Physica A 390 (2011) 3055–3059

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Self-organized critical model for protein folding

M.A. Moret *

Programa de Modelagem Computacional - SENAI - CIMATEC, 41650-010 Salvador, Bahia, Brazil
Departamento de Física - UEFS 44031-460 Feira de Santana, Bahia, Brazil

ARTICLE INFO

Article history:
Received 27 October 2010
Received in revised form 30 March 2011
Available online 13 April 2011

Keywords:
Tsallis statistics
Mass-size exponent
Solvent accessible surface area
Self-organized criticality
Scaling

ABSTRACT

The major factor that drives a protein toward collapse and folding is the hydrophobic effect. At the folding process a hydrophobic core is shielded by the solvent-accessible surface area of the protein. We study the fractal behavior of 5526 protein structures present in the Brookhaven Protein Data Bank. Power laws of protein mass, volume and solvent-accessible surface area are measured independently. The present findings indicate that self-organized criticality is an alternative explanation for the protein folding. Also we note that the protein packing is an independent and constant value because the self-similar behavior of the volumes and protein masses have the same fractal dimension. This power law guarantees that a protein is a complex system. From the analyzed data, q-Gaussian distributions seem to fit well this class of systems.

Proteins are involved directly or indirectly in all biological processes, and their functions range from catalysis of chemical reactions to the maintenance of the chemical potentials across cell membranes. They are synthesized on ribosomes as linear chains of amino acids in a specific order from information encoded within the cellular DNA. To function, it is necessary for these chains to fold into the unique native three-dimensional structure characteristic of each protein. This involves a complex molecular recognition phenomenon that depends on the cooperative action of many nonbonded interactions. As the number of possible conformations for a polypeptide chain is astronomically large, a systematic search to find the native (lowest energy) structure would require an almost infinite length of time, the so-called "Levinthal paradox" [1].

Protein folding is driven by hydrophobic forces [2]. It is well known that the native fold determines the packing but packing does not determine the native fold [3]. This view is corroborated by the widespread occurrence of protein families whose members assume the same fold without having a sequence similarity. However, there are a large number of ways in which the internal residues can pack together efficiently. Recently, some aspects of biological molecules were obtained by their fractal behavior. Thus, self-similarity were uncovered in, e.g., multifractality in the energy hyper-surface of the proteins and a possible alternative explanation of the Levinthal paradox [4], degree of compactness of the proteins [5–7], loss of the accessible surface area of amino acids and hydrophobicity scale [8], among others. Furthermore, the fractal methods corroborate to identify different states of the same system according to its different scaling behaviors, e.g., the fractal dimension is different for structures with (without) hydrogen bonds [4,9], or different long-range correlations in a liquid–vapor-phase transition of the solvent [10]. Then, the correct interpretation of the scaling results obtained by the fractal analysis is crucial for understanding the intrinsic geometry of the systems under study.

In this paper we are mainly interested in investigating the geometric characteristics of 5526 different protein chains deposited in the Brookhaven Protein Data Bank. Our strategy is to compare the fractal dimensions of solvent-accessible surface area, volume and mass of each protein chain. Here we show that solvent-accessible surface area scales with mass. This area scales with number of amino acids [11] with the same nontrivial exponent.

The behavior of the volume and solvent-accessible surface area are distinct. While the fractal dimensions of mass and volume are equal ($\delta_M = \delta_V = 2.47$) the dimension of the solvent-accessible surface area is $\delta_{\text{Area}} = 2.26 \pm 0.09$ [11].

* Corresponding address: Programa de Modelagem Computacional - SENAI - CIMATEC, 41650-010 Salvador, Bahia, Brazil. Tel.: +55 71 82272352.
E-mail address: mamoret@gmail.com.

0378-4371
doi:10.1016/j.physa.2011.04.008

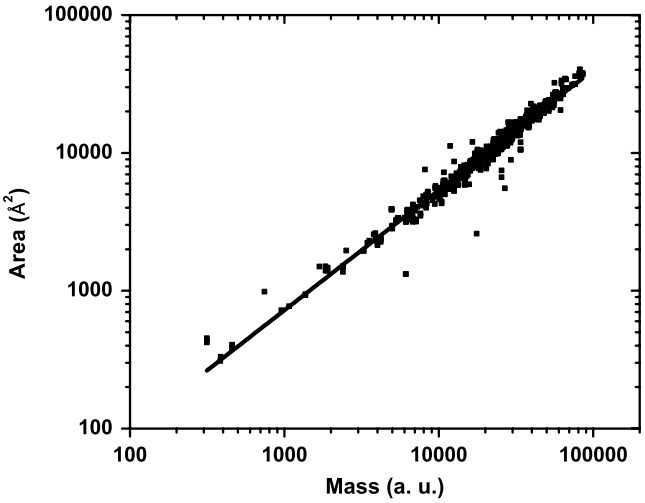

Fig. 1. The solvent-accessible solvent area as function of protein chain mass. We recall that the correlation coefficient of Pearson $R = 0.99$ and F statistic value ≈106 000 to the fitting (black line) and the power law follows $\gamma = 0.87 \pm 0.00$.

Therefore, the atoms that compose the protein are distributed in the macromolecule volume as a fractal object. In this context, the interactions among these atoms are made among the most different scales of the object, like a randomly packed spheres in percolation threshold [7] for volume [5] and due mass-size exponent [12] for the mass among different scales [6]. On the other hand, amino acids over the solvent-accessible surface area behave on average as irreversibly crushed wires [11].

The average packing density has been studied in different approaches. Voronoi tessellation methods were the first approach to analyze protein packing and they have been widely used for examining packing, volume and surface area [13–15]. By using the Voronoi tessellation method the mean interior packing density is 0.74 (van der Waals volume per total volume). Another interesting approach to evaluate the packing of residues is based on a coarse-grained scale. In this method a Monte Carlo algorithm is utilized for superimposing residue clusters collected from known protein structures. A residue cluster is composed of a central residue, and the set of all neighboring residues located within a first coordinate shell. A constant radius is used for defining the first coordinate volume [16]. From this approach it is possible to approximate two-thirds of the protein packing as an fcc geometry on a coarse-grained scale. The remaining one-third refer to residues which are more loosely or randomly packed.

Different models about protein packing were proposed, like the fcc presented above. Some of them are very instigating. The first model proposed protein packing as crystalline structures [13], other one proposed that these structures behave like liquid systems [14] and another suggested that they are like plastic structures [15]. Some different characteristics of proteins can be explained by each one of these models. Thus, the packing of the interior of proteins behaves like a solid structure, because their mean interior packing density ($\rho = 0.74$) is close to the one obtained from crystalline structures. On the other hand, the tertiary structure can be lost if we change a single amino acid, e.g., if we change one alanine to one proline inside the α-helix structure, this secondary structure must be missing. Plastic is an intermediate model between solid and liquid ones.

Recently, some models of packing have been proposed. Thus, replacing spheres with less symmetric objects (such as rods or ellipsoids) introduces new degrees of freedom that alter the packing structure and create new modes of response [17–20]. The density of jammed disordered packing using ellipsoids suggests that the higher density was directly related to the higher number of degrees of freedom per particle and thus the larger number of particle contacts required to mechanically stabilize the packing. Refs. [19,21] observed that the number of contacts per particle $Z \approx 10$ for spheroids, as compared to $Z \approx 6$ for spheres.

Fig. 1 depicts the behavior of solvent-accessible surface area as function of protein mass. The scaling exponent $\gamma = 0.87 \pm 0.00$. We recall that the same scaling exponent is obtained for solvent-accessible surface area as function of volume, as follows.

$$A \propto M^{\gamma} \propto V^{\gamma} \tag{1}$$

where A is the solvent-accessible surface area, M the protein chain mass and V the volume chain mass.

The power law, presented in Eq. (1), is far from the value obtained by Euclidian objects. Thus, the present result strongly indicates that the generation of proteins occurs in scale invariant media, as observed from the Fig. 1. In this sense, amino acids over the solvent-accessible surface area behave on average as irreversibly crushed wires [11]. On the other hand, the packing of amino acids over the surface depend on amino acids distributed inside the protein volume.

Recently, a simple mean field model [11] was proposed. In this context, energy can be stated as $E = E_{el} + E_{sa}$, where $E_{el} \propto R^n$ is the elastic energy, R is the radius of gyration of the protein configuration, n is a scaling exponent and $E_{sa} \propto M^2 \times R^{-3}$ is a self-avoidance energy [11]. From the elastic scaling exponent we observe that the solvent-accessible

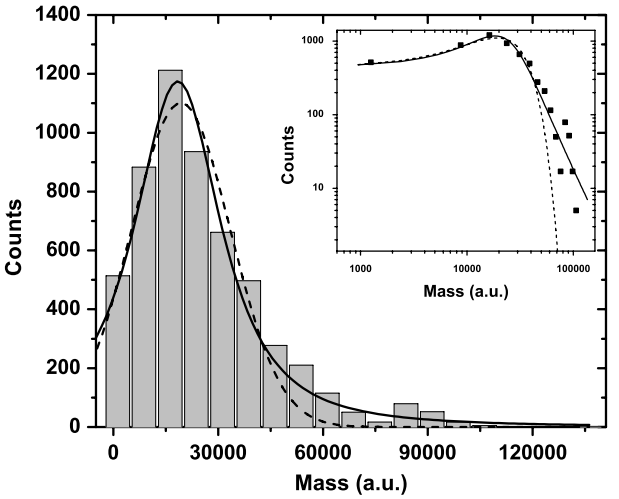

Fig. 2. The q-Gaussian distribution of 5526 mass of proteins (black line) and Gaussian distribution (dashed line). We recall that the correlation coefficient of Pearson $R = 0.98$ and F statistic value $\approx$150 to the q-Gaussian distribution (black line) to $q = 1.74$ as entropic index.

surface area restoring force is greater then the core restoring one. Then, proteins present an elegant and simple behavior. Thus, the inside protein packing behaves like random spheres in the percolation threshold [5–7].

On the other hand, Gaussian distributions do not fit well in fractal objects [22]. Long-tail distributions must represent complex systems like fractals. Long-tail distributions are solutions in a theory known as Tsallis Statistic (TS) [23,24]. We recall that TS has q-Gaussian distributions (long-tail ones) [25] as a solution and it is appropriate to analyze complex systems with long-range interactions.

In recent years, TS has received increasing attention due its success in the description of certain phenomena exhibiting atypical thermodynamical features. For example, the Tsallis formalism has been applied to protein folding [26–28], global optimization [29–31], stellar distributions [32], X-ray binary systems [33], dissipative systems [34], the nonlinear Fokker–Planck equation [35] and many others [36]. But TS was consolidated for complex systems analysis [36–40]. Commonly, TS covers this class of systems since it postulates a nonextensive (nonadditive) entropy S_q such that $S_q(A+B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B)$, where A and B are two independent systems in the sense that $P(x, x')_{A+B} = P(x)_A P(x')_B$. We recall that $P(x)$ is the probability density distribution of the macroscopic variable x, that characterizes the system. In this context, the so-called entropic index q is a measure of the degree of nonextensivity. In BG statistics, the entropy $S = -k \int P(x) \ln P(x) \mathrm{d}x$, gives rise to exponential probability density distributions, $P(x) \propto \exp(-x)$. Yet, within TS formalism, the maximization of the q-entropy

$$S_q = k \frac{\left(1 - \int [P(x)]^q \mathrm{d}x\right)}{(q-1)}, \tag{2}$$

produces power-law distributions called q-exponential distributions, defined by:

$$P_q(u) = A\, e_q^u \equiv A\,[1 - \beta\,(1-q)\,u]^{1/(1-q)}\,, \tag{3}$$

if $1 + (1-q)\,u \geq 0$ and $e_q^u = 0$ otherwise. In the limit $q \to 1$ the usual BG entropy is recovered, $S_1 \equiv S_{\mathrm{BG}}$, and the q-exponential distribution converges to the usual exponential distribution.

Fig. 2 depicts the behavior of protein-chain masses. Thus, as a consequence of the fitting analysis we observe that the protein masses are not Gaussian distributed for Gaussian distributions. A possible fit for masses is the q-Gaussian distribution ($q = 1.74 \pm 0.30$), as shown in Fig. 2 (volumes follow the same q-Gaussian). Also we observe that the dashed line (the Gaussian fitting) vanishes and does not follow the experimental values greater than 4500 a.u.. On the other hand, the black line (q-Gaussian fitting) fits well the experimental values.

However, it is interesting to notice that protein masses present a linear fit with the number of amino acids, with Pearson correlation coefficient $R = 0.99$, $F_{\mathrm{Value}} \geq 67,000$ and Prob $> F_{\mathrm{Value}} = 0$. Then, protein masses (and consequently protein volumes) are extensive variables. In this sense, the q-Gaussian behavior observed (Fig. 2) is only due to the self-similarity that is an inherent characteristic of this type of complex system.

Here, we observe some aspects of protein structure that suggest to us to propose a different way for protein packing. First aspect, protein masses and volumes are packed like random spheres in the percolation threshold [5–7]. In this sense, if we change one amino acid in the sequence the secondary structure can be broken, but the packing remains as random spheres in the percolation threshold. Then, these structures have a great packing observed from the scaling exponent of the volume and mass. Second, solvent-accessible surface area is a hydrophilic region that is in contact with solvent medium and it has

a larger roughness and it behaves on average as irreversibly crushed wires [41,42], as was proposed recently [11]. The third aspect that we taking into account is the q-Gaussian distribution observed from Fig. 2.

The widespread occurrence of different protein families whose members assume the same fold without having a sequence similarity can be explained by the internal residue packing that follows a large number of ways in which random spheres are packed in the percolation threshold (the first aspect of this model). The Amino acids exposed in the solvent-accessible surface area follow a crushed wires like packing. And, since protein packing follows q-Gaussian distribution, then this type of polymer behaves as a complex system. Recently from the self-similar hydrophobic scale [8] self-organized criticality (SOC) [43] in protein packing was suggested [44,45].

In summary, we investigate geometric characteristics of protein chains. We conclude that volumes, masses and solvent-accessible surface areas of globular proteins behave as fractal objects. The protein volume follows a q-Gaussian distribution as a direct consequence of the characteristic power law ($V \propto R^{\delta_V}$) [5]. As protein volumes and protein masses follow power laws with the same exponent ($\delta_V = \delta_M \approx 2.47$) [6] and these variables present a linear fit with the number of amino acids [11] the protein packing density has a constant value on average.

Finally, biological systems are believed to have evolved from simple to complex, from small to large, guided by a multitude of laws of nature. With a gradual increase of the evolving protein chains lengths, secondary structures have been reached when the flexible polypeptide chains present a possible amino acid sequence and a minimum number of amino acids. For example, different amino acids have been found to present weak though definite preference in favor or against being in α-helix structure, and the intrinsic helical propensity of some amino acids has been demonstrated to be position dependent [46]. In this sense, Ala, Glu, Leu, and Met are considered to be good α-helix promoters whereas Pro, Gly, Tyr, and Ser are considered to be poor ones. This behavior suggest to us to propose that the protein folding behaves as an SOC object. Thus, an isolated α-helix becomes a stable secondary structure with 13 or more amino acids [27,47]. In this sense, above a critical number of 13 amino acid residues the enhancement of the hydrogen bond number stabilizes the isolated polypeptide in an α-helix structure. In fact, for long-chain peptides, most of the possible H-bonds of the backbone tend to be formed. Then, the competition between H-bonds and dipole alignment turns the α-helix into a favorable conformation above the critical number of amino acids. This behavior can be viewed as an avalanche that occurs due that the critical number of amino acids. Moreover, the insertion or deletion into an amino acid sequence of a protein can lead to loss or change of the biological activity of the protein structure. All of these collective behaviors can be caused by an amino acid insertion (or deletion).

Acknowledgment

This work received financial support from CNPq (Brazilian federal grant agency).

References

[1] C. Levinthal, J. Chem. Phys. 65 (1968) 44.
[2] J.M. Yon, Cell. Mol. Life Sci. 53 (1997) 557.
[3] E.E. Lattman, G.D. Rose, Proc. Natl. Acad. Sci. 90 (1993) 439.
[4] M.A. Moret, P.G. Pascutti, K.C. Mundim, P.M. Bisch, E. Nogueira, Phys. Rev. E 63 (2001) 020901(R).
[5] J. Liang, K.A. Dill, Biophys. J. 81 (2001) 751.
[6] M.A. Moret, J.G.V. Miranda, E. Nogueira, M.C. Santana, G.F. Zebende, Phys. Rev. E 71 (2005) 012901.
[7] M.A. Moret, M.C. Santana, E. Nogueira, G.F. Zebende, Physica A 361 (2006) 250.
[8] M.A. Moret, G.F. Zebende, Phys. Rev. E 75 (2007) 011920.
[9] J.S. Helman, A. Coniglio, C. Tsallis, Phys. Rev. Lett. 53 (1984) 1195.
[10] G.F. Zebende, M.V.S. da Silva, A.C.P. Rosa, A.S. Alves, J.C.O. de Jesus, M.A. Moret, Physica A 342 (2004) 322.
[11] M.A. Moret, M.C. Santana, G.F. Zebende, P.G. Pascutti, Phys. Rev. E 80 (2009) 041908.
[12] M.A.F. Gomes, J. Phys. A 20 (1987) L283.
[13] F. Richards, J. Mol. Biol. 82 (1974) 1.
[14] J. Finney, J. Mol. Biol. 96 (1975) 721.
[15] B. Honig, J. Mol. Biol. 293 (1999) 283.
[16] Z. Bagci, R.L. Jernigan, I. Bahar, J. Chem. Phys. 116 (2002) 2269.
[17] F.X. Villarruel, B.E. Lauderdale, D.M. Mueth, H.M. Jaeger, Phys. Rev. E 61 (2000) 6914.
[18] S.R. Williams, A.P. Philipse, Phys. Rev. E 67 (2003) 051301.
[19] A. Donev, I. Cisse, D. Sachs, E.A. Variano, F.H. Stillinger, R. Connelly, S. Torquato, P.M. Chaikin, Science 303 (2004) 990.
[20] M. Mailman, C.F. Schreck, C.S. OHern, B. Chakraborty, Phys. Rev. Lett. 102 (2009) 255501.
[21] A. Donev, F.H. Stillinger, S. Torquato, Phys. Rev. Lett. 95 (2005) 090604.
[22] B.B. Mandelbrot, J. Bus. 36 (1963) 394;
B.B. Mandelbrot, The Fractal Geometry of Nature, W. H. Freeman and Co., New York, 1982.
[23] C. Tsallis, J. Stat. Phys. 52 (1988) 479.
[24] E.M.F. Curado, C. Tsallis, J. Phys. A 24 (1991) L69; J. Phys. A 24 (1991) 3187. Corrigenda.
[25] K.C. Mundim, Physica A 350 (2005) 338.
[26] C. Tsallis, G. Bemski, R.S. Mendes, Phys. Lett. A 257 (1999) 93.
[27] M.A. Moret, P.M. Bisch, K.C. Mundim, P.G. Pascutti, Biophys. J. 82 (2002) 1123.
[28] F.P. Agostini, D.D.O. Soares-Pinto, M.A. Moret, C. Osthoff, P.G. Pascutti, J. Comput. Chem. 27 (2006) 1142.
[29] M.A. Moret, P.G. Pascutti, P.M. Bisch, K.C. Mundim, J. Comput. Chem. 19 (1998) 647.
[30] M.A. Moret, P.M. Bisch, F.M.C. Vieira, Phys. Rev. E 57 (1998) R2535.
[31] M.A. Moret, P.G. Pascutti, P.M. Bisch, M.S.P. Mundim, K.C. Mundim, Physica A 363 (2006) 260.
[32] A.R. Plastino, A. Plastino, Phys. Lett. A 174 (1993) 384.
[33] M.A. Moret, V. de Senna, G.F. Zebende, P. Vaveliuk, Physica A 389 (2010) 854.

[34] M.L. Lyra, C. Tsallis, Phys. Rev. Lett. 80 (1998) 53.
[35] A.R. Platino, A. Plastino, Physica A 222 (1995) 347;
B.M. Boghosian, Phys. Rev. E 53 (1996) 4754;
A.R. Platino, A. Plastino, Braz. J. Phys. 29 (1999) 79;
L. Borland, et al., Eur. Phys. J. B 12 (1999) 285;
C. Anteneodo, C. Tsallis, J. Math. Phys. 44 (2003) 5194;
A.R. Plastino, et al., Astrophys. Space Sci. 290 (2004) 275.
[36] C. Tsallis, Braz. J. Phys. 29 (1999) 1;
C. Tsallis, Braz. J. Phys. 39 (2009) 337.
[37] P. Douglas, S. Bergamini, F. Renzoni, Phys. Rev. Lett. 96 (2006) 110601.
[38] B. Liu, J. Goree, Phys. Rev. Lett. 100 (2008) 055003.
[39] R.G. DeVoe, Phys. Rev. Lett. 102 (2009) 063001.
[40] R.M. Pickup, R. Cywinski, C. Pappas, B. Farago, P. Fouquet, Phys. Rev. Lett. 102 (2009) 097202.
[41] M.A.F. Gomes, V.P. Brito, M.S. Araújo, J. Braz. Chem. Soc. 19 (2008) 293.
[42] M.A.F. Gomes, V.P. Brito, A.S.O. Coelho, C.C. Donato, J. Phys. D: Appl. Phys. 41 (2008) 235408.
[43] P. Bak, C. Tang, K. Wiesenfeld, Phys. Rev. Lett. 59 (1987) 381.
[44] J.C. Phillips, Proc. Natl. Acad. Sci. USA 106 (2009) 3107.
[45] J.C. Phillips, Proc. Natl. Acad. Sci. USA 106 (2009) 3113.
[46] M. Petukhov, K. Uegaki, N. Yumoto, S. Yoshikawa, L. Serrano, Protein Sci. 8 (1999) 2144.
[47] K.R. Shoemaker, P.S. Kim, E.J. York, J.M. Stewart, R.L. Baldwin, Nature 326 (1987) 563.