Freezing Transition of Compact Polyampholytes

Vijay S. Pande,* Alexander Yu. Grosberg,† Chris Joerg,‡ Mehran Kardar, and Toyoichi Tanaka
Department of Physics and Center for Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
(Received 15 July 1996)

Polyampholytes (PAs) are heteropolymers with long range Coulomb interactions. Unlike polymers with short range forces, PA energy levels have nonvanishing correlations and are thus very different from the random energy model (REM). Nevertheless, if charges in the PA globule are screened as in a regular plasma, PAs freeze in REM fashion. Our results shed light on the potential role of Coulomb interactions in folding and evolution of *proteins*, which are weakly charged PAs, in particular, making connection with the finding that sequences of charged amino acids in proteins are not random. [S0031-9007(96)01429-9]

PACS numbers: 61.41.+e, 64.70.Dv

The freezing transition of heteropolymers, in which the number of thermodynamically relevant states goes from an exponentially large value [$\mathcal{O}(e^N)$] in the random globule state, to only a few [$\mathcal{O}(1)$] conformations in the frozen state, has attracted a great deal of interest. In addition to providing an interesting problem in the statistical mechanics of disordered materials [1], this system is potentially relevant to the biologically important question of protein folding. Most previous investigations have focused on heteropolymers with short range interactions. Recently, however, there has been renewed theoretical [2–4] and experimental [5,6] interest in polyampholytes (PAs), which are heteropolymers with charged monomers of both signs. It has been shown that, due to screening effects, PAs collapse to compact globules if their net charge is below a critical value [7]. There is also some evidence from exact enumeration studies of short chains [8] that dense globules of neutral PAs may have a freezing transition. However, it is unclear how long range (LR) interactions affect freezing, or whether the formalism developed for globular polymers with short range (SR) interactions remains applicable to the LR case.

The freezing transition of SR heteropolymers is most commonly described by the random energy model (REM) [9], although it is not always applicable even in this case [10]. As the principal underlying assumption of REM is the statistical independence of energies of states (polymer conformations) over disorder (sequence of charges along the chain), we first examine correlation of the energies and then discuss the resulting freezing transition. Our starting point is the Hamiltonian

$$\mathcal{H} = \sum_{I \neq J}^{N} B s_I s_J f(\mathbf{r}_I - \mathbf{r}_J), \qquad (1)$$

where B is a constant, I labels monomers along the chain, and $s(I) \in \pm 1$ is the charge of monomer I. The range of interactions is indicated through $f(r)$, such that $f(r) = \Delta(r)$ for SR interactions, and $f(r) = 1/r^{d-2}$ for Coulomb forces in d dimensional space. Finally, we only consider the case of maximally compact polymers, assuming that maximal density is maintained independently of Coulomb interactions, i.e., by an external box, poor solvent, or internal attractions, such that $R \sim N^{1/d}$.

The simplest characteristics of statistical dependence of energies is the pair correlation between two arbitrary conformations α and β, given by

$$\langle E_\alpha E_\beta \rangle_c \equiv \langle E_\alpha E_\beta \rangle - \langle E_\alpha \rangle \langle E_\beta \rangle = B^2 \mathcal{Q}_{\alpha\beta}, \qquad (2)$$

with $\mathcal{Q}_{\alpha\beta} \equiv \sum_{I \neq J} f(\mathbf{r}_I^\alpha - \mathbf{r}_J^\alpha) f(\mathbf{r}_I^\beta - \mathbf{r}_J^\beta)$. In the familiar case of SR interactions, $\mathcal{Q}_{\alpha\beta}^{\mathrm{SR}} = \sum_{I \neq J} \Delta(\mathbf{r}_I^\alpha - \mathbf{r}_J^\alpha)\Delta(\mathbf{r}_I^\beta - \mathbf{r}_J^\beta)$ is just the number of bonds in common between configurations α and β. Numerical simulations [10] indicate that in many cases the probability distribution for $\mathcal{Q}_{\alpha\beta}^{\mathrm{SR}}$, i.e., $P_{\mathrm{SR}}(\mathcal{Q}) \equiv \sum_{\alpha\beta} \delta(\mathcal{Q} - \mathcal{Q}_{\alpha\beta}^{\mathrm{SR}})$, is sharply peaked at small $\mathcal{Q}$. This happens because one can easily "hide" monomers by moving them only a small distance and decreasing their contribution to $\mathcal{Q}^{\mathrm{SR}}$. Large statistical dependence is thus achieved only for conformations that are closely related. The validity of REM rests on the statistical rarity of such closely related conformations. REM is valid when configurations that are statistically dependent can be ignored in a large N limit.

By contrast, with long range interactions, the relevant parameter for judging statistical dependence is $\mathcal{Q}_{\alpha\beta}^{\mathrm{LR}} = \sum_{I \neq J} [|\mathbf{r}_I^\alpha - \mathbf{r}_J^\alpha| \cdot |\mathbf{r}_I^\beta - \mathbf{r}_J^\beta|]^{-(d-2)}$. While the geometric interpretation of $\mathcal{Q}_{\alpha\beta}^{\mathrm{LR}}$ is not as clear as $\mathcal{Q}_{\alpha\beta}^{\mathrm{SR}}$, it measures the similarity in contributions from monomer pairs (I, J) in conformations α and β to the overall energy. Unlike the SR case, polymeric bonds always keep monomers within the scale of LR interactions. Thus, for two conformations chosen at random, the overlap $\mathcal{Q}_{\mathrm{rand}}^{\mathrm{LR}}$ may not be negligible (even if $\mathcal{Q}_{\mathrm{rand}}^{\mathrm{SR}}$ is). The following scaling argument provides an estimate of the width of the probability distribution $P_{\mathrm{LR}}(\mathcal{Q}) \equiv \sum_{\alpha\beta} \delta(\mathcal{Q} - \mathcal{Q}_{\alpha\beta}^{\mathrm{LR}})$.

First, consider the maximum overlap which occurs (for both LR and SR) when *all* elements are correlated (i.e., $\mathcal{Q}_{\max} = \mathcal{Q}_{\alpha\alpha}$ is the correlation of a configuration with itself). To compute this, we note that for each of

the N monomers there is a contribution from $\mathcal{O}(r^{d-1})$ monomers at a distance r (for *compact* states in d dimensions), resulting in $\mathcal{Q}_{\max} \sim N \int dr\, r^{d-1} f(r)^2$. For SR interactions, this integral is dominated by contributions at a microscopic length scale (set by the interaction range) and we get $\mathcal{Q}_{\max}^{\mathrm{SR}} \sim N$. For LR interactions, while contributions from monomers far away are smaller, there are more of them. For Coulomb interactions in $d \leq 4$, the integral is dominated by the longest distance, and for a polymer of size R, we get $\mathcal{Q}_{\max}^{\mathrm{LR}} \sim NR^d/R^{2(d-2)} \sim NR^{4-d}$.

We can use similar arguments for the overlap between two conformations chosen at random ($\mathcal{Q}_{\mathrm{rand}}^{\mathrm{LR}}$). In fact, for the LR problem, $\mathcal{Q}_{\max}^{\mathrm{LR}}$ and $\mathcal{Q}_{\mathrm{rand}}^{\mathrm{LR}}$ scale identically, as both cases involve $\mathcal{O}(N^2)$ pairs of monomers, each giving a contribution $\mathcal{O}(1/R^{2(d-2)})$, for a total of $\mathcal{Q}_{\max}^{\mathrm{LR}} \sim \mathcal{Q}_{\mathrm{rand}}^{\mathrm{LR}} \sim N^2 R^{2(2-d)}$. Moreover, as the main contribution to $\mathcal{Q}_{\mathrm{rand}}^{\mathrm{LR}}$ comes from far away sites, this residual overlap is only weakly conformation dependent. The existence of a residual overlap changes the problem fundamentally from the SR case: REM is not valid as there is always a statistical dependence in $d < 4$ [11].

Computer simulations support the above arguments. To examine a large range in N, we generated random conformations on a lattice by first choosing a radius R, and then enumerating random paths [12] on the set of lattice sites which are within R. R was varied from 3 to 10 lattice sites, and the following results represent averages over 20 conformations for each R value. Figure 1 shows that the scaling exponents γ defined by $\mathcal{Q} \sim N^{\gamma}$ appear to be the same within error for random pairs of conformations, as well as the overlap of any conformation with itself. Furthermore, the fits agree well with the predictions $\gamma_{\max}^{\mathrm{LR}} = \gamma_{\mathrm{rand}}^{\mathrm{LR}} = 4/3$. By contrast, with SR interactions $\gamma_{\max}^{\mathrm{SR}} = 1$, while $\gamma_{\mathrm{rand}}^{\mathrm{SR}} \approx 0.75$ is distinctly smaller. We also calculated SR and LR overlaps $\mathcal{Q}^{\mathrm{SR}}$ and $\mathcal{Q}^{\mathrm{LR}}$ for 1000 pairs of 64-mer conformations ($d = 3$, cubic lattice). The resulting histograms, with overlaps normalized by the maximal value, are shown in Fig. 2. SR overlaps are peaked at small values, whereas the LR overlaps are peaked closer to unity. Furthermore, the sharpness of the distribution suggests that $\mathcal{Q}^{\mathrm{LR}}$ is approximately independent of the chosen pairs of conformations.

Having demonstrated the residual overlap between energies of conformations with LR interactions, and hence the breakdown of REM, we go on to better characterize the density of states. This will take us a step closer to understanding the freezing of PAs. To describe the density of states, we use the following three characteristics: the annealed energy variance σ_{ann} (the width of the density of states for annealed disorder), the average quenched energy variance σ_{quen} (the width of the density of states for quenched disorder), and the quenched energy correlation function g (the statistical dependence between states). These quantities are given by the formulas

$$\begin{aligned} \sigma_{\mathrm{ann}}^2 &\equiv \langle \overline{(E^2)} \rangle_c = \langle \overline{(E^2)} \rangle - \langle \overline{E} \rangle^2, \\ \sigma_{\mathrm{quen}}^2 &\equiv \langle \overline{(E^2)} \rangle_c = \langle \overline{(E^2)} \rangle - \langle (\overline{E})^2 \rangle, \\ g &\equiv \langle (\overline{E})^2 \rangle_c = \langle (\overline{E})^2 \rangle - \langle \overline{E} \rangle^2, \end{aligned} \tag{3}$$

where $\overline{\cdots}$ and $\langle \cdots \rangle$ denote averaging over conformations and sequences, respectively. Note that these quantities are related by a mathematical identity $\sigma_{\mathrm{ann}}^2 = \sigma_{\mathrm{quen}}^2 + g$.

In the annealed case, the energy variance is $\sigma_{\mathrm{ann}}^2 = B^2 \mathcal{Q}_{\max}$, since, in this case, all possible states can be accessed and thus the width of the energy spectrum must be maximal. This result is also easily extracted from Eq. (2) by averaging over conformations with $\alpha = \beta$. Averaging the same equation over *all pairs* of states α and β, we can find g: For $\mathcal{M}$ conformations, there are $\mathcal{M}$ pairs $\alpha = \beta$ which completely overlap $\mathcal{Q}_{\alpha\beta} = \mathcal{Q}_{\max}$, but this

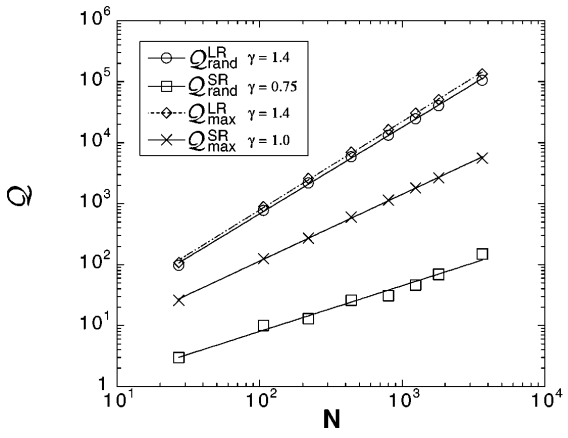

FIG. 1. Scaling of $\mathcal{Q}_{\mathrm{rand}}$ and $\mathcal{Q}_{\max}$ with N for LR and SR interactions ($d = 3$). Power law scaling of the form $\mathcal{Q} \sim N^{\gamma}$ indicates that $\mathcal{Q}_{\mathrm{rand}}^{\mathrm{LR}}/\mathcal{Q}_{\max}^{\mathrm{LR}}$ does not vanish in the thermodynamic limit, whereas $\mathcal{Q}_{\mathrm{rand}}^{\mathrm{SR}}/\mathcal{Q}_{\max}^{\mathrm{SR}}$ does.

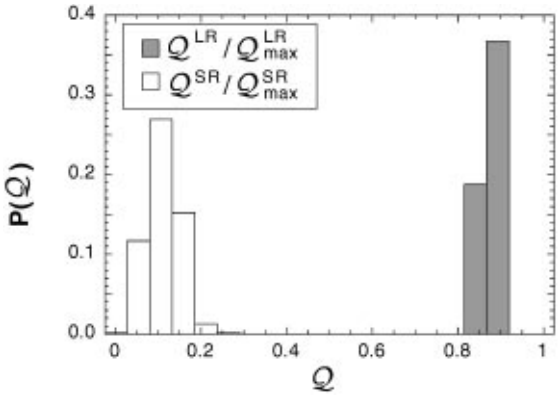

FIG. 2. Probability distributions $P(\mathcal{Q}^{\mathrm{LR}})$ and $P(\mathcal{Q}^{\mathrm{SR}})$, obtained from 64-mers on a cubic lattice. Because of finite size effects, there is some residual overlap in the SR case (here peaked at 0.1). However, we expect that the SR residual overlap vanishes in the thermodynamic limit, while the LR overlap does not.

is overshadowed by the remaining $\mathcal{M}(\mathcal{M}-1)$ pairs with overlap $\mathcal{Q}_{\alpha\beta}=\mathcal{Q}_{\rm rand}$, resulting in $g\approx B^2\mathcal{Q}_{\rm rand}$. In addition to measuring the statistical dependence between states, $g=\langle(\overline{E})^2\rangle_c$ also describes how the mean of the energy spectrum for a given sequence varies between sequences. Finally the width of the energy spectrum for a typical sequence is $\sigma^2_{\rm quen}\equiv\sigma^2_{\rm ann}-g=B^2(\mathcal{Q}_{\rm max}-\mathcal{Q}_{\rm rand})$. This makes sense physically as correlation (anticorrelation) in the energies should narrow (broaden) the width of the energy spectra. Also, we see that when there is no correlation ($g=0$), $\sigma_{\rm ann}=\sigma_{\rm quen}$, as in the REM.

The following picture emerges from the above results. As $\mathcal{Q}^{\rm SR}_{\rm rand}=0$, we have $g=0$ for the SR case above the freezing temperature, and the mean of the energy spectrum does not vary significantly between sequences. Also, the width of the spectrum for a given sequence is large (the maximum possible value, as in the annealed case). The variation of the means of the energy spectra between sequences g is much smaller than the typical width of each spectrum $\sigma^2_{\rm quen}$; thus disorder is not important for SR interactions above freezing. Of course, below the freezing temperature, self-averaging breaks down, and disorder is relevant. By contrast, for LR interactions, $\mathcal{Q}^{\rm LR}_{\rm rand}$ does not vanish and is significant. We thus expect the widths of the energy spectra to be small and the means to vary widely from sequence to sequence.

The results of a computational test of the above scenario, obtained from the exact enumeration of all globular states of 36-mers on a cubic lattice ($d=3$) are presented in Fig. 3. We see that for SR interactions, the means of the spectra are indeed well defined and their width (gray region) is large. For LR interactions, the means are poorly defined, with a variance between sequences which is greater than the widths of individual spectra (error bars).

Is the insight gained above sufficient to analyze the freezing transition in PAs? In general, freezing is

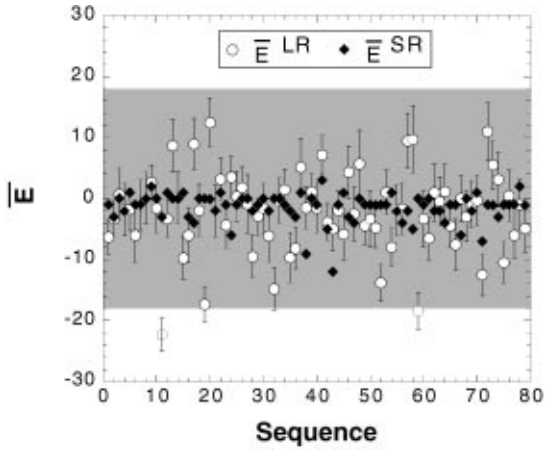

FIG. 3. Mean and width of the energy spectra for 80 sequences of 36-mers, determined by full enumeration over all maximally compact conformations (see text for details).

governed by the low energy tail of the density of states $\rho(E)=\mathcal{M}P(E)$, where $\mathcal{M}$ is the total number of conformations, and $P(E)$ is the single level energy distribution. In the standard REM entropy crisis scenario, the system freezes in a microstate, much like a snapshot, at a temperature T_f at which $\rho_T\sim 1$, where $\rho_T=\rho(E_T)$ is the density of states at the equilibrium energy E_T at the temperature T.

The density of states in the high temperature regime is governed by $\sigma_{\rm ann}$, as can be seen by a high temperature expansion: The partition function $Z={\rm tr}[\exp(-\beta\mathcal{H})]$ is first expanded in powers of $\beta=1/T$, resulting in (after averaging over sequences) $-\beta F=\langle\ln Z\rangle=\ln\mathcal{M}-\beta\langle\overline{E}\rangle+\beta^2\langle\overline{(E^2)}\rangle_c/2+\cdots$. From this expression [and using Eq. (3)], the entropy is calculated as $S(T)=\ln\mathcal{M}-\beta^2\sigma^2_{\rm quen}/2+\cdots$, where (as demonstrated earlier) for Coulomb interactions in $d=3$, $\sigma^2_{\rm quen}\sim e^2N^2/R$, yielding

$$\rho_T\sim\mathcal{M}\exp[-\tfrac{1}{2}(e^2N/TR)^2]. \tag{4}$$

From the structure of the series [3], we expect the high temperature expansion to break down for temperatures $T<T_D\equiv e^2N/R$. This temperature can also be obtained by regarding the polymer globule as a (nonpolymeric) plasma of the same N charges confined within the volume R^3. As the Debye screening length for this plasma is of the order $r_D\sim(TR^3/Ne^2)^{1/2}$, there are two regimes: For $T<T_D$, the plasma is fully screened as $r_D<R$. However, for $T>T_D$, $r_D>R$ and the charges are not screened. The latter regime is meaningless for a regular plasma, but describes the high temperature behavior of the polymer globule. It is not clear that, with the constraints of polymeric bonds, the scaling for a PA should be the same as that for a screened plasma at low temperatures. However, assuming that this is the case, the entropy can be estimated by noting that the plasma is composed of roughly $\mathcal{N}\sim R^3/r_D^3\sim(Ne^2/RT)^{3/2}$ independent Debye volumes. Assuming that the entropy is proportional to $\mathcal{N}$, we finally conclude

$$\rho_T\sim\mathcal{M}\exp[-c(e^2N/TR)^{3/2}], \tag{5}$$

where c is a numerical constant. Note that Eq. (5) indicates a very sharp decrease of the density of states in the low energy tail, proportional to $\exp[-c'(E-\overline{E})^3]$, which reflects the fine tuning of configurations necessary for screening.

Typically the number of conformations of a polymer scales as $\mathcal{M}\sim e^{\omega N}$, with ω of the order of unity. In the limit where the polymer is kept maximally compact by an external box, poor solvent, or internal attractions, such that $R\sim aN^{1/3}$, where a is a monomeric length scale, ω is approximately the entropy of Hamiltonian walks. Freezing, which is signaled by $\rho\sim 1$, can take place in the unscreened regime only for short chains with $N<1/\omega$. (The "apparent" freezing temperature for unscreened polymers grows as $N^{1/6}$.) In this case, a

further decrease of temperature will not lead to screening, of course. For longer chains, we predict freezing at an N-independent temperature of $T_f \sim e^2/(a\omega^{2/3})$ in the screened regime. In this sense, the compact PA freezes in a phase transition that is similar to REM. We stress that this happens despite the unusual scaling of the width of the density of states, $\sigma \sim N^{2/3}$. The distinction between the two behaviors is important for understanding the results of lattice simulations, as it appears that 36-mers are in the short chain regime.

We expect that the nature of the frozen state also depends on T_f/T_D. For freezing in the screened regime ($T_f < T_D$), the system looks much like that of the SR case, i.e., like a disordered version of a salt crystal. For freezing in the unscreened regime ($T_f > T_D$), we expect a smaller degree of antiferromagnetic ordering; consistent with the idea that freezing at a higher temperature leads to a state which is less energetically optimized.

An important class of PAs are *proteins.* In the light of our findings in this work, we make here some concluding remarks about protein folding and evolution. Of the 20 natural amino acids, three are positively charged (Lys, Arg, His), two are negatively charged (Asp, Glu), and the rest are neutral. Nevertheless, it is often assumed that LR interactions are not essential to proteins, as the screening length in biological solvents is often quite small. It is less clear that screening is also effective in compact globular configurations with little or no solvent in their interiors. Furthermore, secondary structural elements such as α-helices effectively reduce the conformational flexibility of proteins. Indeed, the conformation space of small proteins (i.e., 70–90 amino acids) perhaps corresponds to that of lattice 27-mers [13], and small proteins are likely to be in the short chain regime with respect to LR interactions. Thus, while the total charge on a given protein may be small, in solvents with few counterions, this may be sufficient to lead to a REM-violating correlated energy landscape, making the results obtained here relevant. Moreover, for the typical separation of charges in a globular protein (roughly 20 Å), and given a dielectric constant of order 5–10, and $\omega \approx 2$, the characteristic freezing temperature T_f is of the order of (biologically relevant) room temperatures.

We have discussed how the mean of the density of states can vary greatly from sequence to sequence. It appears that a large contribution to this mean comes from the interaction between monomers that are not far apart along the sequence. For example, while next nearest neighbors along the chain can somewhat vary their spatial distance from each other, this will still not break their great contribution to the mean energy. This is why the conformational average energy depends strongly on the correlations between charges quenched along the sequence. For Coulomb interactions, chains with anticorrelated sequences have low mean energies. This is intriguing, considering the recent finding that protein sequences are indeed anticorrelated with respect to their charge [14]. This indicates that perhaps protein evolution was not just dictated solely by the degree of hydrophobicity of monomers (which depends on the degree of charge, not the sign), but by Coulomb effects as well.

The work was supported by NSF (DMR 94-00334). A. Y. G. acknowledges the support of Kao Fellowship. Computations were performed on Project SCOUT (ARPA Contract No. MDA972-92-J-1032). We thank R. Du for a critical reading of the manuscript.

*Present address: Physics Department, University of California, Berkeley, CA 94720-7300.

†On leave from Institute of Chemical Physics, Russian Academy of Sciences, Moscow 117977, Russia.

‡Present address: Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

[1] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987); E. Shakhnovich and A. Gutin, Biophys. Chem. **34**, 187 (1989).

[2] S. F. Edwards, P. R. King, and P. Pincus, Ferroelectrics **30**, 3 (1980); P. G. Higgs and J. F. Joanny, J. Chem. Phys. **94**, 1543 (1991).

[3] Y. Kantor, H. Li, and M. Kardar, Phys. Rev. Lett. **69**, 61 (1992); Phys. Rev. E **49**, 1383 (1994).

[4] A. V. Dobrynin and M. Rubinshtein, J. Phys. II (France) **5**, 677 (1995).

[5] J. Copart and F. Candau, Macromolecules **26**, 1333 (1993).

[6] A. E. English, S. Mafe, J. A. Manzanares, X.-H. Yu, A. Yu. Grosberg, and T. Tanaka, J. Chem. Phys. **104**, 8713 (1996).

[7] Y. Kantor and M. Kardar, Europhys. Lett. **27**, 643 (1994); Phys. Rev. E **51**, 1299 (1995).

[8] Y. Kantor and M. Kardar, Phys. Rev. E **52**, 835 (1995).

[9] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).

[10] V. S. Pande, A. Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).

[11] It is tempting to compare this conclusion with an earlier result [3] that PA interactions are relevant for ideal chains in $d < 4$. This may, however, be just a coincidence as our result here is about energy *correlations* ($N^2/[N^{1/d}]^{d-2}$) for a *compact* chain of the size $\sim N^{1/d}$, while the result of Ref. [3] is obtained by considering the interaction of two half chains, with charge imbalance $N^{1/2}$ each, over the Gaussian distance $N^{1/2}$, i.e., $[N^{1/2}]^2/[N^{1/2}]^{d-2}$.

[12] V. S. Pande, C. Joerg, A. Yu. Grosberg, and T. Tanaka, J. Phys. A **27**, 6231 (1994).

[13] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Shulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).

[14] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12 976 (1994).