Roberts, W.K.: Encephalomyocarditis virus RNA. III. Presence of a genome associated protein. J. Virol. *25:*413–415 (1978).
40 Lee, Y.F.; Nomoto, A.; Detjen, B.M.; Wimmer, E.: A protein covalently linked to poliovirus genome RNA. Proc. natn. Acad. Sci. USA *74:*59–63 (1977).
41 Mayo, M.A.; Barker, H.; Harrison, B.D.: Specificity and properties of the genome-linked proteins of nepoviruses. J. gen. Virol. *59:* 149–162 (1982).
42 Reisman, D.; De Zoeten, G.A.: A covalently linked protein at the 5′ ends of the genomic RNAs of pea enation mosaic virus. J. gen. Virol. *62:* 187–190 (1982).
43 Hari, V.: The RNA of tobacco etch virus: further characterization and detection of protein linked to RNA. Virology *112:*391–399 (1981).
44 Ghosh, A.; Dasgupta, R.; Salerno-Rife, T.; Rutgers, T.; Kaesberg, P.: Southern bean mosaic viral RNA has a 5′ linked protein but lacks 3′ terminal poly A. Nucl. Acids Res. *7:*2137–2146 (1979).
45 Mayo, M.A.; Barker, H.; Robinson, D.J.; Tamada, T.; Harrison, B.D.: Evidence that potato leafroll virus is positive-stranded, is linked to a small protein and does not contain polyadenylate. J. gen. Virol. *59:*163–167 (1982).