PHYSICAL REVIEW E **86**, 050901(R) (2012)

Sequence repeats and protein structure

Trinh X. Hoang,[1] Antonio Trovato,[2] Flavio Seno,[2] Jayanth R. Banavar,[3] and Amos Maritan[2]
[1]*Institute of Physics, Vietnam Academy of Science and Technology, 10 Dao Tan, Hanoi 10000, Vietnam*
[2]*Dipartimento di Fisica e Astronomia "G. Galilei", Università di Padova and CNISM, Unità di Padova, Via Marzolo 8, 35131 Padova, Italy*
[3]*Department of Physics, University of Maryland, College Park, Maryland 20742, USA*
(Received 29 June 2012; published 8 November 2012)

Repeats are frequently found in known protein sequences. The level of sequence conservation in tandem repeats correlates with their propensities to be intrinsically disordered. We employ a coarse-grained model of a protein with a two-letter amino acid alphabet, hydrophobic (H) and polar (P), to examine the sequence-structure relationship in the realm of repeated sequences. A fraction of repeated sequences comprises a distinct class of bad folders, whose folding temperatures are much lower than those of random sequences. Imperfection in sequence repetition improves the folding properties of the bad folders while deteriorating those of the good folders. Our results may explain why nature has utilized repeated sequences for their versatility and especially to design functional proteins that are intrinsically unstructured at physiological temperatures.

DOI: 10.1103/PhysRevE.86.050901 PACS number(s): 87.15.Cc, 87.15.ak

Proteins [1] are short linear heteropolymers made up of amino acids able to adopt a wide range of structures uniquely determined by their sequence. The vast majority of proteins fold in aqueous environments into globular compact structures that are stabilized by the cooperative formation of a unique hydrophobic core.

Not all sequences follow this general pattern, and several proteins fold into conformations whose architecture consists of repeated structural regions [2]. Such repeat proteins are present in 14% of known protein sequences with specific functions generally associated with higher organisms [3]. The presence of sequence repeats (or tandem repeats) is believed to imply the formation of the modular architectures of repeat proteins. Yet sequence repeats evolve quickly and their periodicity may become rapidly hidden at the sequence level, while still evident in the structure, so that sophisticated algorithms have been developed to predict structural periodicity from sequence [4–6].

The sequence-structure relationship in proteins reveals other intriguing subtleties, as several protein sequences have been recently shown to be intrinsically unstructured in physiological conditions, adopting a large variety of different conformations interchanging with each other. These intrinsically unstructured proteins (IUPs) interact with different molecular partners and may adopt relatively rigid conformations only in the presence of natural ligands [7–13].

IUPs misfolding into insoluble fibrillar aggregates have been implicated in several devastating neurodegenerative human diseases [14], such as Alzheimer's, Parkinson's, and Creutzfeldt-Jacob's. Interestingly, tandem repeats are often found in proteins associated with such diseases. Examples include homorepeat expansion, as in Huntington's disease, and the octarepeat domain in the N-terminal fragment of the prion protein [15]. A further recent insight was provided in [16], where it was found that the level of sequence conservation in tandem repeats correlates with their propensity to be intrinsically unstructured. This is even more striking in light of the fact that redesigned repeat proteins employed perfect tandem repeats [17]. Indeed, the majority of perfect tandem repeats are found in the Protein Data Bank within proteins designed *de novo*.

In recent years, we have shown how several aspects of the sequence-structure relationships can be rationalized within a general framework by using simplified models [18–23]. We have proposed that protein-like structures are found as low-energy conformations in a marginally compact "phase" of matter in the proximity of a transition to the swollen phase. The marginally compact phase emerges as a consequence of the geometry and symmetry implied by the self-avoiding tube description of the protein backbone together with the directionality of hydrogen-bond-like interactions. In this way, the limited menu of globular protein folds observed in nature [24] arises independent of sequence specificity, a fact recently verified by means of accurate all-atom molecular dynamics simulations [25]. The role of sequence is then to pick from this menu the best-fit fold for its native state, and sequence design turns out to be relatively straightforward starting from random peptides. We have also modeled the disorder-to-order transition that takes place when IUPs bind to their target.

In this Rapid Communication, we will investigate within the same simplified framework the nuances in the protein sequence-structure relationship brought about by the presence of repetitive sequence patterns. The key specific question we ask is as follows: Can we explain in simple terms the presence of perfect sequence repeats devoid of any structural order?

The polymer model considered in our study [19] mimics proteins within a coarse-grained description where beads are located at the positions of the C_α atoms and constrained to stay along the axis of a self-avoiding tube of thickness $\Delta = 2.5$ Å [26,27]. The model is described in full detail in [23] and entails specific interactions such as bending energy, pairwise hydrophobic contact interaction, and directional hydrogen bonding with energetic and geometrical constraints. We employ a two-letter hydrophobic-polar (HP) amino acid alphabet.

We employ a parallel tempering [28] Monte Carlo (MC) scheme for obtaining the ground state as well as other equilibrium characteristics of the system. For each system, 20 to 24 replicas are considered, each evolving at its own selected temperature T_i. For each replica, the simulation is carried out with standard pivot and crankshaft move sets [29] and the METROPOLIS algorithm for move acceptance. An attempt to

 ©2012 American Physical Society

exchange replicas is made every 100 MC steps. The weighted multiple-histogram technique [30] is used to compute the specific heat of the system.

In order to test the sequence-structure relationship in the presence of repeated sequences, we chose to compare the thermodynamic properties of a set of 28 repeated HP sequences with those of a set of 20 random HP sequences. In all cases, we considered HP sequences consisting of 48 residues. The fraction of H residues in the sequence is kept constant at 0.25. The length of the repeated pattern is 8, so the whole sequence consists of six repeated patterns each including two H residues. There are exactly 28 distinct HP patterns of length 8 with two H residues, and all of them are considered in our study. Repeated sequences sample uniformly different values of $s_{\max}$, the length of the longest stretch of consecutive P residues in the sequence. Flanking polar stretches at the end and at the beginning of the repeat are considered as one stretch (see Fig. 1 for specific examples).

Figure 1 shows some of the ground states obtained for repeated sequences. As expected, sequence repetition results in modular ground state structures that essentially provide the best fit for the regular sequence pattern. However, there are exceptions to this rule (see S7), showing that already at the ground state level sequence repetition does not always imply a modular structure. Intriguingly, ground states of repeated sequences exhibit either α-helical or β-sheet structures, as is indeed found in repeat protein structures [31].

We then investigated thermodynamic folding properties of different sequences, by looking at their specific heat curves as

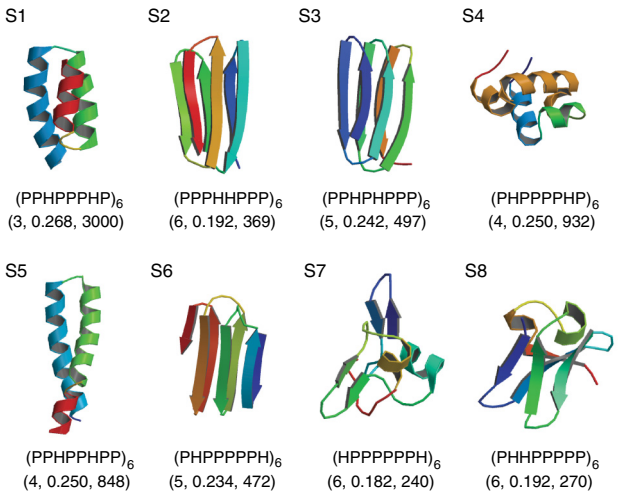

FIG. 1. (Color online) Gallery with ground states of repeated sequences. Ribbon presentations of the ground state conformations are shown for eight sequences (S1–S8) in a coarse-grained model of a protein with tube symmetry, directional hydrogen bonding, and hydrophobic interaction with hydrophobic (H) and polar (P) amino acids (see main text). The corresponding sequences are shown below their ground state conformations as sixfold repeats of eight residue HP patterns. The numbers shown in parentheses are the length of the longest stretch of consecutive P residues in the sequence, $s_{\max}$, the temperature of the maximum of the specific heat, $T_{\max}$, in units of ϵ/k_B, and the value of that maximum, $C_{\max}$, in units of k_B, of the corresponding sequence. The ground states as well as the specific heat are obtained through parallel tempering Monte Carlo simulations and by using the weighted histogram method.

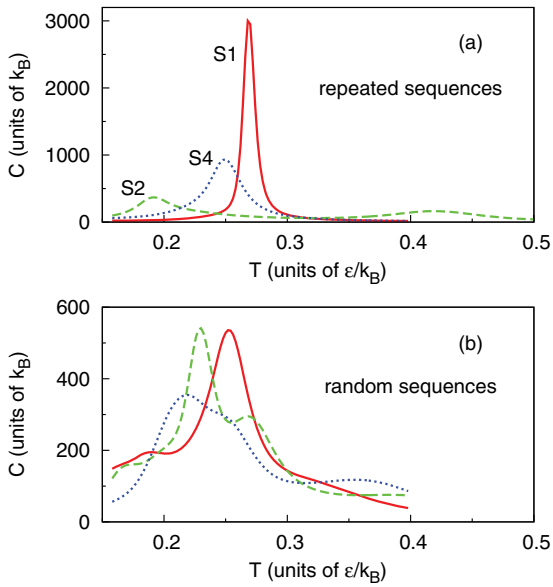

FIG. 2. (Color online) Temperature dependence of the specific heat C of three selected repeated sequences (S1, S2, and S4) (a) and several random sequences (b) with a length of $N = 48$ beads within our coarse-grained protein model. The ground states for the repeated sequences are shown in Fig. 1. Note that sequences S1 and S4 display a single peak in the specific heat curve while sequence S2 has two peaks. Based on the position and the height of the main peak in the specific heat curve, one could classify sequence S2 as a bad folder because it has a significantly lower ground state stability than the two other sequences. The specific heat curves of random sequences (b) typically display a main peak along with a smaller peak or a shoulder at either a higher or lower temperature, manifesting more complex folding properties than a simple two-state folder.

a function of temperature. Some examples are shown in Fig. 2, for both repeated and random sequences.

Good folders are characterized by a high stability and folding cooperativity [32], as signaled by the position ($T_{\max}$) and the height of the specific heat peak ($C_{\max}$), respectively. Good folder sequences are expected to have a single peak in the specific heat, while for bad folders the specific heat may have multiple peaks and/or the main peak may be broader. Note that the main peak of the specific heat signals a folding transition from a denatured high-energy state to a compact low-energy state. However, the low-energy state may or may not correspond to a ground state conformation well separated in energy from other excited states. This also determines the good or bad folding character of the considered sequence. For good folder sequences, the specific heat at low temperatures, C_{low}, should be small. One might expect that the properties are correlated, such that sequences with a high $T_{\max}$ and a high $C_{\max}$ also have a low C_{low}. A similar thermodynamic analysis was employed in previous work [23], showing that designed sequences are indeed better folders than random sequences, within the same HP model.

In order to assess the relation between repeat perfection and intrinsic protein disorder, we need to identify how putative IUPs behave within our model. Our main assumption here is to identify IUPs as those bad folder sequences whose folding

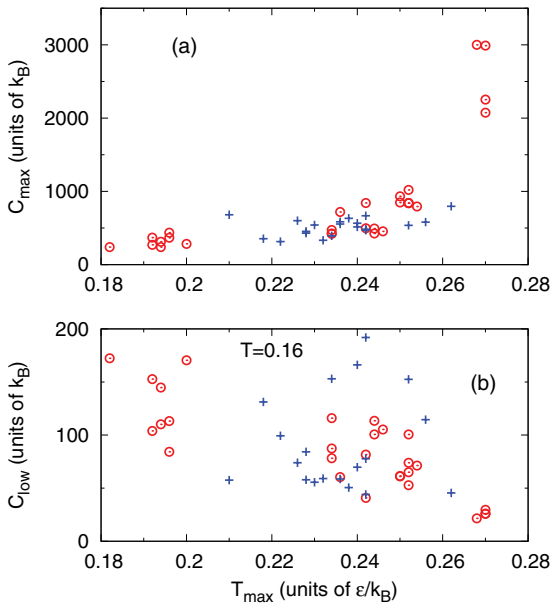

FIG. 3. (Color online) Correlation between different folding properties. The panels show the maximum value of the specific heat, $C_{\max}$ (a) and the specific heat at a low temperature, C_{low} (b), plotted against the temperature of the maximum of the specific heat, $T_{\max}$, for 28 repeated sequences (open circles) and 20 random sequences (crosses) considered in this study. C_{low} is obtained at temperature $T = 0.16\varepsilon/k_B$.

temperature $T_{\max}$ is below the physiological temperature in native conditions, as was already done in Ref. [21].

Figures 3(a) and 3(b) show how different indicators of folding properties correlate with each other for both repeated and random sequences. Strikingly, one can see how repeated sequences cluster into two main groups, well separated by the values of their folding temperature $T_{\max}$. The first (second) group is characterized by $T_{\max} > 0.22$ ($T_{\max} < 0.22$). For repeated sequences $T_{\max}$ and C_{low} are correlated, as somewhat expected. For random sequences, they are uncorrelated.

Even more strikingly, the folding temperatures of the random sequences are never lower than 0.21, so that repeated sequences in the second cluster are worse folders than all the random sequences considered in this study. The p value assigned to the null hypothesis that the mean folding temperature of the second cluster of repeated sequences is the same as for random sequences, computed using the Welch's t test, is equal to 1.3×10^{-12}. On the other hand, if one assumes that the distributions of the folding temperatures of the two groups of sequences are identical then the p value, given by the Mann-Whitney-Wilcoxon test, is 3.2×10^{-7}. Both statistical tests provide overwhelming evidence that the folding temperatures of the repeated sequences in the second cluster tend to be lower than those of the random sequences.

This result cannot be explained by assuming that all random sequences that we tested are good folders, because some of them have high values of C_{low}, even at relatively high folding temperature, signaling that the low-energy conformations sampled at low T are not unique. On the other hand, it is known that a good fraction of random polypeptide libraries do actually exhibit good folding properties [33].

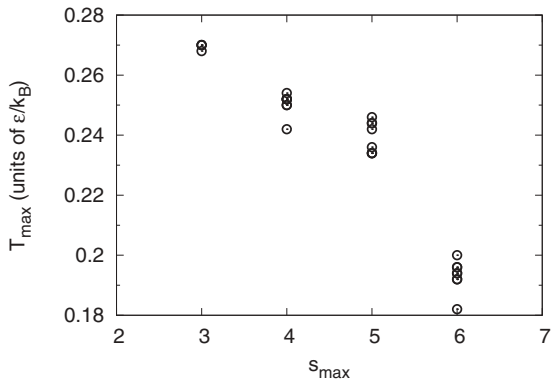

FIG. 4. Correlation between sequence properties and folding properties. The longest stretch of segment containing only P residues in the sequence, $s_{\max}$, plotted against the temperature of the maximum of the specific heat, $T_{\max}$, for 28 repeated sequences (open circles) considered in this study.

Our main conclusion is that the second cluster of repeated sequences is made up of bad folders. Moreover, having identified intrinsic disorder with low folding temperature, bad folders with a repeat sequence will result in an intrinsic disorder on average significantly higher than the one achievable with random sequences. If structural disorder brings some advantage to living organisms, a way to enhance it is to select repeated sequences.

However, not all repeat sequences produce intrinsic disorder. Many of them do actually correspond to very good folders with highly cooperative folding transitions, typically the ones with α-helical ground states (like S1). In principle, we can speculate that for good repeated folders a structure from the presculpted menu can be selected that nicely fits the sequence repetition pattern, whereas for bad repeated folders no such structure exists in the presculpted menu. Figure 4 shows how $s_{\max}$, a simple descriptor of sequence repetition, correlates with the folding temperature. Our results clearly indicate that, for repeated sequences, the longer the longest stretch of segments containing only polar residues in the sequence, the lower is the folding temperature (such a correlation is not present in the case of random sequences). All the putative intrinsically disordered sequences in our study have the longest stretch containing six polar residues. In the simulations, we found that for repeated sequences with long stretches of polar residues, the hydrophobic core is formed at a much higher temperature (as indicated by the appearance of a second peak of the specific heat curve for sequence S2 in Fig. 1) than the folding temperature, leading to noncooperative folding. It has been found [34] that intrinsically disordered proteins have a lower sequence complexity, as measured by Shannon's entropy, than ordered proteins. Strikingly, our simple descriptor $s_{\max}$ is related to sequence entropy as it defines the longest stretch of one type of amino acid. The longer the $s_{\max}$, the lower is the entropy. Thus, our results show a perfect example of how lowering sequence complexity may lead to intrinsic disorder.

Next, we examine how imperfection in the sequence repetition affects the folding properties. We proceed by introducing a slight mutation into the repeated sequences by swapping the

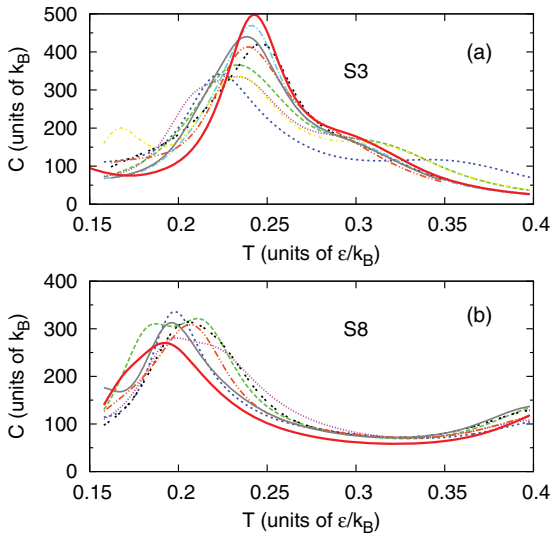

FIG. 5. (Color online) Effect of mutations on the folding properties of repeated sequences. The panels show the temperature dependence of the specific heat C for a moderately good folder, sequence S3 (a), and a bad folder, sequence S8 (b), and their mutated sequences. The specific heat curves are shown as solid lines for the original sequence and discontinuous lines for the mutated sequences, respectively. The mutated sequences are obtained from the original sequence by swapping a randomly chosen H residue with a randomly chosen P residue.

position of a H residue with a P residue, both randomly chosen. Mutations in a good folder decrease the folding temperature, as expected [Fig. 5(a)]. In contrast, Fig. 5(b) shows that mutations in a bad folder increase the folding temperature. This result is in good agreement with the finding by Jorda *et al.* [16] that if repeat perfection is lost, intrinsic disorder is also diminished.

In summary, by using a relatively simple coarse-grained model, we have shown that repetition in a protein's amino acid sequence does not necessarily imply a repetition of structural motifs in the folded structure. A few repeated sequences, whose folded structures are well modulated, have very good folding properties as given by a high thermodynamic stability of the folded state and a cooperative folding transition. These two properties are typically well correlated. On the other hand, a few other repeated sequences exhibit exceptionally bad folding. The folding temperatures of these repeated sequences are much lower than those of random sequences, suggesting that they may be intrinsically disordered at physiological temperatures. Thus, our results explain why sequence repetition is found in many sequences with intrinsic disorder. In agreement with previous experiments, we have shown that increasing imperfection in sequence repetition moves bad folding repeated sequences towards random sequences and thus can destroy the intrinsic disorder. Given that repeat sequences encode much less information than a random sequence and that their folding properties exhibit a wide range of behaviors, repeat proteins can be deployed by nature in a versatile manner.

We are indebted to Paul Smith for helpful comments. This work was supported by NAFOSTED Grant No. 103.01-2010.11.

[1] A. V. Finkelstein and O. Ptitsyn, *Protein Physics* (Academic Press, London, 2002).
[2] M. A. Andrade, C. P. Ponting, T. J. Gibson, and P. Bork, J. Mol. Biol. **298**, 521 (2000).
[3] E. M. Marcotte, M. Pellegrini, T. O. Yeates, and D. Eisenberg, J. Mol. Biol. **293**, 151 (1999).
[4] A. Biegert and J. Soeding, Bioinformatics **24**, 807 (2008).
[5] L. Marsella, F. Sirocco, A. Trovato, F. Seno, and S. C. E. Tosatto, Bioinformatics **25**, I289 (2009).
[6] A. Vo and N. Nguyen, and H. Huang, Bioinformatics **26**, I467 (2010).
[7] P. E. Wright, and H. J. Dyson, J. Mol. Biol. **293**, 7960 (1999).
[8] A. K. Dunker *et al.*, J. Mol. Graphics Modell. **19**, 26 (2001).
[9] A. K. Dunker, and Z. Obradovic, Nat. Biotechnol. **19**, 805 (2001).
[10] H. J. Dyson, and P. E. Wright, Curr. Opin. Struct. Biol. **12**, 54 (2002).
[11] P. Tompa, Trends Biochem. Sci. **27**, 527 (2002).
[12] V. N. Uversky, Eur. J. Biochem. **269**, 2 (2002).
[13] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa, J. Mol. Biol. **338**, 1015 (2004).
[14] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).
[15] A. Y. Yam, C. M. Gao, X. M. Wang, P. Wu, and D. Peretz, PLoS One **5**, 9316 (2010).
[16] J. Jorda, B. Xue, V. N. Uversky, and A. V. Kajava, FEBS J. **277**, 2673 (2010).
[17] E. R. G. Main, A. R. Lowe, S. G. J. Mochrie, S. E. Jackson, and L. Regan, Curr. Opin. Struct. Biol. **15**, 464 (2005).
[18] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Nature (London) **406**, 287 (2000).
[19] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. USA **101**, 7960 (2004).
[20] J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato, Phys. Rev. E **70**, 041905 (2004).
[21] T. X. Hoang, L. Marsella, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. USA **103**, 6883 (2006).
[22] J. R. Banavar, M. Cieplak, T. X. Hoang, and A. Maritan, Proc. Natl. Acad. Sci. USA **106**, 6900 (2009).
[23] J. R. Banavar, T. X. Hoang, F. Seno, A. Trovato, and A. Maritan, J. Stat. Phys. **148**, 636 (2012).
[24] C. Chothia, Nature (London) **357**, 543 (1992).
[25] P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, and A. Laio, PLoS Comput. Biol. **6**, 1000957 (2010).
[26] O. Gonzalez and J. H. Maddocks, Proc. Natl. Acad. Sci. USA **96**, 4769 (1999).
[27] J. R. Banavar, O. Gonzalez, J. H. Maddocks, and A. Maritan, J. Stat. Phys. **110**, 35 (2003).
[28] R. H. Swendsen and J. S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).

[29] In a crankshaft move, the two beads i and j that define the rotation axis are chosen randomly with the constraint $i - j < 6$. In both move sets, the rotation angle is drawn randomly from a Gaussian distribution of zero mean and a dispersion of 4°.

[30] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[31] A. V. Kajava, J. Struct. Biol. **134**, 132 (2001).

[32] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).

[33] A. R. Davidson, K. J. Lumb, and R. T. Sauer, Nat. Struct. Biol. **2**, 856 (1995).

[34] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, Proteins **42**, 38 (2001).