Eur. J. Biochem. *157*, 169 – 180 (1986)
© FEBS 1986

Protein volumes and hydration effects

The calculations of partial specific volumes, neutron scattering matchpoints and 280-nm absorption coefficients for proteins and glycoproteins from amino acid sequences

Stephen J. PERKINS

Kennedy Institute, London, and Laboratory of Molecular Biophysics, Oxford

(Received September 6, 1985/January 30, 1986) – EJB 85 0997

Amino acid sequences, carbohydrate compositions and residue volumes are used to compare critically calculations of partial specific volumes $\bar{v}$, neutron scattering matchpoints and 280-nm absorption coefficients with experimental $\bar{v}$ values for proteins and glycoproteins. The $\bar{v}$ values that are obtained from amino acid densitometry underestimate experimental $\bar{v}$ values by 0.01 – 0.02 ml/g while the $\bar{v}$ values from crystallographic volumes overestimate the experimental $\bar{v}$ values by 0.04 – 0.05 ml/g. An intermediate consensus volume set of amino-acid-residue volumes is proposed in order to predict experimental $\bar{v}$ values using sequence information. The method is extended to carbohydrates and glycoproteins. Neutron scattering matchpoints can be calculated from crystallographic residue volumes on the basis of the non-exchange of 10% of the main-chain NH protons. Crystallographic results on protein-bound water are used to account for the experimental values of $\bar{v}$ and matchpoints. Finally, 280-nm absorption coefficients, $A_{280}^{1\%, 1\,\mathrm{cm}}$, of 5 – 27 are found to be well predicted by the Wetlaufer procedure based on the totals of Trp, Tyr and Cys residues. Average errors are ±0.7, and the experimental $A_{280}^{1\%, 1\,\mathrm{cm}}$ values can be larger than the predicted values by 3%.

A wealth of accurate amino acid compositions for proteins and glycoproteins is presently available due to modern, rapid means of sequencing proteins and nucleic acids. In principal, macromolecular physical properties are more accurately and easily determined from these compositions in comparison to the use of classical biochemical techniques. The calculation of partial relative molecular masses M_r is an obvious example of this. The calculation of macromolecular volumes V and partial specific volumes $\bar{v}$ is required in a range of applications: scattering density, matchpoint and molecular mass control measurements in X-ray and neutron solution scattering; molecular mass determinations by ultracentrifugation; constraints for use in low-resolution modelling of macromolecular shapes; packing analyses of amino acid and carbohydrate residues by solution or crystallographic studies. Starting from the classical 1943 Cohn and Edsall publication, several tabulations of residue volumes for amino acid and carbohydrates have been reported [1 – 7]. These volumes can be summed on the basis of accurate amino acid and carbohydrate compositions to give V and $\bar{v}$. In the present study, new compilations of amino acids and carbohydrate residue volumes are derived from small-molecule crystal studies. These are critically compared with previous compilations [1 – 7]. The ability of these residue volumes to predict partial specific volumes and neutron scattering densities for proteins and glycoproteins is in turn critically compared with experimental data, bearing in mind that the partial volume is the particle volume corrected for hydration, solute binding and electrostriction effects. This clarifies the application of these calculations at a phenomenological level with the use of accurate composition data for not only solution scattering studies, but also for more general biophysical and biochemical applications. It is to be noted that these calculated $\bar{v}$ values are only valid for use in two-component solutions; in multicomponent solutions, such as in the presence of high concentrations of denaturants or electrolytes, considerable changes due to interactions with the solvent or with ligands will occur, and in such cases the calculations will generally fail [8]. Finally, the calculations of $\bar{v}$ and matchpoint data are correlated with protein hydration concepts on the basis of observed protein-bound water molecules from macromolecular crystal studies. This enables the macromolecular volume to be interpreted in molecular terms of (a) apparent changes induced by protein-bound water in densitometric studies and (b) the 'dry' molecular volume as visualised by neutron scattering studies.

Macromolecular concentrations are conveniently determined in a wide range of applications by absorbance measurements at 280 nm on the basis of the absorption coefficients $A_{280}^{1\%, 1\,\mathrm{cm}}$. The ability to calculate $A_{280}^{1\%, 1\,\mathrm{cm}}$ from accurate amino acid sequences would be more straightforward than the use of biochemical procedures, especially in circumstances where biochemical determinations are time-consuming or difficult. In a further examination, experimental and calculated absorption coefficents are compared with reference to accurate amino acid compositions to indicate the utility of this method.

METHODS

Densitometric theory

Volumes obtained from partial specific volumes $\bar{v}_i$ by summation or multiplication are partial volumes. A partial volume is the volume change upon the addition of component i at constant T and P and composition of all other compounds. Partial volumes thus include all volume changes derived from hydration, solute binding in general and electrostriction.

Correspondence to S. J. Perkins, Kennedy Institute, Bute Gardens, Hammersmith, London W6 7DW, England

Amino-acid and carbohydrate-residue partial molecular volumes V_i are obtained from literature sources in units of 10^{-3} nm^3 using the expressions:

$$V_i = \frac{\bar{v}_i M_i}{N_A} = \frac{\bar{V}_i}{N_A}$$

where N_A is Avogadro's number, and for each residue i, $\bar{v}_i$ is the partial specific volume, M_i is the molar mass, and $\bar{V}_i$ is the partial molar volume (ml/mol). The total macromolecular volume V is the sum of its components, i.e.

$$V = \Sigma N_i V_i$$

where N_i is the number of residue i that is present. The total partial specific volume $\bar{v}$ is given by:

$$\bar{v} = \frac{\Sigma \bar{v}_i w_i}{\Sigma w_i} = \frac{N_A \Sigma N_i V_i}{\Sigma N_i M_i} = \frac{\Sigma N_i V_i}{\Sigma N_i M_i}$$

where w_i is the weight fraction of each residue i.

Amino acid volumes from crystallography

Amino acid volumes were derived from literature sources, except for those based on the unit cell dimensions of small-molecule crystal structures. Data for these calculations were taken from 105 reports found in the crystallographic series Structure Reports between 1956–1978 which relate either to the free amino acids, or to crystal forms containing only H_2O, HCl or HBr as additional co-crystallites, or to dipeptides or tripeptides containing only additional Gly residues. The volumes V_i were determined from:

$$V_i = ABC(1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\ \cos\gamma)^{1/2}$$

where A, B, C are the lengths of the unit cell axes and α, β, γ are the unit cell angles. The volumes of H_2O, HCl, HBr and Gly residues were first determined. Comparisons of crystal volumes that differ only be the addition of H_2O gave the H_2O volume of $25.8 \pm 3.8 \times 10^{-3}$ nm^3 (five values). This agrees well with volumes of 23:6 $\pm 3.4 \times 10^{-3}$ nm^3 (seven values) from carbohydrate crystal structures [1], $24.5 \pm 2.3 \times 10^{-3}$ nm^3 (46 values) from the comparison of 187 anhydrous and hydrated inorganic salts [2] and $26.3 \pm 4.5 \times 10^{-3}$ nm^3 as the mean of data on nine crystal forms of ice [2]. The value of 24:5 $\times 10^{-3}$ nm^3 is used in this study since it is the best-determined. Similar calculations gave an HCl volume of $46.4 \pm 4.7 \times 10^{-3}$ nm^3 (22 values), an HBr volume of $58.3 \pm 3.1 \times 10^{-3}$ nm^3 (6 values) and a Gly residue volume of $68.2 \pm 7.8 \times 10^{-3}$ nm^3 (13 values). These data were used to determine the volumes of the 20 free amino acids from the 105 crystal structures. In the case of Gly, this procedure gave a volume of $78.2 \pm 1.8 \times 10^{-3}$ nm^3 (15 values). Comparison with the volume of the Gly residue above gives the volume change on peptide formation as -10.0×10^{-3} nm^3. This is comparable with other values of -12.3×10^{-3} nm^3 [3, 4], -13.9×10^{-3} nm^3 [5] and -11.1×10^{-3} nm^3 [6]. It was used to correct the amino acid crystal volumes before their presentation in Table 1.

Carbohydrate volumes from crystallography

Monosaccharide carbohydrate volumes were derived from the unit cell dimensions of monosaccharides and disaccharides and are summarized in [1]. As above, crystal structures that differed only by the addition of an Me (methyl) group or the presence of H_2O were analysed for this present work by differences using respectively volumes of 28.9×10^{-3} nm^3 [1] and 24.5×10^{-3} nm^3 [2]. Data for the α and β anomers of a monosaccharide were averaged since these were identical within error [1]. From the appropriate differences, estimates of the volume of condensation of two monosaccharides to a disaccharide range of $13-28 \times 10^{-3}$ nm^3. The volume change is larger than that for peptide formation, which is as expected since electrostriction factors are not involved in polysaccharide formation as they are in peptide formation [3–5]. A mean value of 20.7×10^{-3} nm^3 was used [1]. This was used to correct the monosaccharide volumes before presentation in Table 3.

Carbohydrate volumes from densitometry

Densitometry data on saccharides are reported in [7]. From these data, the average monosaccharide, disaccharide and trisaccharide partial molecular volumes are calculated as 184.9 $(\pm 1.4) \times 10^{-3}$ nm^3 (five values), 346.2 $(\pm 5.4) \times 10^{-3}$ nm^3 (six values) and 507.7 $(\pm 4.2) \times 10^{-3}$ nm^3 (four values). The differences between these volumes show that the volume of polysaccharide condensation is $23.5-23.6 \times 10^{-3}$ nm^3. This difference was used to correct the densitometric volumes of the free Glc, Gal and Man carbohydrates to their residue form (Table 3). Since these three volumes are on average 5.4×10^{-3} nm^3 smaller than those derived from monosaccharide crystal structures (Table 3), densitometric volumes for GlcNAc, GalNAc, Fuc and NeuNAc residues were estimated using this difference to correct the crystal volumes of these residues.

Experimental protein $\bar{v}$ values and sequences

Experimental $\bar{v}$ values for proteins and glycoproteins are derived from the densities of the buffer ϱ_{buff} and the solution ϱ_{sol} of the macromolecule, and the macromolecular concentration c [8]:

$$\bar{v} = \frac{1}{\varrho_{buff}}\left[1 - \frac{(\varrho_{sol} - \varrho_{buff})}{c}\right].$$

As an illustration of errors, if $\bar{v}$ is 0.75 ml/g and the error in c is arbitrarily taken as $\pm 4\%$, the resulting error in $\bar{v}$ is ± 0.01 ml/g. An error in density measurement of ± 0.0001 g/ml in either ϱ_{buff} or ϱ_{sol} for $c = 10$ mg/ml also leads to an error of ± 0.01 ml/g in $\bar{v}$. In this context, the density of water changes by ± 0.0001 g/ml if the temperature fluctuates by 0.5 °C at 20 °C. Experimental $\bar{v}$ values are thus sensitive to the accuracy of concentration and density measurements, and to the precision of temperature control during these measurements. For this study, data on 12 protein $\bar{v}$ values were taken from [9] where precautions were explicitly taken. It should be noted that ϱ_{sol} values should strictly be obtained, not from dialysis into buffer solutions, but instead from solutions containing only the protein in question. This is not straightforward unless the solutions are in pure water. Amino acid sequence data are available: ribonuclease A [10]; lima bean trypsin inhibitor (composition only [11]); hen lysozyme [12]; catalase [13]; α-lactalbumin [14]; chymotrypsinogen A and α-chymotrypsin [15]; bovine serum albumin [16]; tubulin [17, 18]; lactate dehydrogenase [19]; carboxypeptidase A [20] and β-lactoglobulin [21].

Data on $\bar{v}$ of glycoproteins were taken from the following sources: α_1 acid glycoprotein [22]; α_2-macroglobulin [23, 24]; immunoglobulin IgM GAL [25]; component C3 of complement [26, 27]; and immunoglobulin IgG3 (human) [28]. Protein sequences and carbohydrate sequences or compositions were taken from the following sources: α_1 acid glycoprotein [29, 30]; α_2-macroglobulin [31, 32]; IgM GAL [33–36], component C3 of complement [27, 37]. Mouse IgGl MOPC21 [38] and human IgGl KOL [39] were used with the carbohydrate data of [40] as representatives of IgG macromolecules.

Neutron scattering matchpoints

The calculation of neutron scattering length densities requires the summation of scattering lengths $\Sigma\, b$ which is divided by the partial volume V of the macromolecule or the solvent [41, 42]. Here, the volume of the macromolecule is not the determining quantity, but rather the total volume and its interaction terms as represented in the partial volume. The matchpoint is that percentage ^2H_2O whose scattering length density corresponds to that of the macromolecule. Matchpoints for proteins and glycoproteins are readily calculated from the $\Sigma\, b/V$ terms for the macromolecule in H_2O and ^2H_2O solvents, where the $\Sigma\, b$ terms allow for the solvent-exchangeable contents [1, 41]. The determination of matchpoints by contrast variation are typical experiments in which multicomponent concepts apply, although in dilute solutions the apparent value of $\bar{v}$ will correspond to the actual $\bar{v}$ value [8].

Experimental neutron scattering matchpoints are obtained from interpolated plots of $\sqrt{I(0)/ctT_s}$ measured as a function of the volume fraction of H_2O and ^2H_2O in the buffer. $I(0)$ is the intensity of scattering at zero scattering angle obtained from Guinier analyses, and the matchpoint corresponds to that volume fraction of ^2H_2O where $I(0)$ is zero. For matchpoint determinations, relative and not absolute concentration measurements c are sufficient. Neutron transmission measurements T_s are usually made simultaneously or close in time to those of $I(0)$. Sample cell thicknesses t are well determined or are held constant in the experiment. The volume fraction of $H_2O/^2H_2O$ in solutions is accurately known from buffer preparation and dialysis procedures and can be verified by transmission measurements. An error analysis based on a matchpoint determination of 40% ^2H_2O using measurements in 0%, 70%, 80% and 100% ^2H_2O buffers shows that as a worst case, errors of $\pm 4\%$ or $\pm 8\%$ in $I(0)$, c or T_s in the 0% ^2H_2O measurement leads to errors of $\pm 0.5\%$ or $\pm 1\%$ ^2H_2O, respectively, in the matchpoint determination. In practical terms, experimental matchpoints are seen to be insensitive to errors, provided that a sufficient number of intensities are recorded on each side of the matchpoint.

Experimental matchpoint data were taken from the following sources: lysozyme [43]; α_1 acid glycoprotein [44]; ribonuclease A [45]; fibrinogen [46]; components $C1r_2C1s_2$, C1q, C1 of complement [47, 48]; components C3, C3c and C3dg of complement [27]; myoglobin [49]. Additional sequences or compositions were taken as follows: fibrinogen [50], $C1r_2C1s_2$ [47], C1q [51, 52] (and K. B. M. Reid, unpublished results).

Calculation of absorption coefficients

Calculated absorption coefficients $A_{280}^{1\%,1\,\mathrm{cm}}$ or A_{cal} were obtained from molar absorption coefficients ε by [53]:

$$A_{\mathrm{cal}} = 10\ \varepsilon/M$$

where

$$\varepsilon = 5550\ \Sigma\mathrm{Trp} + 1340\ \Sigma\mathrm{Tyr} + 150\ \Sigma\mathrm{Cys}$$

where the summations refer to the total numbers of Trp, Tyr and Cys residues in the macromolecule. Experimental absorption coefficients A_{exp} (Fig. 2) were taken from the following sources. 13 values were taken from the Handbook of Biochemistry [12]; where more than one A_{exp} value is cited, the mean is taken after exclusion of unusually high or low values (deviating by over one standard deviation from the mean). The source of seven further A_{exp} values are given in [54]. That for subtilsin Carlsberg is from [55]; those for C-reactive protein and serum amyloid P component were from [56, 57].

Additional sequences were taken as follows: aspartate transcarbamylase, trypsin, insulin, subtilsin BPN, subtilsin Carlsberg, bovine pancreatic trypsin inhibitor [12, 58]; tortoise lysozyme [59]; human lysozyme [60]; components C4 and factor B of complement [61–63]; C-reactive protein and serum amyloid P component [64–67].

RESULTS AND DISCUSSION

The purpose of this work is to determine the precision to which partial specific volumes $\bar{v}$, neutron scattering matchpoints and 280-nm absorption coefficients can be calculated from accurate amino acid compositions of known complete sequences. Volume calculations for $\bar{v}$ and matchpoints depend on the use of accurate residue volumes for amino acids and carbohydrates. Accordingly amino acid residue volumes from four different methods of calculation are compared with one another and with the classical 1943 Cohn-Edsall volumes. The sets of volumes are next tested for their ability to reproduce protein $\bar{v}$ values, and from this a consensus volume set is derived to calculate protein $\bar{v}$ values. With the inclusion of carbohydrate residue volumes, similar tests are carried out for glycoprotein $\bar{v}$ values. Finally, comparisons are made between experimental neutron scattering matchpoints and those predicted from residue volumes, and further comparisons of matchpoints and $\bar{v}$ values are made with water hydration shells observed by protein crystallography. In a similar vein, 280-nm absorption coefficients require the knowledge of accurate molecular masses and accurate contents of Trp, Tyr and Cys residues in the macromolecule. The correlation between calculated and experimental values could thus be examined.

Comparisons of amino acid residue volumes

Amino acid residue volumes are compared in Table 1. The residues are arranged in order of decreasing hydrophobicity to follow the consensus hydrophobicity scale of Eisenberg [68]. The classical 1943 Cohn-Edsall compilation (Table 1) was obtained from densitometric studies of eight free amino acids and molar group summations for the other twelve amino acids. The original value for Cys was 103.3×10^{-3} nm^3 [3]; however, this was based on a calculation error and was corrected later on [69, 70] to give 106.7×10^{-3} nm^3 which is used here (Table 1). Since that time, full residue compilations can be derived by four independent approaches based on densitometry, molar group summations, protein crystal structures, and amino acid crystal structures. Five further compilations from these are given in Table 1.

In the densitometric approach, Zamyatnin [4] summarized data in 1972 for the 20 free amino acids. Of these, 13 were revised further in 1984 [5]. It is necessary to correct the amino acid volumes for peptide bond formation with the attendant

Table 1. *Amino acid residue volumes*

Residue		Cohn & Edsall [3]	Densitometry (Zamyatnin, 1972 [4])	Densitometry (Zamyatnin, 1984 [5]	Molar group summations (method B) [71]	Protein crystal structures [6]	Amino acid crystal structures (Methods)	Consensus volume (average of six sets)
		$\times 10^{-3}$ nm^3						
Hydrophobic	Ile	168.9	166.1	164.6	173.8	168.8 ± 9.8 (69)	170.1 ± 2.1 (2)	166.1 ± 3.4
	Phe	187.9	189.2	187.2	182.5	203.4 ± 10.3 (29)	203.9 ± 2.5 (4)	189.7 ± 7.4
	Val	141.4	139.4	136.8	147.3	141.7 ± 8.4 (91)	142.3 ± 2.9 (9)	138.8 ± 3.6
	Leu	168.9	166.1	164.6	173.8	167.9 ± 10.2 (57)	182.8 ± 7.5 (6)	168.0 ± 4.3
	Trp	228.5	226.9	225.1	236.6	237.6 ± 13.6 (9)	228.9 ± 1.4 (4)	227.9 ± 3.8
	Met	163.1	162.3	161.0	173.7	170.8 ± 8.9 (19)	176.0 ± 1.5 (2)	165.2 ± 1.8
	Ala	87.2	88.3	86.4	92.1	91.5 ± 6.7 (71)	97.1 ± 5.6 (6)	87.8 ± 2.3
	Gly	60.6	59.9	57.8	62.5	66.4 ± 4.7 (60)	68.2 ± 1.8 (15)	59.9 ± 2.2
	Cys	106.7[b]	108.1	107.9	107.9	105.6 ± 6.0 (16)	112.4 ± 2.6 (5)	105.4 ± 5.0
	Tyr	192.1	192.9	190.5	181.5	203.6 ± 9.6 (13)	202.3 ± 4.1 (12)	191.2 ± 8.0
	Pro	122.4	122.2	120.6	132.3	129.3 ± 7.3 (16)	129.0 ± 6.1 (3)	123.3 ± 1.8
Hydrophilic	Thr	117.4	115.7	113.5	127.9	122.1 ± 6.7 (32)	129.0 ± 3.6 (4)	118.3 ± 2.3
	Ser	91.0	88.6	86.2	97.8	99.1 ± 7.4 (46)	103.3 ± 0.7 (5)	91.7 ± 1.8
	His	152.4	152.5	150.1	172.9	167.3 ± 7.4 (8)	158.3 ± 7.7 (10)	156.3 ± 6.1
	Glu	141.4	137.8	128.7	150.3	155.1 ± 11.4 (13)	148.0 ± 2.8 (4)	140.9 ± 5.3
	Asn	117.4	117.3	115.6	123.8	135.2 ± 10.1 (12)	127.4 ± 0.5 (2)	120.1 ± 4.1
	Gln	142.4	143.3	141.9	150.3	161.1 ± 13.0 (5)	147.3 ± 2.5 (2)	145.1 ± 5.1
	Asp	114.6	110.6	108.5	123.8	124.5 ± 7.7 (17)	125.5 ± 1.6 (3)	115.4 ± 2.2
	Lys	174.3	167.9	166.2	187.8	171.3 ± 6.8 (5)	184.5 ± 1.2 (3)	172.7 ± 5.9
	Arg	181.3	172.7	197.3	198.8	202.1 ± 3.2 (3)[a]	192.9 ± 13.6 (4)	188.2 ± 9.6
Mean difference between calculated and experimental protein $\bar{v}$ values (ml/g)		−0.001 ±0.005	−0.011 ±0.005	−0.019 ±0.005	0.036 ±0.006	0.036 ±0.005	0.047 ±0.005	0.000 ±0.005
Correction constant $\Delta V/\Sigma N$ to be added to the above volumes prior to calculation of the consensus volume ($\times 10^{-3}$ nm^3)		0.3	2.2	3.4	−6.5	−6.9	−8.5	=

[a] From [74].
[b] See text.

loss of H_2O and the zwitterionic charges. Volume corrections of -12.3×10^{-3} nm^3 (7.4 ml/mol [3, 4]) or -13.9×10^{-3} nm^3 (8.4 ml/mol [5]) are used, which were derived from model densitometry experiments. The 1984 volumes are accordingly smaller than the 1972 volumes. One uncertainty of this approach thus lies in the value of the volume change of peptide formation to be used. Another lies in the experimental uncertainties in measuring $\bar{v}$ for the free amino acids (Methods); Zamyatnin [5] has also observed that V_i for the charged amino acids are dependent on the measurement conditions, unlike those for the uncharged amino acids.

In the approach based on summations of molar volumes, those for up to 12 distinct chemical groups are tabulated elsewhere [3–5]. More detailed analyses have been carried out by Richards [71] and Finney [72] based directly on the known crystal structures of ribonuclease S and lysozyme, from where volumes are assigned to 16 or 17 distinct groupings. In the present study, the 20 residue volumes of the amino acids were summed from molar volumes derived using methods A and B of Richards [71] and the radical plane and Voronoi methods [72]. Comparisons of the four resulting sets with densitometric and crystallographic-based data (Table 1) showed that the Richards' method B set gave volumes with the smallest deviations of the differences from the other compilations. The other three sets gave residue volumes that decreased in the order as just given. The main uncertainty of this method lies in the error ranges found in the crystal analyses; Richards [71] has pointed out that there are large since uncertainties in the assumed van-der-Waals radii that are required in this method are translated into their cubes at the level of volumes.

In the protein crystallography approach, Chothia [6] analysed 588 amino acid residues that are burried by 95% or more in the interior of proteins, using nine crystal structures where the coordinates had at least been submitted to a preliminary energy refinement. Using the Voronoi procedure [71], volumes were assigned directly to residue types. Less extensive compilations are reported elsewhere [73, 74]. Table 1 shows that the standard deviations are large, ranging over 4–8% of each volume. These variations might be reduced if the coordinates are derived from the most recent techniques of crystallographic refinement.

Finally, crystal structures of the free amino acids can be used to determine volumes. The present availability of sufficient crystal structure determinations, together with the use of structures with Gly (residue), H_2O, HCl and HBr moieties permit all 20 amino acid volumes to be determined by this means [1] (Methods). In Table 1, it is seen that the standard deviations of most volumes are less than 3% of the volume; in five cases, this ranges over 4–7%. The correction for peptide bond formation of -10.0×10^{-3} nm^3 is determined as the difference of the average of 13 and 15 volumes respectively of Gly as the amino acid and as the residue; it should be noted that the uncertainty in this correction can be as high as $(7.8 + 1.8) \times 10^{-3}$ nm^3 (Methods).

Cross comparisons of the six volume sets of Table 1 give the following results. For a given residue volume, the values vary in ranges of $9-29 \times 10^{-3}$ nm^3. The smallest volumes are generally those of Zamyatnin in 1984 [5], followed by the 1972 Zamyatnin [4] and the Cohn and Edsall [3] sets. The largest volumes are those from amino acid crystals (Methods), closely followed by the protein crystal volumes [6] within error. Linear regression analyses between pairs of compilations give similar good correlation coefficients of 0.96–1.00. The 1972 Zamyatnin [4], Chothia [6] and Methods sets tend to be better correlated with each other. The Richards [71] set is the least successful in these correlations as expected from its sensitivity to summation errors. These correlations show that the volume differences between the six sets can be well approximated by a constant difference applied equally to all 20 residues. For the 1972 Zamyatnin–Chothia pair, the mean difference is $10 (\pm 7) \times 10^{-3}$ nm^3; for the 1972 Zamyatnin–Methods pair, this is $10 (\pm 5) \times 10^{-3}$ nm^3.

Further comparisons in Table 1 are based on the subdivision into hydrophobic and hydrophilic subgroups. A slight dependence of the constant difference presumed above between the six volume sets is observed with this subdivision. For the 1972 Zamyatnin–Chothia and the 1984 Zamyatnin–Chothia pairs, the constant difference is $6-8 (\pm 5) \times 10^{-3}$ nm^3 for the hydrophobic subgroup. This is slightly greater at $12-15 (\pm 5) \times 10^{-3}$ nm^3 for the hydrophilic subgroup. An explanation can be proposed from the observation that the densitometric data correspond to solution measurements which include hydration effects, while the crystal data correspond to dry volumes that are unaffected by hydration. That there is a difference for all the free amino acids is consistent with the interaction of solvent with all 20 amino acids in solution at the zwitterionic centre. That this is larger for the hydrophilic subgroup corresponds to the additional interaction of solvent water with the polar sidechains. This is consistent with data on water hydration in protein crystal structures (below). These differences can be compared with a value of $22-33 \times 10^{-3}$ nm^3 for the electrostriction of solvent water at a fully charged group quoted by Cohn and Edsall [3]. Volume differences that correspond to specific electrostriction effects for the charged amino acids are not observed in Table 1. More precise analyses of electrostriction are, however, precluded by the errors inherent in each volume set, the magnitudes of which can be judged from the differences between the Zamyatnin 1972 and 1984 volumes, or the standard deviations of the crystallographic volumes reported in Table 1.

Comparisons of protein $\bar{v}$ values

The $\bar{v}$ values of 12 proteins [9] can be compared with $\bar{v}$ values calculated from the volume sets of Table 1 and the known amino acid sequences. The three comparisons of Fig. 1 are typical of the $\bar{v}$ values calculated from the six volume compilations. It is significant that the predicted and experimental values show similar good correlations in all cases. The six volume sets lead to correlation coefficients of 0.82–0.88 by linear regression. The largest differences between experiment and calculation are seen for ribonuclease A, lysozyme and carboxypeptidase A (Fig. 1). Since $\bar{v}$ determinations can be in error by ±0.01 ml/g (Methods), the possibilities of errors were examined. For ribonuclease A, lysozyme and carboxypeptidase A, other literature values of 0.691–0.707 ml/g, 0.703–0.722 ml/g and 0.723 ml/g, respectively, have been reported [12, 75], in comparison to the present values of 0.696 ml/g, 0.702 ml/g and 0.748 ml/g [9]. The $\bar{v}$ of lysozyme was redetermined using a Paar densitometer to be 0.708 ml/g in 0.15 M NaCl, 0.02 M acetate buffer, pH 4.7 at 20°C. This shows that while some difference is seen, the value of 0.702 ml/g is reproducible within error.

Improved correlation coefficients of 0.94–0.97 are obtained if ribonuclease A, lysozyme and carboxypeptidase A are excluded from the regression analyses. For the proteins remaining, the mean differences between calculated and experimental protein $\bar{v}$ values are given in Table 1. The Cohn-Edsall values gives surprisingly good agreement. The densitometric volumes underestimate $\bar{v}$ by 0.01–0.02 ml/g, in agreement with the analyses of Zamyatnin [4, 5]. The crystallographic volumes overestimate $\bar{v}$ by 0.04–0.05 ml/g, in agreement with other findings [71, 73, 75, 76]. The good and similar correlation coefficients (Fig. 1) show that agreement between the six compilations of Table 1 can be well approximated by applying (as above) a constant volume correction to the residue volumes of each set in Table 1.

A consensus set for amino acid residues is calculated in order to predict protein $\bar{v}$ values in solution. An average of the six compilations (Table 1), after correction by a term $\Delta V/\Sigma N$ to place the residue volumes on a common scale (Table 1), is taken to minimize the effect of any large errors in any given set where ΔV is the total volume difference in nm^3 between the experimental and calculated volumes using the 9 or 12 proteins above and ΣN is the total of amino acid residues (3174 or 3733) in these proteins. Similar $\Delta V/\Sigma N$ values were calculated from the use of 9 or 12 proteins. Since amino acid residues in a protein are either in or out of contact with solvent water, and since the densitometric and crystallographic volumes correspond to these two amino acid environments, the consensus volume set in effect takes an average of these two situations. The mean difference between the experimental and the consensus protein $\bar{v}$ values is 0.000 ± 0.005 ml/g as desired. The correlation coefficient for nine proteins is 0.96, indicating that no further improvement in the agreement between experiment and calculation can be obtained. Nonetheless, as seen from the $\Delta V/\Sigma N$ values, the consensus volumes can be used to reproduce protein $\bar{v}$ values in solution from accurate amino acid compositions more precisely than the sets from densitometry or crystallography sources. Table 1 shows furthermore than there is little

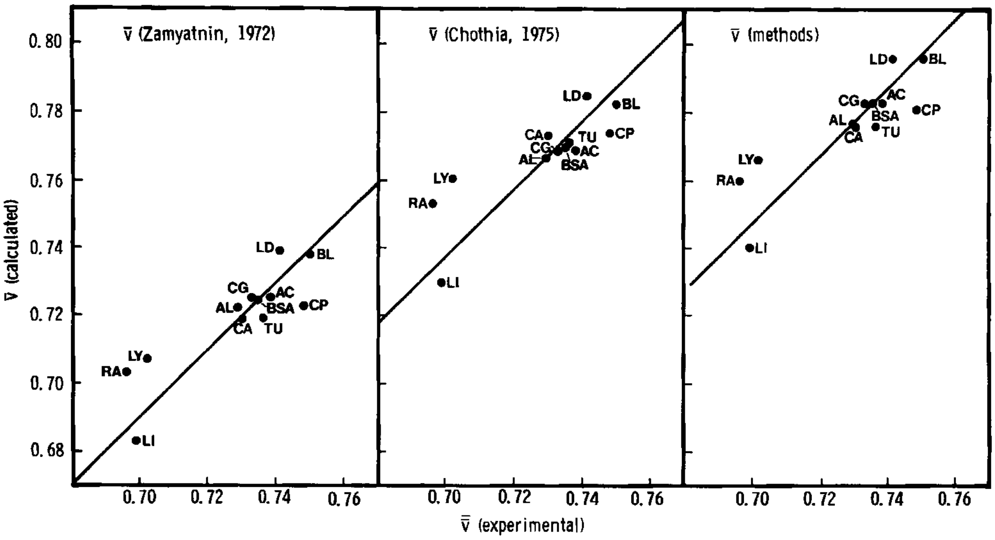

Fig. 1. *Comparison of experimental and calculated* v̄ *values for 12 proteins.* The experimental values are taken from Lee and Timasheff [9]. The protein are: RA, ribonuclease A; LI, lima bean trypsin inhibitor; LY, hen lysozyme; CA, catalase; AL, α-lactalbumin; CG, chymotrypsinogen A; BSA, bovine serum albumin; TU, calf brain tubulin; AC, α-chymotrypsin; LD, beef heart lactate dehydrogenase; CP, carboxypeptidase A; BL, β-lactoglobulin. Slopes of unit are shown in the comparisons with data calculated from Zamyatnin [4], Chothia [6] and this work (Methods). The linear regression coefficients are 0.88, 0.85 and 0.87, respectively, for all 12 proteins, which become 0.96, 0.95 and 0.97 if RA, LY and CP are excluded from the regressions. Specific volumes for metal atoms or heme moieties are not included in the calculation of v̄

difference between the Cohn and Edsall and the consensus sets.

In order to complete the comparisons of amino acid volumes with experimental data, a molecular explanation of the difference between the densitometric and crystallographic residue volumes is required. Residue volumes calculated from densitometry are $2.2-3.4\times10^{-3}$ nm^3/residue smaller than the average protein residue volume, while those from crystallography are $6.9-8.5\times10^{-3}$ nm^3/residue larger (Table 1). Total protein volumes have been described as the sum of three chief contributions, namely those from residue volumes as above, together with the imperfect packing of residues and solvation effects [4, 5]. Since similar volumes are found for amino acid residues and the free amino acids from crystallographic work (Table 1), it is concluded that the packing density of residue within proteins is similar to that of the free amino acid in crystals [6, 71, 73, 75–78]. Thus, packing defects are not of consequence in determining protein volumes. Protein solvation effects, however, can be used to account for the different volumes determined by densitometry and crystallography, where protein-water interactions have been reviewed by several authors [79–89].

It is shown that a volume difference exists between free water molecules and protein-bound water molecules. At 4 °C, the volume of a free water molecule is readily calculated as 29.9×10^{-3} nm^3, and this volume is marginally decreased to 29.5×10^{-3} nm at 50 °C. From this, the average distance between water molecules is calculated as 0.310 nm. The X-ray (and neutron) radial distribution curves of bulk water and their comparison with those of ice in its hexagonal or cubic forms (types I_h and I_c) provides further insights into the average separation between free water molecules. The ordinary ice structure is very open for reason of the tetragonal array of hydrogen-bonded oxygen atoms. Were it not for this hydrogen bonding, the ice structure could in principle be repacked to twice the density of ice I_h [85]. The radial distribution curve of ice reflects the high concentration of neighbours at 0.276 nm, 0.45–0.53 nm and 0.64–0.78 nm [81]. That curve for water is similar, reflecting a largely tetrahedral and open environment around each molecule as in ordinary ice I_h. Calculation of the difference radial distribution curve between ice and water shows however a substantial contribution at 0.35 nm for water [81, 85]. Within 0.37 nm, the first minimum of the radial distribution curve, there are about 5.5 neighbouring water molecules around each molecule, rather than 4 as in ice I_h. One early analysis of the water radial distribution curve [90] did so in terms of hydrogen-bonded framework and non-hydrogen-bonded interstitial water molecules. This framework water has one neighbour at a distance of 0.277 nm and three at 0.294 nm; interstitial water has three neighbours at each of 0.294-nm, 0.330-nm, 0.340-nm and 0.392-nm separation. Since the relative amounts of framework and interstitial water in the model is 4:1, the mean separation between water molecules is again calculated as 0.311 nm. More modern treatments based on Monte Carlo and molecular dynamics simulations [84, 85, 91] confirm these types of structures in the sense that they predict 5.3–5.7 nearest neighbours, depending on the simulation and the intermolecular potential [87]. In conclusion, even though hydrogen-bond distances are maintained at 0.29 nm in free water, the average distance between the water molecules is 0.31 nm, corresponding to a volume of about 30×10^{-3} nm^3/molecule.

Protein-bound water molecules are now considered. Analyses of the volume of bound water in small-molecule crystal structures of hydrated carbohydrates, amino acids and inorganic salts shows that the average values lie in a range of $23.6-25.8\times10^{-3}$ nm^3/molecule [1, 2] (Methods). This is lower than that for free water in reflection of the different nature of the intermolecular interaction within the crystallographic unit cell. It is of interest that, while water in ordinary ice I_h and I_c has a larger volume of 32.3×10^{-3} nm^3/molecule, the average volume found in seven other forms of ice (stable under high pressure) is also 24.6×10^{-3} nm^3/molecule [2, 81].

Table 2. *Crystallographically observed water molecules on protein surfaces*
The number of water molecules observed on the protein surfaces are given, followed (in parentheses) by the mass ratio of water/protein

Protein	Residues	Observed ordered water in 'first shell'	Total observed water	Predicted bound water
Bovine pancreatic trypsin inhibitor [92, 93]	58	43	63 (0.17)	76–93 (0.21–0.26)
Rubredoxin [94]	54	80	123 (0.37)	68–83 (0.20–0.25)
Erythrocrurin [95]	136	111	=	171–223 (0.21–0.27)
Actinidin [96]	220	163	272 (0.20)	289–326 (0.22–0.25)
Penicillopepsin [97]	323	264	319 (0.17)	440–528 (0.24–0.28)
Lysozyme (human) [98]	130	95	140 (0.17)	178–188 (0.22–0.23)
Lysozyme (tortoise) [98]	130	90	128 (0.16)	162–195 (0.20–0.24)

The average distance between water molecules is slightly less at 0.290 nm than in free water. Relatively small average distance changes of 0.02 nm are thus involved in the volume reduction of 5.4×10^{-3} nm^3 since distance-cubed terms are involved. Sufficiently detailed protein crystallographic studies of protein-water interfaces show the existence of well-defined water molecule positions [82, 92–98] (Table 2). Ordered hydrogen-bonded waters are found at surface main-chain peptide CO and NH groups (60–62% of the total in lysozyme) and the remainder are hydrogen-bonded to the side-chains of mostly polar residues. Histogram analyses of protein-water distances in lysozyme and rubredoxin peak at 0.28–0.30 nm [98] and 0.25–0.30 nm [94]. For lysozyme, the mean H_2O–protein O distance is 0.282 ± 0.015 nm and the mean H_2O–protein N distance is 0.296 ± 0.015 nm. These distances are as expected for typical hydrogen-bonding interactions and few cases are found where there are larger separations of the order of 0.35 nm, as noted above [81, 85] for free water. Thus, when water is hydrogen-bonded to a protein surface or within a small-molecule crystal structure, the apparent water molecule volume is reduced through electrostriction and causes the protein volume to decrease.

The effect of the reduced water molecule volume on the total protein volume can be estimated from the number of ordered water molecules seen in the protein crystal analyses (Table 2). The volume difference between the Chothia and Methods compilations and the consensus compilation is calculated for each protein. Using 29.9×10^{-3} nm^3 and 24.5×10^{-3} nm^3 for the volumes of free and bound water molecules, these volume differences are expressed as the equivalent total of bound water required for the decrease in the crystallographic volume (Table 2). The ensuing comparison in Table 2 shows that, with the exception of rubredoxin, the amount of crystallographically observed ordered water in the first hydration shell corresponds to $53 \pm 3\%$ of the volume difference between the consensus and the crystallographic volumes of the proteins. Except for rubredoxin again, the total of observed ordered water corresponds to about 74% of the total difference between the consensus and crystallographic volumes. The discrepancy seen with rubredoxin is attributed to the low water content of the crystal which induces a higher ordering of the water molecules in the crystal. Disordered surface polar sidechains and other residues cannot be analysed in this way [9] and several water molecules internal to the protein have been neglected. Thus, these summations have underestimated the effect of the total bound water. Corrections which allow for this using Table 3 of [95] lead to the expected total of water molecules required for the volume difference.

Since protein hydration shells involve specific water-protein hydrogen bonding in well-defined locations, as opposed to the loose and strong hydrogen-bond interactions of bulk water, this leads to the observation that the apparent hydrated volume of a protein as measured by densitometry is less than the crystallographic volume of the protein. Since different proteins may be associated with different degrees of hydration, the prediction of $\bar{v}$ from accurate composition data is subject to this uncertainty. Interestingly, Bull [99], using equilibrium dialysis with a sucrose hydration probe, has reported larger experimental protein hydrations for ribonuclease A and lysozyme compared to myoglobin, β-lactoglobulin and bovine serum albumin. These larger hydrations are compatible with the reduced experimental values of $\bar{v}$ for ribonuclease A and lysozyme (Fig. 1) noted above. In general however, even though the $\bar{v}$ measurements are performed in dilute buffers and correspond therefore to a two-component system, the possibility of effects arising from the multicomponent nature of protein solutions in mixed buffer solvents cannot be completely ruled out. These might contribute to the $\bar{v}$ discrepancies for ribonuclease A and lysozyme noted above in addition to the volume change associated with protein-bound water.

Only crystallographic amino acid volumes have been considered above. The densitometric amino acid volumes (Table 1) can be interpreted by applying the above arguments in reverse, where the burial of amino acid residues within the folded protein eliminates water molecules bound to the free amino acids and in turn causes the apparent calculated protein volume to be increased relative to that of the experimental value.

Comparisons of carbohydrate and glycoprotein volumes

Two full and two partial compilations of carbohydrate residue volumes are available. A hypothetical set was calculated by Gibbons [100] from the molar group volumes of Cohn and Edsall [3] and these are given in Table3. Monosaccharide crystal structures lead to a second full compilation (Methods). Comparison of these two volume sets (Table 3) shows that they are consistent, apart from a lower volume for sialic acid in the first set where a hypothetical electrostriction correction of -22×10^{-3} nm^3 had been applied. A third, partial, compilation based on polysaccharide crystal structures gave values for Glc and GlcNAc residues, which could be extended to Gal, Man and GalNAc residues (Table 3) [1]. Densitometry experiments with the free Glc, Gal and Man monosaccharides lead to a fourth, partial, set of volumes [7], which for completion was extended to the other four residues

Table 3. *Carbohydrate residue volumes*
Bracketted volumes are estimated (see text)

Source	M_r	Molar volume summations [3, 100]	Monosaccharide crystal structures (Methods)	Polysaccharide crystal structures [1]	Densitometry [7] (Methods)
		$\times 10^{-3}$ nm^3			
Glc	162	164.9	171.9 ± 2.0 (9)	167.3 ± 2.7 (11)	162.7 ± 0.7
Gal	162	164.9	166.8 ± 2.5 (9)	(167.3)	162.2 ± 0.5
Man	162	164.9	170.8 ± 0.4 (3)	163.3 (1)	161.9 ± 0.8
GlcNAc	203	224.5	222.0 ± 2.0 (2)	230.4 ± 3.4 (4)	(216.6)
GalNAc	203	224.5	232.9 (1)	(230.4)	(227.5)
Fuc[a]	146	164.3	160.8 ± 2.9 (2)	=	(155.4)
NeuNAc[a]	290	281.2	326.3 ± 8.8 (2)	=	(320.9)

[a] Terminal positions only.

Table 4. *Comparison of calculated and experimental* $\bar{v}$ *values for glycoproteins*

Glycoprotein	Carbohydrate	Consensus $\bar{v}$ calculation	Experimental $\bar{v}$
	% (w/w)	ml/g	
α_1 Acid glycoprotein [29, 30]	43.1	0.697	0.704 [22]
α_2-Macroglobulin [31, 32]	9.9	0.731	0.739 [23] 0.720, 0.735 [24]
IgM GAL [33–36]	7.8	0.724	0.724 [25]
C3 of complement [27, 37]	4.4	0.737	0.736 [26] 0.73 [27]
IgG3 (human) [38–40]	1.7	0.728 (MOPC21) 0.735 (KOL)	0.725 [28]

by correction of the crystal volumes by -5.4×10^{-3} nm^3 (Methods; Table 3).

Densitometric glycoprotein $\bar{v}$ values are summarized in Table 4. These are compared with $\bar{v}$ values that are calculated from anino acid and carbohydrate compositions or sequences, the consensus amino acid volumes (Table 1) and the carbohydrate volumes of Table 3. Use of the monosaccharide crystal volumes (Table 3) gave the best agreement between experiment and calculation as shown in Table 4. This result shows that hydration effects do not influence apparent carbohydrate volumes in solution as much as amino acid residues are. This is consistent with the reduced number of charged groups and NH · CO peptide moieties in bound carbohydrate chains, since it is these that constitute the principal water binding sites in proteins [98].

Comparisons of calculated and experimental matchpoints

Neutron scattering matchpoints are dependent on scattering length densities, i.e. the total of scattering lengths divided by the total partial volume (Methods). Experimental data for 11 proteins and glycoproteins are compared in Table 5 with predicted matchpoints calculated from volumes based on densitometry, the consensus and crystallography (Tables 1 and 3). These comparisons show that the crystallographic volumes give the best agreement with experiment, where the average difference from the experimental values is $-0.2 \pm 0.7\%$ ^2H_2O. Use of the densitometric volumes gives matchpoints that are higher by $2.6 \pm 1.0\%$ ^2H_2O, while those of the consensus volumes also gives matchpoints that are higher by $1.8 \pm 1.0\%$ ^2H_2O.

Matchpoint calculations are sensitive to assumptions on the exchangeable proton content of the macromolecule since neutron scattering properties are strongly dependent on the ^{1}H nucleus and these are now explored. Hydrogen exchange has been reviewed elsewhere [102, 102]. Exchangeable protons on O and N atoms exchange freely with $H_2O/^2H_2O$ solvent except for those buried within stable secondary structures within the macromolecule. The latter are usefully considered as peptide NH protons. If as hypothesis, the discrepancy between the experimental and calculated matchpoint is attributed in full to the nonexchange of main-chain peptide NH protons, nonexchanged NH peptide levels of 60 ± 15%, 45 ± 15% and 5 ± 10% are required to obtain agreement between the experimental matchpoints and the densitometric, consensus and crystallographic volumes, respectively. These proton contents can be compared with determinations of non-exchanged protons by solution ^{1}H nuclear magnetic resonance (NMR) studies and by neutron protein crystallography. Typically after extensive dialysis or exchange, ^{1}H NMR gives 8% of nonexchange for hyaluronate binding region of proteoglycans, 10 ± 2% for bovine trypsin, 12 ± 4% for α_1 acid glycoprotein, 20% for bovine pancreatic trypsin inhibitor, 24% for hen lysozyme [1, 44, 103, 104] (and my unpublished result). Neutron protein crystallography gives higher extents of non-exchange of 19% (bovine pancreatic ribonuclease A), 20% (bovine pancreatic trypsin inhibitor), 28% (bovine trypsin), 29% (sperm whale oxymyoglobin) and 34% (hen lysozyme) [10, 93, 105–107]. ^{1}H-^{2}H exchange is much slower in the crystal state than in the solution state [93, 104]. Using ^{1}H NMR, 5 of 17 NH proton signals that are assigned to specific non-exchanged NH protons seen in the lysozyme crystal were found to be exchangeable with solvent in solution [107]. Thus, in solution the nonexchanged main-chain proton contents of 19–34% by neutron crystallography will be significantly reduced to values in the range of 8–24% by ^{1}H NMR. The latter corresponds well to the neutron matchpoints calculated using the crystallographic volumes of proteins and glycoproteins and what is seen to be a reasonable estimate of 10% nonexchange of the main-chain peptide NH protons. Arbitrary estimates of the nonexchange of 25% of all labile hydrogens in proteins (which is equivalent to 55% nonexchange of the main-chain NH protons) by other neutron workers, such as [45], are too high, even though these permit similar partial specific volumes to be used for both densitometry and matchpoints. It thus follows from the present

Table 5. *Comparison of calculated and experimental neutron scattering matchpoints for proteins and glycoproteins*
These are arranged in order of diminishing matchpoint calculated from the crystallographic volumes

Macromolecule	Carbohydrate	Predicted matchpoint (assuming 10% nonexchange of the main chain NH protons)			Experimental matchpoint [6] (Methods) value (no. of intensities measured in number of contrasts)
		densitometric [4, 7]	consensus (Results)	crystallographic [6] (Table 3)	
	%				
Lysozyme	0	48.0	47.0	44.4	44.7 (10 in 10) [43]
α_1 Acid glycoprotein	43	46.5	45.5	44.0	44.7 (21 in 6) [44]
Ribonuclease A	0	47.0	46.2	43.7	42.5 (4 in 4) [45]
Fibrinogen	≈0	46.7	45.8	43.3	42.5 (12 in 12) [46]
$\mathrm{Clr_2Cls_2}$	7	45.2	44.4	42.3	43 (3 in 3) [47]
Cl	7	44.7	43.9	41.8	43 (9 in 5) [48]
Clq	7	44.2	43.4	41.4	41.5 (11 in 6) [48]
C3	4	42.9	42.1	40.3	40.4 (36 in 4) [27]
C3c	6	42.9	42.1	40.3	40.7 (11 in 4) [27]
C3dg	0	42.3	41.6	39.7	39.3 (9 in 3) [27]
Myoglobin	0	41.7	41.1	39.4	40.2 (10 in 10) [49]

analyses that the volumetric and neutron techniques lead to different views of partial specific volumes.

A complete analysis of neutron matchpoints requires a molecular consideration of the hydration shell that surrounds the macromolecule. The classical view of volumes as visualised by neutron scattering is that these correspond to the dry volume, i.e. that volume which is inaccessible to solvent water molecules [41, 42]. The results above indicate that this dry volume corresponds well with the crystallographic volume and that the known existence of a hydration shell has been negated by reason of the neutron technique. Calculations of neutron scattering matchpoints based on an added hydration shell equivalent to the number of water molecules reflecting the difference between the consensus and Chothia volumes (as above) are sensitive to assumptions relating to the two water protons. It is of interest that consideration of this protein-bound water in terms of OH moieties and not H_2O moieties in the reduced water volume of 24.5×10^{-3} nm^3 (compared to 29.9×10^{-3} nm^3 as in free water) gives protein and glycoprotein matchpoints close to those from the crystallographic volumes. This concept is compatible with the location of the bound water-oxygen by protein X-ray crystallography (and, by inference, its hydrogen-bonded proton). In consequence the second water proton that is not in contact with the protein is associated with a volume and position uncertainty in as far as the neutron experiment is concerned, and this means that the effective protein-bound water volume is close to 29.9×10^{-3} nm^3 and is not readily distinguishable from free water. In distinction, the densitometric observations of protein $\bar{v}$ values are dominated by the water oxygen atom, since this is 89% of the mass of water. The different observations of volumes by neutron matchpoints and densitometry might thus be accounted for by the property of the two techniques to examine preferentially in that order the hydrogen or the oxygen atom of the hydration shell.

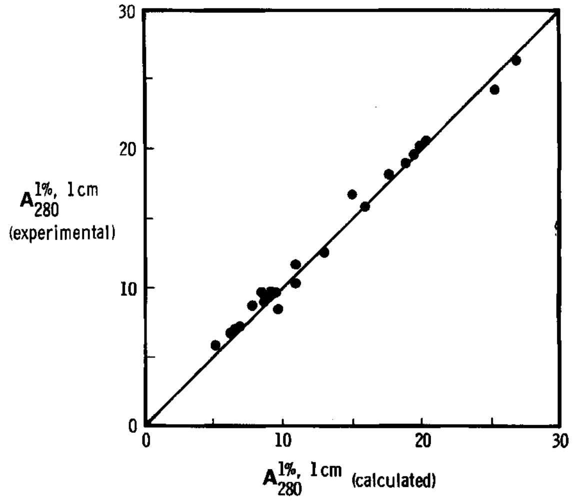

Fig. 2. *Comparison of experimental and calculated absorption coefficients for 23 proteins and glycoproteins of known primary sequence.* The data are compared with a line of slope equal to unity. In order of increasing experimental absorption coefficients these are: aspartate transcarbamylase (5.9), bovine serum albumin (6.8), complement component Clq (6.8), ribonuclease A (7.1), bovine pancreatic trypsin inhibitor (8.2), α_2-macroglobulin (8.6), α_1 acid glycoprotein (8.93), complement component C4 (9.2), subtilsin Carlsberg (9.6), β-lactoglobulin (9.6), complement C3 (9.7), insulin (10.3), subtilsin BPN (11.7), complement factor B (12.7), trypsin (15.9), fibrinogen (16.8), serum amyloid P component (18.2), carboxypeptidase A (19.0), C-reactive protein (19.5), chymotrypsinogen A (20.2), α-lactalbumin (20.5), human lysozyme (24.2), hen lysozyme (26.3)

Comparisons of calculated and experimental absorption coefficients

The calculation of absorption coefficients requires accurate M_r values together with accurate Trp, Tyr and Cys contents and their molar absorption coefficients [53] (Methods). Experimental absorption coefficients $A_{280}^{1\%\ 1\mathrm{cm}}$ or A_{exp} are compared in Fig. 2 with the calculated values A_{cal} for 23 proteins and glycoproteins. The Trp, Tyr and Cys contents of these macromolecules provide a representative survey of absorption coefficients. Thus, the values of A_{exp} range over 5–27. Of the 23 macromolecules (which includes eight glycoproteins), three have no Trp residues, two have no Cys residues, two have very high relative Trp contents, and two have very high relative Cys contents. The Tyr residues are more evenly distributed. A good linear relationship between A_{exp} and A_{cal} is observed, where linear regression analysis gives a correlation coefficient of 0.994 and a best straight line with a slope close to unity:

$$A_{\mathrm{exp}} = 0.960\ A_{\mathrm{cal}} + 0.740.$$

The average difference ($A_{exp} - A_{cal}$) is 0.2 ± 0.7, with the greatest differences being seen for bovine pancreatic trypsin inhibitor (−1.5) and human fibrinogen (1.6) When A_{exp} is less than 10, which corresponds to relatively low Trp contents, the difference ($A_{exp} - A_{cal}$) is slightly increased at 0.5 ± 0.3. conclusion, A_{cal} values from sequence data are able to reproduce A_{exp} values within errors of ±0.7, and the precision of the calculation is improved for macromolecules with high Trp contents. Since the ratio of A_{exp} to A_{cal} is on average 1.03 ± 0.07 (23 values), predicted A_{cal} values should be increased by this factor. Earlier work using six glycoproteins has suggested a factor of 1.06 ± 0.03 [54].

The applicability of the calculations to other macromolecules whose sequences are unknown was found to be less precise. For example, $C\bar{1}$ inhibitor of complement has A_{exp} values of 3.6−4.5 [108, 109], while A_{cal} on the basis of amino acid analyses are 6.1−6.4; this difference can be attributed to errors in the determination of Trp contents. In certain cases, errors in A_{exp} can be large. The C4b binding protein of complement was inferred to have A_{exp} of 9.3, but A_{cal} was determined to be 14.1 from the sequence, and use of the latter determination lead to a much improved molecular mass determination by small-angle X-ray scattering [54, 110]. Finally it is noted that A_{exp} can depend on the conditions of measurement; that for bovine serum albumin was reduced by 0.5 on addition of 3 M guanidine hydrochloride [111]. Large variation of pH can sometimes cause changes in A_{exp} of up to 0.3 [112].

CONCLUSIONS

Two distinct approaches have been employed to calculate protein volumes starting from accurate amino acid compositions. The first approach is based on densitometry of the free amino acids in solution. This yields residue volumes that have been reduced compared to the physical residue volume for reason of bound water molecules of effective volumes 24.5×10^{-3} nm^3/molecule (in comparision to the free water molecule volume of 29.9×10^{-3} nm^3/molecule). The formation of a globular protein structure is concurrent with the removal of part of this bound water from the individual amino acids. Thus the Zamyatnin 1972 and 1984 volumes give predicted protein $\bar{v}$ values that are too low. The second approach is based on the use of directly observed crystallographic volumes (using either crystallographic molar group summations, protein crystal structures or amino acid crystal structures). These correspond to the physical residue volume. Since no allowance for bound water has been included, the predicted protein $\bar{v}$ values are too high. In order to predict $\bar{v}$ values for typical dilute proteins surrounded by bound water molecules in dilute buffer solutions as seen by densitometry, a consensus volume set was derived from the six available compilations. For this, the V_i of each compilation was first corrected by amounts $\Delta V/\Sigma N$ to correspond to the deviation between the calculated $\bar{v}$ from each compilation and the experimental $\bar{v}$ for well-characterized standard proteins. The ensuing consensus volume set is fortituously close to the classical 1943 Cohn-Edsall volumes (corrected for Cys; Table 1) in its ability to calculate protein $\bar{v}$ in solution; in conclusion either set can be used. In terms of molecular structures, where specifically bound water molecules have been directly observed in recent refined protein structures, the difference between the crystallographic and consensus volumes corresponds well to the total number of observed bound water molecules in the protein crystal structure. This is confirmatory evidence of a general electrostrictive effect of the protein hydration shell on the protein volume in solution as measured by densitometry.

The calculation of protein and glycoprotein neutron scattering matchpoints is sensitive to the main-chain NH protons that do not exchange as the $H_2O/^2H_2O$ ratio is varied. Comparisons based on protein NMR and neutron protein crystallography indicate that in solution 8−24% of these NH protons do not exchange (depending on the protein). Comparisons of the predicted matchpoints from the three volume sets (densitometric, consensus, crystallographic) shows that the use of the crystallographic volumes gives the best match with both the estimated nonexchange of peptide protons and the experimental matchpoints. In conclusion, neutron matchpoints are well predicted from the Chothia crystallographic volumes assuming 10% of mainchain NH nonexchange (i.e. 5% of all exchangeable protons). In relation to molecular structures, the hydration shell of the protein apparently has the same effective volume as that of bulk water. Consequently the so-called 'dry' volume of neutron scattering corresponds empirically to the physical macromolecular volume, even though a multicomponent system is under consideration. That densitometry and matchpoints lead to different views of $\bar{v}$ might be due to the different emphasis of each method on the observation of the oxygen atom and the hydrogen atom in hydration shells, respectively.

The above comparisons have been made using experimental densitometric and matchpoint data which, strictly speaking, relate to multicomponent system. These are not necessarily equivalent to very dilute protein or glycoprotein solutions in dilute buffers, i.e. to the pure macromolecule at infinite dilution in pure water. It is possible that a careful reinvestigation of this aspect using a more detailed formulism for multicomponent systems [8, 113] might resolve some of the discrepancies noted in the course of this work.

The comparison of experimental and calculated 280-nm absorption coefficients on the basis of accurate amino acid compositions is shown to be a useful quantitative procedure (Fig. 2). Predicted absorption coefficients should be increased by a factor of 1.03 and the calculation, on average, reproduced experimental values to ±0.7 (1% solutions, 1-cm pathlength). It is useful to calculate these as a control from sequences for comparison with literature values.

S.J.P. acknowledges financial support from a Lister Institute of Preventive Medicine Fellowship, the Science and Engineering Research Council (UK) and the Medical Research Council (UK), and thanks Professor Sir David Phillips and Dr P. J. Artymiuk for useful discussions.

REFERENCES

1. Perkins, S. J., Miller, A., Hardingham, T. E. & Muir, H. (1981) *J. Mol. Biol. 150*, 69−95.
2. Leclaire, A. & Monier, J. C. (1982) *Acta Crystallogr. B38*, 724−727.
3. Cohn, E. J. & Edsall, J. T. (1943) in *Proteins, amino acids and peptides*, pp. 155−176 & 370−381, Reinhold Publ. Corp., New York.
4. Zamyatnin, A. A. (1972) *Progr. Biophys. Mol. Biol. 24*, 109−123.
5. Zamyatnin, A. A. (1984) *Annu. Rev. Biophys. Bioeng. 13*, 145−165.
6. Chothia, C. (1975) *Nature (Lond.) 254*, 304−308.
7. Shahidi, F., Farrell, P. G. & Edward, J. T. (1976) *J. Solution Chem. 5*, 807−816.

8. Casassa, E. F. & Eisenberg, H. (1964) *Adv. Protein Chem. 19*, 287–395.
9. Lee, J. C. & Timasheff, S. N. (1974) *Biochemistry 13*, 257–265.
10. Wlodawer, A. & Sjölin, L. (1983) *Biochemistry 22*, 2720–2728.
11. Kassell, B. (1970) *Methods Enzymol. 19*, 862–871.
12. Sober, H. A. (ed.) (1970) *Handbook of biochemistry*, section C, 2nd edn, The Chemical Rubber Co., Cleveland, OH.
13. Schroeder, W. A., Shelton, J. R., Shelton, J. B., Robberson, B., Apell, G., Fang, R. S. & Bonaventura, J. (1982) *Arch. Biochem. Biophys. 214*, 397–421.
14. Brew, K., Castellino, F. J., Vanaman, T. C. & Hill, R. L. (1970) *J. Biol. Chem. 245*, 4570–4582.
15. Wilcox, P. E. (1970) *Methods Enzymol. 19*, 64–108.
16. Reed, R. G., Putnam, F. W. & Peters, T. (1980) *Biochem. J. 191*, 867–868.
17. Valenzuela, P., Quiroga, M., Zaldivar, J., Rutter, W. J., Kirschner, M. W. & Cleveland, D. W. (1981) *Nature (Lond.) 289*, 650–655.
18. Postingl, H., Krauhs, E., Little, M. & Kempf, T. (1981) *Proc. Natl Acad. Sci. USA 78*, 2757–2761.
19. Eventoff, W., Rossmann, M. G., Taylor, S. S., Torff, H. J., Meyer, H., Keil, W. & Kiltz, H. H. (1977) *Proc. Natl Acad. Sci. USA 74*, 2677–2681.
20. Petra, P. H. (1970) *Methods Enzymol. 19*, 460–503.
21. Pervaiz, S. & Brew, K. (1985) *Science (Wash. DC) 228*, 335–337.
22. Kawahara, K., Ikenaka, T., Nimberg, R. B. & Schmid, K. (1973) *Biochim. Biophys. Acta 295*, 505–513.
23. Branegård, B., Österberg, R. & Sjöberg, B. (1980) *Int. J. Biol. Macromol. 2*, 321–323.
24. Barrett, A. J. (1981) *Methods Enzymol. 80*, 737–754.
25. Wilhelm, P., Pilz, I., Goral, K. & Palm, W. (1980) *Int. J. Biol. Macromol. 2*, 13–16.
26. Tack, B. F. & Prahl, J. W. (1976) *Biochemistry 15*, 4513–4521.
27. Perkins, S. J. & Sim, R. B. (1986) *Eur. J. Biochem. 157*, 155–168.
28. Sjöberg, B., Rosenquist, E., Michaelsen, T., Pap, S. & Österberg, R. (1980) *Biochim. Biophys. Acta 625*, 10–17.
29. Schmid, K., Kaufmann, H., Isemura, S., Bauer, F., Emura, J., Motoyama, T., Ishiguro, M. & Nanno, S. (1973) *Biochemistry 12*, 2711–2734.
30. Yoshima, H., Matsumoto, A., Mizuochi, T., Kawasaki, T. & Kobaka, A. (1981) *J. Biol. Chem. 256*, 8476–8484.
31. Sottrup-Jensen, L., Stepanik, T. M., Kristensen, T., Wierzbicki, D. M., Jones, C. M., Lonblad, P. B., Magnusson, S. & Petersen, T. E. (1984) *J. Biol. Chem. 259*, 8318–8327.
32. Dunn, J. T. & Spiro, R. G. (1967) *J. Biol. Chem. 242*, 5556–5563.
33. Laure, C. J., Watanabe, S. & Hilschmann, N. (1973) *Hoppe-Seyler's Z. Physiol. Chem. 354*, 1503–1504.
34. Watanabe, S., Barnikol, H. U., Horn, J., Bertram, J. & Hilschmann, N. (1973) *Hoppe-Seyler's Z. Physiol. Chem. 354*, 1505–1509.
35. Shimizu, A., Putnam, F. W., Paul, C., Clamp, J. R. & Johnson, I. (1971) *Nat. New Biol. 231*, 73–76.
36. Niedermeier, W., Tomana, M. & Mestecky, J. (1972) *Biochim. Biophys. Acta 257*, 527–530.
37. de Bruijn, M. H. L. & Fey, G. H. (1985) *Proc. Natl Acad. Sci. USA 82*, 708–712.
38. Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M. & Perry, H. (1983) *Sequences of proteins of immunological interest*, US Dept of Health & Human Services, Public Health Service, National Institute of Health.
39. Schmidt, W. E., Jung, H. D., Palm, W. & Hilschmann, N. (1983) *Hoppe-Seyler's Z. Physiol. Chem. 364*, 713–747.
40. Rademacher, T. W., Homans, S. W., Fernandes, D. L., Dwek, R. A., Mizuochi, T., Taniguchi, T. & Kobata, A. (1983) *Biochem. Soc. Trans. 11*, 132–134.
41. Jacrot, B. (1976) *Rep. Progr. Phys. 39*, 911–953.
42. Stuhrmann, H. B. & Miller, A. (1978) *J. Appl. Cryst. 11*, 325–345.
43. Stuhrmann, H. B. & Fuess, H. (1976) *Acta Crystallogr. A32*, 67–74.
44. Li, Z. Q., Perkins, S. J. & Loucheux-Lefebvre, M. H. (1983) *Eur. J. Biochem. 130*, 275–279.
45. Lehmann, M. S. & Zaccai, G. (1984) *Biochemistry 23*, 1939–1942.
46. Marguerie, G. & Stuhrmann, H. B. (1976) *J. Mol. Biol. 102*, 143–156.
47. Boyd, J., Burton, D. R., Perkins, S. J., Villiers, C. L., Dwek, R. A. & Arlaud, G. J. (1983) *Proc. Natl Acad. Sci. USA 80*, 3769–3773.
48. Perkins, S. J., Villiers, C. L., Arlaud, C. J., Boyd, J., Burton, D. R., Colomb, M. G. & Dwek, R. A. (1984) *J. Mol. Biol. 179*, 547–557.
49. Ibel, K. & Stuhrmann, H. B. (1975) *J. Mol. Biol. 93*, 255–265.
50. Doolittle, R. F., Watt, K. W. K., Cottrell, B. A., Strong, D. D. & Riley, M. (1979) *Nature (Lond.) 280*, 464–468.
51. Reid, K. B. M. (1979) *Biochem. J. 179*, 367–371.
52. Reid, K. B. M., Gagnon, J. & Frampton, J. (1982) *Biochem. J. 203*, 559–569.
53. Wetlaufer, D. B. (1962) *Adv. Protein Chem. 17*, 303–390.
54. Perkins, S. J., Chung, L. P. & Reid, K. B. M. (1986) *Biochem. J.*, in the press.
55. Ottesen, M. & Svendsen, I. (1970) *Methods Enzymol. 19*, 199–215.
56. Wood, H. F. & McCarty, M. (1951) *J. Clin. Invest. 30*, 616–622.
57. Haupt, H., Heimburger, N., Kranz, T. & Baudner, S. (1972) *Hoppe-Seyler's Z. Physiol. Chem. 353*, 1841–1849.
58. Konigsberg, W. H. & Henderson, L. (1983) *Proc. Natl Acad. Sci. USA 80*, 2467–2471.
59. Pulford, W. C. A. (1982) D. Phil. Thesis, University of Oxford.
60. Artymiuk, P. J. (1979) D. Phil. Thesis, University of Oxford.
61. Belt, K. T., Carroll, M. C. & Porter, R. R. (1984) *Cell 36*, 907–917.
62. Mole, J. E., Anderson, J. K., Davison, E. A. & Woods, D. E. (1984) *J. Biol. Chem. 259*, 3407–3412.
63. Chan, A. C. & Atkinson, J. P. (1985) *J. Immunol. 134*, 1790–1798.
64. Lei, J. K., Liu, T., Zan, G., Soravia, E., Liu, J. Y. & Goldman, N. D. (1985) *J. Biol. Chem. 260*, 13377–13383.
65. Woo, P., Korenberg, J. R. & Whitehead, A. S. (1985) *J. Biol. Chem. 260*, 13384–13388.
66. Mantzouranis, E. C., Dowton, S. B., Whitehead, A. S., Edge, M. D., Bruns, G. A. P. & Colten, H. R. (1985) *J. Biol. Chem. 260*, 7752–7756.
67. Baltz, M. L., de Beer, F. C., Feinstein, A., Munn, E. A., Milstein, C. P., Fletcher, T. C., March, J. F., Taylor, J., Bruton, C., Clamp, J. R., Davies, A. J. S. & Pepys, M. B. (1982) *Ann. N.Y. Acad. Sci. 389*, 49–75.
68. Eisenberg, D. (1984) *Annu. Rev. Biochem. 53*, 595–623.
69. McMeekin, T. L., Groves, M. L. & Hipp, N. J. (1949) *J. Am. Chem. Soc. 71*, 3298–3300.
70. McMeekin, T. L. & Marshall, K. (1952) *Science (Wash. DC) 116*, 142–143.
71. Richards, F. M. (1974) *J. Mol. Biol. 82*, 1–14.
72. Gellatly, B. J. & Finney, J. L. (1982) *J. Mol. Biol. 161*, 305–322.
73. Finney, J. L. (1975) *J. Mol. Biol. 96*, 721–732.
74. Chothia, C. & Janin, J. (1975) *Nature (Lond.) 256*, 705–708.
75. Liquori, A. M. & Sadun, C. (1981) *Int. J. Biol. Macromol. 3*, 56–59.
76. Chothia, C. (1984) *Annu. Rev. Biochem. 53*, 537–572.
77. Klapper, M. H. (1971) *Biochim. Biophys. Acta 229*, 557–566.
78. Richards, F. M. (1979) *Carlsberg. Res. Commun. 44*, 47–63.
79. Kuntz, I. D. & Kauzmann, W. (1974) *Adv. Protein. Chem. 28*, 239–345.
80. Cooke, R. & Kuntz, I. D. (1974) *Annu. Rev. Biophys. Bioeng. 3*, 95–126.
81. Eisenberg, D. & Kauzmann, W. (1969) *The structure and properties of water*, Clarendon Press, Oxford.
82. Finney, J. L. (1977) *Phil. Trans. R. Soc. Lond. B278*, 3–32.
83. Finney, J. L. (1979) in *Water: a comprehensive treatise* (Franks, F., ed.) vol. 6, pp. 47–122, Plenum Press.
84. Stillinger, F. H. (1980) *Science (Wash. DC) 209*, 451–457.

85. Edsall, J. T. & McKenzie, H. A. (1978) *Adv. Biophys. 10*, 137–208.
86. Edsall, J. T. & McKenzie, H. A. (1983) *Adv. Biophys. 15*, 53–183.
87. Nemethy, G., Peer, W. J. & Scheraga, H. A. (1981) *Annu. Rev. Biophys. Bioeng. 10*, 459–497.
88. Hvidt, A. (1983) *Annu. Rev. Biophys. Bioeng. 12*, 1–20.
89. Rupley, J. A., Gratton, E. & Careri, G. (1983) *Trends Biochem. Sci. 8*, 18–22.
90. Danford, M. D. & Levy, H. A. (1962) *J. Am. Chem. Soc. 84*, 3965–3966.
91. Narten, A. H. & Levy, H. A. (1969) *Science (Wash. DC) 165*, 447–454.
92. Deisenhofer, J. & Steigemann, W. (1975) *Acta Crystallogr. B31*, 238–250.
93. Wlodawer, A., Walter, J., Huber, R. & Sjölin, L. (1984) *J. Mol. Biol. 180*, 301–329.
94. Watenpaugh, K. D., Margulis, T. N., Sieker, L. C. & Jensen, L. H. (1978) *J. Mol. Biol. 122*, 175–190.
95. Steigemann, W. & Weber, E. (1979) *J. Mol. Biol. 127*, 309–338.
96. Baker, E. N. (1980) *J. Mol. Biol. 141*, 441–484.
97. James, M. N. G. & Sielecki, A. R. (1983) *J. Mol. Biol. 163*, 299–361.
98. Blake, C. C. F., Pulford, W. C. A. & Artymiuk, P. J. (1983) *J. Mol. Biol. 167*, 693–723.
99. Bull, H. B. (1981) *Arch. Biochem. Biophys. 208*, 229–232.
100. Gibbons, R. A. (1972) in *Glycoproteins*, part A, 2nd edn (Gottschalk, A., ed.) pp. 31–140, Elsevier, Amsterdam.
101. Woodward, C. K. & Hilton, B. D. (1979) *Annu. Rev. Biophys. Bioeng. 8*, 99–127.
102. Englander, S. W. & Kallenbach, N. R. (1984) *Q. Rev. Biophys. 16*, 521–655.
103. Wüthrich, K. & Wagner, G. (1979) *J. Mol. Biol. 130*, 1–18.
104. Delepierre, M., Dobson, C. M., Howart, M. A. & Poulsen, F. M. (1984) *Eur. J. Biochem. 145*, 389–395.
105. Kossiakoff, A. A. (1982) *Nature (Lond.) 296*, 713–721.
106. Phillips, S. E. V. (1984) in *Neutrons in biology* (Schoenborn, B. P., ed.) pp. 305–322, Plenum Press, New York.
107. Bentley, G. A., Delepierre, M., Dobson, C. M., Wedin, R. E., Mason, S. A. & Poulsen, F. M. (1983) *J. Mol. Biol. 170*, 243–247.
108. Haupt, H., Heimburger, N., Kranz, T. & Schwick, H. G. (1970) *Eur. J. Biochem. 17*, 254–261.
109. Harrison, R. A. (1983) *Biochemistry 22*, 5001–5007.
110. Villiers, M. B., Reboul, A., Thielens, N. M. & Colomb, M. G. (1981) *FEBS Lett. 132*, 49–54.
111. Durchschlag, H. & Jaenicke, R. (1983) *Int. J. Biol. Macromol. 5*, 143–148.
112. Gekko, K. & Timasheff, S. N. (1981) *Biochemistry 20*, 4667–4676.
113. Eisenberg, H. (1981) *Q. Rev. Biophys. 14*, 141–172.