Amino acid hydrophobicity and accessible surface area

M. A. Moret* and G. F. Zebende†
Programa de Modelagem Computacional-Fundação Visconde de Cairu, 40226-900 Salvador, Bahia, Brazil and Departamento de Física-UEFS, 44031-460 Feira de Santana, Bahia, Brazil
(Received 2 December 2005; revised manuscript received 20 October 2006; published 19 January 2007)

It is well known that the hydrophobic effect is the major factor that drives a protein toward collapse and folding. We analyze the variation of the solvent-accessible surface area of amino acids in small fragments of protein ($3 \leqslant N \leqslant 45$). In this way, we look into 5526 protein chains deposited in the Brookhaven Protein Data Bank. The accessible surface area behaves as a power law for $N \geqslant 9$. The comparison between the loss of accessible area and the self-similar behavior gives us a measure of the possibility of an amino acid to have apolar or polar side chain. It is therefore possible to infer about amino acid hydrophobicity, i.e., if one amino acid has a hydrophobic side chain or if it has a hydrophilic one. Furthermore, the present findings indicate that the variation of the accessible surface area describes an alternative hydrophobicity scale.

DOI: 10.1103/PhysRevE.75.011920 PACS number(s): 87.14.Ee, 05.45.Df, 87.15.−v

The study of fractal characteristics of the proteins provides countless results. The fractal analysis uncovered self-similarity in many research fields such as cluster dimension of proteins [1,2], anomalous temperature dependence of the Raman spin-lattice relaxation rates [3,4], relation between the fractal dimension and the number of hydrogen bridges [5], multifractality in the energy hypersurface of the proteins [6], packing of small protein fragments [7], degree of compactness of the proteins [8] as well as the measurement of the average packing density [9], among others. Furthermore, the fractal methods identify different states of the same system according to its different scaling behavior, e.g., the fractal dimension is different for structures with (without) hydrogen bonds [5,6], or different long-range correlations in a liquid-vapor-phase transition of the solvent [10]. In this sense, the correct interpretation of the scaling results obtained by the fractal analysis is crucial to understand the intrinsic geometry (and sometimes dynamics) of the systems under study.

Proteins are involved in virtually every biological process and their functions range from oxygen transport to potential maintenance across cell membranes. They are synthesized on ribosomes as linear chains of amino acids in a specific order. These protein chains necessarily fold into the unique tertiary three-dimensional structure characteristic of each protein (native structure). This process involves a complex molecular recognition phenomenon that depends on the cooperative action of relatively weak nonbonded interactions. In this sense, several empirical rules of the protein folding can be deduced from studies of protein *in vitro* [11]. A global knowledge of several different folding theories [12–18] is crucial to understand the process of protein folding. As the number of possible molecular conformations for a polypeptide chain is astronomically large [6,11,18–22], a systematic search to find the native structure (lowest energy) would require an enormous amount of time [19,22–24]. However, alternative strategies based on structural properties can give powerful insights about the protein folding process.

*Electronic address: moret@cairu.br
†Electronic address: zebende@cairu.br

It is well known that protein folding is driven by hydrophobic forces. Thus, while protein folding occurs, hydrophobic groups are expelled spontaneously from water engendering a sequestered, solvent-shielded core. Thereby, a folded protein is a compact structure where all the amino acids are packed.

The concept of accessible surface area was proposed by the solvent-accessible surface model [25]. Thus, the outermost atoms can be represented as atomic spheres having the appropriate van der Waals radii. The solvent-accessible molecule is enclosed by the surface defined 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.

Packing in proteins was first analyzed by using a Voronoi analysis for proteins in a space-filling model [26]. In this model, each atom is taken to be a sphere with a fixed radius given by the van der Waals interactions. As consequence, the average packing density in proteins is as high as that inside crystalline solids [26–29]. Several important studies of packing in proteins have been carried out [30–36]. For void detection and calculation, there are many numerical [25,37–42] as well as analytical methods [43–46]. Yet, it remains a challenge to identify pockets and to calculate their sizes. Protein compactness has been studied by varied approaches and in recent years several methods were proposed based on, for instance, Delaunay tesselation [47], coarse-grained scale [48], mass-size exponent [8,9,49] among others. The Delaunay triangulation allows us to compute packing densities such as the mean interior packing density: 0.74 given by van der Waals volume divided by the total volume [47]. Another interesting approach to measure the packing of residues is based on a coarse-grained scale [48]. From this, it is possible to approximate two-thirds of the protein packing as a fcc geometry on a coarse-grained scale. The remaining one-third refers to residues that are more randomly packed [48]. The average packing density ($\rho = M/V$) of protein chains can be investigated from the scaling analysis of the mass-size exponent [8]. This analysis shows a tendency to stabilize the average packing density at $\rho = 0.86$ a.u./Å^3 [9].

Our main interest is to investigate hydrophobic aspects

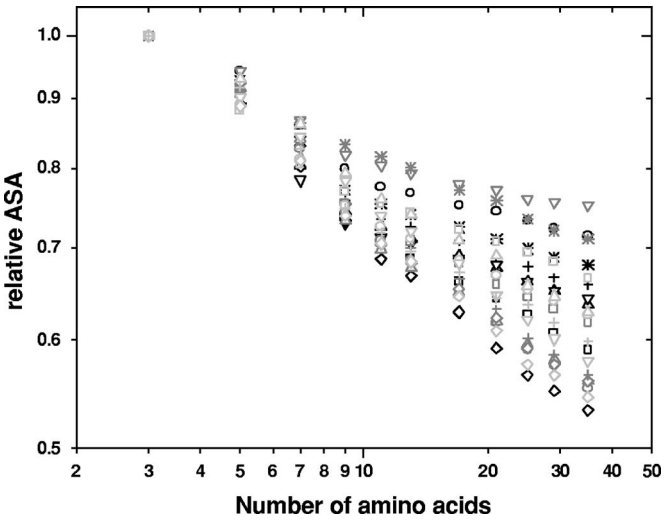

FIG. 1. The accessible surface area (ASA) of side chain in the central residue as function of number of neighbor amino acids for alanine (black square), arginine (black circle), asparagine (black up triangle), aspartic acid (black down triangle), cysteine (black diamond), glutamine (black cross), glutamic acid (black star), histidine (gray square), isoleucine (gray circle), leucine (gray up triangle), lysine (gray down triangle), methionine (gray diamond), phenylalanine (gray cross), Proline (gray star), serine (light gray square), threonine (light gray circle), tryptophan (light gray up triangle), tyrosine (light gray down triangle), valine (light gray diamond), glycine (light gray cross).

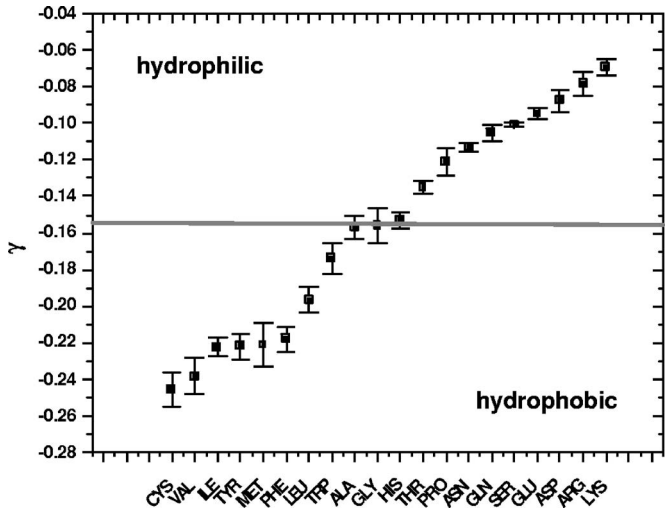

FIG. 2. The exponent γ $(9 \leqslant N \leqslant 35)$ for biological amino acids. The dashed grey line separates the hydrophilic amino acids ($\gamma > -0.155$) from the hydrophobic ones ($\gamma < -0.155$).

associated with the amino acid packing and their consequences. We show that the loss of the solvent-accessible surface area of individual amino acid behaves as a power law, where the exponent defines the hydrophobicity of an amino acid, i.e., it either has an apolar side chain or a polar one.

A way to measure the compactness in small protein segments ($N < 40$ amino acids) is by using Flory exponent [50], which was measured for protein fragments [7]. In our analysis, the loss of solvent-accessible surface area of amino acids was performed according to the method proposed by Richards [26], i.e., by using Voronoy Tesselation. Several protein fragments of length 3 to 45 were randomly extracted of the protein chains. All 5526 protein chains that we used to measure the loss of solvent-accessible surface area of central amino acids possess known structures with well-refined and high-resolution proteins (resolution lower than 2.0 Å). We remark that the same 5526 protein chains were used to measure the mass-size exponent [8] and the average packing density [9]. We measure the solvent-accessible surface area (ASA) of amino acids observing that it behaves as a power law when number of amino acids is greater than eight, as shown in Fig. 1. Figure 1 depicts the behavior of the average solvent-accessible surface area of the side chain in the central residue as a function of the number of neighbor amino acids of protein segments extracted from the protein data bank. In the case of glycine, that lacks a side chain, we measure the loss of the C_α area.

We observe that each amino acid obeys a power law when the number of neighbors increases

$$\mathrm{ASA} \propto N^{\gamma}, \tag{1}$$

where ASA is the solvent-accessible surface area and N is the number of amino acids.

This self-similar behavior allows us to measure if the amino acid has a hydrophilic side chain or hydrophobic one. Thus, from the behavior of the accessible surface area, we notice that for $N \geqslant 9$ a power law representing the loss of solvent-accessible surface area is obtained. The scaling exponent determines the amino acid hydrophobicity. From Fig. 2, we observe that for exponents lower than $\gamma = -0.155$ all amino acids have hydrophobics side chains.

From the above analysis of self-similarity it follows that the accessible surface area of amino acids is entirely correlated with the hydrophobicity, because as fast as the amino acid loses the accessible surface area the more hydrophobic it will be. From amino acids behavior, Fig. 2, we propose another hydrophobicity scale based on the solvent-accessible surface area. Several hydrophobicity scales were proposed and Table I shows some of these scales.

It is clear that there are significant differences between the scales shown in Table I, with respect to the placement of specific amino acids. However, from the obtained results we observe that most of the charged amino acids (LYS, ARG, ASP, and GLU) correspond to the more polar ones. In general, the standard hydrophobicity scales do not present the same property, viz. charged amino acids are the more polar ones (for instance see E, F and G scales [63–65] in Table I). Also, Refs. [63,65] place cysteine as the most hydrophobic residue, as our scale does. The reason for this difference is the fundamentally different methods used for constructing the scales. Our proposed scale and two of the others were constructed by examining proteins with known structures. Scales [63,65] define hydrophobic character as the tendency for a residue to be found inside of a protein, rather than on its surface. In the case of cysteine, because it is involved in disulphide bonds that must necessarily occur inside of a

TABLE I. The present findings (this work) in comparison with different hydrophobicity scales [59–65]. The capital letters (*A*, *B*, *C*, *D*, *E*, *F*, and *G*) are merely headings.

This work	*A* [59]	*B* [60]	*C* [61]	*D* [62]	*E* [63]	*F* [64]	*G* [65]
LYS	ARG	GLU	ASP	ASP	LYS	HIS	LYS
ARG	LYS	ASP	LYS	LYS	ARG	ARG	GLU
ASP	ASP	LYS	GLU	GLU	GLU	LYS	ASP
GLU	GLU	HIS	ARG	ARG	GLN	GLN	GLN
SER	ASN	ARG	GLN	SER	ASP	ASN	ASN
GLN	GLN	GLN	PRO	GLN	ASN	ASP	ARG
ASN	HIS	PRO	HIS	ASN	TYR	TYR	PRO
PRO	PRO	ASN	ASN	PRO	PRO	GLU	SER
THR	TYR	ALA	SER	GLY	THR	PRO	THR
HIS	TRP	THR	GLY	THR	HIS	SER	GLY
GLY	SER	SER	TYR	HIS	SER	THR	ALA
ALA	THR	VAL	THR	ALA	TRP	TRP	TYR
TRP	GLY	GLY	TRP	CYS	GLY	GLY	HIS
LEU	ALA	MET	ALA	MET	ALA	ALA	TRP
PHE	MET	CYS	MET	VAL	MET	MET	MET
MET	CYS	ILE	CYS	LEU	PHE	LEU	LEU
TYR	PHE	LEU	VAL	ILE	LEU	PHE	VAL
ILE	LEU	TYR	PHE	TYR	VAL	CYS	ILE
VAL	VAL	PHE	LEU	PHE	ILE	VAL	PHE
CYS	ILE	TRP	ILE	TRP	CYS	ILE	CYS

globular structure, cysteine comes up as extremely hydrophobic.

In summary, 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, the stage should have been reached when the flexible polypeptide chains would frequently make loops with the ends coming in contact. This loop closure phenomenon, well known in polymer physics [51], is characterized by the optimum contour length of the loops, about 25 to 30 residues in the case of proteins [52–55]. These standard sized loops were, indeed, discovered [52–55] to be a major building block in proteins. Here we note that small fragments of proteins with $N>35$ break the power-law pattern.

The apolar amino acids tend to associate forming a hydrophobic core. In this process, a protein collapse and packing turns the geometry of these macromolecules into a fractal object. The fractal dimension ($\delta=2.47$) is close to the dimension of randomly packed spheres at the percolation threshold ($\delta=2.50$) [56–58]. Therefore, this scenario suggests that the amino acids are packed in a similar way to random spheres in the percolation threshold [9]. On the other hand, the γ exponent indicates the loss rate of the solvent-accessible surface area of amino acids. Hence, the behavior of the γ exponent supplies a new hydrophobicity scale.

Concluding, the understanding of protein folding is of interest in itself and important for the analysis of many events, like those involved in cellular regulation. The notion of how changes in the solvent-accessible surface area of amino acids could be important. For instance, the sequence to information analysis used by the various genome projects is important to the design of proteins with functions. Also, it is important to the development of new therapeutic strategies for the treatment and prevention of human diseases that are associated with the failure of proteins to fold correctly.

The authors wish to thank Pablo Vaveliuk and Valter de Senna for suggestions and very useful discussions.

[1] J. Feder, T. Jossang, and E. Rosenqvist, Phys. Rev. Lett. **53**, 1403 (1984).
[2] T. Jossang, J. Feder, and E. Resenqvist, J. Chem. Phys. **120**, 1 (1984).
[3] J. T. Colvin, and H. J. Stapleton, J. Chem. Phys. **82**, 4699 (1985).
[4] G. C. Wagner, J. T. Colvin, J. P. Allen, and H. J. Stapleton, J. Am. Chem. Soc. **107**, 5589 (1985).
[5] J. S. Helman, A. Coniglio, and C. Tsallis, Phys. Rev. Lett. **53**, 1195 (1984).
[6] M. A. Moret, P. G. Pascutti, K. C. Mundim, P. M. Bisch, and E. Nogueira, Phys. Rev. E **63**, 020901(R) (2001).
[7] T. P. Creamer, R. Srinivasan, and G. D. Rose, Biochemistry **34**, 16245 (1995).

[8] M. A. Moret, J. G. V. Miranda, E. Nogueira, M. C. Santana, and G. F. Zebende, Phys. Rev. E **71**, 012901 (2005).
[9] M. A. Moret, M. C. Santana, E. Nogueira, and G. F. Zebende, Physica A **361**, 250 (2006).
[10] G. F. Zebende, M. V. S. da Silva, A. C. P. Rosa, A. S. Alves, de J. C. O. Jesus, and M. A. Moret, Physica A **3**,42, 322 (2004).
[11] J. Yon, Mol. Cell **53**, 557 (1997).
[12] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[13] A. Sali, E. I. Shakhnovich, and M. Karplus, Nature (London) **369**, 48 (1994).
[14] K. A. Dill, Curr. Opin. Struct. Biol. **3**, 99 (1993).
[15] K. A. Dill, Biochemistry **29**, 7133 (1990).
[16] M. Karplus, and D. L. Weaver, Nature (London) **260**, 404 (1976).
[17] M. Karplus, and D. L. Weaver, Protein Sci. **3**, 650 (1994).
[18] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[19] M. A. Moret, P. G. Pascutti, P. M. Bisch, and K. C. Mundim, J. Comput. Chem. **19**, 647 (1998).
[20] M. A. Moret, P. M. Bisch, K. C. Mundim, and P. G. Pascutti, Biophys. J. **82**, 1123 (2002).
[21] M. A. Moret, P. M. Bisch, E. Nogueira, and P. G. Pascutti, Physica A **353**, 353 (2005).
[22] F. P. Agostine, D. D. O. Soares-Pinto, M. A. Moret, C. Osthoff, and P. G. Pascutti, J. Comput. Chem. **27**, 1142 (2006).
[23] M. A. Moret, P. M. Bisch, and F. M. C. Vieira, Phys. Rev. E **57**, R2535 (1998).
[24] M. A. Moret, P. G. Pascutti, P. M. Bisch, M. S. P. Mundim, and K. C. Mundim, Physica A **363**, 260 (2006).
[25] B. Lee, and F. Richards, J. Mol. Biol. **55**, 379 (1971).
[26] F. Richards, J. Mol. Biol. **82**, 1 (1974).
[27] J. Finney, J. Mol. Biol. **96**, 721 (1975).
[28] B. Honig, J. Mol. Biol. **293**, 283 (1999).
[29] D. Voet, and J. Voet, *Biochemistry*, 2nd ed. (Wiley, New York, 1995).
[30] C. Chothia, Nature (London) **254**, 304 (1975).
[31] A. Rashin, M. Iofin, and B. Honig, Biochemistry **25**, 3619 (1986).
[32] S. J. Hubbard, and P. Argos, Protein Sci. **3**, 2194 (1994).
[33] S. Hubbard, K. Gross, and P. Argos, Protein Eng. **7**, 613 (1994).
[34] M. Gerstein, and C. Chothia, Proc. Natl. Acad. Sci. U.S.A. **93**, 10167 (1996).
[35] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam, Proteins **33**, 18 (1998).
[36] J. Liang, H. Edelsbrunner, and C. Woodward, Protein Sci. **7**, 1884 (1998).
[37] A. Shrake, and J. Rupley, J. Mol. Biol. **79**, 351 (1973).
[38] R. Voorintholt, M. Kosters, G. Vegter, G. Vriend, and W. Hol, J. Mol. Graphics **7**, 243 (1989).
[39] J. Pascual-Ahuir, and E. Silla, J. Comput. Chem. **11**, 1047 (1990).
[40] C. Ho, and G. Marshall, J. Comput.-Aided Mol. Des. **4**, 337 (1990).
[41] J. Delaney, J. Mol. Graphics **10**, 174 (1992).
[42] G. Kleywegt, and T. Jones, Acta Crystallogr., Sect. D: Biol. Crystallogr. **50**, 178 (1994).
[43] M. Connolly, J. Appl. Crystallogr. **16**, 548 (1983).
[44] T. Richmond, J. Mol. Biol. **178**, 63 (1984).
[45] K. Gibson, and H. Scheraga, Mol. Phys. **62**, 1247 (1987).
[46] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam, Proteins **33**, 1 (1998).
[47] J. Liang, and K. Dill, Biophys. J. **81**, 751 (2001).
[48] Z. Bagci, R. L. Jernigan, and I. Bahar, J. Chem. Phys. **116**, 2269 (2002).
[49] M. A. F. Gomes, J. Phys. A **20**, L283 (1987).
[50] P. J. Flory, *Statistical Mechanics of Chain Molecules*, (Wiley, New York, 1969).
[51] J. Shimada, and H. Yamakawa, Macromolecules **17**, 689 (1984).
[52] I. N. Berezovsky, A. Y. Grosberg, and E. N. Trifonov, FEBS Lett. **466**, 283 (2000).
[53] E. N. Trifonov, and I. N. Berezovsky, J. Mol. Biol. **307**, 1419 (2001).
[54] E. N. Trifonov, and I. N. Berezovsky, FEBS Lett. **527**, 1 (2002).
[55] E. N. Trifonov, and I. N. Berezovsky, Curr. Opin. Struct. Biol. **13**, 110 (2003).
[56] D. Stauffer, Phys. Rep. **54**, 1 (1979).
[57] D. Stauffer, *Introduction to Percolation Theory* (Taylor and Francis, London, 1985).
[58] T. Vicsek, *Fractal Growth Phenomena*, 2nd ed., (World Scientific, Singapore, 1992).
[59] J. Kyte, and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).
[60] S. H. White, and W. C. Wimley, Annu. Rev. Biophys. Biomol. Struct. **28**, 319 (1999).
[61] T. Hessa, H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White, and G. von Heijne, Nature (London) **433**, 377 (2005).
[62] T. P. Hopp, and K. R. Woods, Proc. Natl. Acad. Sci. U.S.A. **78**, 3824 (1981).
[63] J. Janin, Nature (London) **277**, 491 (1979).
[64] C. Chothia, Nature (London) **248**, 338 (1974).
[65] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, Science **229**, 834 (1985).