Scaling Law for the Radius of Gyration of Proteins and Its Dependence on Hydrophobicity

LIU HONG, JINZHI LEI
Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, People’s Republic of China 100084

Received 7 June 2008; revised 31 October 2008; accepted 4 November 2008
DOI: 10.1002/polb.21634
Published online in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: The scaling law between the radius of gyration and the length of a polymer chain has long been an interesting topic since the Flory theory. In this article, we seek to derive a unified formula for the scaling exponent of proteins under different solvent conditions. The formula is obtained by considering the balance between the excluded volume effect and elastic interactions among monomers. Our results show that the scaling exponent is closely related to the fractal dimension of a protein’s structure at the equilibrium state. Applying this formula to natural proteins yields a 2/5 law with fractal dimension 2 at the native state, which is in good agreement with other studies based on Protein Data Bank analysis. We also study the dependence of the scaling exponent on the hydrophobicity of a protein chain through a simple two-letters HP model. The results provides a way to estimate the globular structure of a protein, and could be helpful for the investigation of the mechanisms of protein folding. © 2008 Wiley Periodicals, Inc.

Keywords: molecular modeling; proteins; structure; structure property relations

INTRODUCTION

It is known that a protein can refold to its native structure from a denatured state under physiological conditions, i.e., water solution with appropriate pH values, ions, temperature, pressure etc., in which a natural protein can fold spontaneously. But the underlying mechanism is still unknown and has been one of the basic intellectual challenges in modern molecular biology.1 In the study of protein folding, the radius of gyration \( R_g \) is an important quantity, because of its straightforward physical significance, and easy to be measured in experiments. Radius of gyration is often used to describe the compactness of a protein, as well as the folding process from the denatured state to the native state. Experimentally, the time evolution of the radius of gyration of a protein can be measured by small-angle X-ray scattering. Significant changes in the radius of gyration are observed in several proteins after a fast pH jump.2 This article will study the dependence of the radius of gyration on the protein length under different solvent conditions that can be altered by pH jump.2 This article will study the dependence of the radius of gyration on the protein length under different solvent conditions that can be altered by pH jump.2

For natural proteins under physiological conditions, we find a scaling law of \( R_g \propto N^{2/5} \) by exploiting Protein Data Bank (PDB), which consists with previous studies.3–5 However, the origin of this 2/5 law is still not clear. In this article, we seek a simple explanation for the 2/5 law. Because proteins are long chain molecules, the differences between amino acids can be neglected when considering the statistical average properties of the large-scale structures, such as the radius of gyration and the scaling exponent. So we can apply Flory’s original idea for homopolymers to the present study. However, the interaction
forms used in Flory's original expressions need to be modified in order to incorporate the specialities of proteins. By connecting the scaling exponent with the fractal dimension of a protein's equilibrium conformation under different solvent conditions, we will derive a unified formula for the scaling exponent. As two extreme cases, our formula reobtains the 1/3 law for homopolymers in poor solvent conditions with dimension 3, and Flory's 3/5 law for homopolymers in good solvent conditions with dimension 1. For general natural proteins in physiological condition, the 2/5 law yields a fractal dimension 2, which shows good agreement with former studies as well as self-similarity analysis of protein structures.3–6

Hydrophobic interaction is an important driving force in protein folding. Hence, the dependence of a protein’s compactness on its hydrophobicity is of great interest, and will be helpful for predicting the globular structure of a protein from its amino acid sequence. In this article, we will study how a protein’s compactness depends on its hydrophobicity. By considering the equivalence between the protein-solvent coupled systems, we establish the connection between the process of varying the hydrophobicity of a chain with the process of varying the polarity of the solvent condition. This enables us to obtain a correlation between the scaling exponent of a protein at the folded state and its hydrophobicity.

This article is organized as follows: Section “Scaling exponent for protein under physiological conditions” shows the statistical results for the scaling law of the radius of gyration. In section “Theoretical approaches”, we obtain a unified formula by modifying Flory’s theory to consider the balance between the exclude volume effect and the elastic interactions under different solvent conditions. In Section “Dependence of hydrophobicity”, the dependence of the protein compactness on its hydrophobicity is studied. The last section includes the conclusion part.

**SCALING EXPONENT FOR PROTEIN UNDER PHYSIOLOGICAL CONDITIONS**

From the Flory theory7–10 the radius of gyration of a homopolymer at the equilibrium state depends on the chain length through a scaling law $R_g \propto N^\nu$, where the exponent $\nu$ depends on the solvent conditions. For example, we have $\nu \approx 1/3$ under poor solve conditions, and $\nu \approx 3/5$ under good solvent conditions.

However, proteins are heteropolypeptides made up of 20 different kinds of amino acids, which can be either hydrophobic or hydrophilic. If all residues are hydrophobic, which corresponds to poor solvent conditions, the protein will be highly compressed by solvent pressure and has a scaling exponent $\nu \approx 1/3$. On the other hand, if all residues are hydrophilic, corresponding to good solvent conditions, the protein will be extended with $\nu \approx 3/5$ to form as many hydrogen bonds with surrounding water molecules as possible. Thus the solvent condition for a general natural protein under physiological conditions should be between good and poor solvent condition. A statistical study of over 37,000 native structures of proteins in PDB yields a scaling law with exponent $\nu \approx 2/5$ (Fig. 1), which is in agreement with early studies.3–5

A more refined result shows the scaling law for proteins with different secondary structure categories, namely all-α, all-β, α/β mixed structures (with more than 50% amino acids in secondary structure), and unstructured proteins (with less than 20% amino acids in secondary structure) (Fig. 2). We can see that the scaling exponents $\nu$ are all about 2/5 despite the varieties in secondary structures. This suggests that the scaling exponent $\nu \approx 2/5$ is independent of a protein’s detailed structure, and there may exist a unified mechanism for the scaling law. Further detailed discussions will be given in Section “Theoretical approaches”.

We next examine the dependence of the scaling exponent on the hydrophobicity. The hydrophobicity ($h$) of a protein is defined as the fraction of amino acids that are hydrophobic. Here, we adopt the Kyte-Doolittle scale to define the hydrophobic amino acids.11 The amino acids with positive Kyte-Doolittle values are regarded as hydrophobic; and those with negative values are hydrophilic. Accordingly, there are Eight hydrophobic amino acids (I, V, L, F, C, M, A, G), and 12 hydrophilic ones (T, S, W, Y, P, H, E, N, Q, D, K, R). According to the Kyte-Doolittle scale, the hydrophobicity of natural proteins varies mainly from 0.25 to 0.75, and exhibits a Gaussian-like distribution $N(0.5, 0.054)$ (Fig. 3). The dependence of the scaling exponent with respect to the hydrophobicity is given in Figure 4. It shows that the scaling exponent is almost unchanged with $\nu \approx 2/5$ when $0.4 < h < 0.6$. For small values of $h$ ($h < 0.4$), $\nu$ increases to $\sim 3/5$, as $h$ approaches 0.2; while for large values of $h$ ($h > 0.6$), $\nu$ decreases to $\sim 1/3$. Therefore, proteins with lower hydrophobicity have a larger scaling exponent than those with higher hydrophobicity. This is justified since
higher hydrophobicity would result in more compact conformations due to stronger solvent pressure. Theoretical formulation for fitting the statistical data in Figure 4 will be discussed in Section “Dependence on hydrophobicity”.

The scaling exponent $1/3 \leq \nu \leq 3/5$ indicates that folded proteins are more compact than polymers in good solvent, but looser than highly compressed polymers in poor solvent. Moreover, most natural proteins with hydrophobicity $0.4 < h < 0.6$ possess almost the same scaling exponent $\nu \approx 2/5$. These results indicate the structural specificity of natural proteins is distinct from usual homopolymers. On the one hand, proteins are compact because of hydrophobic interactions, which is essential to maintain the stability of the hydrophobic core in water solution. On the other hand, proteins need to be flexible to ensure their biological activities in living cells. They are usually not well-packed and contain many internal cavities for the presence of rigid secondary structures. Furthermore, large proteins are often made up of several separate domains, such that their overall geometrical packing may deviate substantially from a spherical shape.  

THEORETICAL APPROACHES

In this section, we will extend the Flory theory to obtain a unified formula for the scaling law of
proteins under different solvent conditions. The validity of the current approach is based on the fact that proteins are long chain molecules. Since we only consider the statistical average effect of an ensemble of proteins, the detailed differences of the small-scale structures, such as amino acids, do not affect the overall properties of the large-scale structures in the current study.

The original idea proposed by Flory for calculating the size of a homopolymer was to consider the balance between two dominant effects.\textsuperscript{7–10} In poor solvent conditions, the polymer chain is highly compressed by the solvent pressure. Thus the excluded volume effect ($\propto N^2/R_g^3$) and the three-body repulsive interaction ($\propto N^3/R_g^6$) constitute the main parts of the total energy. Minimizing this energy with respect to $R_g$ gives the 1/3 law. Although in good solvent conditions, the polymer monomers are well separated by solvent molecules. The dominant interactions are the excluded volume effect and the entropy effect ($\propto R_g^2/N$), which imply the well-known 3/5 law. However, in the case of natural proteins under physiological conditions, many other interactions are relevant, including the hydrophobic interactions due to solvent pressure, hydrogen bonding between neighboring amino acids, and Van der Waal's interactions, etc.\textsuperscript{1,10,13} Instead of studying these complicated interactions individually, we will consider their overall effect in the framework.
of elastic energy, which enables us to obtain a unified formula for the scaling exponent. Detailed calculations are given below.

First, according to the Flory theory, the excluded volume effect is repulsive and contributes an energy term as\textsuperscript{7–10}

\[ E_{\text{rep}} = k_B T v \frac{N^2}{R_g^2}, \]  

(1)

where \( v \) is single residue’s effective volume.

Next, we consider the attractive effect between the residues. At the equilibrium state, the protein is compact due to the hydrophobic interactions from the surrounding water molecules, and is stabilized by the hydrogen bonding between the residues. We treat the folded protein as an elastic body, and apply the elastic energy to describe the overall effect of the attractions.

Let \( d_{ij} \) be the distance between two contact residues \( i \) and \( j \), and \( d_0 \) be the default distance between them, which corresponds to the distance with minimum elastic energy \( \varepsilon_0 \). At the equilibrium state, the distance between any two contact residues limits to a small variance around the default distance \( d_0 \). Thus, we can apply the harmonic approximation to expand the elastic energy to the second order,

\[ E_{\text{ela}} = \frac{1}{2} \sum_{i,j=1}^{N} \chi_{ij} \left[ \varepsilon_0 + \frac{1}{2} \kappa (d_{ij} - d_0)^2 \right] \]  

(2)

where \( \varepsilon_0 < 0 \), and \( \kappa \) is the Hooke coefficient. \( \chi_{ij} \) equals to 1 if residues \( i \) and \( j \) are contact; and equals to 0 otherwise.

For each residue, we define its average contact number \( \bar{n}_i \) and root-mean-square contact distance \( d_i \) as

\[ \bar{n}_i = \sum_{j=1}^{N} \chi_{ij}, \quad d_i^2 = \frac{1}{\bar{n}_i} \sum_{j=1}^{N} \chi_{ij} d_{ij}^2. \]  

(3)

In the current study, we only consider the statistical average effect of the amino acids along the chain and disregard their difference. Thus, we can assume that \( \bar{n}_i \) and \( d_i \) are independent of the index \( i \). For simplicity, the subscript \( i \) will be omitted in following discussions. Moreover, from the harmonic approximation, \( d_0 \) is given by

\[ d_0 = \frac{1}{\bar{n}} \sum_{j=1}^{N} \chi_{0j} d_{0j}. \]  

(4)

Now, we can rewrite \( E_{\text{ela}} \) as

\[ E_{\text{ela}} = \frac{1}{4} \bar{n} N \kappa d_0^2 - \frac{1}{4} \bar{n} N \kappa d_0^2 + \frac{1}{2} \bar{n} N e_0. \]  

(5)

Since the last two terms are not concerned with the compactness, and do not affect our final results. We drop them and obtain

\[ E_{\text{ela}} = \frac{1}{4} \kappa \bar{n} N d_0^2 \]  

(6)

Here the root-mean-square contact distance \( d \) depends on \( R_g \) and \( N \), and measures the compactness of a protein structure.

For many chain molecules, their equilibrium structures reveal a self-similarity in the sense of number density distribution,\textsuperscript{14–19} which reads

\[ n/N \propto (r/R_g)^\alpha, \]  

(7)

where \( n \) is the average number of monomers within the distance \( r \) from any given monomers; and \( \alpha > 0 \) refers to the fractal dimension of a conformation. In the case of natural proteins under physiological conditions, an examination of proteins from PDB yields a fractal dimension \( \alpha \approx 2 \) (Fig. 5).

Applying eq. 7 to the neighborhood of a monomer with radius \( d \), we obtain

\[ (\bar{n} + 1)/N = (d/R_g)\nu. \]  

(8)

Thus the root-mean-square contact distance is given by

\[ d = (\bar{n} + 1)^{1/\nu} R_g/N^{1/\nu}. \]  

(9)

Substituting \( d \) into eq. 6, we obtain the elastic energy as

\[ E_{\text{ela}} = \frac{1}{4} \kappa \bar{n} (\bar{n} + 1)^{2/\nu} \frac{R_g^2}{N^{2/\nu - 1}}. \]  

(10)

Now the total energy is given by

\[ E = E_{\text{rep}} + E_{\text{ela}} = k_B T v \frac{N^2}{R_g^3} + \frac{1}{4} \kappa \bar{n} (\bar{n} + 1)^{2/\nu} \frac{R_g^2}{N^{2/\nu - 1}}. \]  

(11)

In the above argument, we neglect many other effects,\textsuperscript{7–10} such as the entropy effect \( (E_{\text{entropy}} = \gamma R_g^2/N \propto N^{3/2}) \), three-body repulsive interaction \( (E_{\text{three-body}} = \mu N^3/R_g^3 \propto N^{3-6}) \), etc. However, in the region we are interested \( (\nu \in (\frac{1}{3}, 1/2)) \), or \( \alpha \in (1, 3) \), it is easily to check that \( \lim_{N \to \infty} E_{\text{entropy}}/E_{\text{ela}} = 0 \), and
Thus the hydrophobicity, which is defined as the fraction of hydrophobic amino acids in a chain, is important for the compactness of a protein in water solution. If all amino acids are hydrophobic, corresponding to poor solvent conditions, the protein is highly compressed by solvent pressure and has a scaling exponent $\nu \approx 1/3$. On the other hand, if all amino acids are hydrophilic, which corresponds to good solvent conditions, the protein will be extended with $\nu \approx 3/5$. For natural proteins, their hydrophobicity has a Gaussian-like distribution with mean value about 0.5 (Fig. 3). And the corresponding scaling exponent, as shown in previous section, is $\nu \approx 2/5$. To determine how the scaling exponent depends on the hydrophobicity, we will start with a simple two-letters HP model based on the Kyte-Doolittle scale, which has been described in Section “Scaling exponent for protein under physiological condition.”

Consider an ensemble of proteins with the same hydrophobicity in certain solvent condition. And we denote the state of this protein-solvent coupled system as $X(h, p)$, where $h$ is the hydrophobicity of the protein, and $p$ is the polarity of the solvent. Therefore, the scaling exponent of a protein with hydrophobicity $h$ in a solvent with polarity $p$ can be written as $\nu(h, p)$. Furthermore, we assume that each protein can only be in one of three following states: the extended state $X_1$ (good solvent) with scaling exponent $\nu_1 = 3/5$, the native state $X_2$ (water solution) with $\nu_2 = 2/5$, and the compressed state $X_3$ (poor solvent) with $\nu_3 = 1/3$. Then the average scaling exponent is assumed to be a linear combination of the exponents for proteins in above three states.

$$\nu = \nu_1 \frac{[X_1]}{[X_T]} + \nu_2 \frac{[X_2]}{[X_T]} + \nu_3 \frac{[X_3]}{[X_T]},$$

where $[X_i]$ stands for the concentration of proteins in state $X_i$, and $[X_T] = [X_1] + [X_2] + [X_3]$. In the real case, there may be many other intermediate states, but we omit them for simplicity.

It is widely known that the hydrophobicity of a protein is closely related to the polarity of its surrounding solvent. Increasing the polarity of the solvent leads to higher hydrophobicity of the protein. Likewise, decreasing the polarity of solvent leads to an increase of the hydrophilicity. Thus, the dependence of a protein’s compactness on its hydrophobicity is equivalent to the dependence on the polarity of the solvent. And the latter process can be treated with a chemical process by adding polar or nonpolar molecules to the solvent.

Let us first consider the process of adding nonpolar molecules. As a result, the protein-solvent coupled system changes from state $X_2$ to state $X_1$ according to following chemical process

$$X_2 + mN \xrightarrow{k_2^+} X_1.$$

**DEPENDENCE ON HYDROPHOBICITY**

From previous discussions, the hydrophobic interaction is a main attractive force in protein folding. Thus the hydrophobicity, which is defined as the fraction of hydrophobic amino acids in a chain, is important for the compactness of a protein in water solution. If all amino acids are hydrophobic, corresponding to poor solvent conditions, the protein is highly compressed by solvent pressure and
Here, \( N \) stands for nonpolar molecules; and \( m \) is the number of nonpolar molecules required to change the system from state \( X_2 \) to \( X_1 \). Let \( [X_1] \) be the concentration of proteins in state \( X_1 \), and \([N]\) be the concentration of nonpolar molecules. When above reaction reaches equilibrium, we have

\[
[X_1]/[X_2] = (K_1[N])^m, \tag{16}
\]

where \( K_1 = (k_1^+ / k_1^-)^{1/m} \) is the association constant. Let \( X_T = [X_1] + [X_2] \) be the total concentration of proteins, then the average scaling exponent is given by

\[
\nu = \frac{[X_1]}{X_T} + \frac{[X_2]}{X_T} = \frac{v_1(K_1[N])^m + v_2}{(K_1[N])^m + 1}. \tag{17}
\]

Similarly, consider the process of adding polar molecules to change the protein-solvent coupled system from state \( X_2 \) to state \( X_3 \). Thus

\[
X_2 + nP \rightleftharpoons X_3. \tag{18}
\]

Here \( P \) stands for polar molecules. And \( n \) polar molecules are required to change the system from state \( X_2 \) to \( X_3 \). At the equilibrium state, we have

\[
[X_3]/[X_2] = (K_2[P])^n, \tag{19}
\]

where \( K_2 = (k_2^+ / k_2^-)^{1/n} \) is the association constant too. Let \( X_T = [X_2] + [X_3] \), we have the average scaling exponent

\[
\nu = \frac{[X_2]}{X_T} + \frac{[X_3]}{X_T} = \frac{v_2 + v_3(K_2[P])^n}{1 + (K_2[P])^n}. \tag{20}
\]

So the polarity of the solvent or the hydrophobicity of a protein is directly correlated to the concentration of the polar or nonpolar molecules. The real dependence between the polarity and the hydrophobicity may be very complex, and beyond the scope of our current study. Here we assume a most simple linear relationship

\[
\begin{align*}
[N] = 1 - 2h, & \quad 0 \leq h \leq 0.5 \\
[P] = 2h - 1, & \quad 0.5 < h \leq 1. \tag{21}
\end{align*}
\]

The validity of eq. 21 is left to future studies.

Taking \( K_1 = 2.5, K_2 = 3.2, m = n = 8 \), we obtain following equation for \( \nu(h) \), which fits well with the statistical data of natural proteins in Fig. 4.

\[
\nu(h) = \begin{cases} 
0.6 - \frac{0.2}{1 + (2.5 \times (1 - 2h))^8}, & 0 \leq h \leq 0.5 \\
0.33 - \frac{0.07}{1 + (3.2 \times (2h - 1))^8}, & 0.5 < h \leq 1. \end{cases} \tag{22}
\]

From eq. 22 and Figure 4, we can see that the scaling exponent of a protein shows a step function like dependence on its hydrophobicity. When the hydrophobicity \( h \) belongs to the regions \((0.2, 0.4)\) and \((0.6, 0.8)\), the scaling exponent is very sensitive to the hydrophobicity. Even a small variation of the hydrophobicity may cause a large structure change of the chain. But when the hydrophobicity \( h \) in the intervals \([0.2, 0.4], [0.4, 0.6]\) and \([0.8, 1]\), their corresponding scaling exponents are almost unchanged with \( \nu = 3/5, 2/5 \) and \( 1/3 \), respectively. For most natural proteins, their hydrophobicity concentrates in the region \([0.4, 0.6]\) (Fig. 3). This property shows the specificity of the protein, which is a consequence of natural selection.

CONCLUSION

In summary, we have derived a unified formula for the scaling law between the radius of gyration and the length of a protein. This formula demonstrates that the scaling exponent is generally correlated with the fractal dimension of a chain’s equilibrium conformations under certain solvent condition. Our result covers Flory’s theory for homopolymers in good and poor solvents as two extreme cases, as well as natural proteins under physiological conditions. The influence of a protein’s hydrophobicity on the scaling exponent has also been studied through a simple two-letters HP model, which shows good agreement with the statistical data from PDB. These results indicate that proteins may share some common principles with general chain molecules, despite their specialized nature as heteropolypeptides.

The authors thank C. C. Lin and Kerson Huang for their guidance and many useful comments. Authors thank Wen-An Yong and Weitao Sun for their helpful discussions.

REFERENCES AND NOTES


