Nonextensive statistical mechanics applied to protein folding problem: kinetics aspects

J. P. Dal Molin,* Marco Antonio Alves da Silva, I. R. da Silva, and A. Caliri
Universidade de São Paulo, FCFRP, Departamento de Física e Química,
Av. do Café S/N, Monte Alegre, 14040.903, Ribeirão Preto, SP, Brasil

(Received on 23 December, 2008)

A reduced (stereo-chemical) model is employed to study kinetic aspects of globular protein folding process, by Monte Carlo simulation. Nonextensive statistical approach is used: transition probability p_{ij} between configurations $i \to j$ is given by $p_{ij} = [1+(1-q)\Delta G_{ij}/k_BT]^{1/(1-q)}$, where q is the nonextensive (Tsallis) parameter. The system model consists of a chain of 27 beads immerse in its solvent; the beads represent the sequence of amino acids along the chain by means of a 10-letter stereo-chemical alphabet; a *syntax* (rule) to design the amino acid sequence for any given $3D$ structure is embedded in the model. The study focuses mainly kinetic aspects of the folding problem related with the protein folding time, represented in this work by the concept of *first passage time* (FPT). Many distinct proteins, whose native structures are represented here by compact self avoiding (CSA) configurations, were employed in our analysis, although our results are presented exclusively for one representative protein, for which a rich statistics was achieved. Our results reveal that there is a specific combinations of value for the nonextensive parameter q and temperature T, which gives the smallest estimated folding characteristic time $\langle t \rangle$. Additionally, for $q = 1.1$, $\langle t \rangle$ stays almost invariable in the range $0.9 \leq T \leq 1.3$, slightly oscillating about its average value $\overline{\langle t \rangle} = 27 \pm\sigma$, where $\sigma = 2$ is the standard deviation. This behavior is explained by comparing the distribution of the folding times for the Boltzmann statistics ($q \to 1$), with respect to the nonextensive statistics for $q = 1.1$, which shows that the effect of the nonextensive parameter q is to *cut off* the larger folding times present in the original ($q \to 1$) distribution. The distribution of natural logarithm of the folding times for Boltzmann statistics is a triple peaked Gaussian, while, for $q = 1.1$ (Tsallis), it is a double peaked Gaussian, suggesting that a log-normal process with two characteristic times replaced the original process with three characteristic times. Finally we comment on the physical meaning of the present results, as well its significance in the near future works.

Keywords: Protein folding problem, Stereo-chemical model, Monte Carlo simulation, q-Gaussian, Protein folding time

1. INTRODUCTION

Globular proteins are peptide chains identified by their amino acid sequence. Under physiological condition, each protein naturally folds over itself in a compact and precise $3D$ structure (its native structure) [1]; remarkably, such proteins are soluble in polar solvents [2, 3]. The number of different proteins found in the nature is very large [4]; only in the human body about 10^5 different types of proteins are found; actually, they correspond to half of the dry weight of the human body. On the other hand, the set of structural patterns, named as protein folds [1], from which all known protein structures can be ensembled, is much smaller; there are little more the one thousand registered folds. A globular protein works properly only when it is meet in its native structure, which is reached through a process named protein folding [1]; the folding problem is a fundamental process in molecular biology without yet a complete explanation [6–9].

Here the protein system, comprehended by the chain and its solvent, is characterized by its heterogeneity (chain plus solvent) and complex energetic interactions (intra-chain and solvent-chain interactions) under the effects of its nanoscopic size. Therefore, the use of simplified (or reduced) models is imperative in order to handle the folding process in a general form, through fundamental physics [10]. An important characteristic of such minimalist approach is that a significant amount of data can be generated, so that statistical analysis becomes viable, for instance, on kinetic processes, facilitating the use of experimental results. Therefore, a minimalist model, namely the stereo-chemical model, is employed to study kinetic aspects of the protein folding problem [11–13]. Due the number of distinct aspects of the protein folding problem the stereo-chemical model can be applied, and it can be considered as a *reduced parallel world* regarding the globular protein system.

The model is treated by Monte Carlo (MC) simulation with the nonextensive statistical approach; the transition probability p_{ij} between configurations $i \to j$ is given by $p_{ij} = [1+(1-q)\Delta G_{ij}]^{1/(1-q)}$ [14, 15]. Our main objective in this work is to analyze the behavior of the folding characteristic time as a function of the temperature T and the parameter q; the folding time for each protein and for each run, is taken as the first passage time. A representative native structure was utilized to illustrate our results, but many distinct structures were also considered. Our work is organized as follows: a brief description of the stereo-chemical model and the motivations to use the q-exponential function are given in the next section; in the third section the results are presented and, finally, the last topic is dedicated to comments and perspectives for future works.

2. THE STEREO-CHEMICAL MODEL

The physical system of interest is a protein molecule in solution, which is represented here as one single chain with 27 beads immersed in its solvent (water). The beads, representing the residues (amino acids) along the chain, occupy consecutive and exclusive sites of a three-dimensional infin-

*Electronic address: jpdm@fcfrp.usp.br

[1] Folds are resultant forms - for example, α-helix, β-sheets, turns and others - of complex interaction among the elements presents in the folding process [5].

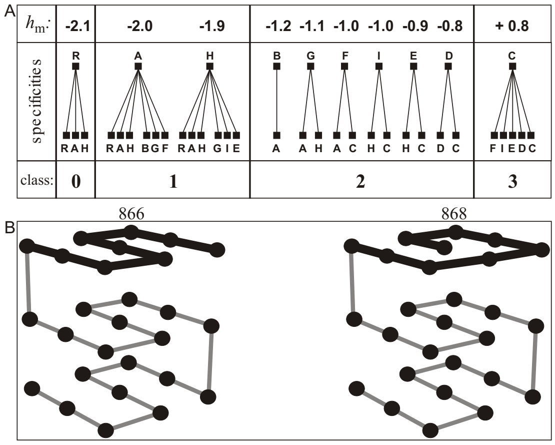


FIG. 1: Steric specificities and hydrophobic level (A). Connected letters indicate that these residue types are allowed to contact as lattice first neighbors. On B, two maximally compacted (CSA) configurations representing distinct native structures.

ity cubic lattice, while the remainder sites, named as *solvent sites*, are filled out by *effective solvent molecules*, which interact explicitly with the chain. Interactions that count for the present model are intended to represent the inter-residue sterical specificities and the hydrophobic effect. Therefore, additionally to the trivial excluded volume restriction for solvent and chain elements, the intra-chain interactions present specific pairwise steric constraints, and the chain interacts with the solvent according to the hydrophobic level of their residues, as summarized in Fig.1*A* [11, 12].

As native structures, we consider configurations from a *catalogue* with 51.704 distinct compact self avoiding (*CSA*) structures, which form cubes 3x3x3 for a chain with 27 residues. For each protein in the *catalogue* there is a corresponding specific sequence of residues, taken from a repertory of ten distinct types, that is: different hydrophobic levels and steric specificities define a 10-letter alphabet, see Fig.1*A* [12]. A syntax, or rule to design the protein, given each particular CSA structure (native), is also provided for the stereochemical model [13].

The system is described in the context of microcanonical ensemble. Its configurations (chain and solvent) all have the same weight $p = 1/\Gamma_o$, but note that for the same chain configuration there is a huge number of distinct configurations for the solvent. So, if the chain is considered *frozen* in a particular configuration, say configuration a, the system entropy is reduced by an amount δS_a; therefore a weight $p_a = e^{\delta S_a/k}$ is associated with that configuration a, (k: Boltzmann constant). Next, the transition probability $W(a \rightarrow b)$ of configuration a to configuration b is given by detailed balance: $W(a \rightarrow b) = e^{-\Delta S_{ba}/k}$, where $\Delta S_{ba} = \delta S_b - \delta S_a$. Finally, as there are no variations on the system internal energy, we write $W(a \rightarrow b) = e^{\Delta G_{ba}/kT}$ [16].

Formally, the interactions between parts of the system are: $(i)-$ intra-chain interactions: excluded volume and specific restrictions between first neighbors, Fig.1*A*; $(ii)-$ solvent-solvent interactions: excluded volume; $(iii)-$ chain-solvent: excluded volume and hydrophobic effect. The latter, in spite of being due to the direct interaction between each residue and the solvent, in practice, for lattice models in which the solvent density does not fluctuate, the chain-solvent interactions can be exactly represented by the addictive hydrophobic potential, $g_{m,n} = (h_m + h_n)$, between first neighbors, Fig.1*A*. The hydrophobic level h_m represents the free energy change to transfer the residue type-m from a polar environment (solvent) into an apolar medium (protein interior). Thus, for a given chain configuration a, the free energy of the system can be considered as $G = G_o + \sum g_{m,n}(a)$, where G_o is a constant. Details about target configurations and 10-letter alphabet are found in [11–13].

The finite nature of the system (nanometric structures) and the complex interactions between its parts (which can generate a rough free energy landscape), suggest that thermal fluctuations are preponderant factors of the system's behavior. Therefore, the Boltzmann weight $e^{-\Delta G_{ba}/kT}$ [17] is exchanged by the q-exponential, namely $[1 + (1-q)\Delta G_{ba}/kT]^{1/1-q}$ [14] (which embeds the Boltzmann factor for $q \rightarrow 1$), in order to examine the effect of the nonextensive parameter q on the kinetic characteristics of the folding process. Specifically, we perform many independent Monte Carlo experiments, which generate large sets of folding times for different temperatures T, and for different values of nonextensive parameter q. More general motivations for using Tsallis statistics include the fact that: $(i)-$ the protein folding process is a stochastic process out of thermodynamical equilibrium [18, 19] and $(ii)-$ there is evidence that, to thermostatistically approach living systems, the nonextensive statistical mechanics can be a convenient choice [15].

3. RESULTS

In order to get the precise meaning of the results presented here, we first describe, with some detail, the computational experiments carried out in this work. First, in the real experimental approach, one considers a collection of N open denatured protein molecules of the same type, in diluted solution; the proteins are by identified by labels $1, 2, ..., i, ..., N$. So, once folding conditions have been established, protein 1 finds the native structure in the time t_1, protein 2 at t_2, and so on; that is, t_i is the *folding time* of the protein i. Then, after some time $\langle t \rangle$, half of the proteins, that is $N/2$ proteins, reach the native state, and the amount $\langle t \rangle$ is identified as the folding *characteristic time* of that type of protein [2]. Accordingly, in our MC simulations for each particular protein, the concept of first passage time (FPT) is adopted as the folding time, that is: the time spent, in units of MC steps for a particular MC simulation, to find the native structure by the first time. Therefore, N independent simulations of a given protein produce a set $\{t_i\}$ of N folding times, with which one determines the folding characteristic time $\langle t \rangle$ for that protein, such as in an real experiment. If the folding process is multi-exponential time dependent, $\langle t \rangle$ is found by solving numerically the transcendental equation $N/2 = \sum_i m_i e^{-\langle t \rangle / t_i}$, for given $\{t_i\}$ and $\{m_i\}$, constrained with $\sum_i m_i = N$. At last, the folding rate for that

[2] Note that, one could exchange the factor $1/2$ by $1/e$.

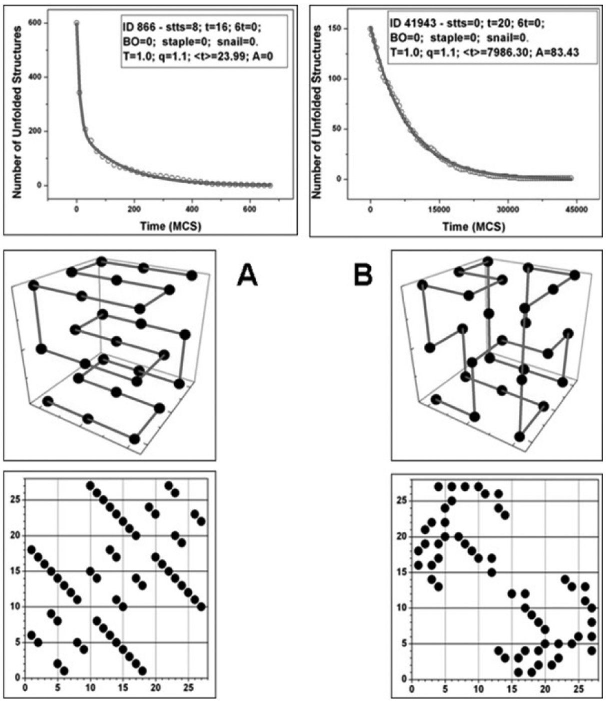


FIG. 2: Folding characteristic time $\langle t\rangle$ for two distinct native structure (A: CSA 866; B: CSA 41943). There are strong correlation between folding times (up side) and structural patterns of the native structure, as shown in the center and down side of figure; CSA structures and Contact Maps, respectively.

protein is then defined as $\tau = 1/\langle t\rangle$.

The structure identified by the label 866 was chosen to represent our results regarding the behavior of folding characteristic time $\langle t\rangle$ for two reasons. First, similarly to real proteins, it is rich in secondary structure-like (indeed it presents the largest possible content of secondary structure among all CSA configurations of *catalogue*), as shown in Fig.2*A*, through its contact map. Second, as a consequence of its structure peculiarity, it presents the smallest $\langle t\rangle$ which implies in an optimized CPU time. The results are summarized in Fig.3, which show the characteristic time $\langle t\rangle$ as a function of temperature T (arbitrary units) in the model-significative interval $0.9 \leq T \leq 1.5$ and of the nonextensive parameter q, also in the interval $0.9 \leq q \leq 1.5$. For each pair (T,q), six hundred distinct MC simulations were performed. The smallest $\langle t\rangle$ are obtained for specific combinations of T and q, producing a *valley* in which the folding characteristic time is always smaller than 50 (MC units of time) Fig.3*A*. However, as shown in Fig.3*B*, only for $q = 1.1$, one gets $\langle t\rangle < 50$ for practically every temperature in the interval $0.9 \leq T \leq 1.4$.

The behavior of $\langle t\rangle$ when q changes with T fixed, seems to be similar to that in which q is fixed at $q \rightarrow 1$ (Boltzmann) and T changes. However only for $q = 1.1$ the folding characteristic time stays almost invariable in the range $0.9 \leq T \leq 1.3$; see Fig.3; indeed $\langle t\rangle$ slightly oscillates about its average value $\overline{\langle t\rangle} = 27 \pm\sigma$, where $\sigma = 2$ is the standard deviation; see TABLE I.

That behavior is explained by Fig.4: when the distribution

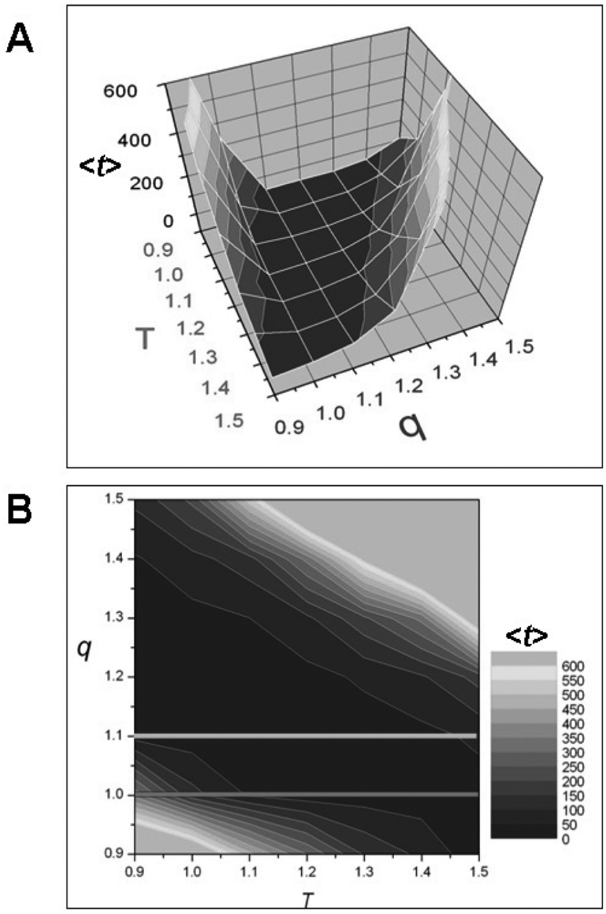
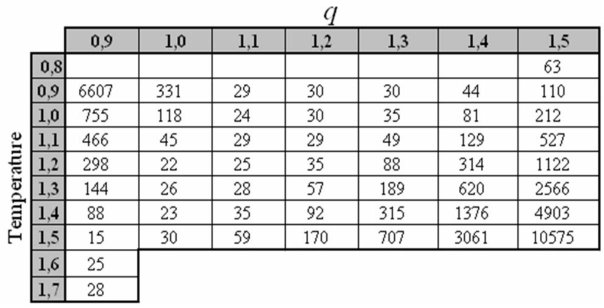


FIG. 3: A linear relation between T and q roughly describes a *valley* in which $\langle t\rangle$ is always smaller than 50; part *A*. For $q = 1.1$, $\langle t\rangle$ obtains its smaller values, which are smaller than 50 for practically the entire range of T, as shown in part *B*; see also TABLE I. The two lines in part B emphasize the results for Boltzmann ($q \rightarrow 1$) and Tsallis ($q = 1.1$) statistics. Note also that the distance between the lines that confine the values $\langle t\rangle < 50$, part A, reduces systematically as T increases and q decreases.



	q						
Temperature	0,9	1,0	1,1	1,2	1,3	1,4	1,5
0,8							63
0,9	6607	331	29	30	30	44	110
1,0	755	118	24	30	35	81	212
1,1	466	45	29	29	49	129	527
1,2	298	22	25	35	88	314	1122
1,3	144	26	28	57	189	620	2566
1,4	88	23	35	92	315	1376	4903
1,5	15	30	59	170	707	3061	10575
1,6	25						
1,7	28						

TABLE I: Folding characteristic time as a function of temperature T and nonextensive parameter q. Errors (standard deviation of the averages) are estimated as being about 6%.

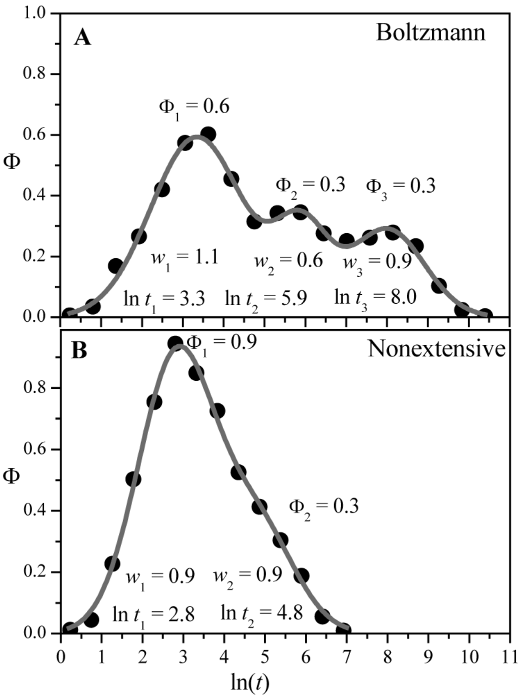


FIG. 4: Distribution of the natural logarithm of the folding time (structure 866) for $T = 1.0$ and $q \to 1$ (Boltzmann statistics, figure *A*) and $q = 1.1$ (Tsallis statistics, figure *B*). Tsallis weight seems to allow stronger fluctuations, efficiently removing the system from eventual stereo-energetic traps.

of natural logarithm of the folding times for the Boltzmann statistics ($q \to 1$, Fig.4*A*) is compared with the nonextensive statistics for $q = 1.1$. (Fig.4*B*), one can see that the effect on varying q is to *cut off* the larger folding times from the distribution. Indeed, the original distribution (Fig.4*A*, Boltzmann) can be seen as a triple peaked Gaussian, suggesting a log-normal process with three characteristic times. In the same way, for $q = 1.1$ (Fig.4*B*, Tsallis), it also suggests a log-normal process, but now with two characteristic times. That is, using the Tsallis statistics, the third peak that appears in the Boltzmann distribution is *cut off.* However, the two other processes are essentially preserved with their respective characteristic frequencies accordingly increased. Differences on the mean values of the folding time -actually $\ln(t)$ as shown in Fig.4, and on the width of the distributions, with respect to the two statistics, possibly are due to the fact that the number of simulations is insufficient to determine those results accurately; for both cases ($q \to 1$, and $q = 1.1$), about 10^4 independent MC experiments were carried out.

4. COMMENTS AND PERSPECTIVES

The folding transition for globular two-state proteins occurs in a relatively small temperature range ΔT; out of this domain the folding characteristic time increases significantly. Experimentally, this fact is observed by a rapid change of the equilibrium constant $K_{\mathrm{eq}} = [N]/[U]$, where $[N]$ and $[U]$ are the folded (native) and unfolded protein concentration, respectively, when the system temperature is out of the range ΔT. However, in the interval ΔT, the folding reaction is robust, with K_{eq} practically constant [20] as it is reproduced when the nonextensive statistics is used, specifically for $q = 1.1$. Indeed, the folding time distribution is broad because the free energy landscape is rough, trapping the chain in local free energy minima or some steric entanglement. Tsallis weight seems to be efficient to quickly remove the system from such traps, and so folding routes with longer folding times are cut off from the distribution; Fig.4. A computational consequence of this results is that simulations become much more faster, opening the possibility to trace a strategy to extend the study of the effect of Tsallis statistics to others kinetic matters of the folding problem, such as on the correlation between the folding rates and global structural parameters of native structures (contact order, for instance) [21], and on the stability of the native state.

Acknowledgements

We thank CNPq and FAPESP (grant No. $2007/04220-4$) for funding.

[1] C.B. Anfinsen, Science **181**, 223 (1973).
[2] D. Chandler, Nature **417**, 491 (2002).
[3] L.F.O. Rocha, M.E. Tarragó Pinto, and A. Caliri, Braz. J. Phys. **34**(1), 90 (2004).
[4] C. Chothia, Nature **357**, 543 (1992).
[5] A.M. Lesk, G.D. Rose, Proc. Natl. Acad. Sci. USA **78**, 4304 (1981).
[6] E.E. Lattman, G.D. Rose, Proc. Natl. Acad. Sci. USA**90**, 439 (1993).
[7] L.S. Itzhaki, P.G. Wolynes, Curr. Opin. Struct. Biol. **18**, 1-3 (2008).
[8] A.C. Clark, A.B.B. **469**, 1-3 (2008).
[9] K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).
[10] T. Head-Gordon, S. Brown, Curr. Opin. Struct. Biol. **13**, 160 (2003).
[11] R.A. da Silva, M.A.A. da Silva, A. Caliri, J. Chem. Phys. **114**(9), 4235 (2001).
[12] M.E.P. Tarragó, L.F.O. Rocha, R.A. da Silva, A. Caliri, PRE **67**, 031901-1 (2003).
[13] I.R. Silva, L.M. dos Reis, A. Caliri, J. Chem. Phys. **123**(15), 154906-1, (2005).
[14] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
[15] C. Tsallis, Phys. Life. Rev. **3**, 1 (2006).
[16] L.F.O. Rocha, M.A.A. da Silva, A. Caliri, Phys. Lett. A **220**, 178 (1996).
[17] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, J. Chem. Phys. **21**(6), (1953).
[18] N. Go, J. Stat. Phys. **30**(2), 413 (1983).
[19] E.M. Popov, IUBMB Life **47**(3), 443 (1999).
[20] M.L. Scalley, D. Baker, Proc. Natl. Acad. Sci. USA **94**, 10636 (1997).
[21] K.W. Plaxco, K.T. Simons, D. Baker, J. Mol. Biol. **277**, 985 (1998).