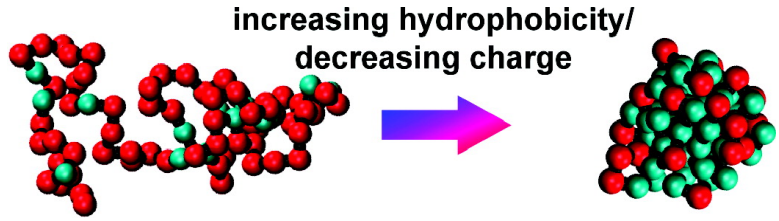
Article

Natively Unfolded Protein Stability as a Coil-to-Globule Transition in Charge/Hydropathy Space

Henry S. Ashbaugh, and Harold W. Hatch

J. Am. Chem. Soc., **2008**, 130 (29), 9536-9542 • DOI: 10.1021/ja802124e • Publication Date (Web): 25 June 2008



J|A|C|S
ARTICLES

Published on Web 06/25/2008

Natively Unfolded Protein Stability as a Coil-to-Globule Transition in Charge/Hydropathy Space

Henry S. Ashbaugh* and Harold W. Hatch

Department of Chemical and Biomolecular Engineering, Tulane University, New Orleans, Louisiana 70118

Received March 21, 2008; E-mail: hanka@tulane.edu

Abstract: In the absence of experimental assignments, the empirical charge/hydropathy correlation for the prediction of natively unfolded protein sequences (Uversky, V. N.; Gillespie, J. R.; Fink, A. L. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 415−427) provides perhaps the most intuitive description of gross polypeptide conformation. The success of this correlation rests on an essential chain length independence of the boundary line between expanded and compact conformations, conversely stabilized by highly charged/weakly hydrophobic residues or weakly charged/highly hydrophobic residues, respectively. We present extensive simulation results for coarse-grained polypeptides over a wide range of sequence hydrophobicities, charges, and lengths. A coil-to-globule transition in sequence composition space analogous to the charge/hydropathy correlation is observed. A near sequence length independent stability boundary is only found when counterions for the charged peptides are explicitly included, as a result of counterion condensation stabilization of repulsive electrostatic interactions on the globule surface. The observed counterion adsorption is shown to be in quantitative agreement with theoretical condensation predictions. We argue that alternate functionalities, beyond charge and hydrophobicity, empirically known to correlate with conformational disorder can be incorporated into our minimalist polypeptide model to study the interplay between independent predictors of unfolded sequences.

Introduction

The central tenet of protein form and function is that individual polypeptides fold and self-assemble into three-dimensional structures that dictate their biological role.[1,2] Increasing evidence has amassed, however, that many proteins display functions that *require* conformational disorder.[3–5] Examples include tumor suppressor p53,[6,7] which regulates gene expression through disordered DNA binding domains, and phenylalanine/glycine-nucleoporins,[8,9] which direct all cellular traffic between the nucleus and cytosol through the nuclear pore complex. Analysis of representative prokaryotic and archea genomes find ∼5% of their sequences code for proteins with long unstructured or partially unstructured domains, compared to ∼30% for eukaryotic genomes![10,11] It may then be inferred that unstructured proteins not only possess functions but also are evolutionarily beneficial for complex multicellular organisms.

Disordered proteins have been identified through a range of experimental assays including circular dichroism measurements of secondary structural content, scattering, and sedimentation determinations of hydrodynamic dimensions, protease digestion sensitivity, X-ray crystallography, and heteronuclear multidimensional NMR.[8,12] In the absence of experimental assignments, a significant effort has been directed toward developing empirical sequence based bioinfomatic tools for disorder prediction, and a number of web based applications have become available, *e.g.*, PONDR,[13] FoldUnfold,[14] and DisEMBL.[15] Data mining has shown that many amino acids can be grouped into larger families that accurately correlate with disorder.[16] Alternatively, average sequence properties like flexibility[17] (characterized by uncertainties in crystallographic α-carbon positions[18]), complexity[13] (characterized by Shannon's information entropy analysis of amino acid sequence populations[19]), contacts[20] (characterized

(1) Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Principles of biochemistry*, 2nd ed.; Worth Publishers: New York, 1993.
(2) Petsko, G. A.; Ringe, D. *Protein structure and function*; Sinauer Associates: Sunderland, MA, 2004.
(3) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321–331.
(4) Fink, A. L. *Curr. Opin. Struc. Biol.* **2005**, *15*, 35–41.
(5) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obrdovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
(6) Lee, H.; Mok, K. H.; Mudhandiram, R.; Park, K.-H.; Suk, J.-E.; Kim, D.-H.; Chang, J.; Sung, Y. C.; Choi, K. Y.; Han, K.-H. *J. Biol. Chem.* **2000**, *275*, 29426–29432.
(7) Bell, S.; Klein, C.; Müller, L.; Hansen, S.; Buchner, J. *J. Mol. Biol.* **2002**, *322*, 917–927.
(8) Denning, D. P.; Patel, S. S.; Uversky, V. N.; Fink, A. L.; Rexach, M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2450–2455.
(9) Shulga, N.; Goldfarb, D. S. *Mol. Cell. Biol.* **2003**, *23*, 534–542.
(10) Dunker, A. K.; Obrdovic, Z. *Nat. Biotechnol.* **2001**, *19*, 805–806.
(11) Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K. *Biochemistry* **2005**, *44*, 1989–2000.
(12) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(13) Romero, P.; Obrdovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 38–48.
(14) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Lobanov, M. Y. *Bioinformatics* **2006**, *22*, 2948–2949.
(15) Linding, R.; Jensen, L. J.; Diella, F.; Bork, P.; Gibson, T. J.; Russell, R. B. *Structure* **2003**, *11*, 1453–1459.
(16) Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. *FEBS Lett.* **2004**, *576*, 348–352.
(17) Radivojac, P.; Obrdovic, Z.; Smith, D. F.; Zhu, G.; Vucetic, S.; Brown, C. J.; Lawson, J. D.; Dunker, A. K. *Protein Sci.* **2004**, *13*, 71–80.
(18) Rhodes, G. *Crystallography made crystal clear: A guide for users of macromolecular models*; Academic Press: San Diego, CA, 1993.
(19) Wootton, J. C.; Federhen, S. *Comput. Chem.* **1993**, *17*, 149–163.

10.1021/ja802124e CCC: $40.75

by residue packing densities), and hydrophobicity[11] (characterized by relative side chain solubilities[21]) have likewise been found to be strong predictors. Improved correlations can be derived by combining metrics,[11,22] although it has been noted that variables like flexibility and contacts are not necessarily independent of one another.[17,23] The correlation between these parameters and protein conformational stability is largely empirical, rather than based on more fundamental physics. Moreover, it has been shown that artificial, fitted parameters that are independent of the physical characteristics of the amino acids may even provide an improved correlation.[20]

Perhaps the most intuitively appealing predictor for unfolded protein stability is the empirical charge/hydropathy correlation developed by Uversky, Gillespie, and Fink.[12,24] They showed that when the mean net absolute charge of a polypeptide at neutral pH is plotted against the mean side chain hydropathy, measured on the Kyte−Dolittle hydrophobicity scale,[21] a stability boundary line between compact (natively folded or globular) and expanded (coil-like or premolten globular) conformations is observed. This correlation is reminiscent of a phase transition in a noncanonical state variable space, where the unfolded state is stabilized by low hydrophobicity and high net charge, while the folded state is stabilized by high hydrophobicity and low net charge. This stability boundary was demonstrated for proteins with a broad range of sequence lengths from 49 to 1827 residues. Near the boundary, however, exceptions can be found, suggesting an interplay between charge/hydropathy[11] and alternate disorder predictors that are not readily decoupled experimentally.

Dill developed a theoretical framework for exploring the stability of globular proteins modeled as random heteropolymers of hydrophobic and uncharged hydrophilic monomers.[25,26] Above a critical fraction of hydrophobic sites, a coil-to-globule collapse transition for uncharged proteins is predicted, comparable to the zero charge intercept of the charge/hydropathy correlation. Charge effects have been incorporated in this model through a linearized Poisson−Boltzmann approach to examine the effects of pH and added salt.[27,28] A significant result of that work was that the maximal stability of the folded state is predicted to occur at the isoelectric point, although additional stability can be gained through the burial of nontitratable groups within the protein interior. Alternate theoretical efforts to include charge effects on the coil-to-globule transition have focused on weakly charged homopolyelectrolytes.[29,30] Recent simulations have demonstrated that since the folded state is typically lower in free energy than the unfolded state, folded protein stability can be achieved by selective mutations that neutrally affect the folded state while raising the free energy of the unfolded state.[31,32] Thus, an understanding of unfolded proteins can play a role in the design of stable protein folds[33,34] and the interpretation of sequence fitness from directed evolution.[35,36] The explicit simulation of natively unfolded proteins in water is problematic, however, since the number of waters required to hydrate a random coil grows more quickly than the number of residues, making the simulation of large numbers of coil-like sequences of even modest length computationally prohibitive. As a result, little simulation work has been performed to explicitly study the stability of extended disordered polypeptides. Alternatively, coarse-grained models which wash out aqueous molecular degrees-of-freedom can be implemented.[37]

Motivated by the physically appealing charge/hydropathy correlation for natively unfolded protein stability,[24] we present extensive simulation results for the conformational equilibrium of coarse-grained polypeptides. Our model incorporates interactions among hydrophobic, polar, and cationic residues. Molecular dynamics simulations were performed on nearly 25 000 random sequences over a broad range of sequence compositions and lengths. We show that this minimalist model displays a coil-to-globule stability boundary analogous to the empirical charge/hydropathy correlation. Analysis of the structure of compact globular states shows a sequence length dependence of the adsorption of counterions, which stabilizes the globule as the surface area to volume ratio decreases. This charge adsorption, in turn, mutes the sequence length dependence of the coil-to-globule stability boundary in agreement with the empirical correlation. An analytical counterion condensation theory description is developed and shown to successfully predict the observed surface adsorption, underscoring the role of counterions in stabilizing the globular state. We believe this is the first demonstration of a minimalist simulation model that captures the hallmarks of natively unfolded protein stability and argue that additional functionalities may be introduced to study the interplay between orthogonal factors that similarly correlate with protein disorder.

Theoretical Methods

Molecular Dynamics Simulations. The charge/hydropathy correlation for protein conformational stability implies a minimal coarse-grained three amino acid alphabet consisting of hydrophobic, uncharged polar, and cationic polar residues. Additional functionalities can be easily introduced, such as anionic residues and chain flexibility, but presently we consider only the minimal requirements to model natively unfolded protein conformational stability. Water is only considered implicitly. Interactions between nonbonded residues are modeled as

$$\phi(r) = \phi_{\text{repulsive}}(r) + \lambda\phi_{\text{attractive}}(r) \tag{1a}$$

where

$$\phi_{\text{repulsive}}(r) = \begin{cases} \phi_{\text{LJ}}(r) + \varepsilon & r \le 2^{1/6}\sigma \\ 0 & r > 2^{1/6}\sigma \end{cases} \tag{1b}$$

and

(20) Garbuzynskiy, S. O.; Yu, M.; Galzitskaya, O. V. *Protein Sci.* **2004**, *13*, 2871–2877.
(21) Kyte, J.; Doolittle, B. F. *J. Mol. Biol.* **1982**, *157*, 105–132.
(22) Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. *Proteins: Struct., Funct., Bioinf.* **2007**, 66.
(23) Halle, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1974–1979.
(24) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 415–427.
(25) Dill, K. A. *Biochemistry* **1985**, *24*, 1501–1509.
(26) Dill, K. A.; Alonso, D. O. V.; Hutchinson, K. *Biochemistry* **1989**, *28*, 5439–5449.
(27) Stigter, D.; Dill, K. A. *J. Phys. Chem.* **1989**, *93*, 6737–6743.
(28) Stigter, D.; Dill, K. A. *Biochemistry* **1990**, *29*, 1262–1271.
(29) Khokhlov, A. R.; Kramarenko, E. Y. *Macromolecules* **1996**, *29*, 681–685.
(30) Kramarenko, E. Y.; Khokhlov, A. R.; Yoshikawa, K. *Macromolecules* **1997**, *30*, 3383–3388.
(31) van Gunsteren, W. F.; Burgi, R.; Peter, C.; Daura, X. *Angew. Chem., Int. Ed.* **2001**, *40*, 352–355.
(32) Smith, P. E.; Lei, H. X. *Biophys. J.* **2003**, *85*, 3513–3520.
(33) de Alba, E.; Jimenez, M. A.; Rico, M. *J. Am. Chem. Soc.* **1997**, *119*, 175–183.
(34) Anil, B.; Craig-Schapiro, R.; Raleigh, D. P. *J. Am. Chem. Soc.* **2006**, *128*, 3144–3145.
(35) Giver, L.; Gershenson, A.; Freskgard, P.-O.; Arnold, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12809–12813.
(36) Miyazaki, K.; Wintrode, P. L.; Grayling, R. A.; Rubingh, D. N.; Arnold, F. H. *J. Mol. Biol.* **2000**, *297*, 1015–1026.
(37) Bright, J. N.; Woolf, T. B.; Hoh, J. H. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 131–173.

$$\phi_{\text{attractive}}(r) = \phi_{\text{LJ}}(r) - \phi_{\text{repulsive}}(r) \quad (1c)$$

In these equations, $\varphi_{\text{LJ}}(r)$ is the Lennard−Jones potential with a well depth and diameter of ε and σ, respectively. The parameter λ moderates the strength of intersite attractive interactions, assuming values of 0, 1, and 2.5 for polar−polar, polar−hydrophobic, and hydrophobic−hydrophobic interactions, respectively. Attractive interactions were truncated and shifted to zero for separations beyond 3σ. A specified fraction of the polar monomers were assigned a cationic charge, imparting a polyelectrolyte character to the chains. Bonded residue interactions were modeled as $\varphi_{\text{bonded}}(r) = K(r - b)^2$, with $K = 800\varepsilon/\sigma^2$ and $b = 1\sigma$.

In one set of calculations, explicit anionic counterions for the charged residues were included with no additional electrolyte. The counteranions were modeled as nonbonded polar monomers with a charge opposite that of the cationic monomers, *i.e.*, $q = \pm 1e$. In these simulations, electrostatic interactions were evaluated using Ewald summation with conducting boundary conditions and a bulk dielectric constant D.[38] The periodic simulation box had a side length of $L = 80\sigma$, which is greater than twice the maximum radius-of-gyration of the simulated sequences, and thereby below the polymer overlap concentration. In a second set of calculations, counterions were treated implicitly and lumped into a screened electrostatic interaction between cationic monomers, $\varphi_{\text{electrostatic}}(r) = e^2 \exp(-\kappa r)/(4\pi\varepsilon_0 D r)$, in an infinite solvent.[39] The Debye screening length, κ^{-1}, was chosen to match the ionic strength of comparable simulations with explicit counterions determined by the net concentration of charged monomers and counterions. While this assumption is not strictly correct, the calculated values of κ are large and the long-range interactions deviate little from the bare $1/r$ charge interaction over the radii-of-gyration of the simulated polypeptides.

Molecular dynamics simulations were performed in the canonical ensemble at a temperature of $T = 2\ \varepsilon/k_B$, which is comparable to an ambient temperature of 300 K for an assumed well depth of $\varepsilon = 0.3$ kcal/mol, the well depth of united-atom methane−methane interactions in the OPLS parameter set.[40] Following this value of ε, the value of $\lambda = 2.5$ for hydrophobic/hydrophobic interactions assigned above mimics the strength of water mediated methane/methane interactions.[41] The collection of terms $(4\pi\varepsilon_0 D)^{-1}$ was chosen to be equal to $2.75\ \varepsilon\sigma/e^2$, which follows from an aqueous dielectric constant of $D \approx 80$ and a mean peptide residue diameter of $\sigma = 5$ Å. The temperature was controlled using Andersen's stochastic thermostat.[42] The equations of motion were integrated using a time step of $\delta t = 0.002(m\sigma^2/\varepsilon)^{1/2}$, where m is the mass of the individual monomers and counterions.

Simulations were conducted for fixed polypeptide sequence lengths, N, at specified fractions of hydrophobic, $\langle H\rangle = N_{\text{hydrophobic}}/N$, and cationic, $\langle q\rangle = N_+/N$, residues. Sequence lengths of $N =$ 50, 100, 150, 200, 250, 300, and 350 residues were considered with charge fractions of $\langle q\rangle = 0.00$, 0.05, 0.10, 0.15, and 0.20. In cases where $\langle q\rangle N$ yielded a fractional number of charged residues (*e.g.*, $N = 150$ and $\langle q\rangle = 0.05$), the number of charged residues was rounded up ($N = 50$) or down ($N = 150$, 250, and 350) to the closest integer. Anywhere from 14 to 16 values of the mean hydrophobicity from $\langle H\rangle = 0$ to $1 - \langle q\rangle$ for each fixed charge fraction were examined. Since our coarse-grained polypeptides lack a genome, 20 to 30 random sequences were generated for each N, $\langle H\rangle$, and $\langle q\rangle$ to average out sequence specific effects. Simulations were repeated with explicit and implicit counterions. For a given polypeptide sequence, 10^6 molecular dynamics steps were performed for equilibration, followed by 5×10^6 steps for the evaluation of averages. A total of 24 830 random sequences were simulated.

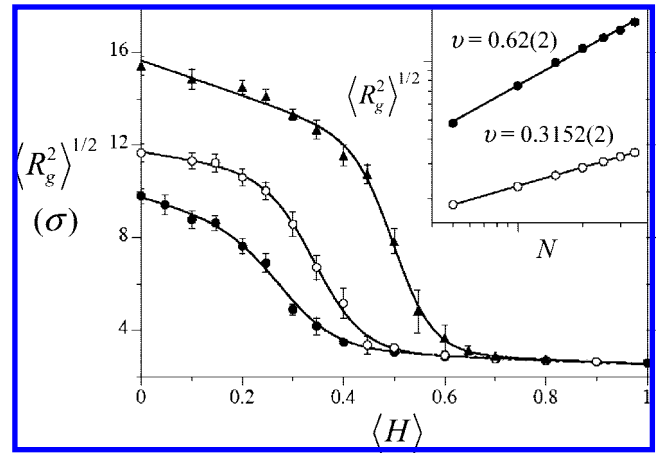


Figure 1. Radii-of-gyration of 150-mer polypeptide chains as a function of fractional hydrophobicity with charge fractions of $\langle q\rangle = 0.0$ (solid circles), 0.1 (open circles), and 0.2 (solid triangles). The points indicate simulation results averaged over 20 random polypeptide sequences. Error bars indicate 1 standard deviation. The lines indicate the fit of the simulation results to eq 2. The inset figure shows the radius-of-gyration as a function of sequence length for the uncharged chain ($\langle q\rangle = 0.0$) for fractional hydrophobicities of $\langle H\rangle = 0.0$ (filled circles) and 1.0 (open circles). The lines show fits to power law scaling $\langle R_g^2\rangle^{1/2} \propto N^{\upsilon}$ characterizing the conformational state of the polypeptides.

Results and Discussion

Charge/Hydropathy Driven Coil-to-Globule Transition. Our coarse-grained polypeptides undergo a dramatic reduction in chain dimension at a collapse transition with increasing hydrophobicity, as determined by the radius-of-gyration, $\langle R_g^2\rangle^{1/2}$. Representative examples are shown in Figure 1 for 150-mer sequences of varying charge with explicit counteranions. For the uncharged sequence, the polypeptides undergo a conformational collapse at a hydrophobic fraction of $\langle H\rangle \approx 0.3$. With increasing sequence charge, the transition fraction shifts to greater hydrophobicities. The radius-of-gyration is essentially independent of charge for the completely collapsed polypeptides ($\langle H\rangle \rightarrow 1$) for a given sequence length. The expanded conformations ($\langle H\rangle \rightarrow 0$), on the other hand, depend sensitively on sequence charge, swelling with increasing charge. These observations are hallmarks of a coil-to-globule transition in hydrophobicity space. To further confirm this is a coil-to-globule transition, we have analyzed the scaling of the radius-of-gyration with chain length for the uncharged chains (Figure 1 inset). For the case of no hydrophobic monomers, $\langle H\rangle = 0$, the scaling exponent is 0.62, indicative of self-avoiding walk/coil-like conformations.[43,44] Charging the polypeptides results in chain swelling and even larger scaling exponents, although we expect this swelling to go down in the presence of added salt. When the sequences are purely hydrophobic, $\langle H\rangle = 1$, the scaling exponent is 0.32, indicative of compact globules.[43] This globular scaling holds for the collapsed polyelectrolytes as well.

(38) Frenkel, D.; Smit, B. *Understanding molecular simulation: From algorithms to applications*, 2nd ed.; Academic Press: San Diego, 2001.
(39) Hill, T. L. *An introduction to statistical thermodynamics*; Dover Publications: New York, 1960.
(40) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.
(41) Ghosh, T.; Garcia, A. E.; Garde, S. *J. Am. Chem. Soc.* **2001**, *123*, 10997–11003.
(42) Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.
(43) Rubinstein, M.; Colby, R. H. *Polymer physics*; Oxford University Press: New York, 2003.
(44) We note that unfolded proteins display a scaling exponent of ~0.5 experimentally,[12] indicative of random walk statistics rather than self-avoiding walk statistics. We ascribe this difference with our scaling exponent of 0.62 to our simplified treatment of water and the polypeptide. Either way, the interpretation that the polypeptide is an expanded coil stands.

Toward either extreme of the hydrophobicity scale, $\langle R_g^2\rangle^{1/2}$ is approximately linear with $\langle H\rangle$ whether the chains are charged or uncharged, although the lines at either extreme are different (Figure 1). Any empirical relationship that describes the coil dimension with hydrophobicity must capture the transition between the two linear regimes. We propose then that $\langle R_g^2\rangle^{1/2}$ over the range of $\langle H\rangle$ can be fitted to the relationship

$$\langle R_g^2\rangle^{1/2} = (\alpha_{\text{globule}}\langle H\rangle + \beta_{\text{globule}})\, f(\Lambda^{-1}(\langle H\rangle - \langle H\rangle^*)) + (\alpha_{\text{coil}}\langle H\rangle + \beta_{\text{coil}})[1 - f(\Lambda^{-1}(\langle H\rangle - \langle H\rangle^*))] \quad (2)$$

where the coil and globular regions far away from the transition are modeled as linear functions of $\langle H\rangle$, with slopes and intercepts given by $\alpha_{\text{coil/globule}}$ and $\beta_{\text{coil/globule}}$. The transition between these two regions is described by the Fermi function, $f(x) = 1/[1 + \exp(x)]$, where the critical transition fraction is $\langle H\rangle^*$ and the width/cooperativity of the transition zone is Λ. Fits of eq 2 to the simulation results at fixed N and $\langle q\rangle$ are excellent (see Figure 1 for example).

Stability diagrams for the coarse-grained polypeptides with and without explicit counterions are obtained by plotting the fitted values of $\langle H\rangle^*$ versus $\langle q\rangle$ for a range of polymer lengths (Figure 2). As may be intuitively expected, globular states are generally favored by a high mean hydrophobicity and a low mean charge, while coil-like states are conversely favored by lower mean hydrophobicities and a higher mean charge. When counterions are explicitly included in our simulations (Figure 2a), the coil-to-globule stability boundary is only modestly dependent on sequence length, becoming essentially independent of length for sequences of 150 residues and longer. Moreover, for polypeptides 150 residues and longer, $\langle H\rangle^*$ is linearly dependent on $\langle q\rangle$. These observations are in excellent qualitative agreement with the experimental charge/hydropathy conformational stability boundary line[12,24] and support the conclusion that our coarse-grained model captures the essential physics underlying the gross polypeptide dimensions. The minor chain length dependence we observe for our shorter sequences ($N < 150$) potentially contributes to the experimentally observed blurring of the charge/hydropathy predictor near the stability boundary.[11] Winnowing our analysis to only consider longer sequences, which is similar to the inclusion of only unfolded sequences 50 amino acids or longer in the original development of the charge/hydropathy correlation,[24] would yield a sharper boundary. A direct mapping between the experimental stability boundary and that of our coarse-grained polypeptide is uncertain, however, given our minimalistic treatment of the amino acids and water.

The origin of the weak chain length dependence of the stability diagram can be probed by considering the case in which the counterions are lumped into a Debye-screening length (Figure 2b). At low mean charge ($\langle q\rangle \leq 0.1$), the coil-to-globule stability boundaries with and without counterions are essentially identical. The stability boundary frays with increasing $\langle q\rangle$ when counterions are treated implicitly, with longer chains requiring more hydrophobic units to stabilize the globular state. Indeed for the longest ($N = 350$), most highly charged chain ($\langle q\rangle = 0.2$) examined, a stable globule is not even observed for the most hydrophobic chain simulated ($\langle H\rangle = 0.8$). In this case, the transition hydrophobicity recorded in the figure ($\langle H\rangle^* = 0.67 \pm 0.04$) was obtained by linearly extrapolating results for shorter sequences as $1/N$. We anticipate that a collapse transition will not be observed for even longer sequences when explicit counterions are neglected.

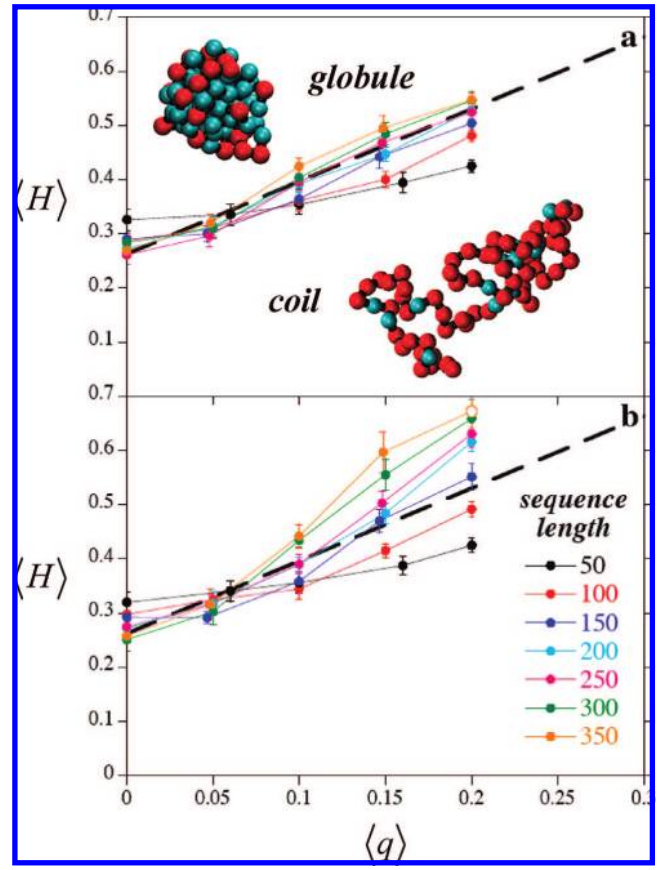


Figure 2. Polypeptide conformational stability diagram in charge/hydropathy space. Diagrams (a) and (b) show results for simulations performed with and without explicit counteranions, respectively. The points indicate results for simulations of chains of differing lengths (see legend). Error bars indicate 1 standard deviation in the transition hydrophobicity. Globules are stable in the region above the lines, while unfolded coil-like states are stable below the boundary. The bold dashed line (same line in (a) and (b)) indicates a linear fit to simulations with explicit counteranions for chain lengths of $N \geq 150$. The open point on (b) for chain length $N = 350$ and charge fraction $\langle q\rangle = 0.2$ indicates the extrapolated transition point determined by extrapolating our results (as N^{-1}) for shorter chain lengths with the same charge fraction.

Globule Structure and Counterion Condensation. The sequence length dependence of the stability boundary can be traced to the charge density of the globular surface. In particular, the volume of a globule is proportional to the number of residues ($V \propto N$) making the surface area proportional to the number of residues to the $^2/_3$'s power ($A \propto N^{2/3}$). At a fixed charge fraction, the net number of cationic charges is equal to $N_+ = \langle q\rangle N$. Given that the charges predominantly reside on the globular surface (see Figure 3b below), it follows that the surface area per unit charge is $\gamma = A/N_+ \propto N^{-1/3}$ and the characteristic separation between charges on the surface ($\gamma = \ell^2$) is $\ell \propto N^{-1/6}$. Therefore, as the sequence length increases for fixed $\langle q\rangle$, the separation between surface cations decreases and electrostatic repulsion increases. The globule thereby becomes progressively unstable, blowing itself apart with increasing N. Explicit counterions, however, can condense on the surface to moderate the repulsion between charges, stabilizing the globule and weakening the dependence of the stability boundary on chain length.

Counterion adsorption can be directly observed from the globule density profiles. In Figure 3, we have plotted the hydrophobic monomer, cationic monomer, and anionic counterion density profiles as a function of the normalized distance

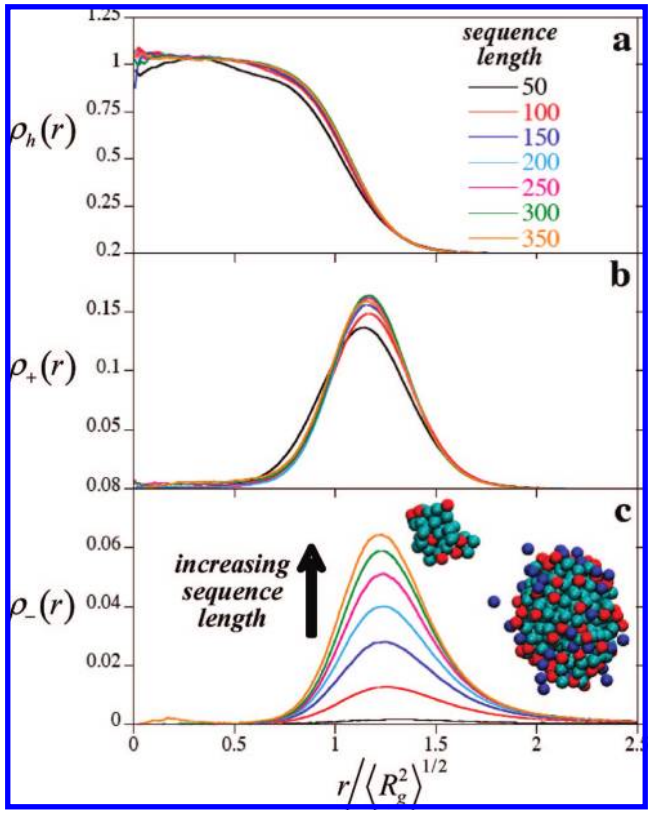


Figure 3. Density of hydrophobic monomers (a), cationic monomers (b), and anionic counterions (c) as a function of distance (normalized by the radius-of-gyration) from the center-of-mass of a collapsed globule with hydrophobic and charge fractions of $\langle H\rangle = 0.8$ and $\langle q\rangle = 0.2$. Results for chains of varying length are indicated by the legend. The inset picture to (c) shows globules 50 and 350 residues in length along with those counteranions within 3σ of the globule surface. Hydrophobic residues, cationic residues, and anionic counterions are colored cyan, red, and blue, respectively.

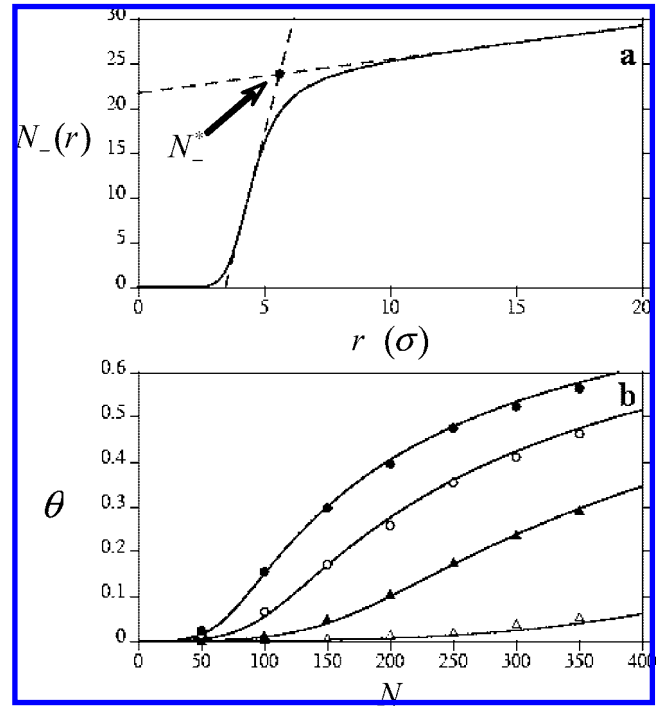


Figure 4. Anion binding to charged globule surfaces. (a) Average number of neighboring counterions, $N_-(r)$, as a function of distance from the center of the 250-mer globule with $\langle H\rangle = 0.8$ and $\langle q\rangle = 0.2$, indicated by the sold line. The adsorbed number of anions, N_-^*, indicated by the solid point, is determined by the intersection of the two dashed lines. The slopes of the dashed lines are evaluated by the first maximum and following minimum in $dN_-(r)/dr = 4\pi r^2\,\rho_-(r)$. (b) Adsorbed fraction of counterions, $\theta = N_-^*/\langle q\rangle N$, as a function of chain length for collapsed globules with a hydrophobicity of $\langle H\rangle = 1 - \langle q\rangle$. The open triangles, solid triangles, open circles, and solid circles indicate simulation results for $\langle q\rangle$ equals 0.05, 0.10, 0.15, and 0.20, respectively. The error bars are smaller than the symbols. The lines indicate the predictions of eq 5 with $R = (^5/_3)^{1/2}\langle R_g^2\rangle^{1/2}$, $\delta = 0.519\sigma$, and $L = 80\sigma$. The radii-of-gyration were assumed to follow the scaling relationship $\langle R_g^2\rangle^{1/2} = \gamma N^{\upsilon}$ with γ and υ equal to 0.55σ and 0.313 for $\langle q\rangle = 0.05$, 0.56σ and 0.312 for $\langle q\rangle = 0.10$, 0.57σ and 0.311 for $\langle q\rangle = 0.15$, and 0.56σ and 0.318 for $\langle q\rangle = 0.20$ as determined from simulation.

from the globule centers-of-mass for the most hydrophobic sequences of the most highly charged polypeptides examined. In general, the hydrophobic monomers aggregate at the center of the globule, stabilizing the compact structure, with quantitatively similar density profiles for all chain lengths (Figure 3a). As noted above, the cationic residues primarily reside on the globule surface, minimizing their repulsive interactions with one another (Figure 3b). As with the hydrophobic residues, the density profiles of the cationic monomers are quantitatively similar for all chain lengths. There is no agreement among the counteranion density profiles, however (Figure 3c). Notably, the density of counterions in contact with the globule surface is a strongly increasing function of sequence length, indicative of condensation.[45–47] The preference of the anions for the globule surface can be visualized in simulation snapshots, showing no anions in the vicinity of the 50-mer globule while a cloud of counterions enshrouds the 350-mer (inset picture to Figure 3c).

The degree of counterion adsorption can be quantified by integrating the counterion density profiles. In particular, the number of counterions residing within a distance r of the center-of-mass of a globule is

$$N_-(r) = \int_0^r \rho_-(\zeta)\, 4\pi\zeta^2\, d\zeta \qquad (3)$$

In general, $N_-(r)$ undergoes a dramatic increase near the globule surface, followed by a gentler increase at further separations (Figure 4a). The knee in $N_-(r)$ between these two regimes is taken to describe the boundary between adsorbed and bulk solution counterions. The number of adsorbed counterions, N_-^*, is evaluated by linearly extrapolating between the adsorbed and free solution regimes to determine their intersection (Figure 4a, see figure caption for description). The fraction of bound counterions, $\theta = N_-^*/\langle q\rangle N$, so determined as a function of N and $\langle q\rangle$ for globules with the highest hydrophobicity at a given charge is shown in Figure 4b. In keeping with the proposed counterion adsorption mechanism, the fraction of bound counterions is an increasing function of N and $\langle q\rangle$. Comparing the binding fractions in Figures 2 and 4b, we observe the differences between the simulations with and without explicit counterions increase as the degree of counterion binding increases. This observation suggests counterion condensation plays little role in globular stability at low binding fractions, but the moderating role of the anions is necessary to observe the sequence length independence of the stability boundary.

(45) Manning, G. S. *J. Chem. Phys.* **1969**, *51*, 924–933.
(46) Manning, G. S. *J. Phys. Chem. B* **2007**, *111*, 8554–8559.
(47) Manning, G. S. *Macromolecules* **2007**, *40*, 8071–8081.

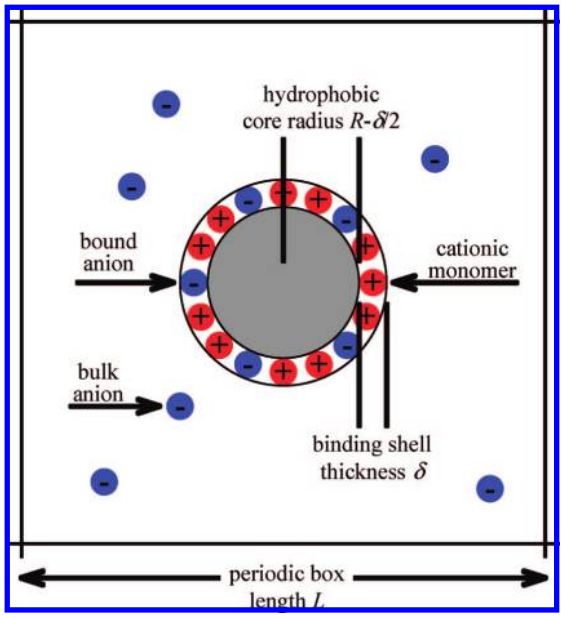


Figure 5. Representation of the system described by the anion binding model. This model is inspired by the monomer and counterion distributions shown in Figure 3. The polypeptide globule resides within a periodic simulation box of length L on a side. The hydrophobic core of the globule extends from the center of the globule to $R - \delta/2$. Charged moieties are assumed to be excluded from this central region. The cationic monomers are confined within the binding shell from $R - \delta/2$ to $R + \delta/2$. The space beyond $R + \delta/2$ is considered the bulk solution. Counteranions are assumed to be able to sample the binding shell and bulk solution.

Counterion Condensation Model. Following the physical picture depicted in Figure 5, we have developed a continuum dielectric anion binding model. Here we outline the binding free energies. More details for the derivation of this model are provided in the appendix (Supporting Information).

The degree of counterion adsorption can be modeled by assuming the bound anions and cationic residues reside within a shell of thickness δ centered on a spherical surface R defining the globule, in line with the density profiles in Figure 3. The anions uniformly explore the surface condensation shell ($R - \delta/2 < r < R + \delta/2$) or the bulk solution volume beyond the globule ($r > R + \delta/2$), while the cationic residues are confined to the shell. The electrostatic free energy associated with binding anions to the cation shell to reduce the mean net charge to $(1 - \theta)\langle q\rangle Ne$, the sum of the anionic and cationic charges, in a continuum dielectric is

$$\Delta G_{\mathrm{elec}}(\theta) = (\theta^2 - 2\theta)\frac{\langle q\rangle^2 N^2 e^2}{8\pi\varepsilon_0 DR}\frac{(1 - \delta/6R + \delta^2/12R^2 + \delta^3/120R^3)}{(1 + \delta^2/6R^2 + \delta^4/144R^4)} \quad (4a)$$

This contribution is increasingly favorable for the binding of anions to the globule. The translational free energy of the anions is

$$\Delta G_{\mathrm{trans}}(\theta) = k_{\mathrm{B}}T(1 - \theta)\langle q\rangle N \ln(1 - \theta) + k_{\mathrm{B}}T\theta\langle q\rangle N \ln\left\{\frac{\theta[L^3 - 4\pi(R + \delta/2)^3/3]}{4\pi R^2\delta + \pi\delta^3/6}\right\} \quad (4b)$$

The first and second terms in this expression correspond to the bulk and adsorbed anions, respectively. This contribution can be either favorable or unfavorable for anion binding and ultimately determines the extent of anion binding. The above equations assume no added salt, but they can be readily modified to include the screening and translational contributions of added electrolyte.[27,28] Minimizing the total free energy ($\Delta G_{\mathrm{bind}}(\theta) = \Delta G_{\mathrm{trans}}(\theta) + \Delta G_{\mathrm{elec}}(\theta)$) with respect to the binding fraction ($\partial\Delta G_{\mathrm{bind}}(\theta)/\partial\theta = 0$) yields

$$(1 - \theta)\frac{\langle q\rangle Ne^2}{4\pi\varepsilon_0 DkTR} = \ln\left\{\frac{[L^3 - 4\pi(R + \delta/2)^3/3]\theta}{(4\pi R^2\delta + \pi\delta^3/6)(1 - \theta)}\right\} \quad (5)$$

Assuming spherical geometry, the globular surface is at $R = (^5/_3)^{1/2}\langle R_g^2\rangle^{1/2}$, which is in excellent agreement with the maxima in the cationic residue and anion densities (Figure 3b and c). The shell thickness is then left as a single adjustable parameter. An optimal quantitative description of counterion adsorption is obtained for $\delta = 0.519\sigma$ (Figure 4b). This shell thickness compares well with the width of the cationic monomer shell densities (as determined by the second moment of the cationic residue distributions), which range from 0.47σ to 0.93σ for sequences from 50 to 350 residues in length. Varying δ over this measured range of cation distribution widths leads to only a mild perturbation in the model predictions. These observations thereby support the physical picture encapsulated by eq 5.

In contrast to condensation at a cylindrical or planar interface, adsorption on a sphere is diminished with counterion dilution.[46] This is captured by the incorporation of the simulation volume in the translational free energy (L^3 in eq 4b). As a result, the predicted stability boundary will fray with increasing simulation box size, tending toward the observed chain length dependent boundary when counterions are excluded (Figure 2b). To this end, we have performed test simulations for 300-mer polypeptides, $\langle q\rangle = 0.2$, and explicit counterions in boxes with dimensions of 100σ and 120σ. While the radii-of-gyration of the coil-like states determined from these simulations show systematic increases with box size, the fitted $\langle H\rangle^*$ are the same as that reported in Figure 2a within the simulation error. We conclude that the stability boundary is insensitive to dilution over the range studied. Moreover, this box size condensation dependence would be mitigated by added electolyte, as is the experimental case with pH regulating buffers and salts.

Conclusions

The incidence of natively unfolded protein sequences has been empirically correlated with a wide range of predictors, each providing distinct physical pictures of the factors governing the polypeptide dimensions. These competing views, however, can lead to alternative criteria for identifying and interpreting the functions of unfolded proteins. A physics based approach for recognizing unfolded sequences should provide a more fundamental route for judging the merits of these diverse disorder predictors. We have presented extensive molecular dynamics simulation results for the conformational equilibrium of coarse-grained polypeptides as a function of sequence hydrophobicity, charge, and length. Our model polypeptides display a coil-to-globule transition with varying sequence hydrophobicity and charge comparable to the empirical charge/hydropathy predictor,[12,24] demonstrating that a minimal, polymer physics based model can capture the elements of gross protein conformation. When counterions are only treated implicitly in our calculations the boundary between coil-like and globular states was found to be sequence length dependent as a result of destabilizing repulsions between surface cationic residues. Explicit inclusion of counterions, however, permits surface adsorption to moderate these repulsive interactions, thereby leading to a near "universal" stability boundary in agreement with the empirical charge/hydropathy correlation. Counterion adsorption is quantitatively

described by an analytical condensation model. We anticipate that further salt addition, as in the ubiquitous use of buffers, will dull the impact of electrostatic interactions and sharpen the coil-to-globule transition boundary.

The success of our polypeptide model at describing the interplay between charge and hydrophobicity in determining the boundary between compact and expanded protein conformations does not rule out contributions from factors like sequence complexity and chain flexibility. Additional functionalities, like anionic residues, bond stiffness, and torsional constraints, may be readily incorporated into our model. Furthermore, the sequences considered here are random, encoding no information regarding preferences for the coil or globule states. Sequence conformational specificity and nonrandomness, for instance, could be incorporated into this model following Khoklov's globule sequence templating procedure.[48] The present coarse-grained model can thereby serve as a base case, to which additional factors can be systematically added and isolated, to gain deeper insight into cross interactions between distinct unfolded protein predictors and provide a physical basis for empirical sequence-based correlations.

Acknowledgment. H.S.A. would like to thank Professors Michael Paulaitis, Tom Woolf, Dilip Asthagiri, Georges Belfort, and Sanat Kumar for helpful conversations regarding natively unfolded proteins. We also thank the Tulane Provost's Fund for Undergraduate/Faculty Engagement for supporting H.W.H.

Supporting Information Available: Appendix. This material is available free of charge via the Internet at http://pubs.acs.org.

JA802124E

(48) Khokhlov, A. R.; Khalatur, P. G. *Phys. Rev. Lett.* **1999**, *82*, 3456–3459.