# Excluded Volume Contribution to the Osmotic Second Virial Coefficient for Proteins

**Brian L. Neal and Abraham M. Lenhoff**
Center for Molecular and Engineering Thermodynamics, Dept. of Chemical Engineering
University of Delaware, Newark, DE 19716

The excluded volume represents a large positive contribution to the osmotic second virial coefficient due to the large size of protein molecules, a coefficient often used to characterize the thermodynamic properties of dilute protein solutions. Since this contribution is often the dominant one, it is important that it be evaluated accurately, but in the past it has usually been estimated by approximating protein molecules as ideal geometric structures such as spheres and spheroids. In this study a numerical approach is used to evaluate the excluded volume contribution more accurately by making use of crystallographic structural data for globular proteins. The numerical results are significantly larger than those obtained using the idealized shapes, with the main reason for the discrepancy being the effective surface roughness rather than the overall molecular shape. A reasonable empirical approximation that emerges from the results is that the excluded volume is roughly 6.7 times the molecular volume, compared to 4 times the volume, which is the result for spheres.

The extent to which predictive procedures can be applied to development of processes involving proteins is dictated by the level at which the physicochemical properties of aqueous protein solutions are understood. Of particular interest is the development of molecular thermodynamic models for the properties of dilute protein solutions, which can be used to predict osmotic pressures as well as the stability of protein solutions and conditions for protein precipitation. The utility of such information for identifying optimal conditions for protein crystal growth has also been described (George and Wilson, 1994). Within the broader context of biotechnology, molecular thermodynamic models for protein solutions are essential for the development of rational design procedures for downstream processing operations. In this regard they would have the same utility as corresponding models for smaller molecules in more traditional unit operations, most notably vapor-liquid equilibrium models for hydrocarbons in distillation column design.

Most solution thermodynamic models used to date for proteins are based on particular formulations of the potential of mean force. In the dilute solution limit, this is used within the context of the McMillan-Mayer virial expansion as a means to characterize the deviations from nonideality, particularly in the case of osmotic pressure correlations. The McMillan-Mayer dilute solution theory (McMillan and Mayer, 1945) was developed in an effort to apply the same concepts to solutes in liquids as had previously been employed in the study of nonideal gases. The virial expansion for the osmotic pressure of a solution:

$$\Pi = kT(c + B_2c^2 + B_3c^3 + \ldots) \tag{1}$$

is used to characterize deviations from ideal solution behavior resulting from the interaction of two, three or more bodies. The first term of this expansion is the van't Hoff equation for ideal solutions. The second virial coefficient $B_2$, which is a measure of the two-body interactions, may be written as (Zimm, 1946):

$$B_2 = -\frac{1}{2V}\int[g(r)-1]d\{1\}d\{2\} \tag{2}$$

where integration is over the positions and orientations of the two molecules under consideration. The pair correlation function $g(r)$ is a measure of the solvent-averaged effect of interaction of two solute molecules on one another. In the absence of energetic interactions arising from electrostatic, dispersion, or other forces, the function $g(r)$ takes on a value of 0 when the two bodies under consideration overlap and a value of 1 otherwise, assuming that the bodies may be considered rigid and impenetrable. For this limiting case the integral characterizes the volume in space associated with one molecule that is unavailable for occupation by a second molecule, that is, the excluded volume. The term "excluded volume" is sometimes also applied to proteins in the sense of the volume of a single protein molecule from which solvent molecules are excluded (for example, Colonna-Cesari and Sander, 1990; Connolly, 1992). Our concern, however, is with the solute rather than the solvent-excluded volume, hence the consideration of pairs of solute molecules in Eq. 2, while the solvent is treated as a continuum.

Because of the relatively large size of the solute (protein) molecules, the excluded volume makes a large, positive con-

tribution to the second virial coefficient, and is often considered to be the dominant contribution (Vilker et al., 1981; Haynes et al., 1992). Thus accurate estimation of the excluded volume is imperative if the thermodynamic properties are to be predicted with confidence.

The first approximation to facilitate calculation of the second virial coefficient is to treat a protein molecule as a rigid sphere of equivalent volume, which yields:

$$B_2 = 4v_m = \frac{2}{3}\pi d^3 \tag{3}$$

where $v_m$ is the molecular or van der Waals volume and $d$ is the sphere diameter, that is, the center-to-center distance at contact. However, the applicability of a sphere model for the excluded volume of proteins has been questioned, and an alternative choice, namely a spheroid approximation, suggested (Vilker et al., 1981). The excluded volume for spheroids is given by (Isihara and Hayashida, 1953):

$$B_2 = 4v_m\left[\frac{1}{4} + \frac{3}{16}\left(1 + \frac{\sin^{-1}\epsilon}{\epsilon(1-\epsilon^2)^{1/2}}\right)\left(1 + \frac{1-\epsilon^2}{2\epsilon}\ln\frac{1+\epsilon}{1-\epsilon}\right)\right] \tag{4}$$

where $v_m$ is $1/6\ \pi ab^2$ for a prolate spheroid and $1/6\ \pi a^2 b$ for an oblate spheroid, $\epsilon^2 = (a^2 - b^2)/a^2$, and $a$ and $b$ are the major and minor axes respectively. The spheroid approach leads to a larger excluded volume contribution to the second virial coefficient: for prolate spheroids with a major to minor axis length ratio of two, the second virial coefficient is 11% greater than that for spheres of the same volume. Use of the spheroid approximation yielded better agreement with experimental osmotic pressure data for bovine serum albumin (BSA) than did the sphere approximation (Vilker et al., 1981). However, since considerable uncertainty exists regarding the accuracy of the other contributions to the potential of mean force, which can be similar in magnitude to the excluded volume, it is not yet clear whether the spheroid approximation provides an adequate representation in general.

A continuing effort is clearly warranted to evaluate and improve methods for calculating the potential of mean force, and such work is in progress in several research groups. In particular, it is necessary to determine whether the detailed molecular structure of the protein must be accounted for, or whether simpler representations such as the hard sphere and spheroid models discussed above are adequate in the sense that details of shape represent only a small perturbation to simple geometric representations of excluded volume. This article investigates this situation by making use of crystallographic structures of a number of proteins to account for differences in their detailed shapes.

## Procedure

Equation 2 is a general formulation of the second virial coefficient expression. Since protein molecules are anisotropic, this expression must be applied to them in its most general form, that is, in the specification of the molecular orientations and positions. This can be done most conveniently by recognizing that the nature of the integration is such that the integrand is nonzero only when the molecules overlap. The most straightforward approach is then to fix the orientation of one molecule by placing its volumetric center at the origin, a convention that allows one to integrate over the volume of that molecule and to work with the relative positions and orientations of the two molecules. The second protein molecule facilitates sampling of all positional and orientational space through the translation of its center to positions specified by a distance $r_{12}$ and two angles $\theta$ and $\phi$, and rotation relative to the first as specified by the three Euler angles. Equation 2 then becomes:

$$B_2 = \frac{1}{16\pi^2}\int_0^{2\pi}\int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi}\int_0^{\infty} \times [g(r)-1]r_{12}^2 dr_{12}\sin\theta d\theta d\phi d\alpha \sin\beta d\beta d\gamma \tag{5}$$

where

$$g(r) = 1 \qquad r_{12} > d$$
$$g(r) = 0 \qquad r_{12} \le d$$

and $d$ is the center-to-center distance at contact. Integration over $r_{12}$ yields

$$B_2 = -\frac{1}{48\pi^2}\int_0^{2\pi}\int_0^{\pi}\int_0^{2\pi}\int_0^{2\pi}\int_0^{\pi} \times d^3\sin\theta d\theta d\phi d\alpha \sin\beta d\beta d\gamma \tag{6}$$

where the integrand is now a function of $d$, which depends on the remaining five variables, the positional and orientational angles. Note that $r_{12}$ and $d$ are defined in terms of the center-to-center distance; the volumetric centers, which fit most naturally, were used in our calculations, but given the initial formulation (Eq. 2), the origin may be chosen arbitrarily.

The key aspect of the evaluation of the integral in Eq. 6 is the determination of the value of $d$ for a given relative position and spatial orientation of two protein molecules. The nature of the problem is similar to that involved in protein docking calculations that are used to examine such issues as complementarity in protein-protein binding (Blanco, 1991; Cherfils et al., 1991; Shoichet and Kuntz, 1991), but in the interests of computational economy we have developed our own methods for this application as follows. To obtain the most accurate representation of the protein molecules' geometry, an atomistic approach was utilized to generate values for $d$. The coordinates of the proteins' atoms (excluding hydrogen) were obtained from the Brookhaven Protein Data Bank (Bernstein et al., 1977), and Bondi radii were assigned to each atom (Bondi, 1968). Surface atoms were identified using the programs PQMS and TRB of the Molecular Surface Package (Connolly, 1985, 1993). PQMS rolls a probe sphere, usually chosen of radius 1.4 Å to simulate a water molecule, over the surface of the protein molecule, and TRB identifies which of the protein's atoms were in contact with the probe. All subsequent calculations were restricted to these surface atoms in the interest of computational economy.

For any given set of angles, one protein molecule was located with its center at the origin, and a second, identical

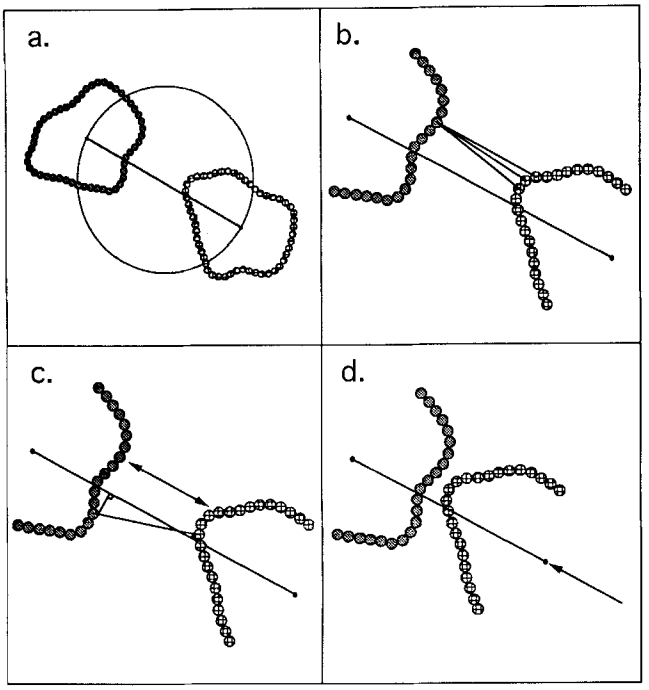


**Figure 1. Numerical procedure for determining center-to-center distance at contact for protein molecules.**

(a) Consider only surface atoms within a sphere with a diameter formed by the line segments connecting the centers of the two molecules; (b) distance between atom pairs formed from opposite molecules is computed; (c) line segment between the centers of the closest atom pair's surfaces is projected onto the line segment connecting the centers of the molecules; (d) molecules are moved together along the center-center line segment by the length of the projection.

molecule was constructed and rotated at the origin to specified Euler angles $\alpha$, $\beta$, and $\gamma$. This second molecule was then translated in a direction specified by the two angles $\theta$ and $\phi$ to a point sufficiently far from the first molecule at the origin that no overlap existed between the two. For this set of angles, $d$ was determined as follows. A sphere was constructed by rotating the line segment connecting the centers of the two molecules (Figure 1a). The surface atoms lying within this sphere were identified, and subsequent operations limited to these atoms, thereby further reducing the computational load. An iterative procedure was then performed to find the closest center-to-center distance $d$ of the molecules before overlap (Figure 1b). At each iteration the distance of closest approach between the two molecules was determined by identifying the most closely spaced pair of atoms (one from each molecule) (Figure 1c). The molecules were then moved together by a distance given by the projection of the line connecting the closest atom pair onto the line connecting the molecules' centers, which defines the direction along which the second molecule was translated (Figure 1d). The iterative procedure was continued until the molecules were less than a specified distance from contact. Five or fewer iterations of this procedure were generally sufficient to determine the center-to-center distance at contact to less than 0.1 Å. One exception, though, arose in those unusual situations where the direction of translation was orthogonal or nearly orthogonal to the line connecting the centers of the closest atoms; in these cases the translation step was taken to be equal to the distance of closest approach.

The above procedure was implemented as the central component of the evaluation of the multidimensional integral in Eq. 6. Because of the high dimensionality of the integral and the irregular nature of the integrand, the Monte Carlo integration subroutine D01GBF (NAG Fortran Library Mark 16, Numerical Algorithms Group, Downers Grove, IL) was utilized. This integration technique was fairly efficient for the present problem, and although it generally provides limited accuracy, it was more than adequate for present purposes with all of the results having estimated errors of less than 1%.

## Results and Discussion

Excluded volumes were computed for nine small to moderately sized globular proteins. Table 1 shows standard measures of protein size, along with the results of the numerical computations and corresponding excluded volume results based on simpler analytical approximations.

The different methods for estimating the excluded volume all yield results that are correlated with molecular size, but

**Table 1. Protein Structural Data and Excluded Volume Values**

| Protein Data Bank File* | Molec. Wt.** | Molec. Vol. (Å$^3$)† | Spheroid Dimens. (Å)‡ | Excluded Volume (Å$^3$) | | |
|---|---|---|---|---|---|---|
| | | | | Numerical | Sphere | Spheroid |
| 3CYT | 12,028 | 12,553 | – | 85,592 | 50,212 | – |
| 1RN3 | 13,691 | 14,820 | 38×25×25 | 108,765 | 59,279 | 52,110 |
| 1LYZ | 14,314 | 15,737 | 45×30×30 | 106,792 | 62,948 | 88,606 |
| 1MBO | 17,814 | 19,574 | 25×44×44 | 133,955 | 78,298 | 110,298 |
| 1TPO | 23,306 | 26,937 | 50×40×40 | 169,432 | 107,748 | 169,788 |
| 2CGA | 25,667 | 29,750 | 50×40×40 | 200,307 | 119,001 | 169,788 |
| 1CAC | 28,767 | 32,374 | 47×41×41 | 230,739 | 129,496 | 166,297 |
| 5PEP | 34,467 | 39,918 | – | 277,009 | 159,672 | – |
| 2HHB | 64,450 | 75,977 | 70×55×55 | 494,983 | 303,909 | 450,408 |

*Brookhaven Protein Data Bank coordinate files: 3CYT, tuna heart cytochrome C; 1RN3, bovine pancreas ribonuclease A; 1LYZ, hen egg white lysozyme; 1MBO, sperm whale oxymyoglobin; 1TPO, bovine pancreatic $\beta$-trypsin; 2CGA, bovine pancreatic chymotrypsinogen A; 1CAC, human erythrocyte carbonic anhydrase C; 5PEP, porcine pepsin; 2HHB, human deoxyhemoglobin.
**Calculated using amino acid composition.
†Calculated using Molecular Surface Package (Connolly, 1985, 1993).
‡From Creighton (1993).

there are appreciable differences in the actual values. In most cases the numerical computations yield the largest values, followed by the spheroid and sphere approximations respectively; differences are typically of order 20% for numerical vs. spheroid and for spheroid vs. sphere. These differences are too large to be explained by approximation errors. Firstly, the estimated errors in the numerical results are all less than 1%. Secondly, uncertainties in equivalent sphere parameters are fairly small. Molecular volumes calculated using molecular weights and partial specific volumes (Creighton, 1993) generally differ by less than 10% from those using crystallographic data and the Molecular Surface Package (Connolly, 1993), which are the ones listed in the table.

While the results in Table 1 suggest that the spheroid approximation represents a substantial, if not completely successful, improvement over the sphere approximation, closer examination indicates that this is not necessarily the case. The bulk of the improvement is in fact due to the crystal structure based spheroid dimensions suggested by Creighton, which lead to molecular volumes larger than those listed in Table 1. Adjustment of the dimensions to satisfy both the volume and axis ratio constraints leads to results much closer to those based on the sphere approximation.

A somewhat different illustration of the inadequacy of accounting only for overall shape comes from previous studies of molecular surface areas, although based on a fairly small set of data. The surface areas (for example, Richards, 1977) are computed by rolling a probe sphere across the van der Waals surface of the molecule to capture the convoluted nature of the surface. Janin (1976) and Teller (1976) showed that Chothia's (1975) computed accessible surface areas of relatively small globular proteins are about 1.95 times the equivalent sphere areas. LeMaster and Richards (1977), on the other hand, determined the inertial ellipsoids of six of those proteins, and found the surface areas to be smaller by a factor of 1.7 than the corresponding accessible surface areas. Thus, the inertial ellipsoid should have an area about $1.95/1.7 = 1.15$ times that of the equivalent sphere, and again this difference is too small for the overall shape to account for the difference in excluded volume. A descriptor for characterizing shape in more detail (Abad-Zapatero and Lin, 1990) does not show a significantly improved correlation.

The failure of shape to explain the discrepancy is not surprising. As mentioned earlier, the spheroid approximation is no more than 11% greater than the sphere approximation for axis ratios of 2 or less, so that the shape effect as described using the spheroid model cannot explain the discrepancy except for very high axis ratios. The large improvement noted by Vilker et al. (1981) is a consequence of their considering axis ratios as high as 4 for BSA. Since crystallographic data are unavailable for BSA, it is not possible to examine this molecule in more detail, but such a high axis ratio is certainly unusual for globular proteins.

Thus the overall shape of the protein molecule appears not to be the major factor determining the excluded volume increments. An alternative explanation is the presence of protruding groups that contribute a degree of roughness to the surface of the molecule, which in turn limits the distance of closest approach. Development of a more versatile approximate model for estimating excluded volume requires a more general characterization of the surface roughness. An empirical approach is shown in Figure 2, in which the numerically calculated excluded volume is plotted against the molecular volume. The data fit a straight line quite well, with the slope and the intercept very highly correlated. However, for unit slope the correlation is of the form $B_2 = 6.68\ v_m$, which is the line shown in Figure 2. The value of the coefficient implies that for the purpose of excluded volume calculations, the effective molecular volume is about 1.7 times the actual molecular volume (see Eq. 3).

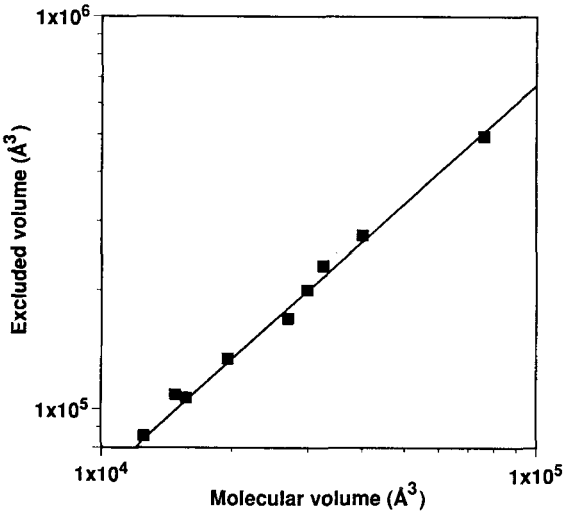


**Figure 2. Relation of computed excluded volume to molecular volume.**

Line shows the relation $B_2 = 6.68\ v_m$.

A model to explain this relationship requires quantification of surface roughness. One measure of roughness is the molecular surface area discussed earlier, which can be characterized in different ways. One is to use the surface area as the basis for estimating the fractal dimension $D$ of the protein surface. Values between 2 and about 2.5 are obtained (Lewis and Rees, 1985; Abad-Zapatero and Lin, 1990), but there is considerable variation with local position on the surface, with probe sphere size and among molecules. Of particular note is that fractal dimensions different from 2 are seen only in a narrow window of probe sizes between about 1 and 3.5 Å, that is, for probe spheres outside this range the surface is smooth (Lewis and Rees, 1985). Since the roughness that determines the increment in size of the smooth molecular envelope over the equivalent sphere is of relatively large scale, the upper end of the window, that is, near 3.5 Å, provides one measure, although it is based on just three relatively small proteins.

To relate the roughness scale to the excluded volume results, one approach is suggested by a simple idealized model employing small, close-packed, hemispherical protuberances on a smooth surface (Lee, 1984). This model captures the surface area increase quite well, but the result is independent of the size of the hemispheres, that is, the scale of the roughness, so surface area data do not directly provide the scale of the roughness. However, if the actual molecular volume is equal to that of a sphere of radius $R$, then a roughness scale $\Delta R$ would lead to an enclosing sphere larger in volume by a factor $(1 + \Delta R/R)^3$. The invariance of the volume ratio sug-

gested by Figure 2 indicates then that the roughness scale should increase with molecular size roughly as $\Delta R = 0.19\ R$, a plausible conclusion, albeit one not directly predicted. For the proteins considered here, the increment in effective radius would correspond to about 3-5 Å, a measure consistent with small groups of atoms on the protein surface and somewhat larger than the water probe sphere used in the numerical calculations.

Alternative routes to accounting for surface roughness would require determining the volume of a smooth surface enclosing the whole molecule. A geometrically simple surface of this kind can be found by the algorithm of Taylor et al. (1983) to produce an "equivalent ellipsoid," while a tighter fit is likely to be obtained by using a large probe sphere in Connolly's (1983, 1985, 1993) Molecular Surface Package. In both cases, though, the computational load is appreciable, and knowledge of the molecular structure is required.

It should be emphasized that neither thermal vibrations nor hydration effects have been included in the above analysis, both of which would increase the effective molecular size. An increased molecular size, usually attributed to hydration, is sometimes used as a correction to yield better agreement with experiment for protein properties such as intrinsic viscosity (Edsall and Mehl, 1943), but the need for such corrections may be explained instead by the surface roughness examined here.

In summary, we have shown that generally accepted methods for estimating the excluded volume of proteins from the molecular size are likely to result in serious underestimates (as much as about 40%), mainly because of surface roughness. Our calculations suggest that using a value of 6.7, rather than 4, times the molecular volume should lead to a good approximation of the true value. Similar efforts to characterize the importance of structural details in estimates of other contributions to the potential of mean force are in progress.

## Acknowledgment

This research was supported by the National Science Foundation under grant no. BCS-9210401. We are grateful to M. E. Paulaitis for useful discussions and comments on the manuscript.

## Notation

$B_n$ = $n$th virial coefficient
$B_n$ =
$c$ = concentration
$k$ = Boltzmann constant
$T$ = absolute temperature
$V$ = volume

### *Greek letters*

$\alpha, \beta, \gamma$ = Euler angles
$\epsilon$ = spheroid eccentricity
$\Pi$ = osmotic pressure
$\phi, \theta$ = angular variables in spherical coordinate system

## Literature Cited

Abad-Zapatero, C., and C. T. Lin, "Statistical Descriptors for the Size and Shape of Globular Proteins," *Biopolymers*, **29**, 1745 (1990).

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, "The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures," *J. Mol. Biol.*, **112**, 535 (1977).

Blanco, M., "Molecular Silverware. I. General Solutions to Excluded Volume Constrained Problems," *J. Comp. Chem.*, **12**, 237 (1991).

Bondi, A., *Physical Properties of Molecular Crystals, Liquids and Glasses*, Wiley, New York (1968).

Cherfils, J., S. Duquerroy, and J. Janin, "Protein-Protein Recognition Analyzed by Docking Simulation," *Proteins*, **11**, 271 (1991).

Chothia, C., "Structural Invariants in Protein Folding," *Nature*, **254**, 304 (1975).

Colonna-Cesari, F., and C. Sander, "Excluded Volume Approximation to Protein-Solvent Interaction, " *Biophys. J.*, **57**, 1103 (1990).

Connolly, M. L., "Solvent-Accessible Surfaces of Proteins and Nucleic Acids," *Science*, **221** 709 (1983).

Connolly, M. L., "Molecular Surface Triangulation," *J. Appl. Cryst.*, **18**, 499 (1985).

Connolly, M. L., "Shape Distributions of Protein Topography," *Biopolymers*, **32**, 1215 (1992).

Connolly, M. L., "The Molecular Surface Package," *J. Mol. Graphics*, **11**, 139 (1993).

Creighton, T. E., *Proteins; Structures and Molecular Properties*, 2nd ed., W. H. Freeman and Co., New York (1993).

Edsall, J. T., and J. W. Mehl, "Translational Diffusion of Amino Acids and Proteins," *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, E. J. Cohn and J. T. Edsall, eds., Chap. 18, Reinhold Publishing, New York (1943).

George, A., and W. W. Wilson, "Predicting Protein Crystallization from a Dilute Solution Property," *Acta Cryst.*, **D50**, 361 (1994).

Haynes, C. A., K. Tamura, H. R. Korfer, H. W. Blanch, and J. M. Prausnitz, "Thermodynamic Properties of Aqueous $\alpha$-Chymotrypsin Solutions from Membrane Osmometry Measurements," *J. Phys. Chem.*, **96**, 905 (1992).

Isihara, A., and T. Hayashida, "Theory of High Polymer Solutions: II. Special Forms of Second Osmotic Coefficient," *J. Phys. Soc. Japan*, **6**, 46 (1951).

Janin, J., "Surface Area of Globular Proteins," *J. Mol. Biol.*, **105**, 13 (1976).

Lee, B., "Increase in Surface Area Due to Atomic Scale Roughness," *Biochim. Biophys. Acta*, **800**, 309 (1984).

LeMaster, D., and F. M. Richards, Unpublished results cited in Richards (1977).

Lewis, M., and D. C. Rees, "Fractal Surfaces of Proteins," *Science*, **230**, 1163 (1985).

McMillan, W. G., and J. E. Mayer, "The Statistical Thermodynamics of Multicomponent Systems," *J. Chem. Phys.*, **13**, 276 (1945).

Richards, F. M., "Areas, Volumes, Packing, and Protein Structure," *Ann. Rev. Biophys. Bioeng.*, **6**, 151 (1977).

Shoichet, B. K., and I. D. Kuntz, "Protein Docking and Complementarity," *J. Mol. Biol.*, **221**, 327 (1991).

Taylor, W. R., J. M. Thornton, and W. G. Turnell, "An Ellipsoidal Approximation of Protein Shape," *J. Mol. Graphics*, **1**(2), 30 (1983).

Teller, D. C., "Accessible Area, Packing Volumes and Interaction Surfaces of Globular Proteins," *Nature*, **260**, 729 (1976).

Vilker, V. L., C. K. Colton, and K. A. Smith, "The Osmotic Pressure of Concentrated Protein Solutions: Effect of Concentration and pH in Saline Solutions of Bovine Serum Albumin," *J. Colloid Interface Sci.*, **79**, 548 (1981).

Zimm, B. H., "Applications of the Methods of Molecular Distribution to Solutions of Large Molecules," *J. Chem. Phys.*, **14**, 164 (1946).

*Manuscript received July 5, 1994, and revision received Oct. 3, 1994.*