International Journal of Modern Physics C
Vol. 20, No. 12 (2009) 1981–1988

World Scientific
www.worldscientific.com

GEOMETRIC STRUCTURAL ASPECTS OF PROTEINS AND NEWCOMB–BENFORD LAW

M. A. MORET*,†,‡, V. de SENNA*,§, M. C. SANTANA† and G. F. ZEBENDE*,†,¶

*Programa de Modelagem Computational — SENAI — Cimatec
Salvador, Bahia, Brazil

†Departamento de Física — UEFS 44031-460 Feira de Santana
Bahia, Brazil

‡mamoret@gmail.com
§vsenna@gmail.com
¶zebende@uefs.br

Received 8 June 2009
Accepted 28 August 2009

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 behavior of the numbers in 5526 protein structures present in the Brookhaven Protein Data Bank. The first digit of mass, volume, average radius and solvent-accessible surface area are measured independently and we observe that most of these geometric observables obey the Newcomb–Benford law. That is volume, mass and average radius obey the Newcomb–Benford law. Nevertheless, the digits of the solvent-accessible surface area do not agree with the Newcomb–Benford law. The present findings indicate that the hydrophobic effect is responsible for the anomalous first digit behavior of solvent-accessible surface areas.

Keywords: Hydrophobicity; protein packing; complex systems.

PACS Nos.: 0.6.30.Ak, 05.40.Aa, 05.45.Tp, 89.75.Da, 96.60.Rd, 97.80.-d.

1. Introduction

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 relatively weak 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. Recently, significant progress has been made towards solving this problem, the so-called "Levinthal paradox"[1] and obtaining an understanding of the mechanism of folding.[2–7] The new insights have come about through advances in experimental studies[8–17] and theoretical approaches that simulate the folding process with simplified models in different contexts.[18–31]

Protein folding is driven by hydrophobic forces.[32] Lattman and Rose[33] analyzed globular proteins and concluded that the native fold determines the packing but packing does not determine the native fold. This view is corroborated by the widespread occurrence of protein families whose members assume the same fold without having a sequence similarity. Evidently, there are a large number of ways in which the internal residues can pack together efficiently. As a consequence of steric constraints in compact polymers, helical and sheets structures appear.[2,18] Exhaustive simulations of the conformations indicate that the proportion of helices and sheets increases dramatically with the number of intrachain contacts.[32] Recently, the fractal dimension of proteins was analyzed[34,28,30] and it is observed that this type of biological polymers pack like random spheres in the percolation threshold.

Several aspects of the protein folding process were examined recently. The current "new" view[3,16,18,19] is that proteins are able to find their native states in the observed time because a bias in their energy surface reduces the number of configurations that are sampled in the folding process,[5,6] relative to the astronomic number envisaged in the Levinthal paradox.[1] Equally important, the transition region, from which folding to the native state is fast, includes a large number of configurations.[19] A focus on the overall energy or free energy surface (the "energy landscape") replaces the specific folding pathways suggested by Levinthal[1] with a distribution of the folding trajectories over multiple pathways. Although experimental data have provided specific information on no more than a few competing pathways,[12] each of these may well involve broad ensembles of structures except in the neighborhood of the native state. For instance, the folding funnel theory[3] proposes to describe the thermodynamics and kinetic behavior of the transformation of unfolded molecule to the native state, from the number of native contacts. The folding funnel theory shows that any polypeptide chain explores the folding routes toward the native structure through intermediates consisting of populations of partially folded species whose number decrease as the protein navigates down to the minimum of the energy landscape.[3] In this sense, there are many different ways for the collapsed globule to reach the native state, in accordance with the new view of protein folding. It thus appears that even for small α-helical proteins a wide range of mechanisms that encompass both the "old" and "new" views are possible.[5,6] On the other hand, if another set of parameters are used in the molecular dynamics, a few number of states are sufficient for one structure to collapse and fold into the native one.

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 measure solvent-accessible surface area, average radius and mass of each protein chain. These intrinsic characteristics of the protein structures are responsible for the explanation of several aspects of those molecules, like the high compactness of those molecules obtained by fractal strategies.[34,28,30] We also measure the solvent-accessible surface area as function of the number of amino acids.

2. Newcomb−Benford Law

We study the behavior of the numbers in the geometric observable variables by using the Newcomb–Benford law. Newcomb–Benford law, also known as the Law of Anomalous Numbers, has been known for a long time. It was first proposed by the Canadian born astronomer and mathematician Simon Newcomb in 1881[35] and popularized 57 years later by the physicist Frank Benford.[36] In its general form the law specifies that for any base $b > 1$, $Prob(mantissa(\text{base } b) \leq t/b) = \log_b t$ for all $t \in [1, b)$. As a direct consequence for base 10 systems it implies, somewhat unexpectedly, that

$$Prob(\text{first significant digit } d) = \log_{10}\left(1 + \frac{1}{d}\right) \tag{1}$$

where $d = 1, 2, \ldots, 9$ are the first significant digit.

Many explanations have been put forward for Newcomb–Benford's law and the interested reader is referred to Refs. 37–42. It has been noted by these authors that base invariance, scale invariance, geometric laws and random samples from random distributions all give rise to this first digit phenomenon. Reference 43 examine survival distributions that obey Newcomb–Benford law and they observe that increasing the value of the shape parameter in these distributions usually give a better fit to the law. Reference 44 examines the X-ray of astrophysical sources and observes that these objects obey the Newcomb–Benford law, presenting long tail distributions.

3. Results

We recall that this analysis was carried out over geometric characteristic aspects of 5526 protein chains deposited in the Brookhaven Protein Data Bank (PDB). All 5526 protein chains possess known structure with well-refined and high-resolution proteins (resolution lower than 2.0 Å). Finally, these same protein chains were used to measure the mass-size exponent[28] and the average packing density.[30]

Figure 1 shows the first digit behavior of protein-chain masses. Thus, as a consequence of application of Newcomb–Benford law, this analysis (Fig. 1) shows that the protein masses are not Gaussian distributed for Gaussian distributions do not follow Newcomb–Benford distribution. As the protein volume is proportional to the protein mass, and these two variables have the same fractal dimension ($V \propto R^{\delta=2.47}$

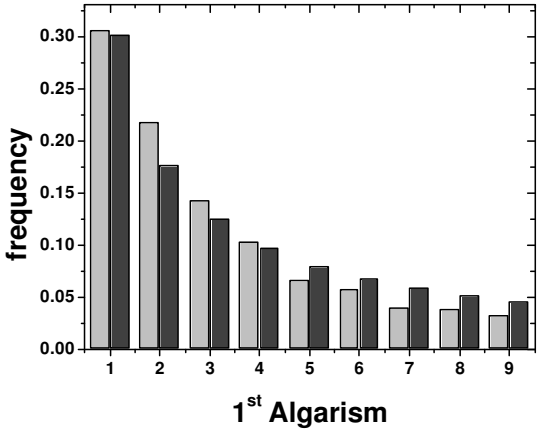

Fig. 1. First digits from 5526 mass of proteins (light gray) and Newcomb–Benford law (light gray). We recall that the p-value = 0.25.

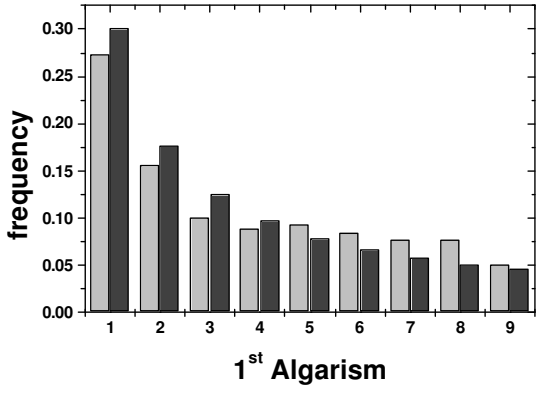

Fig. 2. First digits from the average radius (light gray) and Newcomb–Benford law (gray). We recall that the p-value = 0.04 for this data.

and $M \propto R^{\delta=2.47}$),[28,34] we conclude that protein volume also follows the Newcomb–Benford law.

Again, we notice that the first digit of the average radius of the sample of 5526 protein chains, as shown in Fig. 2 follows Newcomb–Benford law.

In order to extend our analysis, we look into the behavior, regarding Newcomb–Benford law, of the number of amino acids and solvent-accessible surface area of protein chains. The concept of accessible surface area was proposed by the solvent-accessible surface model, proposed by Lee and Richards[45] and it has found many applications in the study of proteins.[46–53] In this sense, the outer surface of a protein molecule can be identified by using ideas of Lee and Richards[45] on rolling a watersized probe sphere over a molecule. The surface of a macromolecule can be defined to be the part of the molecule that is accessible to solvent. The solvent

molecule (water) is represented by a sphere of radius 1.4–1.7 Å, called the probe sphere.

The solvent-excluded volume is that volume of space that the probe is excluded from by collisions with the atoms of the molecule. This volume is bounded by the molecular surface. All surface and volume methods have a dependence on factors such as atomic radii, the probe radius, the quality of atomic coordinates, and whether explicit hydrogen atoms have been included. For protein and nucleic acid molecules, the hydrogen atom positions are generally not known, and heavy atoms with hydrogens are approximated by a single sphere, whose van der Waals radius is augmented by one- to three-tenths of Å. Of course, all molecules have thermal motions, which are not modeled by these static, geometrical methods. The accessible surface is the trace of the probe sphere center as it rolls over the molecule of interest. The contact surface is that part of the van der Waals surface of the atom that can be touched by a probe sphere. The re-entrant surface is the inward facing part of the probe sphere as it is touching more than one atom. Together, the contact surface and the re-entrant surface form a continuous sheet called the molecular surface.

Thus, the outermost atoms can be represented as atomic spheres having the appropriate van der Waals radii ($r = 1.4$ Å). The solvent-accessible molecule is enclosed by the surface swept out by the center of the probe ball spheres. In this case, the protein surface is defined by the sum of the van der Waals radii of the outermost atoms plus the solvent probe spheres.

The analysis was performed according to the method proposed by Richards.[47] We measure the solvent-accessible surface area of 1825 protein chains. In Fig. 3 we can see the first digit behavior of these solvent-accessible surface areas and number

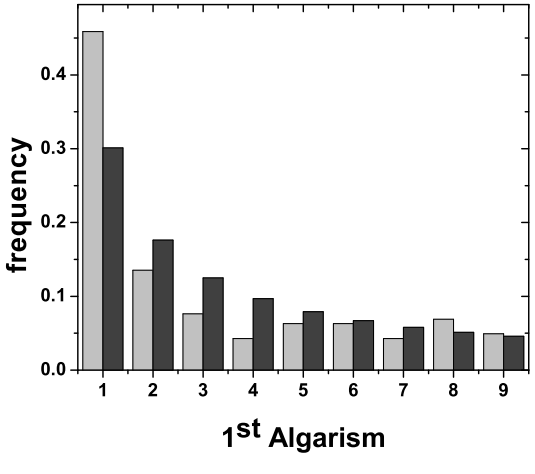

Fig. 3. First digits from the solvent-accessible surface area (light gray) and Newcomb–Benford law (gray), where the p-values = 0.

of amino acids. It is clear that there are large discrepancies with respect to the Newcomb–Benford law.

4. Discussions

This is a novel methodology to study proteins. The obtained results for geometric properties of proteins show that mass, volume and radius have a nonhomogeneous structure, which are distributed over the geometry. Furthermore, different regions in these variables present similar profiles, when they are mapped using different scales.[7,28,30] Different systems present this statistically self-similar behavior.

The fact that the solvent-accessible surface area is in disagreement with the Newcomb–Benford law is in itself very interesting. Usually, measures of area or indeed any geometric variables follow this law. Analyzing the solvent-accessible surface area we notice that if the packing of amino acid residues is based on a coarse-grained scale,[54] it is possible to approximate two-thirds of the protein packing as a fcc geometry on this coarse-grained scale. The remaining one-third refers to residues of the accessible surface area that are more randomly packed.[54] These randomly packed amino acid residues compose the accessible surface area and they directly interact with the water-solvent first shell.

In summary, we investigate geometric characteristics of protein chains. We conclude that most geometric aspects of proteins follow Newcomb–Benford law. The main exceptions are solvent-accessible surface areas that are in disagreement with the Newcomb–Benford law. The solvent-accessible surface area directly interact with the dipole potential from the water-solvent first shell. As this interaction is represented by a sphere of radius 1.4 Å, re-entrant surfaces below 1.4 Å are neglected due to these voids in area do not have a biophysical sense if $R < 1.4$ Å. This interaction can supply an explanation as to why the surface area is in disagreement with Newcomb–Benford law. We recall that in the bulk some other potentials are present, e.g. Coulomb and Lennard–Jones ones. Then, this result leads to propose that hydrophobicity changes the accessibility of the surface area. Therefore hydrophobic effects seem to be responsible for the Newcomb–Benford law failure to apply to the solvent-accessible surface area.

Finally, we study globular proteins and these structures can be composed for one protein chain or more chains. In this paper we analyzed proteins with one monomer and protein chains of the other cases (dimer, trimer, etc.) only. Nevertheless, this result can be generalized to proteins independently of number of monomers.

Acknowledgment

This work received financial support from CNPq (Brazilian federal grant agency) and FAPESB (Bahia state grant agency).

References

1. C. Levinthal, *J. Chem. Phys.* **65**, 44 (1968).

2. K. A. Dill, *Biochemistry* **29**, 7133 (1990).
3. P. G. Wolynes, J. N. Onuchic and D. Thirumalai, *Science* **267**, 1619 (1995).
4. J. M. Yon, *Cell. Mol. Lif. Sci.* **53**, 557 (1997).
5. Y. Zhou and M. Karplus, *J. Mol. Biol.* **293**, 917 (1999).
6. Y. Zhou and M. Karplus, *Nature* **401**, 400 (1999).
7. M. A. Moret *et al.*, *Phys. Rev. E* **63**, 020901(R) (2001).
8. F. M. Richards, *Sci. Am.* **264**, 34 (1991).
9. C. M. Dobson, *Curr. Opin. Struct. Biol.* **2**, 6 (1992).
10. G. A. Elöve *et al.*, *Biochemistry* **31**, 6876 (1992).
11. D. T. Haynie and E. Freire, *Proteins* **16**, 115 (1993).
12. T. Kiefhaber *et al.*, *Protein Sci.* **1**, 1162 (1992).
13. S. E. Radford, C. M. Dobson and P. A. Evans, *Nature* **358**, 302 (1992).
14. S. Khorazanizadeh *et al.*, *Biochemistry* **32**, 7054 (1993).
15. P. A. Evans and S. E. Radford, *Curr. Opin. Struct. Biol.* **4**, 100 (1994).
16. R. L. Baldwin, *J. Biomol. NMR* **5**, 103 (1995).
17. D. T. Clarke *et al.*, *Proc. Natl. Acad. Sci. USA* **96**, 7232 (1999).
18. K. A. Dill, *Curr. Opin. Struct. Biol.* **3**, 99 (1993).
19. A. Sali, E. I. Shakhnovich and M. Karplus, *Nature* **369**, 248 (1994).
20. V. S. Pande *et al.*, *Curr. Opin. Struct. Biol.* **8**, 68 (1998).
21. M. A. Moret *et al.*, *J. Comput. Chem.* **19**, 647 (1998).
22. M. A. Moret, P. M. Bisch and F. M. C. Vieira, *Phys. Rev. E* **57**, R2535 (1998).
23. P. G. Pascutti *et al.*, *J. Comp. Chem.* **20**, 971 (1999).
24. D. J. Wales and H. A. Scheraga, *Science* **285**, 1368 (1999).
25. A. Hiltpold *et al.*, *J. Phys. Chem. B* **104**, 10080 (2000).
26. M. A. Moret *et al.*, *Biophys. J.* **82**, 1123 (2002).
27. M. A. Moret *et al.*, *Phys. A* **353**, 353 (2005).
28. M. A. Moret *et al.*, *Phys. Rev. E* **71**, 012901 (2005).
29. F. P. Agostini *et al.*, *J. Comput. Chem.* **27**, 1142 (2006).
30. M. A. Moret *et al.*, *Phys. A* **361**, 250 (2006).
31. M. A. Moret *et al.*, *Phys. A* **363**, 260 (2006).
32. D. Voet and J. Voet, *Biochemistry*, 2nd edn. (Wiley, New York, 1995).
33. E. E. Lattman and G. D. Rose, *Proc. Nat. Acad. Sci.* **90**, 439 (1993).
34. J. Liang and K. A. Dill, *Biophys. J.* **81**, 751 (2001).
35. S. Newcomb, *Amer. J. Math.* **4**, 39 (1881).
36. F. Benford, *Proc. Amer. Philos. Soc.* **78**, 551 (1938).
37. R. A. Raimi, *Sci. Am.* **221**, 109 (1969).
38. R. A. Raimi, *Am. Math. Mon.* **83**, 521 (1976).
39. B. Buck, A. C. Merchant and S. M. Perez, *Eur. J. Phys.* **14**, 59 (1993).
40. T. P. Hill, *Statist. Sci.* **10**, 354 (1995).
41. T. P. Hill, *Am. Scientist* **86**, 358 (1998).
42. L. Pietronero, E. Tosatti and A. Vespignani, *Phys. A* **293**, 297 (2001).
43. L. M. Leemis, B. W. Schmeiser and D. L. Evans, *Am. Statistics* **54**, 236 (2000).
44. M. A. Moret *et al.*, *Int. J. Mod. Phys. C* **17**, 1597 (2006).
45. B. Lee and F. Richards, *J. Mol. Biol.* **55**, 379 (1971).
46. C. Chothia, *Nature* **248**, 338 (1974).
47. F. Richards, *J. Mol. Biol.* **82**, 1 (1974).
48. C. Chothia, *Nature* **254**, 304 (1975).
49. J. Finney, *J. Mol. Biol.* **96**, 721 (1975).
50. J. Janin, *Nature* **277**, 491 (1979).
51. G. D. Rose *et al.*, *Science* **229**, 834 (1985).

52. T. Hessa *et al.*, *Nature* **433**, 377 (2005).
53. M. A. Moret and G. F. Zebende, *Phys. Rev. E* **75**, 011920 (2007).
54. Z. Bagci, R. L. Jernigan and I. Bahar, *J. Chem. Phys.* **116**, 2269 (2002).