# Proteins: Coexistence of Stability and Flexibility

Shlomi Reuveni,[1] Rony Granek,[2] and Joseph Klafter[1]
[1]School of Chemistry, Tel-Aviv University, Tel-Aviv 69978, Israel
[2]Department of Biotechnology Engineering, Ben-Gurion University, Beer Sheva 84105, Israel

(Received 6 March 2008; published 19 May 2008)

We introduce an equation for protein native topology based on recent analysis of data from the Protein Data Bank and on a generalization of the Landau-Peierls instability criterion for fractals. The equation relates the protein fractal dimension $d_f$, the spectral dimension $d_s$, and the number of amino acids $N$. Deviations from the equation may render a protein unfolded. The fractal nature of proteins is shown to bridge their seemingly conflicting properties of stability and flexibility. Over 500 proteins have been analyzed ($d_f$, $d_s$, and $N$) and found to obey this equation of state.

DOI: 10.1103/PhysRevLett.100.208101 PACS numbers: 87.14.E−, 05.40.−a, 05.45.Df, 87.15.hp

Two seemingly conflicting properties of native proteins, such as enzymes and antibodies, are known to coexist. While proteins need to keep their specific native fold structure thermally stable, the native fold displays the ability to perform flexible motions that allow proper function [1–4]. This conflict cannot be bridged by compact objects which are characterized by small amplitude vibrations and by a Debye density of low frequency modes. Recently, however, it became clear that proteins can be described as fractals: namely, geometrical objects that possess self-similarity [5–9]. Adopting the fractal point of view to proteins makes it possible to describe within the same framework essential information regarding topology and dynamics [10,11] using three parameters: the number of amino acids along the protein backbone $N$, the spectral dimension $d_s$, and the fractal dimension $d_f$.

Based on a generalization [12] of the Landau-Peierls instability criterion [13], we derive a relation between the spectral dimension $d_s$, the fractal dimension $d_f$, and the number of amino acids along the protein backbone:

$$\frac{2}{d_s} + \frac{1}{d_f} = 1 + \frac{b}{\ln(N)}. \tag{1}$$

The spectral dimension $d_s$ governs the density of low frequency normal modes of a fractal or protein. More precisely, denoting the density of modes $g(\omega)$, the scaling relation $g(\omega) \sim \omega^{d_s - 1}$ holds for low frequencies. Describing the mass fractal dimension $d_f$ is most convenient using a three-dimensional example. Draw a sphere of radius $r$ enclosing some lattice points in space and calculate their mass $M(r)$, increase $r$ and calculate again. Do this several times and if $M(r)$ scales as $r^{d_f}$ the exponent $d_f$ is called the fractal dimension. For a regular 3D lattice both $d_s$ and $d_f$ coincide with the usual dimension of 3. For proteins, however, it is usually found that $d_s < 2$ and $2 < d_f < 3$, leading to an excess of low frequency modes and a more sparse fill of space [5,6,9,14,15]. The parameter $b$ weakly depends on temperature and interaction parameters as discussed later.

Equation (1) is obeyed by a large class of proteins regardless of their source or function. It should be noted that every protein in our set, bearing an entirely different sequence, is in fact a *different* physical system. Thus Eq. (1) describes the universal common behavior of these different systems, as opposed to the case of a single underlying system, e.g., a Gaussian or swollen polymer chain, studied at many different sizes. We suggest that deviations from this "equation of state" for protein topology may render a protein unfolded. The fractal character implies large amplitude vibrations of the protein that could have led to unfolding. By selecting a thermodynamic state that is "close" to the edge of stability against unfolding, nature has solved the thermostability conflict. Nature's solution might be incorporated when planning biologically inspired catalysts.

We are led to relation (1) from two different independent pathways. The first approach utilizes the Gaussian network model (GNM) [16]. The melting of a protein is treated in this approach in a way similar to the melting of a solid crystal [17], with an additional assumption: surface residues initiate the melting process in proteins. Another approach that leads to relation (1) is motivated by the viewpoint of a folded protein as a collapsed polymer. It introduces a non-Lindemann criterion and a bond-bending Hamiltonian rather than the GNM Hamiltonian used in the first approach.

The GNM considers proteins to be elastic networks whose nodes correspond to the positions of the $\alpha$ carbons in the native structure, and the interactions among nodes are modeled as homogeneous harmonic springs. An interaction between two nodes exists only if the nodes are separated by a distance less than $R_c$, a distance known as the interaction cutoff. The cutoff distance is usually taken in the range 6–7 Å, based on the radius of the first coordination shell around residues observed in the Protein Data Bank (PDB) structures [18,19]. The GNM is defined by the harmonic potential energy:

$$V_{\mathrm{GNM}} = \frac{\gamma}{2}\sum_{i,j}\Delta_{ij}(\delta\vec{r}_i - \delta\vec{r}_j)^2. \qquad (2)$$

Here $\gamma$ is the springs force constant and is assumed to be homogeneous and $\delta\vec{r}_i$ is the displacement with respect to the equilibrium position $\vec{R}_i^0$ of the $i$th $C_\alpha$ atom. $\Delta_{ij}$ is the network connectivity matrix with the following entries: $\Delta_{ij} = 1$ if $i \neq j$ and the distance $|\vec{R}_i^0 - \vec{R}_j^0|$ between two $C_\alpha$ atoms in the native conformation is smaller than $R_c$, $\Delta_{ij} = 0$ otherwise. The spectrum of the elastic network is given by the set of eigenvalues $\{\omega_0^2, \omega_1^2, \ldots, \omega_{N-1}^2\}$ of the Kirchhoff matrix, $\Gamma_{ij} = -\Delta_{ij} + \delta_{ij}\sum_{k\neq i}\Delta_{ik}$. The only information required to implement the method is the knowledge of the native structure. GNM has been widely applied because it yields results in agreement with x-ray spectroscopy and NMR experiments [16,20].

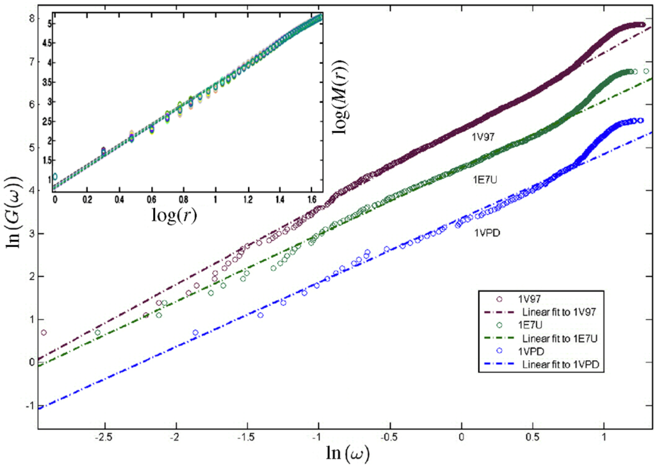

FIG. 1 (color online). Calculating the spectral dimension $d_s$ for PDB codes: 1V97 ($N = 2594$, $d_s = 1.78$), 1E7U ($N = 872$, $d_s = 1.56$), and 1VPD ($N = 279$, $d_s = 1.49$). For each protein, we found the set of vibrational eigenfrequencies $\{\omega_0, \omega_1, \ldots, \omega_{N-1}\}$ that characterize the elastic network it forms when modeled by the GNM and plotted $\ln[G(\omega)]$ vs $\ln(\omega)$. In this example $R_c = 6$ Å and $G(\omega)$ is the cumulative density of modes defined as $G(\omega) = \int_0^\omega g(\omega')d\omega'$. All obtained modes are shown. The low frequency regions of $G(\omega)$ clearly exhibit a power law behavior; i.e., the scaling relation $G(\omega) \sim \omega^{d_s}$ holds for low frequencies. Dashed lines indicate best fits to these regions, $\omega \in [0.109, 1.67]$, $\omega \in [0.119, 1.6]$, and $\omega \in [0.243, 0.888]$ for 1V97, 1E7U, and 1VPD, correspondingly; the slopes correspond to the spectral dimensions. Inset: Calculating the mass fractal dimension $d_f$ for PDB code 1V97 ($N = 2594$, $d_f = 2.64$), $d_f$ was taken to be the average mass fractal dimension obtained by choosing the origin to be each and every one of the ten $C$-$\alpha$ atoms closest to the protein's center of mass. For a given origin, $d_f$ was estimated via a power law fitting to $M(r)$, dashed lines indicate best fits. The data points as well as the best fits for different origins overlap significantly; we take the average slope to be the fractal dimension.

In order to test the validity of Eq. (1), we calculated the spectral and fractal dimensions for a data set of 543 proteins; see Fig. 1. Calculations were preformed on known protein structures, all structures were downloaded from the PDB [21]. The proteins that were chosen may differ in function and/or source organism and represent a wide length scale ranging from 100 to 3000 residues. Statistical analysis of the data gathered reveals satisfying agreement with Eq. (1). Fitting our data with Eq. (1) yields the following best-fit parameters: $b = 2.80$ for the cutoff $R_c = 7$ Å and $b = 3.97$ for a slightly different cutoff $R_c = 6$ Å. Despite the diversity in the sample data both cases yield significant correlation coefficients: 0.64 for $R_c = 7$ Å and 0.55 for $R_c = 6$ Å. In what follows we use the latter cutoff. Testing the validity of our predictions further, we tried fitting the data with the equation $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{\ln(N)}$, which is a modification of Eq. (1) with the unity replaced by a parameter "$a$" on the right-hand side. The results are shown in Fig. 2. Allowing a free constant fitting parameter enabled us to confront theory with practice since our prediction is $a = 1$. A similar relation between $N$ and $d_s$ was suggested and tested on a small set of proteins in [6]. In that study a peculiar offset in the observed value of a constant fitting parameter, predicted to be exactly unity, was reported. We believe to have explained the reason for this offset and by doing so we were led to Eq. (1). The results shown in Fig. 2 indicate that the value of the parameter "$a$" is indeed close to 1. We also checked the validity of Eq. (1) for proteins all originating from the same creature. We thus "sliced" the data according to various sources (human, *E. coli*, etc., ...) in order to gain further insight into the relation between the source organism and the fitting parameters. The results of this analysis are summarized in Table I. Of special interest are proteins

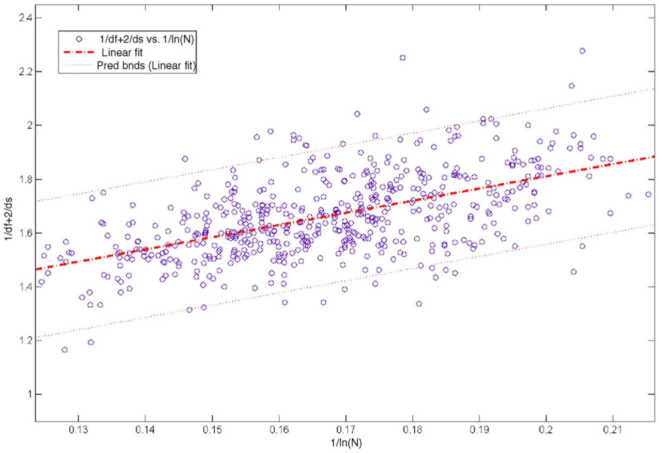

FIG. 2 (color online). Fitting the data gathered for 543 proteins with the equation $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{\ln(N)}$. Here, the spectral dimension was calculated for $R_c = 6$ Å. The best-fit parameters are $a = 0.90$ and $b = 4.53$, the correlation coefficient is 0.55. Prediction bounds are for a confidence level of 95%.

TABLE I. Fitting the data from various creatures with the equation $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{\ln(N)}$. Here, the spectral dimension was calculated for $R_c = 6$ Å, and c.c is the correlation coefficient. It is apparent from the table that when allowing a constant fitting parameter its value remains close to 1; this is true for both the set as a whole and for the overwhelming majority of creatures we analyzed.

| Source | Proteins | $a$ | $b$ | c.c. |
|---|---|---|---|---|
| All | 543 | 0.9 ± 0.09 | 4.53 ± 0.57 | 0.55 |
| Mesophiles | 432 | 0.91 ± 0.1 | 4.45 ± 0.61 | 0.57 |
| *E. coli* | 40 | 1.05 ± 0.25 | 3.66 ± 1.53 | 0.62 |
| *Bacillus subtilis* | 40 | 0.66 ± 0.42 | 6.01 ± 2.49 | 0.62 |
| *Bos taurus* (cattle) | 36 | 1 ± 0.30 | 3.71 ± 1.79 | 0.59 |
| *Homo sapiens* (human) | 44 | 1.13 ± 0.43 | 3.21 ± 2.54 | 0.36 |
| *Mus musculus* (mouse) | 37 | 1.18 ± 0.40 | 3.11 ± 2.26 | 0.43 |
| *Rattus norvegicus* (rat) | 36 | 0.86 ± 0.47 | 5.12 ± 2.71 | 0.55 |
| *Saccharomyces cerevisiae* (yeast) | 38 | 0.81 ± 0.47 | 5.05 ± 3.08 | 0.55 |
| *Salmonella typhimurium* | 28 | 0.59 ± 0.50 | 6.45 ± 3.08 | 0.64 |
| Hyperthermophiles | 111 | 0.87 ± 0.25 | 4.8 ± 1.52 | 0.51 |
| *Pyrococcus* | 44 | 0.99 ± 0.42 | 3.9 ± 2.57 | 0.42 |
| *T. maritima* | 49 | 0.95 ± 0.46 | 4.46 ± 2.70 | 0.44 |
| *A. aeolicus* | 20 | 0.73 ± 0.40 | 5.84 ± 2.36 | 0.77 |

originating in hyperthermophiles [22]. Surprisingly, such proteins that were included in the analyzed data, Fig. 2 and Table I, appear to fulfill Eq. (1).

We now describe the physics behind Eq. (1) and the alternative routes leading to it. In a paper generalizing the Landau-Peierls instability, Burioni *et al.* [12] showed that for $d_s < 2$ the mean square displacement (MSD) $\langle \delta r^2 \rangle$ of a structural unit (in the GNM, a single amino acid) in a system composed of $N$ elements diverges in the limit $N \rightarrow \infty$ as

$$\langle \delta r^2 \rangle \propto \frac{k_B T}{\gamma} N^{(2/d_s)-1}. \tag{3}$$

It is clear that with $d_s < 2$, $\langle \delta r^2 \rangle$ grows indefinitely with $N$. Letting $p$ be the ratio between the number of surface residues and the total number of residues in a protein and $q = 1 - p$ we write $\langle \delta r^2 \rangle_{\text{total}} = p \langle \delta r^2 \rangle_{\text{surface}} + q \langle \delta r^2 \rangle_{\text{bulk}}$. For this equation to hold for every $N$, both terms on the right-hand side must scale as the left-hand side, i.e., as in Eq. (3). Since by definition $p$ is directly proportional to the surface to volume ratio of a protein, we obtain

$$p \propto \frac{S}{V} \propto \frac{1}{R_g} \propto \frac{1}{N^{1/d_f}}, \tag{4}$$

where $R_g$ is the gyration radius of the protein [5,23]. At very low temperatures the MSDs of surface residues and of bulk residues are of the same order of magnitude. As temperature increases, the MSD values grow, and since surface residues are those prone to interactions with the solvent, it is reasonable to assume that melting starts when MSD values of surface residues reach a certain threshold. Denoting this threshold $\langle \delta r^2_{\text{melting}} \rangle_{\text{surface}}$, letting $T_m$ represent the melting temperature and utilizing the scaling law of $p \langle \delta r^2 \rangle_{\text{surface}}$, we obtain the following approximation:

$$\frac{k_B T_m}{\gamma} N^{(2/d_s)+(1/d_f)-1} \propto \langle \delta r^2_{\text{melting}} \rangle_{\text{surface}}. \tag{5}$$

Rearrangement leads to Eq. (1), where the constant $b$ depends on the parameters $\langle \delta r^2_{\text{melting}} \rangle_{\text{surface}}$, $\gamma$, and $T_m$. This dependence, however, is logarithmic and thus very weak, allowing a comparison among different proteins without computation of the specific parameters.

A different route to Eq. (1) is to start with a tensorial elasticity model rather than the scalar elasticity (Born) model described by the GNM. Here we use the bond-bending potential, previously studied for percolation [24,25]:

$$V = \frac{\gamma}{2} \sum_{ij} \Delta_{ij} [(\delta \vec{r}_i - \delta \vec{r}_j) \cdot \hat{r}_{ij}]^2 + \frac{B}{2} \sum_{jik} \Delta_{ij} \Delta_{ik} (\delta \theta_{jik})^2, \tag{6}$$

where $\delta\theta_{jik}$ is the angle between bonds $\langle ij \rangle$ and $\langle ik \rangle$, and $\hat{r}_{ij}$ is the unit vector along the bond $\langle ij \rangle$. We note that the first term is essentially the anisotropic network model discussed by Atilgan and co-workers [26] and describes the stretch-compress penalty, and the second term describes bond-bending penalty. When the bond-bending potentials are effectively softer than stretch-compress potentials ($B \ll \gamma R_g^2$), a very likely situation in proteins, the density of low frequency modes is dominated by bond-bending behavior and $g(\omega) \sim \omega^{d_E - 1}$, where $d_E$ is the bond-bending spectral dimension equivalent to the spectral dimension $d_s$. For percolation clusters $d_E < 1$, and this is expected also for other fractals.

Next consider the variance of fluctuations in the distance between two tagged points on the protein that are distanced $R_g$ apart. This may be evaluated in a similar way to the one described in [10,27], as $\langle \vec{x}^2(R_g)\rangle \sim N^{(2/d_E)-1}$. Importantly, if $d_E < 1$ and $d_f > 2$, this diverges with increasing $N$ faster than $R_g^2 \sim N^{2/d_f}$. We postulate that melting occurs when the magnitude of these fluctuations reaches the protein size, namely, when $\langle \vec{x}^2(R_g)\rangle \sim R_g^2$. This leads to

$$\frac{2}{d_E} - 1 - \frac{2}{d_f} = \frac{\text{const}}{\ln(N)}, \tag{7}$$

an equation that resembles Eq. (1) with $d_s$ replaced by $d_E$.

In order to find $d_E$ one has to solve for the eigenfrequencies of the bond-bending Hamiltonian. To circumvent this difficulty, we use relations that have been derived for percolation clusters, assuming that they hold for other fractals and therefore at least approximately for protein networks [24,25]. The spectral and bond-bending spectral dimensions have been shown to obey the relations [24,25,28] $d_s = \frac{2d_f}{2-d+d_f+t/\nu}$ and $d_E = \frac{2d_f}{d_f+2+1/\nu}$, where $t$ is the percolation conductivity exponent $\sigma \sim (p-p_c)^t$ and $\nu$ is the percolation correlation length exponent $\xi \sim (p-p_c)^{-\nu}$. From these two relations we find, for $d = 3$,

$$\frac{2}{d_E} - 1 - \frac{2}{d_f} \propto \frac{2}{d_s} - 1 + \frac{1}{d_f}. \tag{8}$$

Using Eqs. (7) and (8) leads again to Eq. (1). The relation $\frac{2}{d_s} + \frac{1}{d_f} = 1 + \frac{b}{\ln(N)}$ and the general inequalities $1 \leq d_s \leq d_f \leq 3$ lead to the following effective bounds on $d_s$ and $d_f$: $1 \leq d_s \leq \frac{3}{1+b/\ln N} \leq d_f \leq 3$. Interestingly, the latter bounds permit values of $d_s$ greater than 2. This does not pose any conflict since the Landau-Peierls instability is controlled in this bond-bending model by $d_E$ rather than $d_s$.

One may wonder what will happen if a protein is forced to strongly deviate from Eq. (1) and how artificial deformations of the protein fold may lead to a breakdown of criterion (1). Strong deformations of the protein fold may actually happen *in vivo* as part of a natural process. A possible example is GroEL, a protein chaperon that is required for the proper folding of many proteins. Recent molecular dynamics simulations demonstrate the unfolding action of GroEL on a protein substrate [29,30]. Our work provides a theoretical framework that may help understand GroEL induced unfolding. In addition our work opens new possibilities for nanoscale and biologically inspired engineering of catalysts, emphasizing the importance of internal motion.

We thank Tal Shapiro and Noam Grimberg for assistance with data analysis. R. G. and J. K. acknowledge support from The Israel Science Foundation. R. G. is a member of the Ilse Katz Center for Meso and Nanoscale Science and Technology and the Reimund Stadler Minerva Center for Mesoscale Macromolecular Engineering.

[1] D. Joseph, G. A. Petsko, and M. Karplus, Science **249**, 1425 (1990).
[2] P. J. Steinbach *et al.*, Biochemistry **30**, 3988 (1991).
[3] B. F. Rasmussen, A. M. Stock, D. Ringe, and G. A. Petsko, Nature (London) **357**, 423 (1992).
[4] A. Henzler-Wildman *et al.*, Nature (London) **450**, 913 (2007).
[5] M. B. Enright and D. M. Leitner, Phys. Rev. E **71**, 011912 (2005).
[6] R. Burioni, D. Cassi, F. Cecconi, and A. Vulpiani, Proteins: Struct. Funct. Genet. **55**, 529 (2004).
[7] S. Lushnikov, A. Svanidze, and I. Sashin, J. Exp. Theor. Phys. Lett. **82**, 30 (2005).
[8] H. J. Stapleton, J. P. Allen, C. P. Flynn, D. G. Stinson, and S. R. Kurtz, Phys. Rev. Lett. **45**, 1456 (1980).
[9] R. Elber and M. Karplus, Phys. Rev. Lett. **56**, 394 (1986).
[10] R. Granek and J. Klafter, Phys. Rev. Lett. **95**, 098106 (2005).
[11] Nature (London) **437**, 172 (2005).
[12] R. Burioni, D. Cassi, M. P. Fontana, and A. Vulpiani, Europhys. Lett. **58**, 806 (2002).
[13] R. Peierls, Helv. Phys. Acta **7**, Suppl. 2, 81 (1934).
[14] T. Haliloglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[15] S. Ciliberti, P. DeLosRios, and F. Piazza, Phys. Rev. Lett. **96**, 198103 (2006).
[16] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[17] Y. Zhou, M. Karplus, K. D. Ball, and R. S. Berry, J. Chem. Phys. **116**, 2323 (2002).
[18] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[19] I. Bahar and R. L. Jernigan, J. Mol. Biol. **266**, 195 (1997).
[20] L.-W. Yang, E. Eyal, C. Chennubhotla, J. JunGoo, A. M. Gronenborn, and I. Bahar, Structure **15**, 741 (2007).
[21] F. C. Bernstein *et al.*, J. Mol. Biol. **112**, 535 (1977).
[22] I. N. Berezovsky and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **102**, 12 742 (2005).
[23] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University, Ithaca, New York, 1979).
[24] S. Feng, Phys. Rev. B **32**, 5793 (1985).
[25] I. Webman and G. S. Grest, Phys. Rev. B **31**, 1689 (1985).
[26] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[27] S. Alexander, Phys. Rev. B **40**, 7953 (1989).
[28] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1992), 2nd ed.
[29] A. Van der Vaart, J. Ma, and M. Karplus, Biophys. J. **87**, 562 (2004).
[30] G. Stan, G. H. Lorimer, D. Thirumalai, and B. R. Brooks, Proc. Natl. Acad. Sci. U.S.A. **104**, 8803 (2007).