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

By A. J. GIBBS* AND G. A. McINTYRE†

* Department of Microbiology, John Curtin School of Medical Research,
Australian National University, Canberra, Australia
† Division of Mathematical Statistics, Commonwealth Scientific and
Industrial Research Organization, Canberra, Australia

(Accepted 9 June 1970)

SUMMARY

A method is described for assessing the possible size of a protein from its amino
acid composition. This method is then used to re-examine published data and to
estimate the sizes of the proteins in the particles of various plants viruses. Some of
the results obtained agree with the published reports, others do not.

INTRODUCTION

The problem. The particles of each plant virus are usually all of one architectural type, and
the protein subunits of these particles are usually a single type of protein.

The proteins of different viruses usually have different amino acid compositions, but
different estimates of the composition of the protein of any one virus usually agree closely,
and if the viruses are classified by their amino acid composition (Fig. 3 and 4 of the review by
Gibbs (1969)), a classification is obtained in which the groupings are similar to those ob-
tained by classifying the same viruses in terms of other properties such as nucleic acid
composition, particle morphology, vector type, etc. By contrast, reports on the sizes of these
proteins may vary widely, and if the viruses are classified according to these sizes (Fig. 1), the
usual groupings are not obtained. This discrepancy is difficult to understand as the sizes of
these proteins are mostly derived from determinations of their amino acid composition;
the minimum integral number of amino acid residues in the protein is calculated from its
composition, and other tests are made to check whether the protein contains one or more
times this minimum number.

Because of the discrepancy between the classifications based on protein composition and
size, we have re-examined current methods for calculating the size of a protein from its
composition. We describe here an alternative method which has some advantages over
others, including that described by Ozawa & Tanaka (1968). We have used this method to
examine published data on the proteins of plant virus particles, and find that some of the
results given by our method agree with those previously published, others do not. The
general problem of estimating molecular composition to satisfy integral values was considered
by Hammersley (1950).

Protein size estimation. In the most widely used method for estimating the minimum
integral number of amino acid residues in a protein, the amino acid composition of the
protein is estimated, and from this the relative number of residues of each amino acid is

* Present address: Rothamsted Experimental Station, Harpenden, Herts., England.
calculated. These estimates are then multiplied by a series of factors, chosen so that the relative number of residues of, usually, the least abundant amino acid (the ‘key’ amino acid) is converted in turn to integers between 1 and, say 5. The most suitable factor is judged to be that which converts the relative numbers of all the amino acids most nearly into integers, and this is used to calculate the most likely minimum integral number of residues. The fault of this method is that it takes no account of statistical errors and assumes that the relative amount of the ‘key’ amino acid is known with absolute accuracy, whereas errors are possible when estimating every amino acid, including the ‘key’ amino acid. Obviously, the

Table 1. Analysis of the amino acid composition of proteins; the relationship between the mean and standard deviation of replicate analyses of various proteins

<table>
<thead>
<tr>
<th>Protein analysed‡</th>
<th>Equation (1)</th>
<th>Equations (2)</th>
<th>Value of constants, when data are fitted to linear equations†</th>
<th>The line fitting data to equation (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b</td>
<td>Could pass through origin</td>
<td>Has significant positive slope</td>
</tr>
<tr>
<td>1. Carboxypeptidase</td>
<td>-0.0043 (0.14)</td>
<td>0.0201 (3.92**)</td>
<td>0.0194</td>
<td>Yes</td>
</tr>
<tr>
<td>2. Carnation ringspot virus</td>
<td>0.0913 (1.81)</td>
<td>0.0049 (0.61)</td>
<td>0.0183</td>
<td>Yes</td>
</tr>
<tr>
<td>3. Cucumber mosaic virus</td>
<td>5.5475 (1.95)</td>
<td>0.0099 (1.55)</td>
<td>0.0214</td>
<td>Yes</td>
</tr>
<tr>
<td>Hydrolysed 24 hr</td>
<td>5.5056 (2.75)</td>
<td>0.0237 (2.68)</td>
<td>0.0264</td>
<td>Yes</td>
</tr>
<tr>
<td>Hydrolysed 72 hr</td>
<td>5.4105 (0.10)</td>
<td>0.0315 (2.99**)</td>
<td>0.0264</td>
<td>Yes</td>
</tr>
<tr>
<td>4. Flaggelin</td>
<td>0.1368 (0.33)</td>
<td>0.0211 (3.34**)</td>
<td>0.0239</td>
<td>Yes</td>
</tr>
<tr>
<td>5. Phage λ lysozyme</td>
<td>0.1846 (1.89**)</td>
<td>0.0050 (0.72)</td>
<td>—</td>
<td>No</td>
</tr>
<tr>
<td>6. Satellite virus</td>
<td>0.1291 (3.45**)</td>
<td>0.0053 (0.71)</td>
<td>—</td>
<td>No</td>
</tr>
<tr>
<td>7. Southern bean mosaic virus</td>
<td>0.5907 (3.88**)</td>
<td>0.0765 (3.67**)</td>
<td>—</td>
<td>No</td>
</tr>
<tr>
<td>8. Squash mosaic virus</td>
<td>0.5358 (8.43**)</td>
<td>0.0206 (2.85**)</td>
<td>—</td>
<td>No</td>
</tr>
<tr>
<td>Top component</td>
<td>0.2674 (1.59)</td>
<td>0.0426 (1.78)</td>
<td>0.0261</td>
<td>Yes</td>
</tr>
<tr>
<td>Middle component</td>
<td>0.0222 (0.35)</td>
<td>0.0246 (0.07**)</td>
<td>—</td>
<td>No</td>
</tr>
</tbody>
</table>

† Data fitted to linear equations: (Standard deviation) = a + b (mean) (1); (standard deviation) = b (mean) (2). The value of constant b in equation (2) is given only when constant a in equation (1) does not differ significantly from zero. In parentheses is the variance ratio r; asterisks indicate the probability that a or b are not zero, one or two asterisks indicate P = 0.1-0.01 and P < 0.01, respectively.


size of a protein would be estimated more accurately if the amount of every amino acid, weighted according to its associated error, was used in the calculation. Ideally, the error associated with each amino acid would be estimated separately and used in the calculation, but, as will be shown below, at least 6 to 10 independent analyses of protein composition are needed to assess accurately the errors for individual amino acids, and this is not always possible.

Statistical errors in estimating amino acid composition. For general use and, as in this paper, when using the results of amino acid analyses with unknown errors, it would be most useful if one could generalise about the errors involved. It is likely, for instance that the error in the estimation of each amino acid was related to its amount in the protein. To examine this and other possibilities, data on analyses of 12 different proteins (Table 1) were tested statistically to determine whether the mean and standard deviations of the estimates of their amino acid composition fitted one of the equations:

\[
\text{Standard deviation} = a + b \times \text{mean} \\
\text{Standard deviation} = b \times \text{mean} \\
\text{Standard deviation} = a + b \times \text{mean} + c \times \text{mean}^2
\]

where a, b and c are constants.
Size and composition of plant virus protein

The results for equations (1) and (2) are given in Table I; the constant c of equation (3) did not contribute significantly to fitting the data of any of the analyses. The 12 sets of data gave lines that best fitted the linear equations in three ways: (i) for four proteins a differed significantly from zero and the line did not pass through the origin, and for two of these proteins b also differed significantly from zero and standard deviations changed with mean; (ii) for four proteins b differed significantly from zero, but a did not and standard deviation increased proportionally with mean; (iii) for four proteins neither a nor b differed significantly from zero.

These differences could be the result of differences between the proteins being analysed but seem not to be related to the analyser as three of the proteins giving different relationships between mean and standard deviation were analysed by one group (Kalmakoff & Tremaine, 1967; Tremaine, 1966; Tremaine & Stace-Smith, 1968). It is, however, noticeable that in the earliest analyses both a and b differed significantly from zero, but only b in the most recent.

The data from which Table I was calculated, were also examined for evidence of the relative precision with which individual amino acids were estimated; the sign and amount by which individual amino acids deviated from the fitted lines was recorded. The proportion of deviates that were negative for different amino acids was not greater than would be expected by chance (χ² = 13·3 with 17 degrees of freedom). The estimates of alanine, glutamic acid and valine were less variable than average, whereas the estimates of arginine, methionine and tyrosine were more variable than average.

These results suggest that the most useful generalization is that the error in the estimates of the amount of each amino acid is proportional to the amount of that amino acid in the protein (equation (2) and condition (ii)). The computer programme (FITMOL) described in this paper is based on this generalization and has a statistical basis that is suitable for at least half the analyses whose results were used in the calculations for Table I. For the others, FITMOL is no less suitable than alternative methods.

The errors associated with each amino acid could be used individually. However, these errors must be estimated from several independent analyses of the protein because the variance of the mean of n estimates is distributed as σ²/n(χ²/(n−1)) with n−1 degrees of freedom, so that the reliability of the estimate of the variance is poor for small n. With σ²/n = 1, estimates of σ²/n at various percentiles are

<table>
<thead>
<tr>
<th>Percentile</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>80</th>
<th>90</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0·004</td>
<td>0·016</td>
<td>0·064</td>
<td>0·455</td>
<td>1·642</td>
<td>2·706</td>
<td>3·841</td>
</tr>
<tr>
<td>3</td>
<td>0·052</td>
<td>0·106</td>
<td>0·223</td>
<td>0·693</td>
<td>1·609</td>
<td>2·305</td>
<td>2·995</td>
</tr>
<tr>
<td>4</td>
<td>0·117</td>
<td>0·195</td>
<td>0·335</td>
<td>0·789</td>
<td>1·548</td>
<td>2·083</td>
<td>2·605</td>
</tr>
<tr>
<td>6</td>
<td>0·229</td>
<td>0·322</td>
<td>0·469</td>
<td>0·870</td>
<td>1·458</td>
<td>1·847</td>
<td>2·214</td>
</tr>
<tr>
<td>10</td>
<td>0·364</td>
<td>0·463</td>
<td>0·598</td>
<td>0·920</td>
<td>1·360</td>
<td>1·631</td>
<td>1·880</td>
</tr>
</tbody>
</table>

Unless individual errors are calculated from at least six replicate analyses, or there are other good reasons for differential weighting, it appears that equation (2) is acceptable and that a method such as FITMOL may be used.

METHODS

The FITMOL program, which can be obtained from us, written in CDC or ASA Fortran IV, evaluates how well a given amino acid composition fits each ‘protein’ (to an integral number of each amino acid residue) within a stipulated size range. From the amino
acid composition an ‘integer deviate index’ (IDI) is calculated for every protein size within the stipulated range. The IDI is defined as:

$$C \left( \sum_{i=1}^{N} \left( M \times X_i - \text{the nearest integer to } M \times X_i \right)^2 \right)$$

where $X_i$ is the proportional molar amount of the $i$th of $N$ amino acids and $M$ is the protein size (number of residues) being tested. $C$ is a constant chosen to give a satisfactory magnitude.

---

**Fig. 1.** The reported sizes of the proteins in the particles of various plant viruses.

In the centre columns are the acronyms of the viruses arranged in the order of the reported sizes of their protein subunits: PVX(a), potato virus X (Tremaine & Goldsack, 1968); WCMV, white clover and mosaic virus (Miki & Knight, 1967); PVS, potato virus S (Tremaine & Goldsack, 1968); TMV(a) and (b) tobacco mosaic virus, ribgrass and type strains (Fraenkel-Conrat, 1968); TCV, turnip crinkle virus (Symons et al. 1963); SoMV, sowbane mosaic virus (Kado, 1967); TRSV, tobacco ring spot virus (Stace-Smith et al. 1965); TRV, tobacco rattle virus (Semancik, 1966a); CCMV, cowpea chlorotic mottle virus (Bancroft et al. 1968); BrMV, brome mosaic virus (Stubbs & Kaesberg, 1964); SqMV, squash mosaic virus (Tremaine & Goldsack, 1968); TYMV(a) and (b), turnip yellow mosaic virus, different strains (Symons et al. 1963); WCuMV, wild cucumber mosaic virus (Symons et al. 1963); BBMV, broad bean mottle virus (Yamazaki & Kaesberg, 1963); PEMV, pea enation mosaic virus (Shepherd et al. 1968); BPMV, bean pod mottle virus (Tremaine & Goldsack, 1968); PVX(b) (Miki & Knight, 1968); ToRSV, tomato ringspot virus (Tremaine & Stace-Smith, 1968); SoBMV, southern bean mosaic virus (Tremaine, 1966); CuMV, cucumber mosaic virus (van Regenmortel, 1967); AMV, alfalfa mosaic virus (Hull et al. 1969); TNV, tobacco necrosis virus (Lesnaw & Reichmann, 1969); CaRSV, carnation ring spot virus (Kalmakoff & Tremaine, 1967); SV, satellite virus (Reichmann, 1964); CaMV, CuNV, TBSV, carnation mottle, cucumber necrosis and tomato bushy stunt viruses (Tremaine & Goldsack, 1968); PVX(c) (Shaw & Larson, 1962); PVX(d) (Shaw et al. 1962).

On the left is a dendrogram representing a classification of the viruses computed from the sizes of their protein subunits by the ‘Class’ programme of Lance & Williams (1967); non-metric coefficient of similarity; flexible sorting ($\beta = -0.25$). Horizontal scale gives the computed measure of dissimilarity.

On the right are groupings of these viruses based on a wide range of their properties (Gibbs, 1969).
Size and composition of plant virus protein

to the sum; in FITMOL $C = 0.1$. When, as sometimes happens, the sum of the nearest integers to $M \times X_i$ does not equal $M$ but is for example $M - 1$, then the program identifies which amino acid can be increased by one residue with the least increment to the IDI. Similarly, if the sum of nearest integers is $M + 1$ the same method is used to determine which amino acid to decrease by one.

$X_i$ is used as the term in the denominator in equation (4), so that as $M$ increases, IDI oscillates about a constant mean, whereas if $M \times X_i$ had been used the mean IDI would decrease asymptotically to zero.

When using FITMOL to search for the fundamental size of the protein, the lower limit of $M$ is usually set so that at that limit the least abundant amino acid would be present as $0.85$ residues, i.e. $85 \times$ the molar percentage of the least abundant amino acid, and the upper limit at about five or six times this.

The classification illustrated by the dendrogram in Fig. 1 was computed by the ‘Class’ program of Lance & Williams (1967); Williams, Lambert & Lance, (1966).

RESULTS

Trial data

The way in which different amino acids contribute to the IDI is illustrated in Fig. 2, which shows the IDI values over a small range of $M$ for a hypothetical protein with five different amino acids and a total of 23 residues.

At each value of $M$ the total IDI is the sum of the contributions of the five amino acids. In general, in a protein with $X$ residues, the minima for an amino acid with $p$-residues will be at multiples of $X/p$. For example, in Fig. 2, the amino acid with two residues ($X_i = p/X = 2/23$) will contribute minimally at $0$, $11.5$, $23$, $34.5$ and $46$, since at these trial values of $M$ the values of $M \times X_i$ are exactly, $0$, $1$, $2$, $3$ and $4$, and the deviate from the nearest integer is nil. The maxima occur midway between the minima and have ordinates proportional to $1/p^2$. The IDI values given in Fig. 2 lie on cusps formed by segments of second degree parabolas.

Errors in the estimation of amino acids may obscure the true minima. If (Fig. 2), the amino acids present as two and three residues had been estimated experimentally to constitute $(2.1)/23$ and $(2.9)/23$, of the protein respectively, the IDI at a protein size of 36 would be less than that at 23 and a false estimate of the size of the protein would be obtained. At successive multiples of the fundamental size the interference between contributions to the IDI by inaccurately estimated amino acids with small numbers of residues increases, so that there is a diminishing range between maxima and minima IDI as $M$ increases.

The limitations and sensitivity of FITMOL were examined with data which approximate more closely to real proteins. A hypothetical protein with the 18 different amino acids present as $1, 1, 2, 2, 3, 4, 6, 6, 7, 7, 10, 11, 11, 14, 15, 17, 18, 22$, (total 157) residues, was used in the following tests:

1. The accuracy with which the relative molar compositions (3 decimal places) was supplied to FITMOL was varied by adding random error with either $3$ or $10$ % coefficient of variation. Fig. 3 shows that the minimum found by FITMOL was close to the actual protein size even when $10$ % error had been added, though it was closer with the more ‘accurate’ data. The ratio of the maximum to minimum IDI was much greater with the ‘accurate’ data than with the ‘inaccurate’ and, furthermore, the ‘accurate’ data, but not the ‘inaccurate’, gave a harmonic IDI minimum at twice the actual size of the protein.

2. The relative amounts of different amino acids in the hypothetical protein (157 amino acids residues, $3$ % random error, 3 decimal places) were altered systematically. Fig. 4
shows the results of a FITMOL analysis of a protein which has two amino acids present as single residues (line A), and similar analyses of the protein altered so that the least abundant amino acid is represented by either one, two, three, four or five residues (lines B, C, D, E, and F respectively). There is a distinct minimum IDI near the size of the protein only when at least one amino acid is represented by one or two residues.

Fig. 2. FITMOL analysis of test data. Contributions to IDI (integer deviate index; for explanation see text) of the five different amino acids in a hypothetical protein of 23 residues (no error added). Upper line shows the total IDI and lower lines the contributions of the individual amino acids with 2, 3, 5, 6 and 7 residues. For simplicity the protein size tested (M) is shown as continuous and not integral. The arrows at side indicate the maximum contribution to the IDI of the particular amino acid.

These results suggest how FITMOL analyses may be evaluated qualitatively. When there is a large consistent ratio in the maximum to minimum IDIs over a range of protein sizes, and when there are clear harmonics in IDIs at regular intervals, it is likely that the amino acid composition being tested has been accurately determined, and that the protein perhaps contains one or more amino acids present as single residues. However, when there is a small and diminishing ratio of the maximum to minimum IDI and the minima are at irregular intervals, the results must be interpreted with caution.
Proteins from plant virus particles

(a) Viruses with rod-shaped or filamentous particles. Tobacco mosaic virus [cryptogram†: R/I: 2/5: E/E: S/] is the best studied of these. All reports agree that the protein subunits of most strains of this virus contain 158 amino acid residues. Fig. 5 shows the results of FITMOL analyses of data on the composition of the protein subunits of some strains reported by Rees & Short (1965) and Paul et al. (1965). These analyses agree completely with the published reports and give examples of FITMOL analyses with unequivocal results; there is, for each set of data, a large ratio in the maximum to minimum IDI, and a clear harmonic minimum at twice the fundamental subunit size.

The data for tobacco rattle virus [R/I: (0.6 - 1.3) + 2.3/5: E/E: S/Ne] do not give such a clear result. Tremaine & Goldsack (1968) calculated from analyses made by Semancik (1966a) that the subunits of two strains of the virus contain 178 amino acids residues (excluding

† Names and cryptograms from Martyn (1968), Gibbs (1969).
cysteine and tryptophan). However, FITMOL analyses of the data give minimum IDI at 190–195 amino acid residues (Table 2).

Offord & Harris (1965) analysed the proteins of two European isolates of the virus and reported that both contain 218 residues/subunit including 1 cysteine and 2 tryptophan residues/subunit.

The composition of the protein of barley stripe mosaic virus [R/1: 1/4.5: E/E: S/] has been reported by Gumpf & Hamilton (1968), but FITMOL analysis of this data gives no indication of the subunit size (Table 2).

Potato virus X [R/1: */6: E/E: S(Fu)] has been studied by many workers. Reports on the
Size and composition of plant virus protein

The size of its subunits illustrate clearly how different conclusions can be drawn from similar, or even the same, data. Shaw & Larson (1962) calculated from their results that the subunits contain 402 amino acid residues. Shaw, Reichmann & Hatt (1962) reported the results of further analyses and calculated that the subunits contain 463 residues, basing this on the composition of the protein and on its sedimentation coefficient in 2 M-guanidine; it is now known that the latter may not completely dissociate the subunits. Tremaine & Goldsack (1968) re-examined the data of Shaw et al. (1962) and calculated that the subunits contain 103 residues. Miki & Knight (1968) made a detailed analysis of the protein and calculated that the minimum integral number of amino acid residues in each subunit is about 105, but concluded from tryptic peptide and N-terminal amino acid analyses, that each subunit contains twice the minimum number of amino acid residues. Fig. 6 shows FITMOL analyses of all these sets of data. All sets of data give an IDI minimum at 103 to 107 residues and another between 210 and 238 residues, and all give further minima between either 320 and 360 residues or 410 and 480 residues. The average of the minima near 200 residues given by all four sets of data is about 225 residues, which is close to the size of 210 residues reported by Miki & Knight (1968), and 215 residues calculated from X-ray diffraction data by Wilson & Tollin (1969).

Fig. 5. FITMOL analyses of actual data for strains of tobacco mosaic virus: Crotalaria mosaic strain grown (1) in bean and (2) in tobacco (Rees & Short, 1965); (3) the Odontoglossum ring spot strain (Paul et al. 1965).

In Fig. 5 to 8, the IDI maxima are truncated (broken lines) and the maximum IDI values are indicated.
White clover mosaic virus [R/I: */5: E/E: S/(Ap)], which is serologically related to potato virus X, has also been analysed by Miki & Knight (1967), who calculated that its protein subunits contain 133 amino acid residues. FITMOL analysis of their data gives no minimum IDI at 133 residues (Table 2).

### Table 2. Results of FITMOL analyses of the published amino acid composition of the proteins in various plant viruses

<table>
<thead>
<tr>
<th>Virus</th>
<th>Published size (residues/subunit)</th>
<th>IDI minima* (residues/subunit)</th>
<th>Agreement†</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barley stripe mosaic</td>
<td>ND†</td>
<td>99, 133, 188, 200, 244</td>
<td>+</td>
<td>Gumpf &amp; Hamilton, 1968</td>
</tr>
<tr>
<td>Broad bean mottle</td>
<td>190</td>
<td>101, 193, 279, 377</td>
<td>+</td>
<td>Yamazaki &amp; Kaesberg, 1963</td>
</tr>
<tr>
<td>Broad bean mottle</td>
<td>194</td>
<td>101, 195-200, 285, 387</td>
<td>+ +</td>
<td>Miki &amp; Knight, 1965</td>
</tr>
<tr>
<td>Bean pod mottle</td>
<td>201</td>
<td>84, 167, 284</td>
<td>-</td>
<td>Tremaine &amp; Goldsack, 1968</td>
</tr>
<tr>
<td>Belladonna mottle</td>
<td>ND</td>
<td>113, 191, 253, 305, 365</td>
<td>-</td>
<td>Junkulowa et al. 1968</td>
</tr>
<tr>
<td>Brome mosaic</td>
<td>189</td>
<td>199, 387</td>
<td>+</td>
<td>Stubbs &amp; Kaesberg, 1964</td>
</tr>
<tr>
<td>Carnation ringspot</td>
<td>247</td>
<td>197, 259-375, 569, 681</td>
<td>+</td>
<td>Kalmakoff &amp; Tremaine, 1967</td>
</tr>
<tr>
<td>Cowpea chlorotic mottle</td>
<td>183</td>
<td>184, 372</td>
<td>++ +</td>
<td>Bancroft et al. 1968</td>
</tr>
<tr>
<td>Cucumber mosaic</td>
<td>287</td>
<td>155, 215, 286, 365</td>
<td>+ + +</td>
<td>van Regenmortel, 1967</td>
</tr>
<tr>
<td>Echtes Ackerbohnenmosaik</td>
<td>ND</td>
<td>96, 166, 260</td>
<td>-</td>
<td>Wittmann &amp; Paul, 1961</td>
</tr>
<tr>
<td>Pea enation mosaic</td>
<td>189</td>
<td>86, 157, 280, 355</td>
<td>-</td>
<td>Shepherd et al. 1968</td>
</tr>
<tr>
<td>Potato virus X</td>
<td>210</td>
<td>107, 210, 317, 337</td>
<td>+ +</td>
<td>Miki &amp; Knight, 1968</td>
</tr>
<tr>
<td>Potato virus X</td>
<td>202</td>
<td>107, 237, 410</td>
<td>-</td>
<td>Shaw &amp; Larson, 1962</td>
</tr>
<tr>
<td>Potato virus X (C)</td>
<td>403 (103)</td>
<td>103, 224, 320-70, 465</td>
<td>-</td>
<td>Shaw et al. 1962</td>
</tr>
<tr>
<td>Potato virus X (RS)</td>
<td>403 (103)</td>
<td>102, 209-35, 330, 435</td>
<td>-</td>
<td>Tremaine &amp; Goldsack, 1968</td>
</tr>
<tr>
<td>Satellite</td>
<td>372</td>
<td>240, 374, 485, 620</td>
<td>-</td>
<td>Reichmann, 1964</td>
</tr>
<tr>
<td>Satellite</td>
<td>209</td>
<td>177, 274, 379, 536</td>
<td>-</td>
<td>Uyemoto &amp; Grogan, 1969</td>
</tr>
<tr>
<td>Satellite (sv1 strain)</td>
<td>208</td>
<td>208, 413</td>
<td>++ +</td>
<td>Rees, Short &amp; Kassanis, 1970</td>
</tr>
<tr>
<td>Satellite (sv2 strain)</td>
<td>207</td>
<td>219, 279, 502</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Southern bean mosaic (bean)</td>
<td>270</td>
<td>268, 539</td>
<td>++ +</td>
<td>Tremaine, 1966</td>
</tr>
<tr>
<td>Southern bean mosaic (cowpea)</td>
<td>270</td>
<td>270, 538</td>
<td>++ +</td>
<td></td>
</tr>
<tr>
<td>Southern bean mosaic (cowpea)</td>
<td>260</td>
<td>122, 253, 362, 501</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Southern bean mosaic (Mexican)</td>
<td>260</td>
<td>115, 287, 378, 495</td>
<td>+ +</td>
<td>Ghabrial et al. 1967</td>
</tr>
<tr>
<td>Southern bean mosaic (type)</td>
<td>260</td>
<td>137, 276, 529</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Sowbane mosaic</td>
<td>176</td>
<td>102, 180, 276, 346</td>
<td>+ +</td>
<td>Kado, 1967; Kado &amp; Black, 1968</td>
</tr>
<tr>
<td>Squash mosaic (top)</td>
<td>189</td>
<td>131, 234, 280, 378</td>
<td>-</td>
<td>Tremaine &amp; Goldsack, 1968</td>
</tr>
<tr>
<td>TMV (Crotalaria str. in tobacco)</td>
<td>158</td>
<td>158, 310-320</td>
<td>++ +</td>
<td>Rees &amp; Short, 1965</td>
</tr>
<tr>
<td>TMV (Crotalaria str. in bean)</td>
<td>155</td>
<td>155, 309</td>
<td>++ +</td>
<td></td>
</tr>
<tr>
<td>TMV (Odontoglossum)</td>
<td>158</td>
<td>159, 315</td>
<td>+ +</td>
<td>Paul et al. 1965</td>
</tr>
<tr>
<td>Tobacco necrosis</td>
<td>197</td>
<td>261, 472</td>
<td>+</td>
<td>Uyemoto &amp; Grogan, 1969</td>
</tr>
<tr>
<td>Tobacco necrosis</td>
<td>312</td>
<td>301, 615</td>
<td>+</td>
<td>Lesnaw &amp; Reichmann, 1969</td>
</tr>
<tr>
<td>Tobacco rattle (B)</td>
<td>178</td>
<td>196, 390-420</td>
<td>+</td>
<td>Tremaine &amp; Goldsack, 1968</td>
</tr>
<tr>
<td>Tobacco rattle (C)</td>
<td>178</td>
<td>180-220, 385</td>
<td>+</td>
<td>Tremaine &amp; Goldsack, 1968</td>
</tr>
<tr>
<td>Tobacco ringspot</td>
<td>177</td>
<td>114, 177, 237, 290</td>
<td>++ +</td>
<td>Stace-Smith et al. 1965</td>
</tr>
<tr>
<td>Tomato bushy stunt</td>
<td>402</td>
<td>236, 480, 661, 894</td>
<td>-</td>
<td>Tremaine &amp; Goldsack, 1968</td>
</tr>
<tr>
<td>Tomato ringspot</td>
<td>207</td>
<td>86, 172, 255, 301, 387</td>
<td>-</td>
<td>Tremaine &amp; Stace-Smith, 1968</td>
</tr>
<tr>
<td>Turnip yellow mosaic</td>
<td>191</td>
<td>191, 325</td>
<td>++ +</td>
<td>Harris &amp; Hindley, 1965</td>
</tr>
<tr>
<td>White clover mosaic</td>
<td>133</td>
<td>68, 165, 227, 301, 361</td>
<td>-</td>
<td>Miki &amp; Knight, 1967</td>
</tr>
</tbody>
</table>

* Principal IDI minimum in bold type.
† Agreement between published size and principal IDI minimum: + + + = ±2 residues; ++ = ±10 residues; + = ±25 residues; - = no agreement.
‡ No published estimate of size.

(b) Viruses with isometric particles. Turnip yellow mosaic virus [R/I: 1.9/37: S/S: S/CI] is perhaps the best studied of these. Symons et al. (1963) studied six strains of the virus and reported that their subunits contain 187 to 190 amino acid residues. Harris & Hindley (1965) confirmed this and estimated by various methods that the subunits contain 189-191 residues. FITMOL analysis of Harris & Hindley's data agrees exactly with their conclusions (Table 2).

Symons et al. (1963) also reported that the subunits of wild cucumber mosaic virus [R/I:
Size and composition of plant virus protein

2.4/35: S/S: S/C1, closely related to turnip yellow mosaic virus, contain 189 or 190 residues. The only other related virus of which the protein has been analysed is belladonna mottle virus [R/1: 3/36: S/S: S/\*], and it is noteworthy that a FITMOL analysis of data reported for this virus by Jankulow et al. (1968) shows a clear minimum IDI at 191 amino acid residues.

The group of viruses related to cowpea mosaic virus [R/1: 1.45 + 2.3/24 + 33: S/S: S/C1] share some properties of the turnip yellow mosaic viruses but seem unrelated. The proteins in the particles of four of the cowpea mosaic viruses have been analysed, these are cowpea mosaic virus (Niblett & Semancik, 1969), bean pod mottle virus [R/1: 2.4/35: S/S: S/C1] (Semancik, 1966b), echtes Ackerbohnenmosaik virus [R/1: */35: S/S: S/\*] (Wittman & Paul, 1961), and squash mosaic virus [R/1: 2.4/35: S/S: S/C1] (Mazzone et al., 1962). Niblett & Semancik (1969) stated that cowpea mosaic virus protein contains 159 to 165 amino acid residues and Tremaine & Goldsack (1968) computed that the subunits of bean pod mottle virus contain 201 residues and those of squash mosaic virus 189 residues: FITMOL analyses of the same data (Table 2) do not confirm Tremaine & Goldsack's calculations but give IDI minima of irregular intervals with small ratios in maxima to minima IDI.

The nematode-borne ringspot viruses form another well-defined group which shares several properties with the cowpea mosaic viruses. Stace-Smith et al. (1965) analysed tobacco ringspot virus [R1/: 1.8/42: S/S: S/\*] and a FITMOL analysis of their results

Fig. 6. FITMOL analysis of data for potato virus X: PVX (1) strain c and PVX (2) the ringspot strain from Shaw et al. (1962); PVX (3) from Shaw & Larson (1962) and PVX (4) from Miki & Knight (1968).
(Table 2) agrees with their conclusions of subunits with 177 amino acids residues. Tremaine & Stace-Smith (1968) also analysed the proteins in the particles of tomato ringspot virus (*/*: */*: S/S: S/Ne), and calculated that the subunits contain 217 residues. However, FITMOL analysis of the data shows only a slight minimum in the IDI at 217 residues but much more pronounced minima at 86, 172 and 255 residues per subunit. FITMOL analyses show that it is likely that these two closely related viruses will have similarly sized protein subunits of around 175 amino acid residues.

![Graph showing FITMOL analysis of data for the brome mosaic viruses](image)

**Fig. 7.** FITMOL analysis of data for the bromoviruses: (1) broad bean mottle virus (Yamazaki & Keesberg, 1963); (2) broad bean mottle virus (Miki & Knight, 1965); (3) brome mosaic virus (Stubbs & Keesberg, 1964); (4) cowpea chlorotic mottle virus (Bancroft et al. 1968).

Another clearly defined group of plant viruses with isometric particles is that related to brome mosaic [R/I: 1/22: S/S: S/*], namely, broad bean mottle virus [R/I: 1.1/22: S/S: S/*], cowpea chlorotic mottle virus [R/I: 1.1/24: S/S: S/*] and brome mosaic virus itself. The IDI minima found by FITMOL analyses of the data on these viruses are all at slightly larger protein sizes than those reported (Fig. 7, Table 2).

Cucumber mosaic virus [R/I: 1/18: S/S: S/Ap] and alfalfa mosaic virus [R/I: 1.3/18: U/U: S/Ap] are aphid transmitted viruses whose particles share some properties with those of viruses of the brome mosaic virus group, though their subunits are reported to be much larger. Van Regenmortel (1967) reported that the subunits of cucumber mosaic virus contain 287 residues, and Hull, Rees & Short (1969) reported that the subunits of alfalfa
mosaic virus contain 297 residues; FITMOL analyses of their data agree with their conclusions (Table 2).

Pea enation mosaic virus [R/1: 1.3/27: S/S: S/As] shows no clear affinities to other viruses whose particles have been chemically analysed. Shepherd, Wakeman & Ghabrial (1968) analysed the protein in its particles, and calculated that each subunit contains 189 amino acid residues; however, FITMOL analysis of the data does not unequivocally agree with their conclusions, for although there is a minimum IDI at 198 residues, there are more pronounced minima at 86, 157, 280 and 355 residues.

The remaining viruses to be discussed are those with rounded isometric particles containing about 20% nucleic acid, and stable in concentrated salt solutions. These viruses have many common properties (Gibbs, 1969) but it is not clear how closely they are related. They include sowbane mosaic virus [R/1: 1.3/17: S/S: S/Di], southern bean mosaic virus [R/1: 1.4/21: S/S: S/Ci], carnation ring spot virus [R/1: 1.4/20: S/S: S/*], tomato bushy stunt virus (pelargonium leaf curl virus) [R/1: 1.5/17: S/S: S/*], tobacco necrosis virus
A.J. GIBBS AND G. A. MCINTYRE

and satellite virus [R/I: 0.4/20: S/S: S/Fu]. FITMOL analyses of the compositions reported for the proteins of the first three of these viruses agree, though not closely, with their reported sizes (Table 2), whereas the FITMOL analysis and size reported for tomato bushy stunt virus protein disagree (Table 2). However, the most interesting of these FITMOL analyses are those of the data for tobacco necrosis and satellite viruses, for they show how FITMOL analyses can be useful for assessing and comparing conflicting reports:

(i) Although satellite virus has been much studied, the size of the protein in its particles is uncertain. The sizes reported for this protein together with FITMOL analyses of the reported amino acid composition of the protein are shown in Fig. 8.

Reichmann et al. (1962) reported that chemical analyses showed the protein to contain about 240 amino acid residues, but further chemical analyses (Reichmann, 1964) indicated 372 residues. The FITMOL analysis of the composition reported by Reichmann shows IDI minima at 131, 240 and 374 residues (Fig. 8).

Uyemoto & Grogan (1969) also made chemical analyses of the protein and calculated that it contained 209 residues. FITMOL analysis of their data gives IDI minima at 177, 274 and 379 residues (Fig. 8) and their tryptic peptide analysis (29 peptides) and amino acid composition suggest that the protein has only 169 to 173 amino acid residues.

The most recent report of chemical analyses is that of Rees, Short & Kassanis (1970), who estimated that the protein in the particles of the SV1 strain of the virus contains 208 residues and that of the SV2 strain 207 residues. FITMOL analyses of their results give for the SV1 strain an exceptionally clear minimum IDI at 208 and 413 residues and, for the SV2 strain, less clear minima at 219, 279 and 502 residues (Fig. 8).

The size of the protein has also been estimated by the polyacrylamide gel electrophoresis method (Weber & Osborn, 1969). Roy et al. (1969) estimated by this method that the protein had a molecular weight of 24,000 to 31,000 (about 210 to 270 residues) and, in further more detailed experiments, Lesnaw & Reichmann (1969) estimated that the protein had a molecular weight around 20,000 (about 175 residues). Thus the size of the protein estimated by the electrophoretic method seems to be less than that obtained by Rees et al. (1970), whose chemical analyses were the most detailed.

(ii) There have been two reports of analyses of tobacco necrosis virus protein. Uyemoto & Grogan (1969) estimated from amino acid and peptide analyses that the subunit protein contained 197 amino acid residues: FITMOL analysis of their results shows a clear minimum at 261 residues (Table 2). By contrast, Lesnaw & Reichmann (1969), working with an URBANA isolate of the virus, reported a quite different amino acid composition and from this, and from acrylamide gel electrophoresis experiments, estimated that it contained 312 residues; a value near the clear minimum at 301 residues obtained in a FITMOL analysis of their results. The nucleotide composition of the two virus isolates used by these authors also differed considerably and, although the strain of the virus used by Uyemoto & Grogan has been reported to be serologically related to other strains of the virus, the URBANA strain has not.

DISCUSSION

Ozawa & Tanaka (1968) have published a computer program which estimates the size of a protein from its composition. However, estimates provided by this program are of doubtful accuracy since the program treats the most abundant amino acid in each set of data as the ‘key’ amino acid and only considers proteins that contain integral numbers of residues of that
Size and composition of plant virus protein

amino acid; the evaluation of all possible protein sizes, as by FITMOL, seem to have definite advantages.

Whatever method of calculation is used, it is clear that the most important data for estimating the size of a protein from its composition are the amounts of the least abundant amino acids in the protein, and unfortunately these are often amino acids, such as cysteine or methionine, which are most difficult to estimate accurately.

The use of the FITMOL program to re-examine some of the data reported for plant virus proteins has shown the reasons for some of the discrepancies and contradictions illustrated by Fig. 1. Many previous reports of protein size have been based on amino acid analyses alone, and this has obvious dangers. However, even with other evidence for the size of the protein, it must be interpreted with caution. Table 2 summarizes the results of FITMOL analyses reported in this paper which show, contrary to the evidence summarized in Fig. 1, that there is no unequivocal example of closely related viruses having proteins in their particles that differ in size by more than a few amino acid residues.

Many of the reports re-examined relied on evidence from tryptic peptide analyses which seem to give misleading results. There are presumably several reasons for this. First it is likely that the uncertainty in estimating the number of tryptic peptides obtained from a protein increases as the number of peptides increase, and some plant virus proteins give as many as 30 tryptic peptides. Secondly, it is not clear whether only carefully cloned virus stocks were used in all studies on these virus proteins. Thirdly, trypsin does not hydrolyse all arginyl or lysyl bonds at similar rates, and some not at all (Hill, 1965); Offord & Harris (1965) obtained 29 tryptic peptides from the protein of tobacco-rattle virus particles but only 20 of these represented unique sequences, and they estimated that the protein contains only nine arginine and 15 lysine residues, and should therefore have given 25 peptides.

The amino acid composition of a protein will not, on its own, give an unequivocal estimate of the size of the protein, but can indicate which sizes are more probable; other independent methods of estimating the size of the protein must be used to choose between the alternatives. In the past, many have estimated the number of tryptic peptides to give the necessary independent estimate of protein size; this method can give misleading results unless every peptide is shown to be unique or unless it is used in conjunction with the method suggested by Laver (1964, 1969). Electrophoresis in dodecyl sulphate-polyacrylamide gels (Weber & Osborn, 1969), may be one of the simplest ways of obtaining an independent estimate of protein size at sufficient accuracy to distinguish between the sizes suggested by a FITMOL analysis.

We are indebted to Miss J. M. Turner for technical help.

Note added in proof: It has been pointed out to us that a method similar to that described in this paper was used by Nyman & Lindskog (1964). Biochemica et biophysica acta 85, 141.

REFERENCES


Virology 35, 87.


Virology 37, 404.


Virology 33, 10.


Journal of Molecular Biology 9, 109.


Size and composition of plant virus protein


(Received 10 October 1969)