Mass fractal dimension and the compactness of proteins

Matthew B. Enright and David M. Leitner*
Department of Chemistry and Chemical Physics Program, University of Nevada, Reno, Nevada 89557, USA
(Received 23 September 2004; published 27 January 2005)

Vibrational dynamics and energy flow in a protein are related by Alexander-Orbach theory to the protein's mass fractal dimension D and spectral dimension $\bar{d}$. Burioni *et al.* [Proteins: Struct., Funct. Bioinf. **55**, 529 (2004)] recently proposed a relation between $\bar{d}$ and protein size based on their computational analysis of a set of proteins ranging from about 100 to several thousand amino acids. We report here values for D computed for 200 proteins from the Protein Data Bank (PDB) ranging from about 100 to over 10 000 amino acids and examine variation of D with protein size. The average D is found to be 2.5, significantly smaller than a completely compact three-dimensional collapsed polymer. Indeed, we find that on average a protein in its PDB configuration fills about three-quarters of the volume within the protein surface. Protein mass is also found to scale with radius of gyration with an exponent of 2.5 for this set of proteins.

DOI: 10.1103/PhysRevE.71.011912 PACS number(s): 87.10.+e, 87.15.−v, 82.35.Lr

I. INTRODUCTION

X-ray crystallographers have long observed that proteins are very compact collapsed polymers. Still, the native structure that is captured in a protein crystal is, while perhaps representative, merely one of many that a protein may find itself in during the course of its function in the living cell. Ligands or water molecules enter and leave the cavities that can be resolved in many proteins. As such, the notion that proteins are simply three-dimensional, extremely compact objects [1] may be too simple. Indeed, the possibility that proteins may be better characterized by fractal geometry rather than as a compact three-dimensional object has been pointed out for some time [1–3]. This appears to be the case for protein surfaces, for which a fractal dimension of 2.1 to 2.4 is widely accepted [1,4]. Nevertheless, the fractal dimension of the protein itself based on several estimates has been argued to lie near 3 [1], though a number of studies also suggest smaller values [3,5–7]. For example, the radius of gyration has been found to scale with protein mass (or number of residues) with a dimension near 2.5 for proteins with more than 300 amino acids [3]. On the other hand, counting algorithms coupled with a series of scaling approximations have yielded a fractal dimension for the protein backbone that may lie closer to 3 [1]. We recently computed the mass fractal dimension for three proteins, cytochrome c, myoglobin, and green fluorescent protein, which are made up of about 100–230 amino acids, and found $D \approx 2.3$ [7]. This result is consistent with dispersion relations and the anomalous subdiffusion that we computed for these proteins [7], which are related to the mass fractal dimension by Alexander-Orbach theory [8]. Since the value of the mass fractal dimension influences protein dynamics and energy flow, a closer look at its value for proteins ranging widely in size seems worthwhile.

In this article, we compute the mass fractal dimension D for a set of 200 proteins whose structures are obtained from the Protein Data Bank (PDB). The number of amino acids, N, of the proteins in this set ranges from $N \approx 100$ to 11 000. We compute D with an approach directly related to its definition as described below, and compare D with the scaling of the radius of gyration with protein size. Both sets of results yield dimensions near 2.5. We find that D for larger proteins, with more than 1000 amino acids, settles around a value near 2.6, while it is smaller for smaller proteins. This result is consistent with our earlier computational study [7] of vibrational energy flow in three proteins mentioned above, where $D \approx 2.3$ was computed for three proteins with N from about 100 to 230. We also examine the extent to which a protein of a given configuration fills the volume within its surface, and find for this set of 200 proteins that roughly 25% and thus a sizable fraction of space within the protein surface is unfilled.

Evidence that protein molecules may be characterized by a fractal-like geometry has appeared in a variety of measurements. For instance, the anomalous temperature dependence seen in spin echo experiments revealed an interesting scaling relation for the vibrational mode density with mode frequency [9,10]. A theoretical underpinning for the variation of the vibrational density of states of a protein with vibrational frequency was provided by Alexander and Orbach [8], who assumed a correspondence between the vibrations of a protein and the vibrations of an object with fractal geometry. The scaling exponent characterizing the variation of the vibrational density with mode frequency of a fractal, $\bar{d}$, called the spectral dimension, is analogous to the Euclidean dimension in the Debye expression for the density of states. The value of the spectral dimension is generally smaller than the fractal dimension of the object, and reflects the connectivity or bonding of the atoms [11]. The spectral dimension for a number of modest-sized proteins was deduced from results of the spin echo experiments to range from 1.3 to 1.6 [9,10]; these values were corroborated by theoretical and computational work on fractal models of proteins [2,12]. In a recent study of 58 proteins ranging from $N \approx 100$ to 3600, Burioni *et al.* computed the spectral dimension directly from the density of states for these proteins [13]. The Gaussian network

*Corresponding author. Email address: dml@chem.unr.edu

model [14–16] was used to account for interactions among protein atoms, an approach that has provided reliable descriptions of the low-frequency vibrations of proteins, as seen by comparing computed and measured thermal fluctuations of C_α atoms [14]. The spectral dimension was found to range from about 1.3 to 2.0, and appears to increase logarithmically with protein size [13].

Correspondence between the vibrational properties of a protein and those of a fractal object provides a useful means to learn about vibrational energy flow in proteins and protein dynamics. Indeed, a number of studies of protein dynamics and energy fluctuations reveal fractal properties [17–20]. Alexander and Orbach derived relations between the spectral dimension, the fractal dimension of the object, and scaling exponents characterizing at least two important and related properties [8]. One of these is how the frequency of a protein's normal modes of vibration varies with wave number (i.e., a dispersion relation) at low frequency; the other describes how vibrational energy spreads in time. Thus, assuming that the vibrations of proteins correspond to those of a fractal object, both the spectral dimension and the mass fractal dimension of the protein are required to predict the dispersion relation for a protein and the diffusion of vibrational energy. The recent analysis by Burioni *et al.* provides a means to estimate the spectral dimension of a protein based on its size [13]. In this article we focus on the mass fractal dimension.

In the following section we describe the method we use to compute the mass fractal dimension D for each protein in our sample of 200 obtained from the PDB. In Sec. III we present results for D. Our calculation reveals that D approaches a value of about 2.6 for larger proteins, with over 1000 amino acids, and generally decreases to about 2.3 for smaller proteins with closer to 100 amino acids. We also discuss a calculation carried out to estimate the fraction of volume within the protein surface for a given protein configuration that is not filled by the protein, which we estimate to be about 25% on average by our method.

II. COMPUTATIONAL METHODS

The mass fractal dimension D is defined by

$$M \sim R^D, \tag{1}$$

where M is mass and R is a length scale. The dimension D can be computed for a single protein by plotting the mass of all atoms contained inside concentric spheres of radius R on a log-log scale. The slope gives D. We have carried out this calculation for 200 proteins ranging from $N \approx 100$ to 11 000 amino acids. The proteins, whose structures have all been obtained from the PDB, are listed in Table I by their PDB code. These 200 proteins include the 58 analyzed in Ref. [13].

Describing how we calculate D in practice is easiest by example. Figure 1 presents a log-log plot of the enclosed mass M of all protein atoms inside a sphere as a function of its radius R. The ten sets of points, where each set appears to fall on a line, have been computed for concentric spheres centered at ten α-carbons, which in this case happen to be the ten nearest to the center of mass of the protein 1MZ5, which has 622 amino acids. Data are shown for R ranging from 5 to 20 Å. Most of the points lie close to straight lines, as we have found to be typically the case. The length scale of this particular protein is significantly larger than 20 Å, but we nevertheless only calculate M for R up to 20 Å to avoid finite-size effects when computing D for the interior of the protein. In fact, to avoid possible finite-size effects when computing D for the interior of the smallest proteins in our sample set, we have computed M for R up to 16 Å for proteins with up to 200 amino acids, and up to 18 Å for proteins with from 200 to 400 amino acids. Nevertheless, the results that we report below are very similar to those that we obtain when we calculate M as a function of R up to 20 Å for all proteins. We shall also present results for D calculated in different regions of the protein, other than the center. In these cases D is obtained by using atoms closer to the surface as centers in our calculation, and the cutoff of 20 Å may not exclude surface atoms. The lower value of $R=5$ Å was chosen after considering 3–8 Å as a lower limit, and fitting lines to these. The largest correlation coefficient was found with 5 Å, since significant deviations from the best-fit line were typically found for points with smaller R. The average value of the slopes of the lines in Fig. 1 gives us an estimate for D for the protein 1MZ5, which we calculate by averaging slopes obtained for such plots using all of the C_α's of the protein backbone as centers.

We note for later discussion that the correspondence between a protein and a fractal object allows us to relate the mass fractal dimension D and the spectral dimension $\bar{d}$ to scaling exponents relating how vibrational energy flow varies with time and how vibrational mode frequency scales with wave number [8]. The spectral dimension $\bar{d}$ is defined by [8]

$$\rho(\omega) \sim \omega^{\bar{d}-1}. \tag{2}$$

The scaling of mode frequency with wave number k then obeys the relation [8]

$$\omega \sim k^{D/\bar{d}}. \tag{3}$$

The variance of a vibrational wave packet spreads in time as [8]

$$\langle R^2 \rangle \sim t^{\bar{d}/D}. \tag{4}$$

For a set of polymers of varying length N (or mass M) we may also take D to describe the scaling of the radius of gyration R_G with, say, M,

$$R_G \sim M^{1/D}. \tag{5}$$

For a given protein configuration taken from the PDB we compute the radius of gyration as

$$R_G = \sqrt{\sum_i m_i \mathbf{r}_i^2 \Big/ \sum_i m_i}, \tag{6}$$

where the sum is over each atom i of mass m_i, and distance $\mathbf{r}_i$ from the center of mass. We shall see that the value of D

TABLE I. List of all protein molecules with their PDB code; number of amino acids, N; radius of gyration (Å) for the PDB coordinates, R_G; mass fractal dimension D; void fraction f_V using 1.5-Å- and 2.0-Å-radius atoms. The PDB code names for the 58 proteins analyzed in Ref. [13] are written with capital letters.

Name	N	R_G	D	f_V (1.5)	f_V (2.0)	Name	N	R_G	D	f_V (1.5)	f_V (2.0)
9RNT	104	12.450	2.300	0.228	0.106	1SOM	528	22.402	2.520	0.274	0.146
1r9h	118	13.490	2.229	0.248	0.125	1E3Q	532	22.809	2.512	0.278	0.150
1r2i	143	14.228	2.322	0.248	0.129	1CRL	534	22.131	2.514	0.277	0.148
1r4v	145	15.980	2.282	0.234	0.115	1AKN	547	23.343	2.495	0.261	0.136
1r67	151	14.295	2.330	0.238	0.115	1r5t	554	23.029	2.536	0.284	0.156
1BVC	153	15.285	2.276	0.222	0.101	2r2f	571	25.535	2.498	0.281	0.155
1rda	155	15.505	2.310	0.237	0.116	1rq4	572	23.701	2.447	0.233	0.114
1rf7	159	15.387	2.332	0.247	0.125	1r1y	574	23.414	2.451	0.234	0.113
1G12	167	14.862	2.379	0.247	0.121	1rps	574	23.696	2.444	0.242	0.121
1rm8	169	15.141	2.384	0.250	0.130	1rq3	574	23.644	2.445	0.238	0.117
3rab	169	15.220	2.375	0.250	0.125	1CF3	581	23.266	2.539	0.288	0.161
1AMM	174	16.587	2.380	0.251	0.125	1rqi	598	24.332	2.517	0.265	0.138
4GCR	185	16.694	2.340	0.254	0.126	1EX1	602	24.922	2.536	0.295	0.166
1KNB	186	18.425	2.359	0.241	0.125	1A14	612	26.164	2.513	0.265	0.141
1CUS	197	15.241	2.433	0.240	0.126	1rfv	615	25.559	2.513	0.282	0.155
1IQQ	200	16.692	2.360	0.234	0.111	1ry2	615	24.082	2.528	0.264	0.138
2AYH	214	16.081	2.406	0.261	0.136	1MZ5	622	27.128	2.510	0.271	0.142
1r5a	214	17.049	2.342	0.245	0.123	1rfz	637	23.458	2.547	0.263	0.138
1rei	214	17.155	2.395	0.280	0.154	1rli	648	25.005	2.522	0.275	0.151
1AE5	223	16.455	2.444	0.254	0.135	1r4l	655	25.141	2.478	0.252	0.129
1r18	223	16.855	2.386	0.248	0.124	1CB8	674	27.508	2.507	0.265	0.138
1rm9	223	16.912	2.393	0.281	0.157	1HMU	674	27.500	2.506	0.270	0.143
1rmm	224	17.081	2.399	0.291	0.166	1r65	680	25.978	2.542	0.282	0.155
1emb	225	17.138	2.403	0.254	0.131	1rib	680	26.047	2.539	0.295	0.167
1rw7	235	16.376	2.429	0.258	0.133	1rsv	681	25.897	2.540	0.286	0.159
1LST	239	17.732	2.397	0.261	0.136	1A47	683	25.545	2.524	0.285	0.157
1rxh	239	16.798	2.437	0.237	0.118	1CDG	686	25.397	2.526	0.291	0.162
1r9c	243	17.709	2.410	0.266	0.142	1DMT	696	26.363	2.487	0.255	0.130
1rjk	250	17.898	2.396	0.250	0.128	1r7i	747	25.725	2.534	0.285	0.158
1rk3	250	17.991	2.394	0.257	0.133	1r31	751	25.851	2.539	0.281	0.153
1r5l	251	17.842	2.387	0.248	0.125	1A4G	780	27.888	2.589	0.300	0.169
1ri1	252	18.049	2.388	0.249	0.126	1kko	802	26.384	2.548	0.292	0.164
1rkh	253	17.966	2.397	0.250	0.125	1rtw	809	28.532	2.508	0.257	0.132
1ray	258	17.473	2.427	0.275	0.148	1ry5	822	28.867	2.446	0.247	0.124
1rxf	264	18.168	2.426	0.255	0.133	1rzh	822	28.767	2.446	0.246	0.122
1rxg	275	18.577	2.434	0.268	0.145	1rgn	823	28.921	2.448	0.248	0.124
1A06	279	19.986	2.376	0.241	0.118	1rqk	824	29.116	2.447	0.250	0.127
1NAR	289	18.337	2.465	0.276	0.151	1rov	834	28.440	2.537	0.271	0.144
1r53	291	19.338	2.398	0.250	0.126	1rj8	840	29.206	2.584	0.285	0.155
1r0t	292	18.170	2.450	0.263	0.139	1kzy	854	31.921	2.439	0.254	0.131
1A48	298	19.823	2.386	0.253	0.128	1rqp	873	27.563	2.558	0.277	0.148
1rjb	298	19.179	2.442	0.255	0.129	1km0	901	31.056	2.542	0.288	0.161
1A3H	300	17.602	2.494	0.274	0.148	1kw2	908	34.845	2.377	0.282	0.153
1rb7	304	18.861	2.462	0.283	0.157	1ktw	914	35.434	2.462	0.230	0.112
1SBP	309	19.408	2.435	0.270	0.144	1rvu	929	28.338	2.569	0.279	0.152
1rft	309	19.028	2.415	0.246	0.124	1kre	950	29.673	2.567	0.277	0.148
1rz5	309	20.937	2.401	0.236	0.115	1rzp	988	27.400	2.579	0.287	0.158

TABLE I. (*Continued.*)

Name	N	R_G	D	f_V (1.5)	f_V (2.0)	Name	N	R_G	D	f_V (1.5)	f_V (2.0)
1rkp	311	19.245	2.406	0.243	0.121	1HTY	1014	29.823	2.602	0.271	0.144
1A5Z	312	19.868	2.381	0.256	0.134	1KCW	1017	28.336	2.435	0.250	0.126
1A1S	313	19.389	2.430	0.264	0.141	1kzg	1032	35.874	2.571	0.253	0.128
1ADS	315	18.947	2.469	0.280	0.154	1ipj	1088	32.933	2.557	0.256	0.131
1rya	320	20.476	2.429	0.264	0.138	1ivx	1238	31.718	2.609	0.295	0.164
2ren	320	19.730	2.444	0.260	0.133	1ktv	1264	38.776	2.445	0.274	0.146
1A40	321	19.931	2.451	0.267	0.139	1ksi	1282	32.392	2.605	0.278	0.150
1A54	321	20.036	2.452	0.260	0.134	3req	1345	33.495	2.566	0.279	0.151
1r6w	321	20.365	2.450	0.254	0.129	1rjw	1356	32.961	2.571	0.275	0.148
1r66	322	18.946	2.464	0.258	0.133	1kr2	1395	34.900	2.507	0.280	0.152
1r6d	324	18.888	2.465	0.243	0.120	1kqo	1398	34.958	2.506	0.270	0.142
1r0r	325	18.085	2.483	0.269	0.141	1kev	1404	32.726	2.585	0.275	0.147
1ryo	325	19.403	2.459	0.259	0.134	1jrq	1437	33.104	2.600	0.238	0.116
1A0I	332	23.332	2.347	0.253	0.128	1kor	1538	34.081	2.596	0.234	0.115
1ri6	333	18.633	2.493	0.276	0.146	1ivh	1548	34.292	2.585	0.242	0.121
1re8	337	20.008	2.448	0.269	0.141	1rx0	1573	34.139	2.580	0.276	0.148
3PTE	347	18.949	2.490	0.280	0.150	1rp7	1602	33.291	2.617	0.285	0.155
1A26	351	20.888	2.402	0.268	0.144	1ky4	1712	35.539	2.559	0.270	0.145
1rl9	356	20.068	2.459	0.258	0.133	1ky5	1720	34.567	2.603	0.289	0.160
1BVW	360	19.205	2.473	0.275	0.149	1re5	1767	35.684	2.585	0.275	0.149
8JDW	360	19.068	2.508	0.261	0.146	1nlz	1804	39.888	2.559	0.289	0.160
1rdq	360	19.892	2.493	0.256	0.130	1k93	1884	40.681	2.485	0.285	0.155
1rgy	360	19.668	2.494	0.270	0.145	1nu1	2105	49.292	2.497	0.253	0.129
1r2v	361	20.113	2.475	0.276	0.151	1kf6	2138	45.405	2.551	0.264	0.139
1r7o	362	19.239	2.477	0.278	0.152	4rub	2348	40.220	2.662	0.288	0.160
1r3q	365	19.807	2.475	0.283	0.155	1KEK	2462	38.567	2.642	0.283	0.154
1rgz	370	19.472	2.505	0.263	0.137	1B0P	2462	38.645	2.644	0.220	0.102
1r5y	385	20.108	2.468	0.269	0.143	1rfm	2680	44.033	2.528	0.251	0.126
7ODC	387	23.524	2.447	0.259	0.133	1ggj	2908	41.297	2.686	0.262	0.137
1OYC	399	20.347	2.493	0.259	0.145	1rxc	2970	41.973	2.641	0.293	0.162
1rom	399	21.328	2.442	0.253	0.131	1ijg	3084	50.787	2.569	0.262	0.138
1A39	401	20.730	2.450	0.266	0.139	1K83	3494	48.090	2.576	0.275	0.148
16PK	415	23.146	2.430	0.264	0.136	1I3Q	3542	48.494	2.571	0.279	0.152
1r61	415	21.957	2.505	0.272	0.146	1I50	3558	48.503	2.572	0.275	0.148
1DY4	441	20.459	2.484	0.267	0.139	1r5u	3602	47.353	2.586	0.260	0.137
1BU8	446	25.039	2.460	0.269	0.140	1fqv	3696	57.863	2.474	0.260	0.137
1r9o	455	22.430	2.461	0.246	0.124	1cw3	3952	54.134	2.658	0.302	0.172
1rxj	471	21.401	2.489	0.252	0.126	1mfr	4104	52.990	2.596	0.282	0.153
1rjp	474	21.651	2.511	0.292	0.162	1kyo	4459	59.476	2.529	0.258	0.133
1rk6	475	21.562	2.513	0.284	0.154	1jro	4840	64.391	2.628	0.281	0.154
1r1k	477	22.769	2.452	0.227	0.108	1f52	5616	53.974	2.631	0.280	0.152
1rty	479	21.578	2.497	0.256	0.133	1fpy	5808	53.636	2.630	0.276	0.151
1AC5	483	22.151	2.453	0.248	0.136	1kyi	5904	63.041	2.589	0.254	0.131
1LAM	484	24.146	2.484	0.279	0.150	1nr7	5952	70.379	2.606	0.260	0.136
1reo	484	23.338	2.477	0.263	0.139	1g0u	6296	59.853	2.623	0.270	0.144
1CPU	495	22.975	2.500	0.292	0.162	1g65	6366	59.827	2.630	0.269	0.143
3COX	500	21.999	2.508	0.283	0.154	1mx9	6378	70.549	2.633	0.277	0.150
1rxy	500	22.364	2.516	0.288	0.160	1ryp	6386	59.833	2.634	0.275	0.146

TABLE I. (*Continued.*)

Name	N	R_G	D	f_V (1.5)	f_V (2.0)	Name	N	R_G	D	f_V (1.5)	f_V (2.0)
1r0s	502	25.096	2.413	0.251	0.129	1kp8	7350	63.853	2.508	0.241	0.119
1r12	502	24.830	2.419	0.255	0.131	1mcz	8384	70.402	2.677	0.289	0.160
1A65	504	21.729	2.539	0.286	0.158	1mt5	8592	65.821	2.609	0.293	0.162
1rkm	517	23.869	2.474	0.269	0.412	1fnt	9110	81.070	2.604	0.239	0.117
2rkm	519	23.242	2.508	0.279	0.151	1hto	11448	80.642	2.645	0.275	0.146

obtained in this way compares well with the average value of D we obtain for the individual proteins in our set.

III. RESULTS AND DISCUSSION

A. Mass fractal dimension

We have already introduced Fig. 1, which presents ten log-log plots of the enclosed mass of all protein atoms inside a sphere of radius R as a function of R, centered at one of the ten nearest C_α's to the center of mass of the protein 1MZ5. The ten lines that best fit the ten sets of points shown in Fig. 1 have an average slope of 2.798 ± 0.197, where the error that we report is two standard deviations (95% confidence limit). The correlation coefficients for the lines that best fit each of the ten sets of data range from 0.9985 to 0.9997.

We now compute in this way the slopes for sets of points obtained using as centers the nearest 10% of all C_α's from the center of mass of the protein. We find the value of D using this inner 10% of C_α's, $D_{10\%}$, to be 2.737 ± 0.249 for 1MZ5. In fact, 1MZ5 is quite typical. Carrying out the same analysis for all 200 proteins in our set, we find that $D_{10\%}$ is 2.761 ± 0.164. If we now choose as centers the next closest 10% of the C_α's from the center of mass of the protein, we find that the points on the $\log_{10} M$ versus $\log_{10} R$ plot for one C_α center similarly lie close to a line. However, the average slope of all of these is somewhat smaller than $D_{10\%}$, in this case 2.673 ± 0.175. Indeed, we find that D usually becomes smaller when we compute its value using concentric spheres that are centered on C_α's closer to the exterior of the protein. Using the outermost 10% of the C_α's as centers for the concentric spheres we find a dimension of 2.215 ± 0.229.

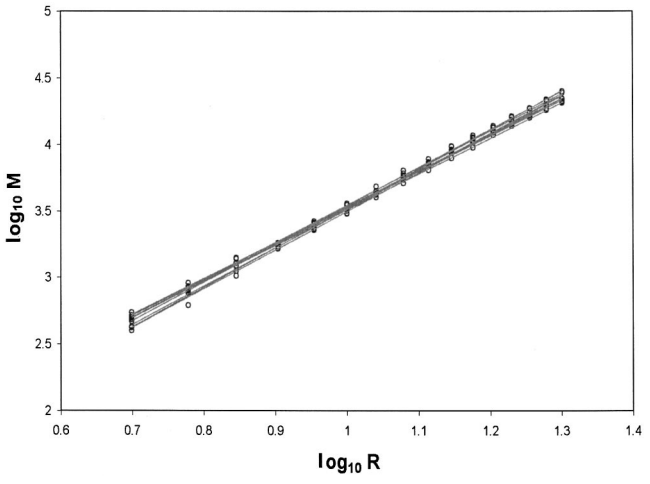

FIG. 1. Plot of $\log_{10} M$ vs $\log_{10} R$ for 1MZ5, where values of M are the masses enclosed by concentric spheres of radius R centered at a backbone atom. Each of the ten sets of points through which lines are fitted corresponds to a center in our calculation, which is one of the ten closest C_α's to the center of mass of the protein. The correlation coefficients for the lines that best fit the data range from 0.9985 to 0.9997.

These trends are shown in Fig. 2(a) for our set of 200 proteins. We thus see that D is not a uniform quantity, but decreases on average toward the exterior of the protein. This is likely due to the greater influence of the surface dimension, which for proteins has been found to be 2.1 to 2.4 [1], on our computed value of D as the calculation is carried out closer to the protein's surface. The influence of the protein surface on the computed values of D is supported by comparing Fig. 2(a) with Fig. 2(b), which is similar to Fig. 2(a) but only includes results for the 63 proteins with 200–400 amino acids. We find that the value of $D_{10\%}$ for the smaller proteins is the same as for the whole set, 2.76, but the average value of D is somewhat smaller for the smaller proteins, 2.43 compared to 2.49. The cutoff radius used in the calculation of $D_{10\%}$ for the proteins in Figs. 2(a) and 2(b) excludes surface atoms. However, more and more atoms near the protein surface are included when the calculation of D is centered at C_α's farther from the center of mass, and this effect is greater for the set of smaller proteins. We note, however, that this trend is not so apparent for the larger proteins, as is illustrated in Fig. 2(c) for the ten largest proteins in our set. The trend is different for the larger proteins because the center of mass often lies outside the denser centers of the individual globules of the quaternary structure.

In Fig. 3 we show how the average value of D that we calculate for each protein varies with protein size. Results are plotted for the calculation of D using all C_α's as centers, and also using only the nearest 10% to the center of mass of the protein, $D_{10\%}$. We compute the value of $D_{10\%}$ to be 2.761 ± 0.164 for all the proteins, a value that does not change much with protein size, as Figs. 2(a) and 2(b) suggest. For example, we find that for proteins with at least 1000 amino acids $D_{10\%}$ is 2.734 ± 0.209. The value of the mass fractal dimension D computed using all C_α's as centers in the calculation is 2.489 ± 0.172 for all proteins. The mass fractal dimension as obtained by averaging its value over each protein molecule appears to depend on the size of the protein. For larger proteins, with at least 1000 amino acids, we find D is 2.584 ± 0.113. For smaller proteins, with $N < 1000$, we find D is 2.456 ± 0.136. Interestingly, $D_{10\%}$ and D appear to converge to similar values for larger proteins, likely due to the

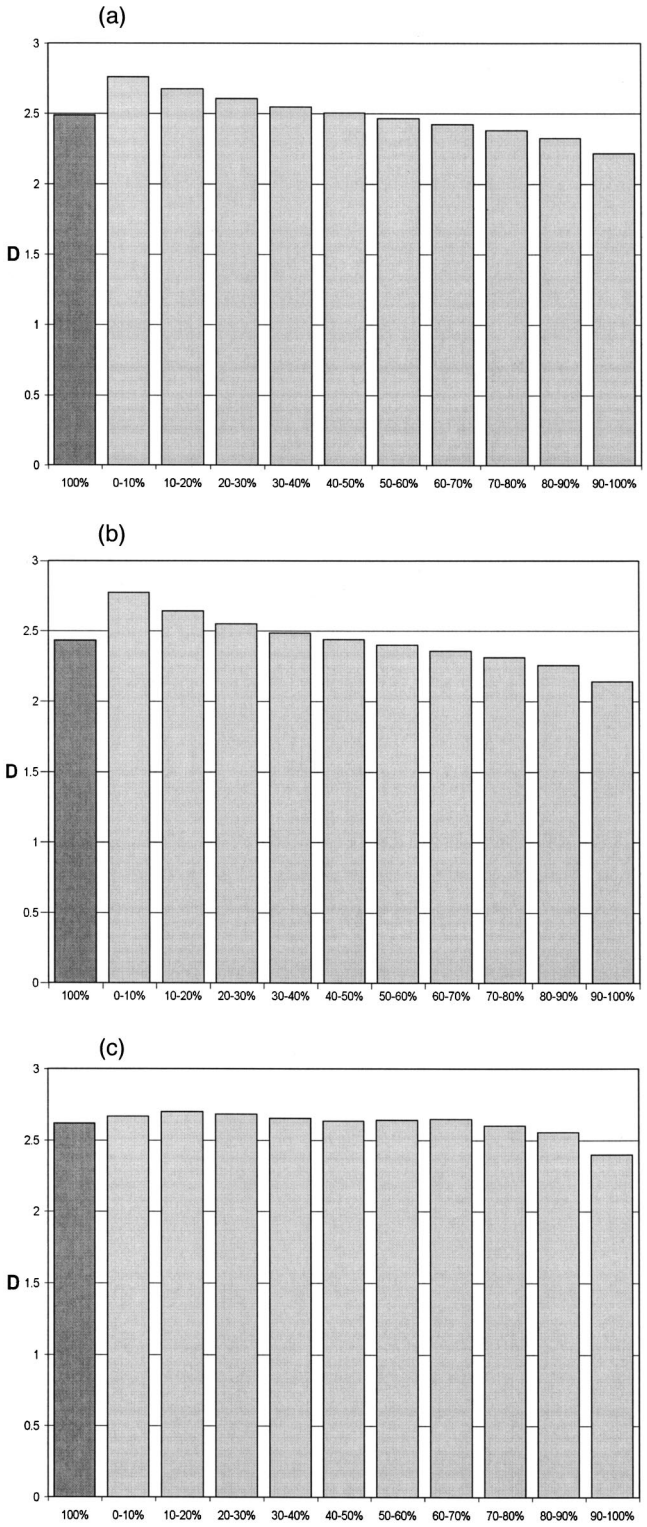

FIG. 2. (a) Average values of the mass fractal dimension D computed for the 200 proteins using as centers in the calculation the nearest 10%, 10–20 %, 20–30 %, etc., of the C_α's from the center of mass of the protein (light gray). Also shown is D computed using all C_α's as centers (dark gray). (b) Same as (a), but for the 63 proteins with 200–400 amino acids. (c) Same as (a), but for the ten largest proteins in the data set.

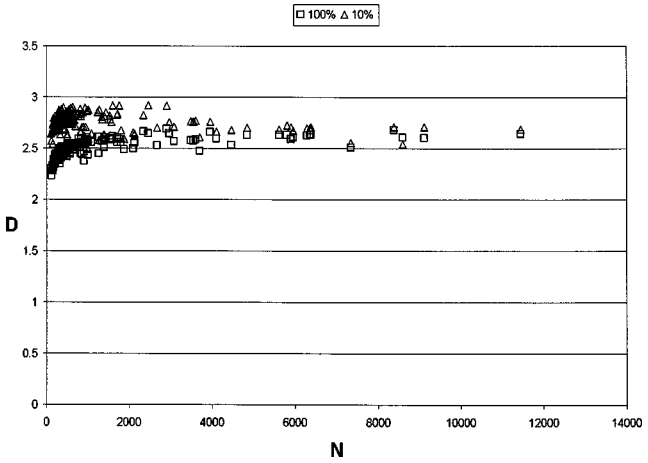

FIG. 3. Plot of the mass fractal dimension D (squares) and its estimate using as centers in the calculation the nearest 10% of all C_α's to the center of mass of the protein, $D_{10\%}$ (triangles), as a function of the number of amino acids of each protein, N. For this set of 200 proteins we find D is 2.489 ± 0.172 and $D_{10\%}$ is 2.761 ± 0.164.

fact that $D_{10\%}$ for the largest proteins is not necessarily larger than that computed in other parts of the protein, as noted above and illustrated in Fig. 2(c).

We plot the radius of gyration R_G versus protein size N in Fig. 4. The slope of the line for this log-log plot is 0.390 and the correlation coefficient is 0.9893. The value of D from the data, which is 1/slope, is thus 2.56, in good agreement with the average value of the mass fractal dimension computed above. In fact, if we switch the x and y axes so that the slope itself now gives us an estimate for D we find from a best fit a value of 2.50. We observe significant dispersion in this plot. Arteca has pointed out that one can select a set of "most compact" proteins, those through which in a plot like that in Fig. 4 one may draw a line with the largest slope [3]. Arteca studied 373 proteins ranging in size from $N \approx 100$ to 900. For

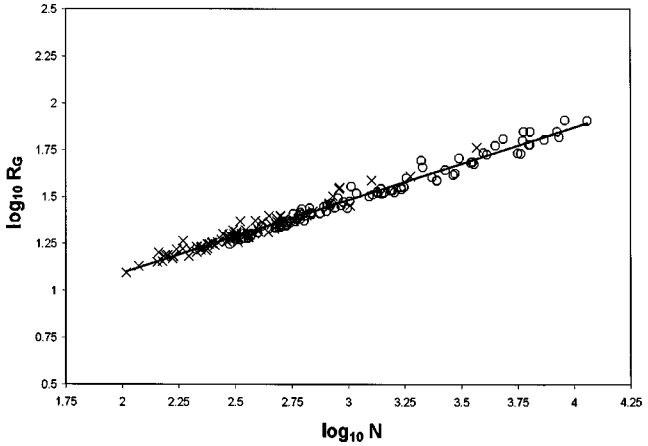

FIG. 4. Plot of $\log_{10} R_G$ vs $\log_{10} N$ for the 200 proteins in the set. Data are labeled as $\times$ for a protein with lower-than-average D, i.e., $D < 2.489$; and $\bigcirc$ for a protein with $D > 2.489$. The best-fit line, with correlation coefficient 0.9893, is drawn through the data. The slope of this line, which can be interpreted as $1/D$, is 0.390, giving an estimate of 2.56 for the dimension.

the most compact proteins with at least 300 amino acids, analysis of R_G versus N gave an average dimension of 2.48 [3]. We also attempt to correlate proteins of relatively high D with relatively small R_G, which could indicate that higher D correlates with a more compact object. In Fig. 4 we plot as $\times$ those proteins with a lower-than-average D, i.e., $D<2.489$, and $\bigcirc$ those proteins with a higher-than-average D, i.e., $D>2.489$. We see clearly that smaller D is found for smaller proteins, as seen already in Fig. 3. We find that 72% of the $\bigcirc$'s lie below the best-fit line in the plot, and so have a relatively small R_G, indicating that a higher-than-average value of the mass fractal dimension indeed correlates with a more compact protein. Similarly, we find that 57% of the $\times$'s lie above the line.

The average result for the mass fractal dimension that we find for this set of 200 proteins, 2.49, agrees quite well with the mass fractal dimensions that we previously computed for cytochrome c, myoglobin, and green fluorescent protein, which we found to be 2.30, 2.36, and 2.42, respectively [7]. These values lie below 2.49, and indeed there is a visible trend in Fig. 3 whereby smaller proteins are characterized by a smaller D, reaching ≈ 2.3 for $N \approx 100$, due to a larger contribution of the atoms near the protein surface in the calculation of D for small proteins. We note that the average value of D that we computed for cytochrome c, myoglobin, and GFP, 2.36, agrees well with the average value of D that we obtained from the spectral dimensions, dispersion relations, and vibrational energy diffusion calculations for these proteins with Eq. (2)–(5), which was 2.25 [7]. Both of these values have an error of ± 0.2. These results are consistent with a correspondence between the vibrational properties of a protein and those of a fractal object. However, the calculations presented above suggest that there is in fact no unique D that characterizes a protein. For the proteins in our sample D typically ranges from 2.75 to 2.25, depending on where in the protein we center our computation of D, and is usually larger as we compute it near the center of the protein and smaller when more of the surface is included. Protein vibrations at low frequency involve atoms throughout the protein. The fact that we find the average D computed for a protein similar to the value of D that we obtain from the vibrational dynamics, using the Alexander-Orbach relations, suggests to us that the average D is the appropriate mass scaling dimension for characterizing properties of protein vibrations.

In addition to the mass fractal dimension, which we report and analyze here, vibrational energy flow in a protein is also influenced by the spectral dimension $\bar{d}$. The spectral dimension has been suggested by Burioni *et al.* based on a computational study of 58 proteins to vary logarithmically with N [13]. For proteins with about 100 amino acids its value lies near $\bar{d} \approx 1.3$ [7,13]. For proteins with more than 1000 amino acids $\bar{d} \approx 2$, which is the largest value that it can have for a harmonic fractal object to remain thermodynamically stable [13]. We find that D is about 2.6 and is largely independent of N for sufficiently large proteins, with more than about 1000 amino acids, and is smaller for smaller proteins, about 2.3 for proteins with about 100 amino acids. We thus conclude that the exponent $a = D/\bar{d}$, which characterizes the variation of vibrational mode frequency with wave number, $\omega \sim k^a$ [Eq. (3)], ranges from about $2.3/1.3 \approx 1.8$ for small proteins ($N \approx 100$) to about $2.6/2 \approx 1.3$ for large proteins ($N > 1000$).

B. Fraction of empty space within the protein surface

The above analysis reveals that proteins are not completely compact objects, but must also have "empty" or "void" space. In this subsection we examine the relative volume of such void space. There is a fair amount of arbitrariness in defining such a quantity. For one thing, we shall calculate the fractional void space within a protein with a fixed configuration, which means we must first establish a protein surface. Then, using a reasonable volume for the protein atoms, we can compute the fraction of space that is filled by them and the remaining void fraction.

We first estimate the surface in a fashion inspired by the "ball rolling" algorithm used in the computation of the surface area and dimension of a protein [1]. We first superpose the protein coordinates with a grid in three dimensions, each point 1 Å from its neighbor. This allows us to approximate the space occupied by the protein by a collection of cubic cells 1 Å on each side. We then identify which cells are "protein" cells and which cells lie outside. We enclose around each protein atom a 3-Å-radius sphere and count as protein cells all of those 1-Å cubic cells whose centers lie within this sphere. In this way we fill the cells belonging to the protein. The use of a 3-Å-radius sphere is of course somewhat arbitrary, but has been used for similar calculations [1]. Smaller spheres give rise to a more porous protein surface; more space that we might reasonably call void would be counted instead as lying outside the protein. A larger sphere would tend to fill in the spaces left by indentations in the protein surface that we would otherwise reasonably decide lie outside the protein; we would then be ultimately designating much of this space as void. We have found, as others have in earlier work on the dimension of protein surfaces [1], that searching for protein atoms in a 3-Å sphere provides a reasonable balance of these effects.

We then have a means to label cells of the grid as "protein" and "outside" cells. The N_P cells that we call "protein" are those enclosed by the protein surface and may be "filled" by a protein atom or may be "void." Cells that we call "outside" are beyond the boundary of the protein, but we emphasize that the surface may be very rugged and is typically pockmarked with deep and narrow craters. A cross-sectional cut near the center of a protein may contain many "outside" cells, as we see in an example below. We must now decide which, and how many, N_F, cells are filled and which and how many, N_V, are void. The void fraction f_V is then given by

$$f_V = \frac{N_V}{N_P} = \frac{N_V}{N_V + N_F}. \tag{7}$$

To estimate the space filled by the protein atoms, we assume each atom is a sphere of radius 1.5 Å. This radius is rather large for a molecule containing C, N, and O atoms, but we must also compensate for the fact that we do not explicitly account for H, so that OH, CH, methyl groups, etc., are all counted as one "atom." In this case a radius of 1.5 Å

FIG. 5. Cross section of the protein 1A4G superposed on a lattice of 1-Å cells, as described in text. White cells are computed to lie outside the protein surface. Black cells contain a protein atom and dark gray cells contain part of the volume of a protein atom. Light gray cells represent the empty or void spaces within the protein surface. The cross section in (a) has been computed with the algorithm described in the text. In (b) and (c) we remove the outer layer of void cells, which are an artifact of our computation of the protein surface and are removed to compute the fraction of void space f_V within the protein surface. In (b) and (c) we use in our computation a sphere of 1.5 and 2.0 Å, respectively, for each protein atom.

seems reasonable. We shall also compare with results using a more conservative radius of 2.0 Å. In any case, our aim is to determine if a substantial region inside the protein in a given configuration can be called void, and it does not matter much if we find that 20% or 30% of the protein's volume is void. We would like to know if the void space estimated in a reasonable way turns out to be, say, 20% or instead 2% of the space within the surface of the protein. We now fill cells whose centers are enclosed by any part of the 1.5-Å- or 2.0-Å-radius sphere representing a protein atom. Such volumes may be cubes 3 or 4 Å on each side, but the volume

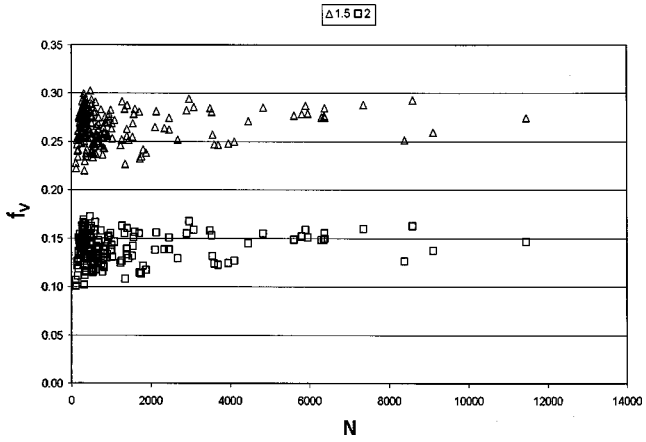

FIG. 6. Plot of the void fraction f_V calculated using as radius for each protein atom a value of 1.5 (triangles) and 2.0 (squares) Å, as a function of the number of amino acids of each protein, N. For this set of 200 proteins we find f_V is 0.265 ± 0.035 using the smaller radius and is 0.139 ± 0.030 with the larger.

around the protein atom may also appear as other shapes built up from 1-Å cubic cells if the center of that cell happens to be enclosed by the spherical shell of the atom. Overlapping atom volumes are possible and not unlikely given the relatively large volume that we ultimately place around each atom.

We illustrate our calculation in Fig. 5, which shows a discretized cross section of the protein 1A4G, which contains 780 amino acids. The white background, as well as some white cells that appear contained in the protein, are all "outside" cells, not counted in estimating the void fraction. That some white cells appear to lie inside the protein is merely due to the display of a cross section, and arise from craters in the established surface above or below the cross section. The black 1-Å cubic cells contain protein atoms. In addition, as described above, adjacent cells are also counted as filled space. These are shown in dark gray. We notice that there are what appear to be islands of dark gray cells in the white region. These result from protein atoms just above or below the cross-sectional cut of the protein. The light gray cells are "void" cells. These lie inside the surface of the protein but outside the cells enclosing protein atoms. We notice that there appears to be a halo of void cells surrounding the protein in Fig. 5(a), which is an artifact of the calculation of the protein surface. We remove all the void cells around the edge in computing the void fraction. The resulting cross section, after removing the layer of void cells from the edge of the protein, is plotted in Fig. 5(b). The same cross section as in Fig. 5(b) is also shown in Fig. 5(c), but this time we compute the filled space using protein atoms that are spheres with a radius of 2.0 Å. The number of light-gray void cells is clearly smaller than in Fig. 5(b) but still fills a sizable fraction of the protein cross section. The void fraction f_V for each protein can be computed with Eq. (7) by counting all of the light gray, void cells, which gives N_V, and the total number of gray and black cells, which gives the total number of protein cells, N_P. For the protein shown in Fig. 5 we obtain $f_V = 0.30$ using spherical atoms with a 1.5 Å radius, and $f_V = 0.17$ using a 2.0 Å radius. Results for all of the proteins are

plotted in Fig. 6. Using the smaller radius, which we consider a reasonable estimate, we find for the 200 proteins in the set that f_V is 0.265 ± 0.035. With the even more conservative 2.0 Å radius we find f_V is 0.139 ± 0.030. We thus find a substantial fraction of space inside the protein is empty, a result that is qualitatively consistent with $D \approx 2.5$ computed above.

IV. CONCLUDING REMARKS

We have computed the mass fractal dimension for a set of 200 proteins ranging from about 100 to about 11 000 amino acids. For proteins with at least 1000 amino acids the dimension is 2.6 and does not appear to vary much with size. The dimension is smaller for smaller proteins, around 2.3 for proteins with about 100 amino acids. The mass fractal dimension for the 200 proteins in our set is 2.489 ± 0.172. This value of D is the same as the value of the scaling exponent for the variation of protein mass with radius of gyration that best fits our data. The value of the mass fractal dimension for each protein is itself an average value over all regions of the protein. Near the center of mass of each protein we find the mass fractal dimension for this set to be 2.761 ± 0.164. It is the average D over the whole protein, about 2.5 for this set, which corresponds most closely to the mass fractal dimensions we have obtained by studying the vibrations of several proteins using the Alexander-Orbach relations.

A mass fractal dimension of 2.5 indicates that a protein is not a completely compact three-dimensional collapsed polymer. We have computed the fraction of volume within the protein surface for each protein in its PDB configuration that is filled and empty. We have indeed found, using reasonable estimates for the protein surface and volume of atoms, that less than 80% of the protein volume is filled, consistent with a mass fractal dimension less than 3.

The computed mass fractal dimensions, together with the recently computed spectral dimensions by Burioni *et al.*[13] for 58 proteins spanning a similar size range and included in our set, allow us to estimate a range of values for scaling exponents characterizing vibrational energy flow in proteins. The variance of a vibrational wave packet spreads in time subdiffusively as $\langle R^2 \rangle \sim t^{d/D}$ [8]. Our results combined with those of Ref. [13] indicate that for proteins with at least 100 amino acids the exponent ranges from ≈ 0.55 for the smaller proteins to ≈ 0.75 for proteins with at least 1000 amino acids.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF CHE-0112631), the Camille and Henry Dreyfus Foundation, and the Research Corporation.

[1] T. G. Dewey, *Fractals in Molecular Biophysics* (Oxford University Press, New York, 1997).
[2] R. Elber, in *The Fractal Approach to Heterogeneous Chemistry*, edited by D. Avnir (John Wiley and Sons, Chichester, 1989), p. 345.
[3] G. A. Arteca, Phys. Rev. E **51**, 2600 (1995).
[4] M. Lewis and D. C. Rees, Science **230**, 1163 (1985).
[5] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).
[6] S.-H. Chen and J. Teixeira, Phys. Rev. Lett. **57**, 2583 (1986).
[7] X. Yu and D. M. Leitner, J. Chem. Phys. **119**, 12673 (2003).
[8] S. Alexander and R. Orbach, J. Phys. (France) Lett. **43**, L625 (1982).
[9] H. J. Stapleton, J. P. Allen, C. P. Flynn, D. G. Stinson, and S. R. Kurtz, Phys. Rev. Lett. **45**, 1456 (1980).
[10] A. R. Drews, B. D. Thayer, H. J. Stapleton, G. C. Wagner, G. Giugliarelli, and S. Cannistraro, Biophys. J. **57**, 157 (1990).
[11] D. ben-Avraham, Phys. Rev. B **47**, 14559 (1993).
[12] R. Elber and M. Karplus, Phys. Rev. Lett. **56**, 394 (1986).
[13] R. Burioni, D. Cassi, F. Cecconi, and A. Vulpiani, Proteins: Struct., Funct., Bioinf. **55**, 529 (2004).
[14] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[15] T. Haliliglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[16] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Phys. Rev. Lett. **80**, 2733 (1998).
[17] A. E. García, R. Blumenfeld, G. Hummer, and J. A. Krumhansl, Physica D **107**, 225 (1997).
[18] D. A. Lidar, D. Thirumalai, R. Elber, and R. B. Gerber, Phys. Rev. E **59**, 2231 (1999).
[19] T. Y. Shen, K. Tai, and J. A. McCammon, Phys. Rev. E **63**, 041902 (2001).
[20] P. Carlini, A. R. Bizzarri, and S. Cannistraro, Physica D **165**, 242 (2002).