Influence of thermal instability on protein sizes

Fabio Cecconi and Angelo Vulpiani
Dipartimento di Fisica and INFM-SMC
Università di Roma La Sapienza
P.le A. Moro 2, 00185 Roma, Italy
email: cecconif@roma1.infn.it

Raffaella Burioni and Davide Cassi
Dipartimento di Fisica, INFM and INFN
Università di Parma
Parco Area delle Scienze 7A, 43100 Parma, Italy
email:burioni@galileo.fis.unipr.it

ABSTRACT
We present an analysis of the role of global topology on the structural stability of folded proteins in thermal equilibrium with a heat bath. For a large class of single domain proteins, we computed the harmonic spectrum within the Gaussian Network Model (GNM) and determined the spectral dimension, a parameter controlling the low frequency behaviour of the density of modes. We find a surprisingly strong correlation between the spectral dimension and the number of amino acids of the protein. Considering that the larger the spectral dimension, the more topologically compact is the folded state, our results indicate that native folds correspond to the less compact structures compatible with thermodynamic stability.

KEY WORDS
Protein native states, Topology, Thermal Fluctuations, Gaussian Network Model, Spectral Dimension.

1 INTRODUCTION

During the folding, a protein evolves from an almost linear structure into final conformation (native state) whose geometrical shape is crucial to the function of the protein itself. Only after the native state is reached without mistakes, a protein becomes active and starts performing its specific biological activity. Besides the biological functions, native state geometry is a key-factor also for the overall folding process [1, 2, 3, 4]. For this reason a great amount of literature has been devoted to the study of the theoretical aspects characterizing the complex networks of interactions between amino acids in folded proteins [5]. For instance, graph theory have been successfully applied to identify flexible and rigid regions of folded conformations [6].

However, the problem of the geometrical arrangement of proteins in their native conformation cannot be regarded as purely static issue. Indeed, a massive accumulation of experimental data collected from X-ray, NMR and neutron spectroscopy, has revealed that native states are rather dynamical structures where amino acids constantly undergo rearrangements around their equilibrium positions. This dynamics, crucially involved in protein functions [7, 8], is usually examined through normal modes analysis (NMA) [9]. However, the study of collective motions of large size proteins is generally computationally expensive for realistic all-atoms Normal Mode simulations [10] and simplified or approximate approaches are usually welcome. Tirion [11] first proposed the possibility to replace, in protein normal mode computations, complicated empirical potentials by harmonic pairwise interactions depending on a single parameter. The approach stems from the observation that low-frequency dynamics is generally insensitive to the finer details of atomic interactions. The success of simple harmonic models in the study of slow vibrational dynamics of large biological macromolecules has been proved in several works [12, 13, 14] and they are considered as a viable alternative to heavy and time-consuming all-atoms NMA. We apply one of these harmonic models, the Gaussian Network Model (GNM) [15] to study the thermal instability of proteins and investigate how it is affected by the global topology of native state. Vibrational thermal instability is a well-known issue in solid state physics. Since the first

classical work by Peierls [16], it has been recognized that the equilibrium with a thermal bath can dramatically influence the allowed topological arrangement of large geometrical structures. The consequence of Peierls instability has concerned up to now low-dimensional crystals: for one and two-dimensional lattices the mean square displacement of a single atom at finite temperature diverges in the thermodynamic limit, i.e. with increasing number of atoms. When such a displacement exceeds the lattice spacing, the topological arrangement of the lattice becomes unstable and the crystal melts. For structures formed by a finite number of units Peierls' instability sets a maximal size, which is negligible for one-dimensional lattices and typically mesoscopic for two-dimensional ones.

However, thermal instability applies not only to crystals but also to structurally inhomogeneous systems, such as, glasses, fractals, polymers and proteins where the problem is much more complex. In particular, we extend the Peierls' arguments to the thermal stability of proteins and we will predict the existence of a critical stability size depending on a global topological parameter (the spectral dimension) and compare our predictions with experimental data [17].

2 Peierls' Instability for Proteins

In a recent paper [18] we have generalized the Peierls' result, showing that a thermodynamic instability appears also in inhomogeneous structures and it is characterized by the spectral dimension $\bar{d}$. The parameter $\bar{d}$ is defined according to the scaling behaviour of the density of harmonic oscillations at low frequencies. More precisely, denoting by $g(\omega)$ the density of modes with frequency ω, then

$$g(\omega) \sim \omega^{\bar{d}-1} \tag{1}$$

for $\omega \to 0$. The spectral dimension is the natural extension of the Euclidean dimension d to non ordered structures, indeed, it coincides with d in the case of lattices, but in general, can assume non-integer values between 1 and 3. The spectral dimension provides a useful measure of the effective connectedness of geometrical structures at large scales, because large values of $\bar{d}$ correspond to high topological connectedness.

The GNM generally describes proteins as elastic networks, whose nodes are the positions of the alpha-carbons (C_α) in the native structure and the interactions between nodes are assimilated to harmonic springs. The information required to implement this model is the knowledge of the native structure only. Two parameters are introduced, the spring constant γ and the interaction cutoff R_0, which, however turn out to be related whenever the model is applied to fit experimental data. The GNM for a chain of N amino acids (residues), is defined by the quadratic Hamiltonian

$$H = \sum_i^N \frac{\mathbf{p}_i^2}{2m} + \frac{\gamma}{2} \sum_{ij} \Delta_{ij} (\delta \mathbf{r}_i - \delta \mathbf{r}_j)^2 \tag{2}$$

the first term is the kinetic energy of the system, $\mathbf{R}_i$ and $\delta \mathbf{r}_i$ indicating the equilibrium position and the displacement with respect to $\mathbf{R}_i$ of the i-th C_α atoms. The model is eventually defined by the contact matrix Δ with entries: $\Delta_{ij} = 1$ if the distance $|\mathbf{R}_i - \mathbf{R}_j|$ between two C_α's, in the native conformation, is below the cutoff R_0, while is 0 otherwise.

The model is exactly solved by a direct diagonalization the matrix $\Gamma_{ij} = -\Delta_{ij} + \delta_{ij} \sum_{l \neq i} \Delta_{il}$ (known as Kirchhoff or valency-adjacency matrix). The protein vibrational spectrum in the GNM approximation is the set of frequencies $\omega_i = \sqrt{\gamma \lambda_i}$ where λ_i is the i-th eigenvalue of Γ. Debye-Waller factors or B-factors, measuring the thermal fluctuations of residues around their native positions, are easily computed in the GNM via the formula

$$B_i(T) = \frac{8\pi^2 k_B T}{\gamma} [\Gamma^{-1}]_{ii}$$

where $[\Gamma^{-1}]_{ii}$ indicates the diagonals entries of the inverse Γ matrix. The B_i's parameters are crucial because are used to fit the experimental B-factors from X-ray crystallography. This fitting connects GNM-theory to experiments and allows to make an optimal adjustments of both the cutoff R_0 and the spring stiffness γ.

The oscillating protein has to be considered in thermal equilibrium with its environment which we describe as a thermal bath at temperature T. The thermal average of the observables are calculated via the usual Boltzmann weight with the quadratic Hamiltonian (2) and amounts to performing Gaussian integration over the degrees of

freedom. In particular the mean square displacement of C_α from their native positions is defined by $\langle r^2 \rangle = 1/N \sum_i \langle |\delta \mathbf{r}_i|^2 \rangle$ reads

$$\langle r^2 \rangle = \frac{1}{N} \sum_i \frac{k_B T}{\gamma \omega_i^2} = \frac{k_B T}{\gamma} \int_{\omega_{min}}^{\omega_{max}} d\omega \frac{\rho(\omega)}{\omega^2} \qquad (3)$$

where, k_B is the Boltzmann constant. Through the density of modes, the sum over the eigenvalues is replaced by an integral in ω between the smallest ω_{min} and the largest ω_{min} frequency. The value of the integral (3), in the large N limit, is dominated by the behavior of the spectral density at low eigenvalues ($\omega \sim \omega_{min}$). Since, for large N, ω_{min} can be shown to scale as $\omega_{min} \sim N^{-2/\bar{d}}$ we obtain, the following asymptotic behaviour: $\langle r^2 \rangle \sim$ const, if $\bar{d} > 2$, while for $\bar{d} < 2$

$$\langle r^2 \rangle \sim \frac{k_B T}{\gamma} N^{2/\bar{d}-1} \, . \qquad (4)$$

therefore $\langle r^2 \rangle$ grows with the size N. For long proteins, the spectral dimension characterizes the scaling of thermal fluctuations with the number of residues. The relevance of $\bar{d}$ in connection with the anomalous density of vibrational modes in proteins has also been considered in refs.[19, 20]. The relationship (4) establishes a rather strong constraint between the spectral dimension and the maximum size N_{max} of a protein can afford. Since, the stability is supposed to fail when the fluctuation $\langle r^2 \rangle^{1/2}$ becomes of the same order of magnitude of the mean distance between non consecutive amino acids (about 7 Å), one can assume that

$$\frac{2}{\bar{d}} = 1 + \frac{b}{\ln(N_{max})}. \qquad (5)$$

The constant b depends on the mean amino acid spacing, on the spring elastic constant γ and temperature T. However, this dependence is expected to be very weak (i.e. only logarithmic) and this allows for a comparison of different proteins without the computation of the specific parameters.

This poses an intriguing question concerning proteins. Indeed, to exploit their biological function proteins must keep a specified geometrical and topological arrangement and cannot afford any, even partial, geometrical large scale fluctuations as it happens, instead, to swollen polymers in a good solvent. This makes the thermodynamical stability problems of crucial importance and suggests a possible correlation between the spectral dimension and the length of protein chains. It should be stressed, however, that equation (5), being based uniquely on thermodynamics stability, can be actually regarded as an upper bound prediction only.

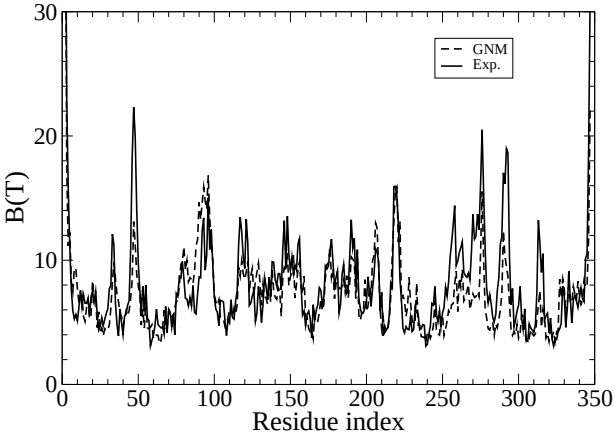

Figure 1: Comparison between experimental B-factors and mean square fluctuations of C_α by GNM, for the structures for the protein 3PTE. Heavy solid line refers to crystallographic data, while dashed line refers to GNM-theory.

We performed GNM harmonic analysis over the dataset of protein native structures, with length ranging from 100 to 3600 residues (Tab.1), downloaded from the Brookheaven Protein Data Bank . The purpose is showing that the correlation predicted by expression (4) exists between the spectral dimension of protein native structures and their length. The value of the interaction cutoff for generating the contact matrix Δ has been set to $R_0 = 7\mathring{A}$, as customary in such kind of studies. However since this choice affects the overall GNM performance, we tested through the correlation coefficient ρ between experimental and theoretical B-factors [17]. Our dataset contains only those protein structures with a coefficient ρ greater than 4.7 (see last column of Table 2) this should, in principle, ensure that GNM correctly reproduces C_α fluctuations for each selected protein. The typical agreement between B-factors from GNM and crystallography is shown in figure 1.

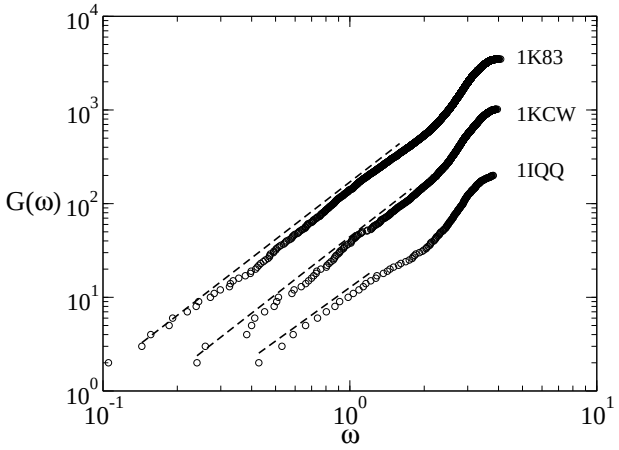

Figure 2: Log-log plot of GNM-harmonic spectrum referred to three proteins with different sizes, 1IQQ (N=200), 1KCW (N=1017) and 1K83 (N=3494). On vertical axis, we report the cumulated distribution $G(\omega)$ of vibrational modes. Low frequencies regions clearly exhibit a power-law behaviour, and dashed lines indicates the best-fits of the power-law whose exponent is the spectral dimension.

The optimal value of the spring constant γ for each protein was obtained through a least-square fitting to the experimental B-factors expressed by formula

$$\frac{k_B T}{\gamma} = \frac{1}{8\pi^2} \frac{\sum_i B_i X_i}{\sum_i B_i^2} \qquad (6)$$

The values of $k_B T/\gamma$, besides being an essential ingredient for the real application of GNM method, are also an indication of the protein global flexibility and allows for a direct comparison among all the considered structure.

The spectral dimension $\bar{d}$ was estimated via a power-law fitting of the low frequency behaviour of the cumulated density of modes $G(\omega)$, namely the integral of $g(\omega)$, which at small arguments, is expected to scale as $G(\omega) \sim \omega^{\bar{d}}$. The GNM vibrational spectrum, for three proteins with sizes, small, medium and large, is plotted in figure 2; the low-frequency regions clearly exhibit the power law behavior whose exponent is the spectral dimension $\bar{d}$. Our statistical analysis for the whole dataset of proteins is summarized in Table 2, where we report, the chain length, the spectral dimension, the estimate for $k_B T/\gamma$, and finally the correlation coefficient.

PDB code	Length	$\bar{d}$	$K_B T/\gamma$	Correl.	PDB code	Length	$\bar{d}$	$K_B T/\gamma$	Correl.
9RNT	104	1.62	1.657	0.474	1DY4	441	1.88	0.785	0.614
1BVC	153	1.56	0.392	0.698	1BU8	446	1.95	0.859	0.632
1G12	167	1.89	0.793	0.584	1AC5	483	1.87	1.091	0.709
1AMM	174	1.71	0.003	0.720	1LAM	484	1.97	0.488	0.583
4GCR	185	1.73	0.001	0.711	1CPU	495	1.92	0.620	0.729
1KNB	186	1.88	1.104	0.699	3COX	500	1.92	0.491	0.670
1CUS	197	1.86	0.914	0.731	1A65	504	2.09	1.042	0.606
1IQQ	200	1.84	0.480	0.626	1SOM	528	2.00	1.585	0.653
2AYH	214	1.86	0.539	0.773	1E3Q	532	1.97	1.577	0.623
1AE5	223	1.93	0.952	0.531	1CRL	534	2.00	0.969	0.652
1LST	239	1.77	0.982	0.647	1AKN	547	1.87	1.737	0.667
1A06	279	1.78	2.184	0.623	1CF3	581	2.01	1.154	0.639
1NAR	289	1.81	0.602	0.696	1EX1	602	2.01	1.193	0.598
1A48	298	1.72	0.664	0.549	1A14	612	2.10	0.865	0.524
1A3H	300	1.90	0.719	0.553	1MZ5	622	2.02	0.750	0.705
1SBP	309	1.74	0.641	0.757	1CB8	674	1.92	1.164	0.630
1A5Z	312	1.74	2.111	0.574	1HMU	674	1.92	0.907	0.684
1A1S	313	1.89	1.068	0.600	1A47	683	2.02	0.646	0.529
1ADS	315	1.79	0.500	0.687	1CDG	686	1.98	1.074	0.593
1A40	321	1.90	0.524	0.546	1DMT	696	1.96	1.204	0.536
1A54	321	1.86	0.601	0.516	1A4G	780	1.98	0.591	0.567
1A0I	332	1.71	1.109	0.826	1HTY	1014	2.07	0.646	0.766
3PTE	347	1.79	0.366	0.840	1KCW	1017	2.05	2.130	0.638
1A26	351	1.82	1.369	0.635	APP1	1021	1.93	0.805	0.576
1BVW	360	1.87	0.652	0.639	1KEK	2462	2.07	1.263	0.730
8JDW	360	1.94	1.293	0.607	1B0P	2462	2.08	0.319	0.810
7ODC	387	1.92	0.859	0.620	1K83	3494	2.01	2.030	0.659
1OYC	399	1.93	1.056	0.697	1I3Q	3542	1.97	2.435	0.758
1A39	401	1.97	1.113	0.656	1I50	3558	1.98	2.236	0.701
16PK	415	1.82	0.630	0.590	—	—	—	—	—

Table 1: Dataset of processed proteins. PDB identifier, length, spectral dimension $\bar{d}$, $k_B T/\gamma$, correlation between theoretical and experimental B-factors.

Figure 3 verifies the prediction drawn form the thermodynamical stability argument and shows the final result of our analysis. We plot the quantity $2/d$ versus $1/\ln(N)$ as suggested by relation (5): indeed, if Eq. (5) holds, we should obtain a straight line crossing the y-axis at 1 for zero abscissa. The error bars cover both the uncertainty due to the fitting procedure and the possible inadequacy (measured by ρ) of the GNM in reproducing experimental B-factors. Our data are well fitted by a straight line,

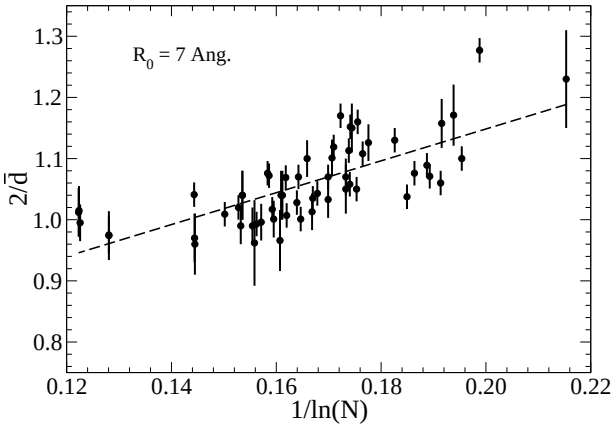

Figure 3: Linear plot showing the dependence of the spectral dimension on protein size. The dashed line, indicating the behaviour 7, is a best fit with a correlation coefficient 0.73.

but, with an offset with respect to the equation (5)

$$\frac{2}{\bar{d}} = a + \frac{b}{\ln(N)}. \qquad (7)$$

The best-fit values of the parameters are $a = 0.63$, $b = 2.61$, with a correlation coefficient 0.73.

3 Conclusions

We applied the Gaussian Network Model (GNM) to investigate the influence of native state topology on thermodynamical stability for a set of folded proteins with sizes ranging from 100 to 3600. The employ of GNM is justified from the fact that such a model correctly accounts for the topological features of the native protein conformations. Our results show that the spectral dimension $\bar{d}$, which is sensitive to the large scale topology of a geometrical structure, is a parameter controlling the low-energy fluctuations of a given protein structure. Then an instability criterion for proteins under thermal fluctuations can be derived through topological considerations only. This criterion, analogous of Peierls' argument developed for ordered crystalline structures, easily predicts a non-trivial logarithmic dependence of the spectral dimension on the length of proteins. We verified that, in the GNM approximation, such a logarithmic dependence is really observed, within statistical and systematic errors, for the whole set of selected proteins. The result expressed by Eq. (7) is in agreement with the upper bound (Eq. (5)), supporting the relevance of topological thermal instability in constraining the protein geometry. Notice that not only the upper bound is satisfied, but the experimental points lie on a straight line parallel to the upper bound one Eq. (5). This suggests a more fundamental role of topological stability: in some sense, a native state tends to arrange in a geometrical structure with the largest $\bar{d}$ compatible to its length and stability constraints. In other words, for any fixed length, it tends to the most swollen state which remains stable with respect to thermal fluctuations.

References

[1] K.W. Plaxco, K.T. Simons and D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, 277(4), 1998, 985-994.

[2] D.E. Makarov, C.A. Keller, K.W. Plaxco, H. Metiu, How the folding rate constant of simple, single-domain proteins depends on the number of native contacts, *Proc. Natl. Acad. Sci. USA*, 99(6), 2002, 3535-3539.

[3] D.S. Riddle, V.P. Grantcharova, J.V. Santiago, E. Alm, I. Ruczinski and D. Baker, Experiment and theory highlight role of native state topology in SH3 folding, *Nature Struct. Biol.*, 6(11), 1999, 1016-1024.

[4] F. Cecconi, C. Micheletti, P. Carloni. and A. Maritan, Molecular dynamics studies on HIV-1 protease drug resistance and folding pathways, *Proteins: Struc. Func. Gen.*, 43(4), 2001, 365-372.

[5] K. Park, M. Vendruscolo, and E. Domany, Toward an energy function for the contact map representation of proteins. *Proteins: Struct. Func. Gen.*, 40(2), 2000, 237-248.

[6] D.J. Jacobs, A.J. Rader, L.A. Kunh and M.F. Thorpe, Protein flexibility prediction using graph theory. *Proteins: Struct. Funct. Genet.*, 44(2), 2001, 150-165.

[7] H. Frauenfelder, G.A. Petsko and D. Tsernoglou, Temperature dependent x-ray diffraction as a probe as of protein structural dynamics, *Nature*, 280(5723), 1979, 558-563.

[8] H. Frauenfelder and B. McMahon, Dynamics and functions of proteins: the search of general concepts, *Proc. Natl. Acad. Sci. USA*, 95(9), 1998, 4795-4797.

[9] M. Karplus and J. McCammon, Dynamics of Proteins: Elements and Functions, *Ann. Rev. Biochem.*, 52, 1983, 263-300.

[10] M. Levitt, C. Sanders and P.S. Stern, Protein normal-mode dynamics; trypsin inhibitor, crambin, ribonuclease, and lysozyme, *J. Mol. Biol.*, 181(3), 1985, 423-447.

[11] M.M. Tirion, Low-amplitude elastic motions in proteins from a single-parameter atomic analysis, *Phys. Rev. Lett.*, 77(9), 1996, 1905-1908.

[12] I. Bahar, A.R. Atilgan, M.C. Demirel and B. Erman, Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability, *Phys. Rev. Lett.*, 80(12), 1998, 2733-2736.

[13] T. Haliloglu, I. Bahar and B. Erman, Gaussian dynamics of folded proteins, *Phys. Rev. Lett.*, 79(16), 1997, 3090-3093.

[14] C. Micheletti, F. Cecconi, A. Flammini, A. Maritan, Crucial stages of protein folding through a solvable model: Predicting target sites for enzyme-inhibiting drugs, *Protein Sci.*, 11(8), 2002, 1878-1887.

[15] I. Bahar, A.R. Atilgan, B. Erman, Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, *Fold. Des.*, 2(3), 1997, 173-181.

[16] R.E. Peierls, Bemerkung über Umwandlungstemperaturen, *Helv. Phys. Acta*, 7, 1934, S81-S83.

[17] R. Burioni, D. Cassi, F. Cecconi and A. Vulpiani, Topological thermal instability and length of proteins, *Proteins: Proteins: Struct. Funct. Bioinf.*, 55(3), 2004, 529-535.

[18] R. Burioni, D. Cassi, M.P. Fontana and A. Vulpiani, Vibrational thermodynamic instability of recursive networks, *Europhysics Lett.*, 58(6), 2002, 806-810.

[19] D. Ben-Avraham, Vibrational normal-mode spectrum of globular proteins, *Phys. Rev. B*, 47(21), 1993, 14559.

[20] R. Elber and M. Karplus, Low frequency modes in proteins: use of effective-medium approximation to interpret fractal dimension observed in electron-spin relaxation measurements, *Phys. Rev. Lett.*, 56(4), 1986, 394-397.