Dihedral-angle Gaussian distribution driving protein folding ☆

P.H. Figueirêdo [a], M.A. Moret [b,c], E. Nogueira Jr. [d], S. Coutinho [a,*]

[a] *Departamento de Física, Universidade Federal de Pernambuco, CEP 50670-901, Recife, Pernambuco, Brazil*
[b] *Programa de Modelagem Computacional, Fundação Visconde de Cairu, CEP 40226-900, Salvador, Bahia, Brazil*
[c] *Departamento de Física, Universidade Estadual de Feira de Santana, CEP 44031-460, Feira de Santana, Bahia, Brazil*
[d] *Instituto de Física, Universidade Federal da Bahia, CEP 40210-340, Salvador, Bahia, Brazil*

Abstract

The proposal of this paper is to provide a simple angular random-walk model to build up polypeptide structures, which encompass properties of dihedral angles of folded proteins. From this model, structures will be built with lengths ranging from 125 up to 400 amino acids for the different fractions of secondary structure motifs, in which dihedral angles were randomly chosen according to narrow Gaussian probability distributions. In order to measure the fractal dimension of proteins three different cases were analyzed. The first contained α-helix structures only, the second β-strands structures and the third a mix of α-helices and β-sheets. The behavior of proteins with α-helix motifs are more compact than in other situations. The findings herein indicate that this model describes some structural properties of a protein and suggest that randomness is an essential ingredient but proteins are driven by narrow angular Gaussian probability distributions and not by random-walk processes.
© 2007 Elsevier B.V. All rights reserved.

The manner in which a protein folds from a random coil into a unique native state in a relatively short time is one of the fundamental puzzles of molecular biophysics. It is well accepted that a unique native three-dimensional structure, characteristic of each protein and determined by the sequence of its amino acids' sequence, dictates protein functions. The folding process should involve a very complex molecular recognition phenomenon depending on the interplay of many relatively weak non-bonded interactions. This would lead to a huge number of possible final conformations under conventional molecular optimization methods based on the search for the minima of the energy hyper-surface. This number, which should increase with the number of the chain's degrees of freedom, however, is severely restricted during the real folding process, excluding relevant portions of the energy landscapes as far as an extended or random conformation is chosen as the initial state [1–10]. On the other hand, if the extreme limit, were considered, where a polypeptide chain departs from its denatured state and in relatively very short period of time finds its unique native

* Corresponding author. Tel.: +55 81 21267633; fax: +55 81 32710359.
E-mail addresses: phf@lftc.ufpe.br (P.H. Figueirêdo), moret@cairu.br (M.A. Moret), enogue@ufba.br (E. Nogueira Jr.), sergio@lftc.ufpe.br (S. Coutinho).

Table 1
Seven possible pairs of dihedral angles and the associated conformations occurring in several amino acid sequences [13]

Φ	Ψ	Conformation
−65	−40	A
−89	−1	C
−117	142	B
−69	140	P
78	20	G
103	−176	E
−83	133	O

The α-helix pair is denoted by A while the β-strand pair is denoted by B.

state after searching amongst the astronomical number of possible configurations, the simulating process for proteins with fifty to five hundred amino acids using approaches such as Monte Carlo and molecular dynamics, becomes impracticable, due to the very high computation cost. Such a contradictory dynamical picture is known as the *Levinthal paradox* [11].

To investigate the role of stochasticity on the final native state, an inverse strategy is proposed, based on a simple angular three-dimensional random-walk model to build up protein backbones with different lengths and distinct percentages of secondary structures. In the proposed model, each step has a fixed radial size l_0 but dihedral Φ and Ψ angles of the protein backbone are chosen according to independent Gaussian probability distributions, following the suggestion given in Ref. [12]. The mean value and standard deviation of each is defined according to the allowed regions of the Φ/Ψ plot of the frequency distribution of dihedral angles, the so-called Ramachandran map. Φ and Ψ mean values were used as proposed in the PRELUDE software package [13]. These values were computed from comparative statistics of the backbone secondary structure for several amino acid sequences. Table 1 indicates the seven possible pairs of (Φ, Ψ) dihedral angles and the associated structures of the main chain backbone, as predicted by this method. These specific angles describe the average conformation of a wide range of proteins with known backbone structures. To simulate structures with a definite percentage of secondary structures f, a characteristic number of steps n is fixed and the growth process within these n steps is divided into two stages:

(1) The first $n \times f$ steps are built according to an angular Gaussian probability distribution, whose mean value is one pair of angles as seen in Table 1, which in turn is associated with a given structure.

(2) The next $n \times (1 - f)$ steps are built according to an angular Gaussian probability distribution, whose mean value at each step is randomly chosen from amongst the seven pairs of angles of Table 1.

For the following n steps rules 1 and 2 are repeated upwards to construct a peptide chain with N amino acids. Therefore, in order to obtain an appropriate choice of the f percentage this stochastic procedure assures that the final peptide main chain follows the Ramachandran map.

Within this simple model structures of the protein backbone were constructed considering only the dihedral angles. All other bonded or non-bonded interactions were not explicitly considered as well as excluded volume and steric effects, which are expected to be taken into account by the appropriate choice of the average values of the Gaussian probability distributions. For this reason it was possible to generate an elevated number of samples of possible protein conformations. An exhaustive number of simulations ($\sim 10^4$) were performed considering three basic cases: (a) $f = 0.6$ with α-helix structures; (b) $f = 0.6$ with β-strand structures and (c) the first $n/2$ steps built with $f = 0.6$ of α-helix structures and the next $n/2$ steps with $f = 0.6$ of β-strand structures, consecutively. Therefore, in the α-case (a) 60% of the amino acids corresponds on average to α-helix structures, in the β-case (b) 60% of the amino acids corresponds to β-strands, while in the mixed-case (c) the whole structure has an average of 30% of α-helix and 30% of β-strands. 10^4 chains of the total size varying from $N = 125$–400, with the number of steps $n = 100$ were generated. For each case described above there was a variation of f in the interval [0, 1], step 0.1. There was also a variation of the standard deviation σ of the Gaussian distribution within a wide range of values from 0 to π.

Fig. 1 shows the average radius of gyration ($\langle R_g \rangle$) as a function of the number of amino acids (N) for the three distinct choices of structures. From this plot however, a power-law behavior pattern can be observed indicating that these structures are self-similar. The corresponding scaling exponent, which somehow describes the compactness of the structure, is calculated by the scaling relation: $R_g \sim N^\nu$, in all cases. The characteristic scaling exponents are $\nu = 0.401 \pm 0.002$ for the α-helix case and $\nu = 0.417 \pm 0.002$ for the β-strands case, with these values falling in
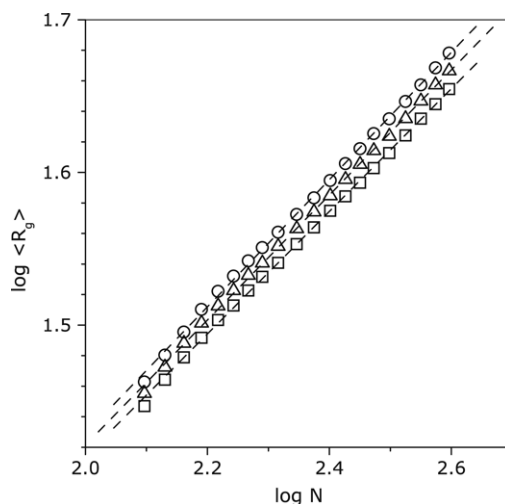
Fig. 1. The *average* radius of gyration as a function of the number of residues obtained from the simulations with $f = 0.60$ for the α-helix structures ($\square$), the mixed structures ($\triangle$) and β-strands structures ($\bigcirc$) with scaling exponent 0.401 ± 0.002, 0.409 ± 0.002 and 0.417 ± 0.002, respectively, obtained from the relation $\langle R_g \rangle \sim N^{\nu}$. Each point results from the average of 10^4 simulations. The dashed lines indicate the linear fitting. The error bars are smaller than the symbol sizes. In all cases $\sigma/\pi = 0.1$ was fixed.
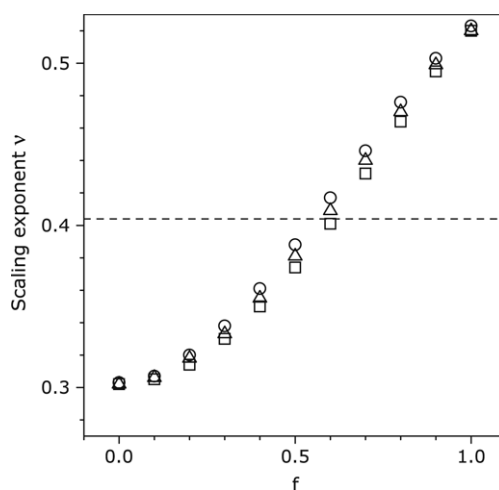

Fig. 2. The dependence of the scaling exponent ν as a function of the percentage f of secondary structures, for α-helices ($\square$), β-strands ($\bigcirc$) and the mixed cases ($\triangle$). The error bars are smaller than the symbol sizes. In all cases $\sigma/\pi = 0.1$ was fixed. The dashed line indicates the experimental value $\nu \simeq 0.405$.

the range of values of real proteins. For the mixed case $\nu = 0.409 \pm 0.002$ were achieved. To further analyze the compactness of structures based on the α-helix, the β-strand and that composed of a mix of α-helices and β-strands the scaling exponent ν was estimated as a function of the percentage of secondary structure f.

In the Fig. 2 it can also observed that for the three cases studied the scaling exponent ν's growth is from $\nu = 0.302 \pm 0.002$ when the percentage of secondary structures is close to zero ($f \simeq 0$), i.e., a complete random structure, up to $\nu_{\max} = 0.520 \pm 0.001$ corresponding to full ordered structures. Slightly different values were observed within the interval $0 < f < 1$ for the structures composed of α-helices, β-strands and the mixed case, the one built with α-helix motifs being more compact than the other cases. It is worth mentioning that the lower limit for the scaling exponent is $1/3$, which would correspond to a fully compact three-dimensional structure commonly observed for globular proteins. However, the plots of Fig. 3 shows that the scaling exponent lies below $1/3$ up to $\sim$0.3, for lower values of the fraction of secondary structure motif ($0 \leq f < 0.3$). Hence, such an interval will correspond to a
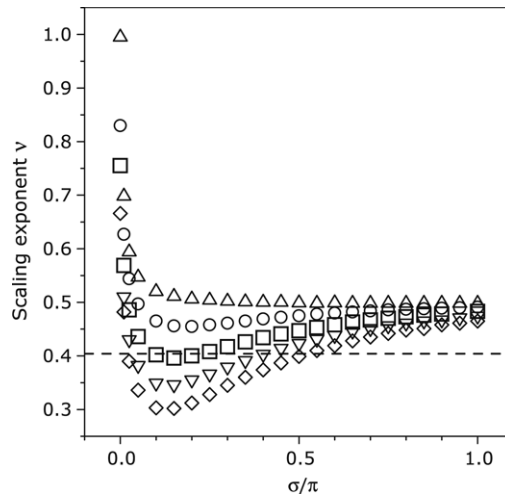
Fig. 3. The dependence of the scaling exponent ν as a function of the variance σ of the Gaussian probability distributions of the dihedral angles (in units of π), for the α-helix structure, considering the value of $f = 0$ ($\Diamond$), $f = 0.40$ ($\triangledown$), $f = 0.60$ ($\square$), $f = 0.80$ ($\bigcirc$) and $f = 1.0$ ($\triangle$). The dashed line indicates the experimental value $\nu \simeq 0.405$.

high fraction $(1 - f)$ of a random structure – built with dihedral angles chosen from Gaussian distributions centered at values randomly chosen at each step from the values given in Table 1. Therefore, it is expected that the excluded volume and steric effects will not play their role and the model fails to reproduce the expected $\nu = 1/3$ limit behavior.

At this point the behavior of the exponent ν against changes in the Gaussian probability distributions variance is explored. The increased dispersion of the (Ψ, Φ) Gaussian distribution corresponds to the increase of randomness in the chain structure, destroying the role of the f percentage of secondary structures. Fig. 3 illustrates the behavior of ν as a function of σ/π of the α-helix case considering the value of $f = 0.0, 0.40, 0.60, 0.80$ and 1.0. For elevated variance values an increase of the ν exponent approaching to 0.5 was observed for all values of f, which is associated with the lack of ordering. On the opposite limit, for vanishing variance values an increase of the ν exponent was observed towards one, which can be easily proved to correspond to the power-law exponent of the linear chain backbone structure for the $f = 1.0$ case. However, between these two limits a minimum value of ν is observed around $\sigma/\pi = 0.15$, independent of the value of f ($\neq 1.0$), corresponding to the maximum compact structures. Therefore, when fitted with the ν value extracted from data, as discussed below, the corresponding optimum value of f is found to be around $f \simeq 0.60$.

1826 different protein chains deposited in the Brookhaven Protein Data Bank (PDB) were also investigated in order to provide a comparison with the obtained simulation results. The number of amino acids was measured as a function of the average radius. Fig. 4 depicts the main characteristics of all systems discussed herein using geometric (dihedral angles) analysis. The figure indicates that several protein chains deposited in the PDB have a self-similar behavior pattern when the average radius ($\langle R \rangle$) is plotted against the number of amino acids (N). In this case the average radius signifies the average distance from the geometric center of all coordinates [10]. It was also noticed that an average value was calculated for radii of chains with the same number of amino acids. This intrinsic characteristic of the protein structures must be responsible for explaining several aspects of these molecules such as the high compactness, which has been discussed in several other different contexts [1,2,14–18]. The ν exponent associated with the average radius obtained is 0.404 ± 0.008, which is in agreement with the recent similar study involving 200 proteins [19]. The volume and mass of proteins with more than fifty amino acids scale with respect to the average radius with exponents $\delta = 2.47 \pm 0.04$ [15] and $\delta = 2.47 \pm 0.03$ [17] respectively. This consequently corresponds to an exponent ($\nu = 1/\delta \approx 0.405$) if we assume that the mass scales to the average radius with exponent one. This exponent would be associated to mixed chains composed of secondary structures, which according to the present simulations vary in an interval of $0.401 \leq \nu \leq 0.417$. Furthermore, the PDB proteins present structures with $\sim$60% of secondary structures [20], which justifies and confirms our initial assumption of $f = 0.6$ for simulated structures shown in Fig. 1. It is worth mentioning that the present model can be generalized for growth chains with distinct percentages of different structures, according to the corresponding protein Ramachandran map.
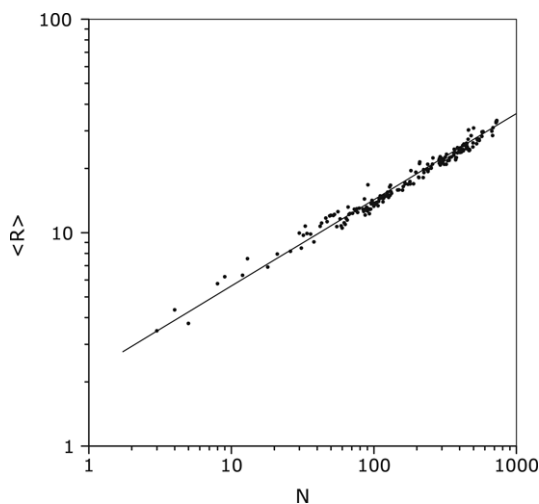
Fig. 4. The behavior of the average radius $\langle R \rangle$ against the number of amino acids N for a set of 1826 proteins chains. The scaling exponent $\nu = 0.40 \pm 0.02$.

Through the approach presented herein the protein folding problem has been investigated assuming that proteins fold in a mixed manner following some directed process while being subject to certain stochastic ingredients. This signifies that the process is neither completely random, as raised by the Levinthal's paradox, nor is it entirely driven by the physical chemistry principles that establish a definite folding pathway. The simple model presented herein focuses on the stochastic aspects of the formation of the secondary structures which is believed to be the earliest relevant precursor event in the folding, as confirmed by other recent experimental evidence [21]. According to rules 1 and 2 of the present model it can be assumed that the formation of a secondary structure (α-helix or β-strands) occurs during a consecutive fraction f of steps governed by a Gaussian probability distribution, whose parameters (mean and standard deviation) are extracted from data associated with the Ramachandran map. These parameters, which characterize a given structure, reflect the physical and chemical processes underlying protein stability. Therefore this fraction f of the chain somehow mimics the interplay between energy stability and entropy. Thus to the extend that the structure reaches a certain size it looses stability and folds randomly, changing the mean value of the Gaussian probability distribution at each step, but still using the possible values extracted from data.

What emerges from this stochastic process is that the narrow Gaussian probability distribution of helical or stranded arrays do provide an insight into the protein folding process, which goes beyond the possibilities of molecular structure or molecular dynamics analysis. Actually, this narrow Gaussian probability distribution supplies a peptide backbone chain with self-similar properties that matches with the one estimated from experimental data (see Fig. 4). Furthermore, Fig. 3 illustrated that if a probability distribution with wide values of the variance (large values of σ/π) is considered to approach a uniform probability distribution, the resulting final structure was less compact than that obtained with the Gaussian distribution process. One further interesting result obtained by the present model is that backbone chains with α-helix motifs are more compact than the β-strands and the mixed case, a result confirmed by the current literature [1,22,23]. The fractal dimension ($\delta = 1/\nu \approx 2.49$) obtained from Figs. 1 and 4 are comparable with that obtained by the volume analysis against radius [15] or by mass-size exponent analysis ($\delta \approx 2.47$) [17–19].

The results of this study indicate that simulated structures are more compact when secondary portions (α-helices and/or β-sheets) are present, than those built with other sets of dihedral angles, as shown in Table 1. The method was systematically compared to other widely used methods of protein folding analysis [1,2,4,6,7]. Several of these methods do not result in a fully consistent assignment of self-similarity of protein structures. The recent work of Huang [24] where a sophisticated conditioned self-avoid-walk model was proposed taking into account the hydrophobic effect and the hydrogen bonding focusing on the physical chemistry mechanisms underlying the protein folding process should be mentioned. Independent of the details of the underlying physical chemistry mechanisms, building protein backbones with the method proposed in the present work suggests that these structures are driven by narrow Gaussian distributions. Thus it is the general conclusion of this work that protein folds like an angular Gaussian-walk and not as a random-walk problem.

References

[1] K.A. Dill, Biochemistry 29 (1990) 7133.
[2] P.G. Wolynes, J.N. Onuchic, D. Thirumalai, Science 267 (1995) 1619.
[3] J.M. Yon, Cell. Mol. Lif. Sci. 53 (1997) 557.
[4] Y. Zhou, M. Karplus, J. Mol. Biol. 293 (1999) 917.
[5] Y. Zhou, M. Karplus, Nature 401 (1999) 400.
[6] M.A. Moret, P.G. Pascutti, P.M. Bisch, K.C. Mundim, J. Comput. Chem. 19 (1998) 647.
[7] D.J. Wales, H.A. Scheraga, Science 285 (1999) 1368.
[8] M.A. Moret, P.G. Pascutti, K.C. Mundim, P.M. Bisch, E. Nogueira Jr., Phys. Rev. E 63 (2001) 020901(R).
[9] M.A. Moret, P.M. Bisch, K.C. Mundim, P.G. Pascutti, Biophys. J. 82 (2002) 1123.
[10] M.A. Moret, P.M. Bisch, E. Nogueira Jr., P.G. Pascutti, Physica A 353 (2005) 353.
[11] C.J. Levinthal, Chem. Phys. 65 (1968) 44.
[12] R.H.A. David Shaw, J.A. Tuszyński, Phys. Rev. E 67 (2003) 031102.
[13] M.J. Rooman, J.-P. Kocher, S.J. Wodak, J. Mol. Biol. 221 (1991) 961.
[14] F. Richards, J. Mol. Biol. 82 (1974) 1.
[15] J. Liang, K. Dill, Biophys. J. 81 (2001) 751.
[16] Z. Bagci, R.L. Jernigan, I. Bahar, J. Chem. Phys. 116 (2002) 2269.
[17] M.A. Moret, J.G.V. Miranda, E. Nogueira Jr., M.C. Santana, G.F. Zebende, Phys. Rev. E 71 (2005) 012901.
[18] M.A. Moret, M.C. Santana, E. Nogueira Jr., G.F. Zebende, Physica A 361 (2006) 250.
[19] M.B. Enright, D.M. Leitner, Phys. Rev. E 71 (2005) 011912.
[20] D. Voet, J. Voet, Biochemistry, 2nd ed., John Wiley & Sons, Inc., New York, 1995.
[21] See B. Honig, J. Mol. Biol. 293 (1999) 283 and references therein.
[22] K.A. Dill, Curr. Opin. Struct. Biol. 3 (1993) 99.
[23] K.A. Dill, H. Chan, Nature: Struct. Biol. 4 (1997) 10.
[24] K. Huang, Biophys. Rev. Lett. 2 (2007) 139.