Coexistence of Flexibility and Stability of Proteins: An Equation of State

Marina de Leeuw1, Shlomi Reuveni2, Joseph Klafter2,3, Rony Granek1*

1 Department of Biotechnology Engineering, Ben-Gurion University of The Negev, Beer Sheva, Israel, 2 School of Chemistry, Tel-Aviv University, Tel-Aviv, Israel, 3 Freiburg Institute for Advanced Studies (FRIAS), Albert-Ludwig-Universität Freiburg, Freiburg, Germany

Abstract

We consider a recently suggested “equation of state” for natively folded proteins, and verify its validity for a set of about 5800 proteins. The equation is based on a fractal viewpoint of proteins, on a generalization of the Landau-Peierls instability, and on a marginal stability criterion. The latter allows for coexistence of stability and flexibility of proteins, which is required for their proper function. The equation of state relates the protein fractal dimension \( d_f \), its spectral dimension \( d_s \), and the number of amino acids \( N \). Using structural data from the protein data bank (PDB) and the Gaussian network model (GNM), we compute \( d_f \) and \( d_s \) for the entire set and demonstrate that the equation of state is well obeyed. Addressing the fractal properties and making use of the equation of state may help to engineer biologically inspired catalysts.

Introduction

Proteins are one of the major components of living cells. They constitute more than half of the cell’s dry weight, and are responsible for the execution of most cellular functions required for life, including among others, catalysis and molecular recognition within and between cells and their surroundings. Understanding the relationships between structure, internal dynamics, and enzymatic activity at the single-molecule level could pave new ways to manipulate individual molecules.

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 large amplitude conformational changes that allow proper function [1]. This conflict cannot be bridged by compact objects which are characterized by small amplitude vibrations [2]. Recently, however, it became evident that proteins can be described as fractals; namely, geometrical objects that possess self similarity [3,4]. Adopting the fractal point of view to proteins makes it possible to describe within the same framework essential information regarding topology and dynamics.

Based on the fractal viewpoint, we have recently derived a universal equation of state for protein topology. The same fractal viewpoint allows describing the near equilibrium dynamics of native proteins. We have recently shown that it leads to anomalous dynamics [5]. For example, the autocorrelation function of the distance between two \( z \)-carbons on a protein is predicted to decay anomalously, first, at short times, as \( 1 - t^\delta \) and later, at long times, as \( t^{-\beta} \), where \( \delta \) and \( \beta \) are exponents that depend on various fractal dimensions. This type of relaxation has been recently observed in single molecule experiments [6,7]. Closely related is the anomalous diffusion of an \( z \)-carbon that is predicted by the fractal model, where the mean square displacement is found to increase as \( \sim t^\delta \). Such a behavior has also been recently observed in molecular dynamics simulations [8].

Natively folded proteins can be characterized by broken dimensions: the fractal and spectral dimensions [2,4,5,9–12]. The mass fractal dimension \( d_f \) describes the spatial distribution of the mass within the protein via the scaling relation \( M(r) \sim r^{d_f} \), where \( M(r) \) is the mass enclosed in a sphere of radius \( r \) [3]. The spectral dimension \( d_s \) governs the density of low frequency vibrational normal modes via the scaling relation \( g(\omega) \sim \omega^{d_s - 1} \), where \( g(\omega) \omega^d \) is the number of modes in the frequency range \( [\omega, \omega + d\omega] \) [13]. While for regular three dimensional (3D) lattices both \( d_f \) and \( d_s \) coincide with the usual dimension of 3, for proteins it is usually found that \( d_f < 2 \) and \( 2 < d_s < 3 \), leading to an excess of low frequency modes and a more sparse fill of space [2,4,12]. Importantly, the regime \( d_f < 2 \) is associated with the so-called Landau-Peierls instability, where the amplitude of vibrations increases with the number of residues \( N \) [14,15]. As this amplitude overcomes a threshold value, it may cause the protein to unfold [2,12].

The Landau-Peierls instability is most readily derived using the density of states. The static mean square displacement (MSD) of an \( z \)-carbon, which is essentially the so-called Bi-factor, averaged over all \( z \)-carbons of the protein, may be expressed as

\[
\langle (\Delta u)^2 \rangle = \frac{k_B T}{m} \int_{\omega_{min}}^{\omega_{max}} d\omega \frac{g(\omega)}{\omega^2}
\]

where \( m \) is the average mass of an amino acid. Since \( g(\omega) \sim \omega^{d_s - 1} \), it follows that if \( d_s < 2 \) the integral diverges with the lower bound

\[
\langle (\Delta u)^2 \rangle \sim \frac{k_B T}{m} \omega_{max}^{d_s - 2}
\]

Citation: de Leeuw M, Reuveni S, Klafter J, Granek R (2009) Coexistence of Flexibility and Stability of Proteins: An Equation of State. PLoS ONE 4(10): e7296. doi:10.1371/journal.pone.0007296

Editor: Roland Dunbrack, Fox Chase Cancer Center, United States of America

Received July 15, 2009; Accepted August 24, 2009; Published October 9, 2009

Copyright: © 2009 de Leeuw et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Rony Granek and Joseph Klafter thank The Israel Science Foundation for financial support. Joseph Klafter acknowledges support from the Excellence Initiative of the German Federal and State Governments. Shlomi Reuveni acknowledges support of the Converging Technologies fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rgranek@bgu.ac.il
The latter depends on the protein radius of gyration $R_g$ and the number of residues $N$ as $\omega_{\text{min}} \sim R_g^{-d/4} N^{-1/8}$. This leads to $\langle (\Delta \nu)^2 \rangle \sim N^{2/d_s - 1}$, which increases with $N$ for $d_s < 2$. In particular, the static MSD of surface residues has been argued to grow as

$$\langle (\Delta \nu)^2 \rangle_{\text{surface}} \sim N^{2/d_s + 1/d_s - 1} \quad (2)$$

We have proposed a marginal stability criterion [16], in which most proteins "exploit" the Landau-Peierls instability to attain large amplitude vibrations, which is required for their proper function, yet maintaining their native fold. Thus proteins are assumed to exist in a thermodynamic state close to the edge of unfolding. Based on this and the Landau-Peierls instability of the surface residues, Eq. (2), a general equation of state has been proposed that relates between the spectral dimension $d_s$, the fractal dimension $d_f$, and the number of amino acids along the protein backbone $N$:

$$\frac{2}{d_s} + \frac{1}{d_f} = 1 + \frac{b}{\ln(N)} \quad (3)$$

where $b$ is a molecular fit parameter depending on the temperature $T$, the GNM spring constant $\gamma$, and the GNM cutoff $R_c$: $b = \ln(\gamma R_g^2/k_B T)$ [12]. It has been shown that this equation is obeyed by about 500 proteins regardless of their source or function [12]. In the present study we check the validity of Eq. (3) for a much larger set of over 5,000 proteins, using a range of statistical methods, and show that also for this very large set Eq. (3) is beautifully fulfilled. This supports the marginal stability criterion that led to this equation.

**Methods**

We have used all data files present in the Protein Data Bank (PDB) [17] and filtered out proteins exceeding 95% sequence identity and proteins that have ligands, RNA, or DNA. We have also removed incomplete data files, files that contained data of the $\alpha$-carbons alone, and also files of proteins smaller than 100 amino acids that are too small to be characterized as fractals. With this screening the set has been reduced to 5793 proteins.

The fractal and spectral dimensions were calculated for all 5793 proteins in similar ways to the procedure described by [12]. Finding the protein center of mass and placing the origin of coordinates at the ten nearest $\alpha$-carbons, the mass was calculated as a function of the distance $r$ on a log-log scale. The fractal dimension $d_f$ has been obtained as the slope of this plot for distances below the protein gyration radius $R_g$, averaged over the ten origin of coordinates, see examples in Fig. 1. It should be noted that when a few alternative locations of an atom are given, only the "A" location (usually the most abundant one) has been used.

To find the spectral dimension $d_s$, we calculate the cumulative density of normal vibrational modes $G(\nu) = \int_0^\nu d\nu' g(\nu')$, representing the number of modes up to a frequency $\nu$. To obtain the vibrational modes, we used a frequently applied elastic model for protein vibrations, the Gaussian network model (GNM) [12,18–23]. Two values were taken for the interaction distance cutoff $R_c$, that describes the range of the interaction between an $\alpha$-carbons pair, $R_c = 6$ Å and $R_c = 7$ Å. Plotting on a log-log scale $G(\nu)$ against the frequency $\nu$, the slope in the low frequency range (containing about 24% of the modes, independent of the protein type or size $N$) defines $d_s$, i.e. $G(\nu) \sim \nu^{d_s}$, see examples in Fig. 2 for the case $R_c = 6$ Å.

To deal with the large number of proteins in this set, both procedures were automated using suitable computer codes. The

---

**Figure 1. Fractal dimension.** The fractal dimension of three selected proteins: 1FTR (1184 amino acids, $d_f = 2.66$), 1UC8 (505 amino acids, $d_f = 2.51$) and 3TSS (190 amino acids, $d_f = 2.50$). The mass $M(r)$ enclosed in concentric spheres of radius $r$ is plotted against $r$ (measured in units of Å) on a log-log scale and the slope determines the fractal dimension, $M(r) \sim r^{d_f}$. The plots of 1FTR and 1UC8 were shifted along the $y$ axis (+1 and +0.5 respectively) for clarity. doi:10.1371/journal.pone.0007296.g001

**Figure 2. Spectral dimension.** The spectral dimension of three selected proteins (same proteins as in Fig. 1): 1FTR (1184 amino acids, $d_s = 1.93$), 1UC8 (505 amino acids, $d_s = 1.73$) and 3TSS (190 amino acids, $d_s = 1.52$). The cumulative density of normal modes $G(\nu)$ is plotted against the frequency $\nu$ (measured in units of the spring natural frequency) on a log-log scale and the slope determines the spectral dimension, $G(\nu) \sim \nu^{d_s}$. doi:10.1371/journal.pone.0007296.g002
automatically calculated spectral dimension values were compared (for the case $R_c = 7 \text{ Å}$) to the manually obtained values for the set of 543 studied in [12]. We found almost vanishing mean of the difference between the two results (0.0034), showing that the error is statistical, and a low standard deviation (0.083), suggesting good agreement between the two methods of calculation.

In order to generally check for correlations between $2/d_s + 1/d_f$ and $1/\ln N$, simple regression was conducted (using SPSS). This shows statistical significance with $p<0.001$ and very high $F$-test values ($F(1,5791) = 3263$ and $F(1,4247) = 2120$ for $R_c = 6 \text{ Å}$, $F(1,5791) = 4059$ and $F(1,3888) = 2314$ for $R_c = 7 \text{ Å}$).

### Results

The results for the whole set appear in the supporting information S1 and are shown in Figs. 3 and 4 (for $R_c = 6 \text{ Å}$ and $7 \text{ Å}$, respectively), where we plot the combination $2/d_s + 1/d_f$ against $1/\ln N$. In order to present the whole set of data, we designed a (smoothed) colored histogram based on a $100 \times 100$ grid, where a pixel color represents the number of proteins associated with the pixel. The data is first fitted to Eq. (3) (dashed lines). This leads to $b = 4.555$ for $R_c = 6 \text{ Å}$ (correlation coefficient $cc = 0.596$), see Fig. 3, and $b = 3.242$ for $R_c = 7 \text{ Å}$ ($cc = 0.605$), see Fig. 4. Using $b = \ln (R_c/\kappa_T)$, with $\kappa_T/\gamma$ in the range $0.5 \text{ Å}^2$ to $2 \text{ Å}^2$, we can estimate $b$ to be in the range 3 to 5. The value of $b$ is within the expected range.

We also fitted the data to an equation resembling Eq. (3) but in which the value ''1'' is replaced by a free parameter $a$:

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

This is done in order to verify if the free fit recovers the value $a = 1$. The results of this fit are also shown in Figs. 3–4 (full lines), and yield $a = 0.884$ and $b = 5.197$ for $R_c = 6 \text{ Å}$ (Fig. 3, $cc = 0.560$), and $a = 0.710$ and $b = 4.841$ for $R_c = 7 \text{ Å}$ (Fig. 4, $cc = 0.642$). Remarkably, the colored histogram shows a ridge roughly centered at the best fitting theoretical lines.

To improve the accuracy of the analyses, a subset was constructed containing only those proteins whose both $d_s$ and $d_f$ values have been determined with a very high precision, such that the squared correlation coefficients for the power-law fits of both $\langle r^2 \rangle$ and $G(r)$ were in the range $R^2>0.99$. Accordingly, this subset for $R_c = 6 \text{ Å}$ (containing 4249 proteins) is not identical to the subset for $R_c = 7 \text{ Å}$ (containing 3890 proteins), see the supporting information S1 for details. The results are presented in Figs. 5–6. Fitting to Eq. (3) (dashed lines) leads to $b = 4.476$ for $R_c = 6 \text{ Å}$ (Fig. 5, $cc = 0.576$) and $b = 3.078$ for $R_c = 7 \text{ Å}$ (Fig. 6, $cc = 0.593$). Fitting the data to Eq. (4) (full lines), yields $a = 0.952$ and $b = 4.747$ for $R_c = 6 \text{ Å}$ (Fig. 5, $cc = 0.577$), and $a = 0.833$ and $b = 4.031$ for $R_c = 7 \text{ Å}$ (Fig. 6, $cc = 0.611$).

Although the data analysis presented in Fig. 3–6 appears complete, it fails to give equal weight to proteins of different sizes. All four different data sets used above are very rich in proteins of small (100–200 residues) and intermediate size, a consequence of

![Figure 3. Full data set for $R_c=6 \text{ Å}$, colored histogram.](image)

![Figure 4. Full data set for $R_c=7 \text{ Å}$, colored histogram.](image)

![Figure 5. High precision data set for $R_c=6 \text{ Å}$, colored histogram.](image)
their abundance in nature, while being poor in large proteins. Yet, the linear regression presented in Figs. 3–6 gives each protein an equal weight. Thus, while the small/intermediate size proteins are spread over a relatively limited range of \(N\), they are overwhelming the linear regression, which is undesirable.

To circumvent this artifact, we have separated the \(x\)-axis (1/\(\ln N\)) into 100 bins. For each bin we calculate the mean value of 2/\(d_z + 1/d_r\). The error of 2/\(d_z + 1/d_r\) for each bin is estimated as the standard deviation of this value. The results are summarized in Figs. 7, 8, 9, 10.

Results from the full set of 5793 proteins are presented in Figs. 7–8. Fitting to Eq. (3) (dashed lines) leads to \(b = 4.580\) for \(R_c = 6\) Å (Fig. 7, \(cc = 0.957\)) and \(b = 3.212\) for \(R_c = 7\) Å (Fig. 8, \(cc = 0.928\)). Fitting the data to Eq. (4) (full lines), yields \(a = 1.026\) and \(b = 4.429\) for \(R_c = 6\) Å (Fig. 7, \(cc = 0.958\)), and \(a = 0.870\) and \(b = 3.977\) for \(R_c = 7\) Å (Fig. 8, \(cc = 0.946\)). Note that all lines pass through almost all error bars, a remarkable result.

---

**Figure 6. High precision data set for \(R_c = 7\) Å, colored histogram.** Same as in Fig. 5 but for the refined subset of increased precision for \(R_c = 7\) Å (3890 proteins).

---

**Figure 7. Full data set for \(R_c = 6\) Å, division into bins.** The values of 2/\(d_z + 1/d_r\) against 1/\(\ln N\) plotted for the full data set (5793 proteins) with \(R_c = 6\) Å. The values of 1/\(\ln N\) were divided into 100 equally sized bins. For each bin we show the average value of 2/\(d_z + 1/d_r\) and the error bar presents its standard deviation. The data was fitted to Eq. (3) (dashed red line) and to Eq. (3) (full black line); the two lines are almost indistinguishable.

**Figure 8. Full data set for \(R_c = 7\) Å, division into bins.** Same as in Fig. 7 but for \(R_c = 7\) Å.

---

**Figure 9. High precision data set for \(R_c = 6\) Å, division into bins.** Same as Fig. 7 but for the refined subset of increased precision for \(R_c = 6\) Å (4249 proteins).

---

**Figure 10. High precision data set for \(R_c = 7\) Å, division into bins.** Same as Fig. 7 but for the refined subset of increased precision for \(R_c = 7\) Å (3890 proteins).
In Figs. 9–10 we present results from the high precision subset of 4249 proteins. Fitting to Eq. (3) (dashed lines) leads to $b = 4.535$ for $R_c = 6 \AA$ (Fig. 9, $cc = 0.941$) and $b = 3.124$ for $R_c = 7 \AA$ (Fig. 10, $cc = 0.937$). Fitting the data to Eq. (4) (full lines), yields $a = 1.065$ and $b = 3.609$ for $R_c = 5.4 \AA$ (Fig. 9, $cc = 0.945$), and $a = 0.917$ and $b = 3.609$ for $R_c = 7 \AA$ (Fig. 10, $cc = 0.946$). Here, as well, all lines pass through almost error bars. This refined analysis gives an even stronger support to Eq. (3).

Discussion

All correlation coefficients mentioned above (Figs. 3–10) are considered excellent. In addition, the values of $a$ are close to the theoretically predicted value $a = 1$, similar to the set of 543 proteins studied by [12]. In particular, the fits of the data to Eq. (4) for all data sets belonging to $R_c = 6 \AA$ (shown in Figs. 3, 5, 7 and 9) yields $a$ values that are remarkably close to 1. The distribution of the data in all four sets appears as a ridge that is roughly centered at the best fitting theoretical lines (Figs. 3–6), and when the binning procedure is being used, all lines pass well through the error bars (Figs. 7–10). We believe that these results strongly confirm the universal behavior described by Eq. (3), thereby supporting the theoretical arguments leading to this equation.

Importantly, $a$ is found to be particularly close to 1 when the binning procedure is introduced, in which we analyze the mean value of $2/d_f + 1/d_f$ for a given $N$, for its dependence on $N$. In these cases we also obtain remarkably good correlation coefficients, significantly better than those obtained without binning. This suggests that, as a group, proteins follow the equation of state, although the error bars indicate that there are other factors present that cause deviations from the equation. These factors could be related to the protein specific structure and/or function.

The distribution of the data in all four sets appears as a ridge that is roughly centered at the best fitting theoretical lines (Figs. 3–6), and when the binning procedure is being used, all lines pass well through the error bars (Figs. 7–10). We believe that these results strongly confirm the universal behavior described by Eq. (3), thereby supporting the theoretical arguments leading to this equation.

To conclude, our analysis confirms the fractal nature of proteins and supports the predicted universal equation of state (3). This suggests that the majority of proteins in the PDB exist in a marginally stable thermodynamic state, namely a state that is close to the edge of unfolding. This could be related to the fact that enzymes require flexibility and large internal motion to function properly [1]. We suggest that Eq. (3) can be used as a tool in the design of artificial enzymes [24]. Interestingly, fractal-like properties have also been suggested to appear in the configuration space of peptides [25].

Supporting Information

Supporting Information S1 A file containing the mass fractal dimension $d_f$ and spectral dimension $d_s$ of all proteins analyzed, divided into the four data sets described in the text: (i) GNM cutoff length 6 Å, (ii) GNM cutoff length 7 Å, (iii) GNM cutoff length 6 Å, a subset with high precision values of $d_f$ and $d_s$, and (iv) GNM cutoff length 7 Å, a subset with high precision values of $d_f$ and $d_s$. Found at: doi:10.1371/journal.pone.0007296.s001 (2.34 MB XLS)

Acknowledgments

We are grateful to Martin Karplus and Dave Thirumalai for illuminating discussions. RG is a member of the Ibe Katz Center for Meso and Nanoscale Science and Technology and the Reimund Stadler Minerva Center for Mesoscale Macromolecular Engineering.

Author Contributions

Analyzed the data: MdL. Wrote the paper: JK RG. Initiated and directed the research: RG. Contributed to data analysis: SR.

References

1. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, et al. (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450(7171): 913–916.
2. Burioni R, Casi D, Cecconi F, Vulpiiani A (2004) Topological thermal instability and length of proteins. Proteins-Structure Function and Bioinformatics 55(3): 329–335.
3. Stauffer D, Aharony A (1994) Introduction to percolation theory. CRC Press.
4. Enright MB, Leitner DM (2005) Mass fractal dimension and the compactness of proteins. Physical Review E 71(1): 011912.
5. Granek R, Klaffer J (2005) Fractons in proteins: Can they lead to anomalously decaying time autocorrelations? Phys Rev Lett 95(9): 098106.
6. Krou SC, Xie NS (2004) Generalized langevin equation with fractional gaussian noise: Subdiffusion within a single protein molecule. Phys Rev Lett 93(18): 180603.
7. Min W, Luo G, Cherayil BJ, Krou SC, Xie NS (2005) Observation of a power-law memory kernel for fluctuations within a single protein molecule. Phys Rev Lett 94(19): 198302.
8. Senet P, Manzuradze GG, Foulie C, Delarue P, Scheraga HA (2008) How main-chains of proteins explore the free-energy landscape in native states. Proceedings of the National Academy of Sciences of the United States of America 105(50): 19708.
9. Stapleton HJ, Allen JP, Flynn CP, Stinson DG, Kurz SR (1980) Fractal form of proteins. Phys Rev Lett 45(17): 1456–1459.
10. Elber R, Karplus M (1986) Low-frequency modes in proteins - Use of the effective-medium approximation to interpret the fractal dimension observed in electron-spin relaxation measurements. Phys Rev Lett 56(4): 394–397.
11. Lushnikov SG, Svanidze AV, Sashin IL (2005) Vibrational density of states of 4249 proteins. Fitting to Eq. (3) (dashed lines) leads to $b = 4.535$ for $R_c = 6 \AA$ (Fig. 9, $cc = 0.941$) and $b = 3.124$ for $R_c = 7 \AA$ (Fig. 10, $cc = 0.937$). Fitting the data to Eq. (4) (full lines), yields $a = 1.065$ and $b = 3.609$ for $R_c = 5.4 \AA$ (Fig. 9, $cc = 0.945$), and $a = 0.917$ and $b = 3.609$ for $R_c = 7 \AA$ (Fig. 10, $cc = 0.946$). Here, as well, all lines pass through almost error bars. This refined analysis gives an even stronger support to Eq. (3).

Supporting Information

Supporting Information S1 A file containing the mass fractal dimension $d_f$ and spectral dimension $d_s$ of all proteins analyzed, divided into the four data sets described in the text: (i) GNM cutoff length 6 Å, (ii) GNM cutoff length 7 Å, (iii) GNM cutoff length 6 Å, a subset with high precision values of $d_f$ and $d_s$, and (iv) GNM cutoff length 7 Å, a subset with high precision values of $d_f$ and $d_s$. Found at: doi:10.1371/journal.pone.0007296.s001 (2.34 MB XLS)

Acknowledgments

We are grateful to Martin Karplus and Dave Thirumalai for illuminating discussions. RG is a member of the Ibe Katz Center for Meso and Nanoscale Science and Technology and the Reimund Stadler Minerva Center for Mesoscale Macromolecular Engineering.

Author Contributions

Analyzed the data: MdL. Wrote the paper: JK RG. Initiated and directed the research: RG. Contributed to data analysis: SR.