Molecules: What Kind of a Bag of Atoms?[†]

Praveen D. Chowdary and Martin Gruebele*

Department of Chemistry, Department of Physics, and Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, Illinois 61801

Received: April 3, 2009; Revised Manuscript Received: June 19, 2009

As discussed by Liang and Dill, Enright and Leitner, and others, proteins are not 3D objects. We study an expanded macromolecular data set ranging from proteins to RNA, lipids, and viruses, and remove surface effects and size bias. Molecules and molecular assemblies with more than 1000 backbone atoms have a volume fractal dimension of $D_v = 2.70 \pm 0.05$ by the embedded sphere method and $D_v = 2.71 \pm 0.04$ by the ensemble method using radius of gyration as the size measure. The much larger $D_v = 2.89 \pm 0.05$ obtained with the average surface radius as the length measure shows that surface corrugation is as extensive as cavity formation. Using a simple "Swiss cheese" model for molecules, we show that the distribution of voids in the interior of molecules cannot be a Boltzmann distribution of void energy as a function of void size. Instead, frustration from imperfect packing builds up with molecular size, allowing larger voids to form in larger molecules. We find that large molecules lie halfway between the extremes of packing for homogeneous objects ($D = 3$) and Apollonian packing, which accounts for packing of a hierarchy of random-sized objects ($D \approx 2.47$).

Introduction

Bob Field once famously mused about whether molecules approach the "bag of atoms" limit when energy is dumped into them. At chemical excitation energies, small organic molecules maintain much order in their vibrational state space and do not get to that limit.[1] Yet in a different sense, large molecules do resemble bags of atoms: they fold back onto themselves, with myriad weak interactions such as hydrogen bonds, van der Waals contacts, or salt bridges lowering the energy of compact structures compared with extended structures. Protein folding is perhaps the most famous example, but any molecule that is large enough will interact with itself via through-space interactions.

As they pack against themselves, large molecules and molecular assemblies contain a distribution of different-sized voids. There is rich literature on their packing.[2−5] Analyses of proteins by the Dill,[2] Leitner,[3] and Zebende[4] groups reveal a mass scaling dimension of $D_m \approx 2.5$ by either embedding variable-size spheres within proteins or plotting mass against a measure of size (diameter, number of residues, etc.) for an ensemble of proteins. In addition, Klafter and coworkers have derived equations connecting the mass dimension to protein size and spectral dimension.[5] In some of the literature, the packing is compared with a hard sphere liquid near the percolation threshold ($D \approx 2.5$).[2,6] Large molecules evidently contain a hierarchy of different-sized voids that makes them less than 3D objects.

Here we analyze a larger number of packed molecular structures ranging in size from small peptides to virus particles and covering macromolecules such as proteins, RNAs, and lipids. We remove both surface intersection bias and sample size bias. We use two methods. The embedded sphere method yields a volume fractal dimension of $D_v = 2.70 \pm 0.05$, slightly larger than previous results. The ensemble method yields $D_v = 2.71 \pm 0.04$ when the radius of gyration (R_g) is used as a size measure, which is in good agreement with the embedded sphere method. The slightly larger than previously estimated values are accounted for by the larger data set and the removal of sample size bias.

We can draw some new conclusions about the approximately fractal nature of void spaces within macromolecules, about molecular surface corrugation, and about the linearity of size measures. We find that smooth size measures (R_g or R_{ext}, the radius of maximal extent defined below) yield similar low dimensions. A length measure that takes into account molecular surface corrugation (the average radius R_{avg} defined below) yields dimensions closer to three. Therefore, molecular voids and surface corrugations obey analogous scaling laws. Our analysis using a "Swiss cheese" model also reveals that the energy of void volumes must depend on macromolecular size; it cannot be simply a function of void size, such as a Boltzmann distribution, whose energy depends on void volume or void surface area. Voids in the interior of a larger structure "know" they are in a large structure. Finally, we re-examine the packing as a function of distance from the center of mass of macromolecules. By using smaller embedded spheres than those used in past work, we find that there is no significant scaling of dimension with closeness to the surface. Rather, very strong surface corrugation leads to underestimates of the bulk fractal dimension near the surface by the embedded sphere method, whereas the dimension is overestimated by the ensemble method.

To understand these results, we propose the following interpretation: molecular structures can be classified by whether they approach the limit of homogeneous packing ($D = 3$) or the limit of Apollonian packing ($D = 2.47$). Homogeneous packing arises when similar-size objects are packed near-optimally (e.g., liquids) or optimally (crystals). Apollonian packing arises when objects with a hierarchical size distribution (fewer large ones than small ones) are packed optimally. How would such a hierarchy of length scales arise in molecules

[†] Part of the "Robert W. Field Festschrift".

* Corresponding author. E-mail: gruebele@scs.uiuc.edu.

10.1021/jp903104p CCC: $40.75 © 2009 American Chemical Society
Published on Web 07/09/2009

because the atoms are all about the same size? Atom connectivity provides the answer. Large molecules are built up along a well-defined hierarchy of length scales, starting with atoms, then on to functional groups and side chains, and on to monomer units, secondary structures (local organization), and finally tertiary structures (global packing). This hierarchy of sizes caused by the constraints of bond connectivity hinders homogeneous packing and drives the molecular packing dimension closer to the Apollonian limit. With $D_v \approx 2.7$, the interior of molecules appears to lie about halfway between these crystalline and Apollonian limits.

Methods

Databases and Molecules Used. Our sample set is composed of a variety of molecular structures including proteins (~91%), viruses (~8%), ribosomes (~0.4%), and lipid layers (~0.4%) covering a wide range of shapes and sizes. All in all, 2752 structures, with the number of atoms ranging from ~200 to ~1 000 000 are included. About 2250 protein structures were picked from the 25% threshold list (October 2008) of PDBSELECT to avoid redundancy.[7] The remaining structures were generated using PDBSELECT for consistent representation of larger molecules. The viruses (listed at VIPERdb)[8−10] and ribosomal structures are from the Protein Data Bank,[7] and the equilibrated lipid structures are from Tielemann and coworkers.[11,12]

Excluded van der Waals Volume. *V* was calculated as follows: The pdb coordinates are superimposed onto a 3D grid with 0.025 nm sized cubic voxels. Every voxel whose center falls within the van der Waals radius of any atomic center is considered to be occupied. *V* is calculated as the sum of the volumes of all occupied voxels. This method takes into account overlapping van der Waals radii of bonded atoms and empty spaces. The results in this article are for backbone atoms with effective van der Waals diameters adopted from VMD.[13] We also studied the effect of adding hydrogen atoms using the autopsf plugin of VMD and found no significant effect on the scaling laws (2.5% variation in dimensions calculated).

Scaling and Definition of the Length Scales $L = R_g$, R_{ext}, R_{avg}. The scaling of the excluded van der Waals volume (*V*) with size is given by

$$V \approx L^{D_v} \tag{1}$$

Here *L* is a size measure (length scale) and D_v is the volume fractal dimension.

We use three different size measures. The radius of gyration R_g, used by Leitner and coworkers,[3] is the mass-weighted root-mean-squared position vector averaged over all atoms in the structure and is given by

$$R_g = \left(\frac{1}{M}\sum_i m_i \left|\mathbf{r}_i\right|^2\right)^{1/2} \tag{2}$$

where m_i is the atomic mass and $\mathbf{r}_i$ is the position vector from the center of mass.

The exterior radius, R_{ext}, is the maximum extent of the molecular structure along the coordinate axes ($q_j = x, y, z$) as used by Dill and coworkers[2]

$$R_{ext} = \frac{1}{6}\sum_{j=x,y,z}(q_j^{max} - q_j^{min}) \tag{3}$$

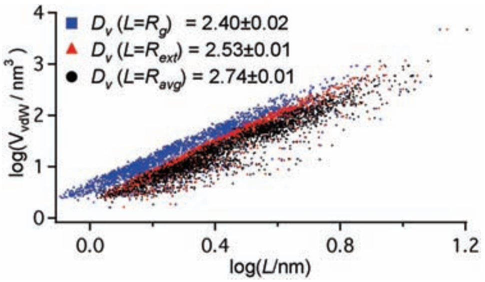

Figure 1. van der Waals volume scaling versus size of 2752 molecules and molecular assemblies, including 2500 proteins from the PDBSELECT database. Three size measures are shown: radius of gyration (blue), average surface radius (red), and exterior radius (black), as defined in the Methods section.

For the purpose of a scaling law exponent, it does not make much difference whether the axes are the principal axes of the molecule.

Finally, we introduce $R_{avg} = \langle \mathbf{r}_i \rangle_{surface}$, the average surface radius defined as the mean distance of all surface atoms from the center of mass. We identified surface atoms by placing the molecule over a sufficiently fine 2D grid in several orientations (*x*-*y*, *y*-*z*, and *x*-*z*), and identifying all of the first and last atoms intersected by lines perpendicular to the grid.

Dimensional Analysis. We computed the volume dimension, D_v, used by Dill[2] by two different approaches. As a check, we also computed the mass dimension, D_m, used by Leitner and Klafter by both approaches[3,5] and found that it always agreed with D_v within 0.01, so we discuss only D_v henceforth.

The first approach is the ensemble method: the van der Waals excluded volume of each structure was plotted against one of the three length measures, *L*, discussed above. A linear fit to the log−log plot was used to obtain D_v from eq 1. The dimension thus obtained is a collective property of the molecular ensemble studied.

In the second approach, we computed D_v using the embedded sphere method of Leitner and coworkers.[3] This allows a dimension to be determined for every individual molecule. A sphere of variable diameter is embedded within the molecule, and the log of the enclosed excluded van der Waals volume is plotted against the log of the embedded sphere radius. A linear fit of mass versus sphere radius yields the fractal dimension $D_v^{(i)}$ of the molecule "*i*". This is then averaged over various atomic centers within the molecule, yielding, for example, $D_{10\%}^{(i)}$ for the 10% of backbone atoms closest to the center of mass of the molecule, $D_{20\%}^{(i)}$ for the closest 20% out from the center of mass, and so on. To avoid surface effects, we restricted the upper limit of the embedded sphere radius to be R_g or even $0.3R_g$, and the lower limit is fixed at 0.5 nm. By increasing the percentage *X* in the subscripts above, shells of atoms closer to the center or to the periphery of the molecule could be studied to look at packing in the interior versus near the surface.

Results

Smooth measures of molecular size ($L = R_g$ or R_{ext}) yield a fractal volume dimension of molecules and molecular assemblies significantly less than 3. Figure 1 shows the fit to the raw data. The logarithm of the van der Waals volume is plotted against the logarithm of the length scale in nanometers. The smallest molecule in this plot is a peptide of 200 atoms. The largest molecular assembly is a virus particle with 812 340 atoms.[11,12] The three measures of size (R_g, R_{ext}, and R_{avg}) yield values of D_v ranging from 2.4 to 2.7. Unfortunately, the raw data strongly

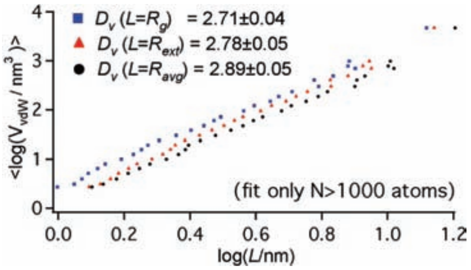

Figure 2. van der Waals volume scaling versus size for binned size distribution, removing any bias toward a certain molecular size. Three size measures are shown: radius of gyration (blue), average surface radius (red), and exterior radius (black), as defined in the Methods section. Only molecules with $N > 1000$ backbone atoms were included in the fits of the dimensions shown.

biases the fit toward medium-size proteins. Furthermore, the increased likelihood of a larger length/volume ratio for molecules with <1000 backbone atoms (see below) tends to push up the dimension for small molecules by this method.

When the size bias is removed, the dimensionality of molecules with respect to a smooth measure of size is still significantly less than 3. We removed the size bias by binning molecules of different logarithmically spaced size ranges into single data points (Figure 2) before fitting. We included only molecules with >1000 backbone atoms in the fitting. The trend already noted for R_g, R_{ext}, and R_{avg} is observed in the binned sample but with dimensions about 0.3 higher than those for the raw sample. The dimensions obtained from R_g and R_{ext} now range from 2.71 to 2.78 within the 1σ uncertainties. As noted in the Methods section, the volume and mass (not shown) dimensions are nearly identical. This result is expected because most of the backbone atoms in the molecular data set are second row atoms of similar mass.

A significant difference in dimension persists for R_{avg} in the binned data (Figure 2). The mass or volume dimension associated with the average surface radius, R_{avg}, approaches 2.9. This implies a scaling of R_{avg} relative to the other length scales, for example

$$\left(\frac{R_{avg}}{R_{avg}^{(0)}}\right) = \left(\frac{R_g}{R_g^{(0)}}\right)^{0.94} \tag{4}$$

Compared with the radius of gyration and the radius of maximal extent, the average surface radius is not quite a 1D measure of size. As discussed below, R_{avg} is reduced by strong surface corrugations (analogous to voids in the interior). The effects of surface corrugation and interior void distribution nearly cancel, yielding a dimensionality close to 3.

Figure 3A shows the volume dimension determined by the alternative method of a sphere of variable radius fully embedded within the molecule. The dimension is plotted against the number of backbone atoms. When sample bias is removed by binning in Figure 3B, the dimensions are ~2.70 ± 0.05 for molecules with $N > 1000$ backbone atoms.

We make the $N > 1000$ cutoff because many smaller molecules have an elongated structure. The circle in Figure 3A highlights this population of small molecules of lower dimension. The effect can also be discerned in Figure 1; there the bottom edge of the distributions is feathered because of a population of molecules with unusually large length-to-volume ratio. Small elongated molecules cause computational artifacts in the fractal bulk volume dimension. By the embedded sphere method, the apparent dimension is too low because even the smallest meaningful embedded spheres intersect the surface. By the ensemble method, the apparent dimension is too high because molecules of unusual length increase the slope below log(L/nm) = 0.2 in Figure 1.

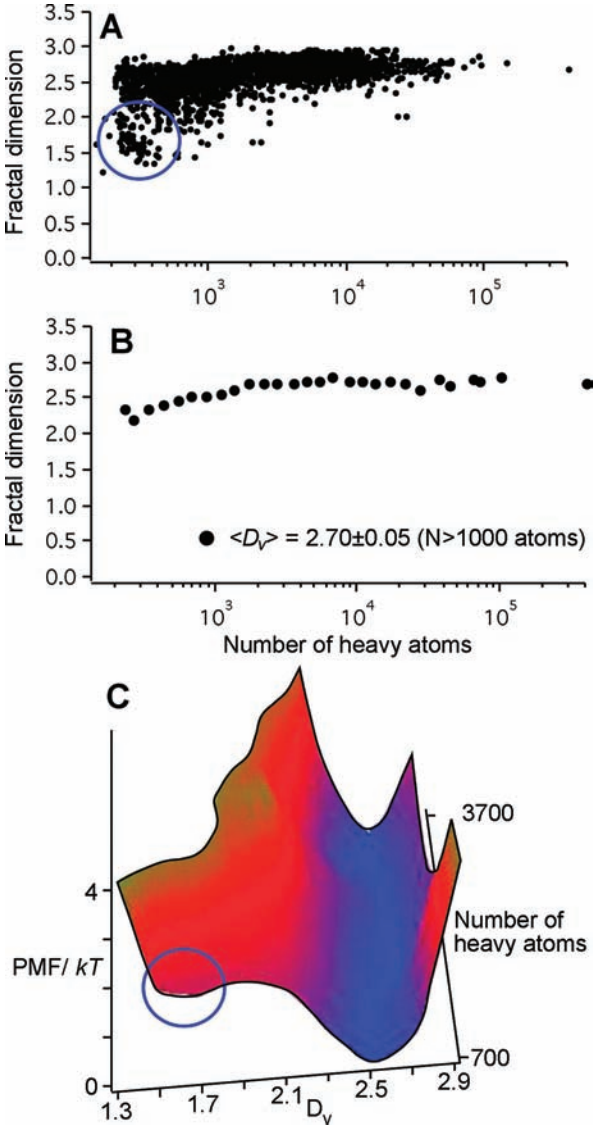

Figure 3. (A) Volume dimension by the embedded sphere method as a function of atom number in the molecule or molecular assembly. The closest 10% of α carbons to the center of mass are used as centers in the dimension calculations. A distinct population of lower dimensionality is circled below $N \approx 1000$ atoms. (B) Binned volume dimensionality. (C) Effective potential as a function of volume dimension and atom number derived from the population according to eq 5.

Figure 3C shows a smoothed plot of a potential of mean force, g, derived from the probability that a molecule with N heavy atoms has a probability P of having volume dimension D_v

$$g(D_v, N) = -kT \ln[P(D_v, N)] \tag{5}$$

A higher potential is associated with the smaller population of extended molecules; nonetheless, a local minimum (circled) occurs for small molecules. The energy function becomes monomodal above about 1000 atoms, with a deep minimum at $D_v = 2.7$.

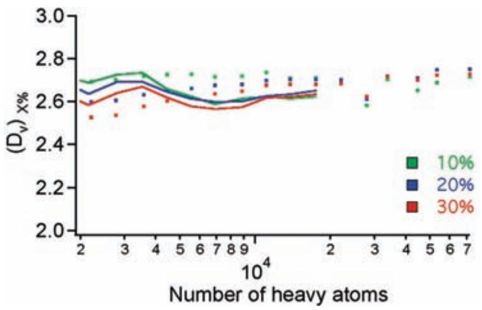

Figure 4. Volume dimension $(D_v)_{X\%}$ calculated by the embedded sphere method for $X = 10-30\%$. The maximal radius of the embedded sphere is set to R_g (dotted) or $0.3R_g$ (solid).

We find that the fractal dimension of molecules is size-independent and decreases at best minimally toward the molecular surface in our large data sample (Figure 4). The use of smaller embedded spheres (<0.3 R_g solid lines, versus R_g markers in Figure 4) reduced sample bias caused by empty spaces from surface corrugation being included in the embedded spheres. The innermost shell produces $D = 2.7$ for all but the smallest molecules, which is in agreement with the ensemble method and Enright and Leitner.[3] As can be seen from the random switching of the 10% (green) with respect to 30% (red) dimensions in Figure 4, size-dependent features such as pores, hinges, and multiple domains have a more pronounced effect than the closeness of the probe sphere to the surface once surface bias is removed.

Discussion

The larger data set with size bias removed supports the previous finding of fractal mass or volume dimensions significantly less than 3.[2−4] By the ensemble method, we obtain somewhat larger values of D than the smaller biased data sets. Specifically, our R_g-derived average value is 2.71 ± 0.04 compared with that reported by the Leitner (2.56)[3] and Zebende (2.47)[4] groups. Similarly, our R_{ext}-derived value is larger than that reported by the Dill group (2.42).[2] With smaller biased data sets, we obtain results identical to those of the previous studies.

The mass and volume fractal dimension trends by the embedded sphere method are similar to those obtained by Enright and Leitner[3] on a set of 200 proteins. Our average $(D_v^{(i)})_{10\%}$ for proteins with at least 1000 residues is 2.70 compared with 2.73 reported for a smaller set of proteins.[3] A trend of decreasing dimension with increasing distance from the molecular center of mass (Figure 4, markers from 10 to 30%) has been previously observed.[3] However, this per se cannot be interpreted as a decrease in packing efficiency near the surface. The convergence of dimensions seen at large molecular sizes in Figure 4 (dotted) as well as in ref 3 suggest that the apparent trend is mainly an effect of surface corrugation. In fact, when we reduce the maximum embedded sphere radii from R_g to $0.3R_g$ so as not to intersect the surface (lines in Figure 4), there is no reliable monotonic trend in the resulting fractal dimension, either with molecular size or with percent from interior for large molecules. Instead, the fluctuations with molecular size in Figure 4 arise from hinges, pores, and other molecular structural features.

Having confirmed the <3 fractal dimensions from previous reports, we consider several new consequences of a fractal dimension $D_v = 2.7$. We discuss in turn: surface energy effects at small molecular size, how R_{avg} scaling connects interior corrugation (voids) and surface corrugation, and the fact that the size distribution of voids cannot depend just on void size. Finally, we propose a simple qualitative explanation for the fractal dimension of molecules in terms of bond connectivity and hierarchical packing.

At small molecular size, significant deviations from optimal packing into a roughly spherical shape are observed: a family of molecules with a high surface-to-volume ratio and an average D as low as 1.5 exists, even though packing into a more compact shape would be more favorable energetically by lowering the van der Waals energy, hydrogen bonding energy, and other interaction energies. These molecules owe their existence as common structures to two factors. They are embedded in a solvent, which reduces surface energy, and smaller size allows stiffness at short persistence lengths to play a role, as observed for the local extended structure favored by unfolded proteins.[14] Therefore, surface tension does not rule molecular shape up to about 1000 atoms (Figure 3A), after which structures tend to be more “baggy” than “tubular”, and outliers of very low dimension become rare.

Interior “corrugation” (voids) and surface corrugation scale nearly the same way with molecular size. The evidence is that smooth size measures (R_g, R_{ext}) yield a volume dimension smaller than 3, whereas a size measure that accounts for surface corrugation (R_{avg}) yields a volume dimension close to 3. Volume and surface area are still nearly related by $A \sim V^{2/3}$, although A is not 2D, and V is not 3D.

A fractal volume dimension significantly less than 3 implies that the probability of finding a void of a given size within a molecule cannot depend just on void size. It has been proposed that void probability scales as a simple Boltzmann factor[15,16]

$$P \approx \mathrm{e}^{-\beta E_{\text{void}}(R_{\text{void}})} \tag{6}$$

with the void probability dependent on only the size of the void carved from the interior of a molecule. If this were true, then the void size distribution would have a characteristic mean value independent of molecular size, and the dimensionality of molecules would have to be 3. $D \approx 2.7$ unambiguously proves that larger molecules have a longer tail of large voids than do smaller molecules. Even the near-spherical viruses, where surface shape does not play a role, nearly produce $D = 2.7$, as does the analysis in Figures 3 and 4 that eliminates molecular surface effects entirely. Therefore, the energy of voids depends on not only the immediate atomic neighborhood from which they were carved but also the overall molecular size.

To test this observation numerically, we simulated a “Swiss cheese” model of molecules. In this model, we start with a solid spherical molecule. Random size and position spherical cavities were punched into the solid all the way from the center to the edge, creating both voids and surface corrugation. By simulating spherical molecules of different sizes with a constant void density, as determined by Liang and coworkers,[2] we checked the scaling of the occupied volume with different length scales. Using a size distribution of voids given by eq 6, the dimension derived from all three length measures is 3.00 ± 0.05 for any choice of energy function (e.g., $E \approx R_{\text{void}}^3$) and prefactor β. When we use scaling that allows larger voids in larger molecules, such as

$$P \approx \mathrm{e}^{-\beta(R_{\text{void}}/R_{\text{ext}})^{\alpha}} \tag{7}$$

(e.g., $\alpha = 1.5$, $\beta = 50$) we obtain scaling laws for R_g and R_{ext} in perfect agreement with Figure 2. By using the actual solid molecular envelopes from our sample set instead of solid spheres to better represent the largest scale molecular surface corrugation, we can also reproduce $D_v(R_{\mathrm{avg}}) > D_v(R_{\mathrm{g}}) \approx D_v(R_{\mathrm{ext}})$.

Why can voids in molecules not be treated like independent, Boltzmann-distributed objects? In a related question, why should molecules not be thought of as a fluid near the percolation threshold, as has been suggested because the percolation threshold dimension also happens to be less than 3?

The answer is that the atoms in the bag are connected into a hierarchy of different-sized groups. The hierarchy ranges from single atoms, to functional groups, all the way up to secondary and tertiary structure elements. When connected molecular pieces of different size come into contact, molecular structure becomes energetically frustrated because not all surfaces can make optimal contacts as a result of the connectivity constraints. Such frustration does not exist in homogeneous materials, such as crystals or fluids assembled from similar-sized objects. Frustration energy builds up with molecular size as more clashing constraints need to be satisfied, resulting in large voids. The consequence of larger voids in larger molecules is a fractal mass or volume dimension substantially less than 3. Beyond the size of the largest hierarchical grouping, the volume dimension has to revert to 3. Perhaps the largest virus in Figure 2 (top right of plot) is finally approaching this limit because it is shifted toward a slightly larger slope than the smaller macromolecules and assemblies.

We can thus consider molecules to lie between two extremes: At one extreme, we have a fluid or crystal of equal-size groups, and the dimension 3 is reached. Diamond would be a good example of such a “molecule”. At the other extreme, we have molecules as a random jumble of groups of many sizes: atoms, functional groups, residues, secondary structures, and finally tertiary folds as we move up the size hierarchy. The packing would then resemble Apollonian packing of dimension 2.47.[17] Apollonian packing corresponds to the densest possible packing of objects with a wide size distribution (In its simplest form, Apollonian packing describes spheres of many sizes.) The actual dimension we find for molecules, $D_v = 2.7$, lies between these extremes. Molecules pack better than completely random-sized objects but not as well as equal-sized unconnected spheres.

Acknowledgment. M.G. was supported by a James R. Eiszner Professorship, and P.D.C. was supported by a James R. Beck fellowship. The computer cluster was funded by an NSF CRIF grant (CHE-0541659). We thank Peter Freddolino for providing the virus structures.

References and Notes

(1) Silva, M.; Jongma, R.; Field, R. W.; Wodtke, A. M. *Annu. Rev. Phys. Chem.* **2001**, *52*, 811.

(2) Liang, J.; Dill, K. A. *Biophys. J.* **2001**, *81*, 751.

(3) Enright, M. B.; Leitner, D. M. *Phys. Rev. E* **2005**, 71.

(4) Moret, M. A.; Miranda, J. G. V.; Nogueira, E.; Santana, M. C.; Zebende, G. F. *Phys. Rev. E* **2005**, 71.

(5) Reuveni, S.; Granek, R.; Klafter, J. *Phys. Rev. Lett.* **2008**, 100.

(6) Moret, M. A.; Santana, M. C.; Nogueira, E.; Zebende, G. F. *Physica A* **2006**, *361*, 250.

(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235.

(8) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. *Structure* **2006**, *14*, 437.

(9) Belnap, D. M.; McDermott, B. M.; Filman, D. J.; Cheng, N. Q.; Trus, B. L.; Zuccola, H. J.; Racaniello, V. R.; Hogle, J. M.; Steven, A. C. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 73.

(10) Basavappa, R.; Syed, R.; Flore, O.; Icenogle, J. P.; Filman, D. J.; Hogle, J. M. *Protein Sci.* **1994**, *3*, 1651.

(11) Anezo, C.; Vries, A. H. d.; Holtje, H. D.; Tieleman, D. P.; Marrink, S. J. *J. Phys. Chem. B* **2003**, *107*, 9424.

(12) Marrink, S. J.; Tieleman, D. P. *J. Am. Chem. Soc.* **2001**, *123*, 12383.

(13) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.

(14) Yang, W.; Larios, E.; Gruebele, M. *J. Am. Chem. Soc.* **2003**, *125*, 16220.

(15) Rashin, A. A.; Rashin, A. H. L. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 547.

(16) Graziano, G. *Chem. Phys. Lett.* **2007**, *434*, 316.

(17) Herrmann, H. J.; Baram, R. M.; Wackenhut, M. *Physica A* **2003**, *330*, 77.

JP903104P