Journal of Chromatography, 472 (1989) 347–356
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 21 430

ACCURACY AND PRECISION IN THE DETERMINATION OF STOKES RADII AND MOLECULAR MASSES OF PROTEINS BY GEL FILTRATION CHROMATOGRAPHY

FRANCESC CABRÉ and ENRIC I. CANELA*

Department of Biochemistry and Physiology, University of Barcelona, Martí i Franquès 1, 08028 Barcelona, Catalonia (Spain)

and

MIGUEL A. CANELA

Department of Applied Mathematics, University of Barcelona, 08007 Barcelona, Catalonia (Spain)

(First received October 18th, 1988; revised manuscript received January 3rd, 1989)

SUMMARY

The accuracy and precision of the estimates of hydrodynamic parameters of globular proteins obtained by inverse regression from gel filtration chromatographic data are discussed. The usefulness of gel filtration chromatography as the basis for a rapid and reliable method for the determination of the Stokes radius and the molecular mass is considered. The discussion is supported by an analysis of the models already proposed in the literature, and is based on the precision of the estimates.

INTRODUCTION

Many analytical procedures have been used for the determination of the relative molecular mass (M_r) and the molecular size of proteins and nucleic acids. Whereas the most accurate of these techniques, *e.g.*, sedimentation velocity and sedimentation equilibrium measurements, viscosity and density determinations and light scattering (see refs. 1–5 and the references cited therein), require selected and expensive instrumentation, gel filtration is a very simple method. Through a mathematical approach, we have checked the quality of the results that can be obtained from gel filtration chromatographic data for globular proteins by using the various models already available.

Gel filtration chromatography can be considered as a transport phenomenon. Although the mechanism of separation of macromolecules by gel filtration is not completely understood[6] it is now well established that the behaviour of proteins in the gel matrix can be better related to their hydrodynamic radius (Stokes radius, R_S) than to their relative mass, M_r[7–9]. Considering the molecules of globular proteins as spheres with a defined hydrodynamic radius[10,11] is a simple assumption which can be very useful in the determination of molecular masses if it is combined with the determination of sedimentation coefficients. This combination is necessary because proteins are

0021-9673/89/$03.50

not truly spherical but have various shapes and extents of hydration, and therefore no unique relationship exists between R_S and M_r[12]. Therefore, it is not reasonable to assume that there exists a method for determining one parameter from the other, valid for any protein, and testing the mathematical models already proposed in the literature, involving any of these two parameters and the chromatographic variables, seems to be a reasonable step.

It is convenient to distinguish between accuracy and precision in the determination. The accuracy is related to bias, *i.e.*, the mean of the deviations from the real value, and the precision is concerned with the reproducibility of the determination[13]. It is possible for a given method to be accurate, *i.e.*, no systematic error is involved, but of low precision; conversely, the same erroneous value could be repeatedly obtained. It is not unusual to find in the literature different estimates of the molecular mass of the same protein even if the same method has been used; specimen purity and calibration technique can account for discrepancies on this magnitude in many instances[13].

In this paper, we discuss seven models that have been already introduced, in the light of real data for nine proteins used for calibration. These models are judged according to the accuracy and precision obtained when determining R_S or M_r by inverse regression from gel filtration chromatographic data.

EXPERIMENTAL

Proteins

The following proteins were used for calibration (see Table I): thyroglobulin (bovine thyroid), ferritin (horse spleen), catalase (bovine liver), and aldolase (rabbit muscle) (all from Pharmacia); albumin (bovine serum) (Serva); and ovalbumin (egg white), chymotrypsinogen A (bovine pancreas), myoglobin (whale muscle) and cytochrome *c* (horse heart) (all from Sigma).

TABLE I

MOLECULAR MASSES AND STOKES RADII OF NATIVE PROTEINS USED FOR CALIBRATION

Standards in aqueous solution obtained from sedimentation equilibrium.

Protein	*Molecular mass*	R_S *(nm)*
Thyroglobulin	670 000[a]	8.60[a]
Ferritin	440 000[b]	6.06[b]
Catalase	230 000[a]	5.23[a]
Aldolase	148 000[c]	4.60[d]
Albumin	67 000[e]	3.55[e]
Ovalbumin	43 500[e]	2.73[e]
Chymotrypsinogen A	23 000[e]	2.24[e]
Myoglobin	17 000[e]	2.08[e]
Cytochrome *c*	13 400[e]	1.65[e]

[a] Potschka[5].
[b] Frigon *et al.*[14].
[c] Righetti *et al.*[15].
[d] Hoorike *et al.*[6].
[e] Mantle[16].

Reagents

Blue Dextran 2000 was purchased from Pharmacia and potassium dichromate from Merck. Distilled water, further purified with a Millipore Milli-Q system, was used throughout.

Gel filtration chromatography

Gel filtration was carried out at 4°C on an Econo-column (Bio-Rad Labs.) of Sephacryl S-300 (Pharmacia) (111 × 1 cm I.D.) equilibrated with 50 m*M* Tris–HCl buffer (pH 8.2)–0.1 *M* NaCl. A 0.8-ml volume of each sample was applied, at a concentration of 3 mg/ml, with elution at a rate of 10 ml/h. The absorbance at 280 nm of the effluent was continuously recorded.

Chromatographic data were expressed in terms of the distribution coefficient, K_D or K_{av}, defined by the equations

$$K_D = \frac{V_e - V_0}{V_i}$$

and

$$K_{av} = \frac{V_e - V_0}{V_t - V_0}$$

where V_e is the elution volume of the protein under study, V_0 is the void volume (elution volume of Blue Dextran 2000, 1 mg/ml), and V_i is the internal volume (given by $V_i = V_t - V_m - V_0$, V_t being the total volume and V_m the matrix volume).

The calibration proteins were chromatographed 3–8 times.

Calibration of the column by Stokes radius

Several equations have been proposed in order to describe the relationship of the distribution coefficient of the protein with R_S. The following models have been considered in this paper.

Model I:

$$\mathrm{erf}^{-1}(1 - K_D) = a + bR_S$$

was used by Horiike *et al.*[17], according to Ackers[18], who assumed that the effective radius of the pores follows a Gaussian distribution. This Gaussian distribution has also been considered by several workers in this context[6,12,19–23].

Model II:

$$K_D^{1/3} = a + bR_S$$

was proposed by Porath[24], and used later by Horiike *et al.*[17].

Model III:

$$(-\log K_{av})^{1/2} = a + bR_S$$

proposed by Laurent and Killander[25], and later by Siegel and Monty[26], is usually used when the measurements are obtained for a variety of proteins at the same gel concentration.

Model IV:

$$\frac{1000}{V_e} = a + bR_S$$

was proposed by Davis[23] as a simplified calibration procedure for gel filtration columns.

Correlation of distribution coefficient and molecular mass

Three models relating the distribution coefficient to the molecular mass have been considered.

Model V:

$$K_{av} = a + b \log M_r$$

has been used by several authors and is usually considered in studies of gel filtration[20].

Model VI:

$$\mathrm{erf}^{-1}(1 - K_D) = a + bM_r^{1/3}$$

was developed by Fish[11].

The sigmoidal model VII:

$$K_{av} = \frac{1}{1 + (M_r/a)^b}$$

can be transformed into a linear model using the function logit $Y = \ln[Y/(1 - Y)]$[20]:

$$\mathrm{logit}\ K_{av} = a + b \log M_r$$

In all the equations a and b are empirical constants for a given chromatographic system, and were estimated by a linear regression discussed below. R_S and M_r are assumed to be free from errors of determination and were therefore taken as control variables. The experimental variables were considered as response variables and written on the left-hand side of the equations.

RESULTS

Preliminary analysis

Let us consider the model

$$y_{ij} = a + bx_i + u_i + \varepsilon_{ij} \qquad 1 \leqslant i \leqslant 9; \qquad 1 \leqslant j \leqslant n_i \qquad (1)$$

where x represents the control variable in any of the seven models introduced above and y is the response variable; n_i replications were made for the ith calibration protein (n_i varies between 3 and 8); y_{ij} is the value of y in the jth replication for the ith protein, and ε_{ij} is the error in the determination of y_{ij}, which is assumed to follow a Gaussian distribution, $N(0,\sigma_i)$; u_i represents the deviation of the ith protein from "ideal" behaviour, *i.e.*, the exact linear model. This deviation has been commented upon in the Introduction and including it in the model allows us to assume that ε_{ij} has a zero mean. Not much can be guessed, at present, about u_i for an individual protein and, more important, nothing can be known from the bare value x_i. Therefore, the best we can do is to make the simplest hypothesis concerning u_i, *i.e.*, that u_i follows a Gaussian distribution $N(0,\sigma)$ (σ independent of x). It is important to distinguish between both types of error, because omission of u_i leads to a model that does not pass the usual test of linearity. Nevertheless, this distinction makes eqn. 1 unmanageable, and a reduction must be made.

Taking means in eqn. 1:

$$\bar{y}_i = a + bx_i + u_i + \bar{\varepsilon}_i \qquad 1 \leqslant i \leqslant 9 \tag{2}$$

and now $\bar{\varepsilon}_i$ is $N(0,\sigma_i/\sqrt{n_i})$. An exploratory analysis of the values of $\sigma^2 + \sigma_i^2/n_i$, carried out by estimating the residual variance after fitting by the ordinary least-squares (OLS) method a linear model to the pairs $(x_i, \bar{y}_i)$, shows that the values σ_i^2/n_i, although different, are small in comparison with σ^2, and thus we are led to a model

$$\bar{y}_i = a + bx_i + w_i \qquad 1 \leqslant i \leqslant 9 \tag{3}$$

where the variance of w_i assumed to be constant, $\bar{y}_i$ can be taken as an estimate of the true value of y for the ith protein (not the expected value corresponding to $x = x_i$, unless the ith protein could be assumed to be "ideal").

Predictions from the models

To check the seven models, the following operations were performed. For each i we considered the sample obtained by omitting the pair $(x_i, \bar{y}_i)$ and fitted a linear model by the OLS method to this sample. Then the model obtained was used to calculate x_i by inverse regression, and this prediction was recorded. We thus obtained nine errors for each model. The predictions, together with the relative errors, in the form of percentages, are presented in Tables II–VIII. We can use now these errors in order to discuss the accuracy and precision of these methods for determining R_S and M_r. The purpose of omitting one pair $(x_i, \bar{y}_i)$ when fitting the models is to avoid the influence of the pair in the prediction of x_i when the prediction is made using a model obtained from a sample in which the pair itself was included (see ref. 27, Chapter 2, for an elementary discussion of this subject).

Let us look first at the models involving R_S (models I–IV). The estimates of R_S obtained by means of these models are negatively biased, the mean of errors being *ca.* -0.7 for all of them. Therefore, these models can be considered as reasonably and similarly accurate.

The precision of a method can be measured in different ways (variance, mean square error, median absolute deviation, etc.). Nevertheless, we are not interested here

TABLE II

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE STOKES RADIUS IN MODEL I

Each predicted value of R_S was obtained by inverse regression from a linear model fitted to the eight other points. The percentage errors are referred to the true values of R_S.

erf^{-1} $(1 - K_d)$	R_S	Predicted R_S	Percentage error
1.2078	86.0	80.73	6.12
0.9437	60.6	64.28	6.07
0.7686	52.3	48.88	6.53
0.7257	46.0	45.92	0.17
0.6244	35.5	38.50	8.45
0.4871	27.3	27.60	1.09
0.4234	22.4	22.70	1.34
0.3778	20.7	18.74	9.46
0.3158	16.5	13.68	17.09

TABLE III

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE STOKES RADIUS IN MODEL II

Details as in Table II.

$K_d^{1/3}$	R_S	Predicted R_S	Percentage error
0.4442	86.0	80.07	6.98
0.5667	60.6	64.80	6.93
0.6519	52.3	49.00	6.31
0.6730	46.0	45.92	0.17
0.7226	35.5	38.32	7.94
0.7888	27.3	27.35	1.83
0.8190	22.4	22.52	0.54
0.8401	20.7	18.73	9.52
0.8685	16.5	13.96	15.39

TABLE IV

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE STOKES RADIUS IN MODEL III

Details as in Table II.

$(-\log K_{av})^{1/2}$	R_S	Predicted R_S	Percentage error
1.0166	86.0	78.76	8.42
0.8463	60.6	64.51	6.45
0.7306	52.3	49.65	5.07
0.7016	46.0	46.76	1.61
0.6322	35.5	39.39	10.96
0.5340	27.3	28.07	2.82
0.4862	22.4	22.67	1.21
0.4489	20.7	17.89	13.50
0.3997	16.5	11.80	28.40

TABLE V

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE STOKES RADIUS IN MODEL IV

Details as in Table II.

$1000/V_e$	*R_S*	*Predicted R_S*	*Percentage error*
25.21	86.0	77.63	9.73
22.28	60.6	66.18	9.20
19.95	52.3	49.64	5.09
19.35	46.0	46.34	0.74
17.96	35.5	38.44	8.28
16.14	27.3	26.92	1.39
15.34	22.4	22.47	0.31
14.79	20.7	18.84	8.99
14.08	16.5	14.54	11.38

TABLE VI

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE MOLECULAR MASS IN MODEL V

Each predicted value of M_r was obtained by calculating log M_r by inverse regression from a linear model fitted to the eight other points and transforming the resulting estimate into an estimate of M_r. The percentage errors are referred to the true values of M_r.

K_{av}	*M_r*	*Predicted M_r*	*Percentage error*
0.0927	670 000	872 048.30	30.16
0.1926	440 000	363 710.70	17.34
0.2928	230 000	189 132.62	17.77
0.3219	148 000	158 330.34	6.98
0.3985	67 000	94 474.86	41.01
0.5187	43 500	37 503.20	13.79
0.5803	23 000	25 380.36	10.35
0.6265	17 000	18 025.14	6.03
0.6921	13 400	10 275.22	23.32

TABLE VII

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE MOLECULAR MASS IN MODEL VI

Each predicted value of M_r was obtained by calculating $M_r^{1/3}$ by inverse regression from a linear model fitted to the eight other points and transforming the resulting estimate into an estimate of M_r. The percentage errors are referred to the true values of M_r.

erf^{-1} $(1 - K_d)$	*M_r*	*Predicted M_r*	*Percentage error*
1.2078	670 000	959 386.11	43.19
0.9437	440 000	332 089.95	24.52
0.7686	230 000	196 804.47	14.43
0.7257	148 000	162 754.97	9.97
0.6244	67 000	102 509.69	53.00
0.4871	43 500	41 021.98	5.70
0.4234	23 000	25 862.72	12.45
0.3778	17 000	16 516.54	2.84
0.3158	13 400	7 246.70	45.92

TABLE VIII

VALUES OF THE CHROMATOGRAPHIC VARIABLE AND THE MOLECULAR MASS IN MODEL VII

Details as in Table VI.

Logit(K_{av})	M_r	*Predicted* M_r	*Percentage error*
−2.2828	670 000	1 850 502.70	176.19
−1.4359	440 000	313 949.36	28.65
−0.8829	230 000	152 276.88	33.79
−0.7452	148 000	127 496.24	13.85
−0.4119	67 000	79 888.69	19.24
0.0750	43 500	36 274.601	16.61
0.3232	23 000	26 574.56	15.54
0.5192	17 000	19 712.90	15.96
0.8119	13 400	12 267.75	8.45

in checking the quality of the models from a purely mathematical point of view, *i.e.*, the goodness of fit, but from a practical point of view, according to the precision of the estimates of R_S which could be obtained using them. In order to stress such an approach, the errors are presented as relative errors, and our discussion will be based on these. This presentation makes the result look worse than when the absolute error/length of interval ratio for the x variable is expressed, which would be the natural way for a linear model. Moreover, the correlations are high, above 0.99 for any of the nine fittings made for each model. However, the main interest here is the usefulness of the model for the determination of R_S, and the approach used here seems to be correct and easy to understand, and any one can draw his or her own conclusions from the results in Tables II–V.

For the models involving the molecular mass (models V–VII), the same analysis was performed. Nevertheless, M_r was transformed to linearize the models and, in spite of the high correlations (always above 0.975), the estimates of M_r show errors whose size is partly due to the change in dimension.

CONCLUSIONS

Gel filtration chromatography is considered to be a rapid and useful technique for the determination of the size and relative molecular masses of proteins[5,12]. Classical physico-chemical methods, such as sedimentation analysis, light scattering and electron microscopy, require very specific instrumentation[13], but gel filtration chromatography has the advantages of being relatively simple and of providing accurate results when the column has been calibrated properly.

The separation mechanism of gel filtration chromatography involves not only the molecular mass but also the shape of the molecules. Potschka[5] suggested that the universal calibration principle for gel filtration chromatography is the viscosity radius, *i.e.*, the molecular volume times a shape function which is defined by the intrinsic viscosity. Nonetheless, the reported differences between the Stokes radius based on the translational frictional coefficient, *i.e.*, calculated for native proteins from the diffusion coefficient with the Stokes–Einstein equation, and that based on the intrinsic viscosity are usually not larger than 10%[28] or are indistinguishable[6,29]. We consider that the use of any Stokes radius for calibration in gel filtration chromatography could lead to good results.

We present here some conclusions from the results of the analysis made on seven models taken from the literature. The technical details have been given in the preceding section. The most obvious fact is that the use of the Stokes radius leads to better results, as could be expected considering what was previously known about the subject.

The models I–IV can be taken as acceptable for the determination of R_S, but some facts deserve attention. The errors obtained for proteins 1 and 9 must be considered, bearing in mind that they come from predictions corresponding to values of x falling outside the interval used in the determination of the parameters. It is interesting that, for all the models checked, the same proteins have either a low (aldolase) or a high (albumin) percentage error. This truly reflects the fact that some proteins behave anomalously with respect to the others. The basis for the difference is probably a greater deviation from a spherical shape (or, less likely, greater hydration) of some proteins. Without considering the error for cytochrome *c*, we do not find significant differences among the four models. Model IV has the advantage of using V_e directly, allowing an easier interpretation, but model I has a suggestive physical explanation, based on the assumption that the pore size of the matrix is Gaussian[18], as mentioned earlier. However, this model is limited as this assumption is not valid except for a particular Gaussian distriution of pore size centred at the origin, and Le Maire *et al.*[12] have shown that when the pore size distribution is calculated using an experimentally determined $K_D = f(R_S)$, the pore site is bimodal and therefore in no way Gaussian.

The importance of robustness in these analyses must be emphasized, because of the risk that the presence of a protein with very far from ideal behaviour could adversely affect the estimates of the parameters. We have already described the cautious approach followed in this work to the analysis of the size of the errors. Unless a deep knowledge of the proteins used for the calibration allows the experimenter to disregard such problems, we consider it advisable to use a robust regression technique in the calculation of the parameters of the model to be used for future determinations.

With respect to models V–VII, our results confirm that the use of M_r as a parameter for the description of the behaviour of the molecule inside the column is not adequate, as has been repeatly stated in the literature. However, if truly spherical proteins, hydrated to the same extent, are used, the errors can be minimized and a direct relationship between M_r and R_S can be achieved. In any event, the size of the errors obtained in this work does not allow us to consider these models as the basis for any precise method of determination of molecular masses of proteins. Nevertheless, they could be used to obtain an approximation of the relative magnitudes of the molecular masses of different proteins, *i.e.*, as a basis for comparative methods.

Finally, it is interesting to emphasize that a combination of R_S and sedimentation coefficient measurements to obtain M_r[26] leads to an error most generally smaller than that which results from a direct determination of M_r by gel chromatography.

ACKNOWLEDGEMENT

The presentation of the results in this paper benefited substantially from the suggestions of the referee.

REFERENCES

1 H. Aburatani, T. Kodama, A. Ikai, H. Itakura, Y. Akanuma and F. Takatu, *J. Biochem. (Tokyo)*, 94 (1983) 1241–1245.

2 H. Hagamoto and K. Yagi, *J. Biochem. (Tokyo)*, 95 (1984) 1119–1130.

3 K. Takeuchi and K. Ishimura, *J. Biochem. (Tokyo)*, 97 (1985) 1695–1708.

4 Y. Morita, H. Iwamoto, S. Aibara, T. Kobayashi and E. Hasegawa, *J. Biochem. (Tokyo)*, 99 (1986) 761–770.

5 M. Potschka, *Anal. Biochem.*, 162 (1987) 47–64.

6 K. Horiike, H. Tojo, T. Yamano and M. Nozaki, *J. Biochem. (Tokyo)*, 93 (1983) 99–106.

7 W. W. Fish, J. A. Reynolds and C. Tanford, *J. Biol. Chem.*, 245 (1970) 5166–5168.

8 K. G. Mann and W. W. Fish, *Methods Enzymol.*, 26 (1972) 25–42.

9 M. le Maire, L. P. Aggerbeck, C. Monteilhet, J. P. Andersen and J. V. Møller, *Anal. Biochem.*, 154 (1986) 525–535.

10 C. Tanford, *Physical Chemistry of Macromolecules*, Wiley, New York, 1961, Ch. 6.

11 W. W. Fish, *Methods Membr. Biol.*, 4 (1975) 189–276.

12 M. le Maire, A. Ghazi, J. V. Møller and L. P. Aggerbeck, *Biochem. J.*, 243 (1987) 399–404.

13 A. J. Rowe, in *Techniques in Protein and Enzyme Biochemistry*, Elsevier/North-Holland, Amsterdam, 1978, B105a, pp. 1–31.

14 R. P. Frigon, J. K. Leypoldt, S. Uyeji and L. W. Henderson, *Anal. Chem.*, 55 (1983) 1349–1354.

15 P. G. Righetti, G. Tudor and K. Ek, *J. Chromatogr.*, 220 (1981) 115–194.

16 T. J. Mantle, in *Techniques in Protein and Enzyme Biochemistry*, Elsevier/North-Holland, Amsterdam, 1978, B105b, pp. 1–17.

17 K. Horiike, H. Tojo, M. Iwaki, T. Yamano and M. Nozaki, *Biochem. Int.*, 4 (1982) 477–483.

18 G. K. Ackers, *J. Biol. Chem.*, 242 (1967) 3237–3238.

19 G. K. Ackers, *Adv. Protein Chem.*, 24 (1970) 343–446.

20 D. Rodbard, in N. Catsimpoolas (Editor), *Methods of Protein Separation*, Vol. 2, Plenum Press, New York, 1976, pp. 145–218.

21 R. Vales, Jr. and G. K. Ackers, *Methods Enzymol.*, 61 (1979) 125–142.

22 R. S. Ehrlich, S. Hayman, N. Ramachandran and R. F. Colman, *J. Biol. Chem.*, 256 (1981) 10560–10564.

23 L. C. Davis, *J. Chromatogr. Sci.*, 21 (1983) 214–217.

24 J. Porath, *Pure Appl. Chem.*, 6 (1963) 233–244.

25 T. C. Laurent and J. Killander, *J. Chromatogr.*, 14 (1964) 317–330.

26 L. M. Siegel and K. J. Monty, *Biochim. Biophys. Acta*, 112 (1966) 346–362.

27 A. C. Atkinson, *Plots, Transformations and Regression*, Clarendon Press, Oxford, (1985).

28 C. Tanford, Y. Nozaki, J. A. Reynolds and S. Makino, *Biochemistry*, 13 (1984) 2369–2376.

29 R. E. Martenson, *J. Biol. Chem.*, 253 (1978) 8887–8893.