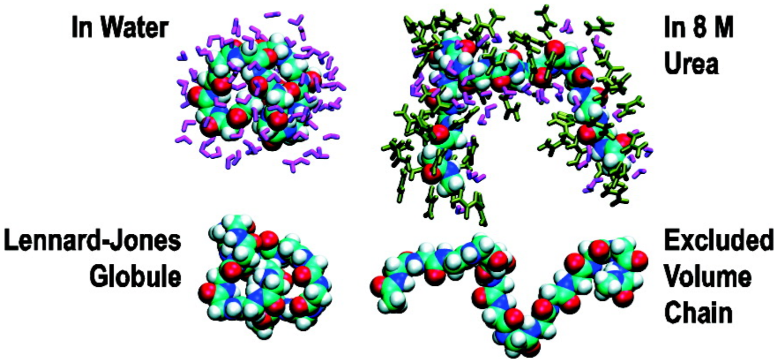
Article

# Role of Backbone#Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins

Hoang T. Tran, Albert Mao, and Rohit V. Pappu

*J. Am. Chem. Soc.*, **2008**, 130 (23), 7380-7392 • DOI: 10.1021/ja710446s • Publication Date (Web): 16 May 2008

**Downloaded from http://pubs.acs.org on February 8, 2009**



Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

J|A|C|S
ARTICLES

Published on Web 05/16/2008

# Role of Backbone−Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins

Hoang T. Tran, Albert Mao, and Rohit V. Pappu*

*Department of Biomedical Engineering and Center for Computational Biology, Washington University in St. Louis, Campus Box 1097, St. Louis, Missouri 63130*

Received November 19, 2007; E-mail: pappu@wustl.edu

***Abstract:*** Intrinsically disordered proteins (IDPs) are functional proteins that do not fold into well-defined three-dimensional structures under physiological conditions. IDP sequences have low hydrophobicity, and hence, recent experiments have focused on quantitative studies of conformational ensembles of archetypal IDP sequences such as polyglutamine and glycine-serine block copolypeptides. Results from these experiments show that, despite the absence of hydrophobic residues, polar IDPs prefer ensembles of collapsed structures in aqueous milieus. Do these preferences originate in interactions that are unique to polar sidechains? The current study addresses this issue by analyzing conformational equilibria for polyglycine and a glycine-serine block copolypeptide in two environments, namely, water and 8 M urea. Polyglycine, a poly secondary-amide, has no sidechains and is a useful model system for generic polypeptide backbones. Results based on large-scale molecular dynamics simulations show that polyglycine forms compact, albeit disordered, globules in water and swollen, disordered coils in 8 M urea. There is minimal overlap between conformational ensembles in the two environments. Analysis of order parameters derived from theories for flexible polymers show that water at ambient temperatures is a poor solvent for generic polypeptide backbones. Therefore, the experimentally observed preferences for polyglutamine and glycine-serine block copolypeptides must originate, at least partially, in polypeptide backbones. A preliminary analysis of the driving forces that lead to distinct conformational preferences for polyglycine in two different environments is presented. Implications for describing conformational ensembles of generic IDP sequences are also discussed.

## 1. Introduction

Intrinsically disordered proteins (IDPs) are functional proteins that do not fold into well-defined, unique three-dimensional structures under physiological conditions (Dunker et al.).[1] IDPs are ubiquitous *in vivo*, and their intrinsic disorder is implicated in a range of regulatory functions, such as signaling, molecular switching, protein trafficking, and protein turnover.[2–7] To answer the question of how disorder is used in function, we need accurate models to describe conformational ensembles of IDPs. Typical IDP sequences have a combination of low overall hydrophobicity, high mean net charge,[8] and in some cases, low sequence complexity.[9,10] Uversky et al.[8] argued that low overall hydrophobicity of IDPs must imply the lack of a driving force for formation of ensembles with compact structures. Recent spectroscopic studies have focused on characterizing conformational ensembles for sequences such as polyglutamine[11] and glycine-serine block copolypeptides,[12] which are archetypal IDPs in that they are devoid of hydrophobic residues and have low sequence complexity. These experiments show that polyglutamine and glycine-serine block copolypeptides prefer to form collapsed structures in aqueous solutions. Mukhopadhyay et al.[13] obtained similar results for the glutamine/asparagine rich N-terminal domain of the yeast prion protein Sup35. These results[11–13] are surprising given the lack of hydrophobic residues in the sequences studied. It is conceivable that the experimental observations reflect unique preferences of polar sidechains and are not generalizable to generic IDP sequences. The present work probes this issue by investigating conformational equilibria for polypeptide backbones in two different solvent environments namely, water and 8 M urea at 298 K. Specifically, we ask if the preference for collapsed states observed for aqueous solutions of polyglutamine,[11] glycine-serine block copolypeptides,[12] and the N-domain of Sup35[13] is reproducible for generic polypeptide backbones or if it arises from interactions unique to the presence of polar sidechains such as side chain-backbone hydrogen bonding.

[1] Fink, A. L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.
[2] Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321–331.
[3] Dunker, A. K. *J. Mol. Graph. Model.* **2001**, *19*, 26–59.
[4] Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
[5] Dunker, A. K.; Brown, C. J.; Obradovic, Z. *Adv. Protein Chem.* **2002**, *62*, 25–49.
[6] Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
[7] Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
[8] Uversky, V. N.; Gillespie, J. R.; Fink, A. L. *Proteins* **2000**, *41*, 415–427.
[9] Sim, K. L.; Creamer, T. P. *Mol. Cell. Proteomics* **2002**, *1*, 983–995.
[10] Weathers, E. A.; Paulaitis, M. E.; Woolf, T. B.; Hoh, J. H. *Proteins* **2007**, *66*, 16–28.
[11] Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16764–16769.
[12] Moglich, A.; Joder, K.; Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12394–12399.
[13] Mukhopadhyay, S.; Krishnan, R.; Lemke, E. A.; Lindquist, S.; Deniz, A. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2649–2654.

**10.1021/ja710446s CCC: $40.75**

We use polyglycine as a model system for generic polypeptide backbones, or more precisely, as a model system for poly secondary-amides. The motivation comes from the transfer free energy model,[14,15] which has been important for understanding the classical hydrophobic effect and driving forces for collapse transitions in proteins.[16–19] The free energy of hydration for *N*-methylacetamide (NMA) at 298 K is ca. −10 kcal/mol,[20] indicating that the transfer of NMA from the gas phase into water is highly favorable. NMA is a secondary amide and a model compound analog of the peptide unit. Extrapolation from the transfer free energy model suggests that polyglycine, which is a poly secondary-amide, should prefer structures that maximize the interface with the aqueous solvent, i.e., water should be a good solvent for generic polypeptide backbones. In a good solvent, chains prefer interactions with the surrounding solvent and mixing occurs between the chain and solvent on all length scales. As a result, ensemble averaged radii of gyration ($R_g$) scale as $N^{0.59}$ with chain length ($N$).[21] Conversely, in a poor solvent, an ensemble of compact conformations is preferred to minimize interactions between chain and solvent.[21] If polyglycine prefers collapsed structures in water, then we can conclude that water is a poor solvent for this system of molecules and that the driving force for collapse in polyglutamine and in glycine-serine block copolypeptides originates, at least partially, from the intrinsic tendencies of polypeptide backbones in water.

## 2. Materials and Methods

We present results from molecular dynamics and Monte Carlo simulations for the sequences Ac-(Gly)-Nme, Ac-$(\text{Gly})_{15}$-Nme, and Ac-$(\text{Gly-Ser})_8$-Nme, respectively. Here, Ac denotes the acetyl group and Nme stands for *N*-methylamide. For brevity, we refer to the three molecules as $G_1$, $G_{15}$, and $(GS)_8$, respectively.

**Design of Molecular Dynamics Simulations.** We report results from nine independent sets of molecular dynamics simulations. For seven of the nine sets of simulations, the peptides were modeled using the all-atom OPLS-AA/L force field;[22] the 3-site TIP3P model[23] was used for water molecules, and the OPLS-AA force field[24] was used to model urea molecules.[25] Two sets of simulations were performed to assess the dependence of our results on the choice of force field; these simulations were carried out using the GROMOS 53A6[26] force field and the SPC water model.[27]

Hu et al.[28] analyzed conformational equilibria for glycine and alanine dipeptides using a hybrid quantum mechanics/molecular mechanics (QM/MM) approach. They modeled intrapeptide interactions using the self-consistent charge density functional tight binding method, whereas peptide-solvent and solvent-solvent interactions were described using standard molecular mechanics models. They compared their results to those obtained using a range of molecular mechanics forcefields, including OPLS-AA,[24] and none of these agreed with the conformational distributions calculated using the QM/MM approach. They also noted that conformational distributions calculated with different molecular mechanics forcefields did not agree with each other. The findings of Hu et al. cause concern and make it necessary that we justify our choice of forcefields for the simulation results reported in this work. The justifications are as follows: First, the OPLS-AA/L forcefield used here is an improved version of the OPLS-AA forcefield used by Hu et al.[28] Specifically, Kaminski et al.[22] refined the backbone torsional parameters to achieve very good agreement with the gas phase potential energy surfaces calculated using high-level quantum mechanical methods. These refinements yield conformational distributions that are in line with those obtained by Hu et al. for alanine and glycine dipeptides (data not shown). Second, the specific question we are trying to answer is relatively insensitive to the details of differences in conformational distributions at the level of individual residues. We are interested in global properties of polyglycine, and we will show (see Results section) that the qualitative results we obtain are robust and invariant to the choice of forcefield we use.

For simulations with 8 M urea, the cosolutes were built using the OPLS atom type definitions for urea C, O, N, and H atoms as defined in the GROMACS OPLS-AA force field definition file, and the corresponding atomic sizes, atomic charges, and bond stretching, angle bending, and torsional parameters were used. Our choice of the OPLS-AA force field for urea requires discussion. Recently, Kokubo and Pettitt[29] carried out a comparative analysis of the Kirkwood-Buff force field developed by Weerasinghe and Smith[30] to the OPLS-AA force field for urea. They concluded that the parameters of Weerasinghe and Smith provide consistent agreement with experimental data for density and diffusion coefficients in urea-water mixtures. The OPLS-AA parameters were also found to be reasonable, although not as accurate as the Kirkwood-Buff force field. In this work, we used the OPLS-AA parameters for urea to maintain consistency with the force field paradigm used for water molecules and peptides.

All molecular dynamics simulations were performed using version 3.3.1 of the GROMACS simulation package.[31] Cubic boxes with periodic boundary conditions were used. The equations of motion were integrated using the leapfrog method and a time step of 1 fs. The two bond lengths and one bond angle in each water molecule were constrained to values prescribed by the TIP3P model using the SETTLE algorithm of Miyamoto and Kollman.[32] For nonbonded interactions, we employed 10 Å spherical cutoffs for van der Waals and for short-range Coulomb interactions. Long-range Coulomb interactions (10−14 Å) were recalculated every 10 steps, as were neighbor lists. The reaction field method[33] was used as a correction term for polar interactions beyond 14 Å. In all of our simulations, the peptides are concatenations of electro-neutral groups. Similarly, water and urea molecules are also electro-neutral. In such systems, long-range electrostatic interactions reduce to dipole−dipole interactions, which are both convergent and decay more rapidly than charge−charge interactions. This is the justification for our use of the reaction field as opposed to Ewald sums for treatment of long-range corrections. Our choice is unlikely to have artifacts, and we gain in computational efficiency.[34]

[14] Bolen, D. W. *Methods* **2004**, *34*, 312–322.
[15] Auton, M.; Bolen, D. W. *Biochemistry* **2004**, *43*, 1329–1342.
[16] Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.
[17] Alonso, D. O. V.; Dill, K. A. *Biochemistry* **1991**, *30*, 5974–5985.
[18] Pace, C. N. *J. Mol. Biol.* **1992**, *226*, 29–35.
[19] Xu, J. A.; Baase, W. A.; Baldwin, E.; Matthews, B. W. *Protein Sci.* **1998**, *7*, 158–177.
[20] Wolfenden, R. *Biochemistry* **1978**, *17*, 201–204.
[21] Rubinstein, M.; Colby, R. H. *Polymer Physics*; Oxford University Press: New York, 2003.
[22] Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.
[23] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
[24] Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.
[25] Duffy, E. M.; Severance, D. L.; Jorgensen, W. L. *Isr. J. Chem.* **1993**, *33*, 323–330.
[26] Chris Oostenbrink; Alessandr Villa Alan, E.; Mark; Gunsteren, W. F. V. *J. Comput. Chem.* **2004**, *25*, 1656–1676.
[27] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren; W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, Holland, 1981; p 331.
[28] Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct. Funct. Genet.* **2003**, *50*, 451–463.
[29] Kokubo, H.; Pettitt, B. M. *J. Phys. Chem. B* **2007**, *111*, 5233–5242.
[30] Weerasinghe, S.; Smith, P. E. *J. Phys. Chem. B* **2003**, *107*, 3891–3898.
[31] Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
[32] Miyamoto, S.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 952–962.
[33] Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486–1493.

We used the isothermal−isobaric ensemble for all molecular dynamics simulations. The Berendsen thermostat[35] (time constant 0.1 ps) and manostat[35] (time constant 1 ps and compressibility of $4.5 \times 10^{-5}$ bar$^{-1}$) were used to maintain a temperature of 298 K and a pressure of 1 atm, respectively. In simulations with 8 M urea, the ratio of the number of water molecules to the number of urea molecules was held fixed at 4.5:1, to simulate concentrations of 8 M.

One of our objectives was to construct potentials of mean force (PMFs) as a function of radius of gyration ($R_g$) for both $G_{15}$ and $(GS)_8$ in water and 8 M urea, respectively. To achieve this goal, we combined molecular dynamics simulations with umbrella sampling.[36] To analyze conformational distributions in terms of parameters other than $R_g$, we used multiple replica molecular dynamics (MRMD) simulations[37] for $G_{15}$ in both water and 8 M urea, respectively.

**Umbrella Sampling.** The goal was to construct PMFs, $W(R_g)$, for $G_{15}$ and $(GS)_8$ in both water and 8 M urea. For each sequence in a given environment, we carried out 11 independent simulations and in each simulation a harmonic potential of the form $U_{\rm rest} = (k/2)(R_g - R_g^0)^2$ was applied to restrain the radius of gyration to a target value of $R_g^0$. The force constant $k$ was set to be 1000 kJ/nm$^2$, or 2.39 kcal/Å$^2$, whereas the values for $R_g^0$ were 5−15 Å, in increments of 1 Å. The values chosen for $R_g^0$ cover the range of plausible values for $R_g$ for $G_{15}$ and $(GS)_8$ and allows overlap of $R_g$ distributions between adjacent windows. The initial peptide conformations for a given simulation were chosen from a random self-avoiding distribution (see below) such that the starting conformation had the same $R_g$ value as the target $R_g^0$. The peptide was then soaked in a pre-equilibrated box of water or 8 M urea. This was followed by steepest-descent energy minimization to remove steric clashes and equilibration runs for 10 ns in the isothermal−isobaric ensemble. Finally, for each restraint value, we carried out production simulations, each of length 50 ns.

Data from the restrained simulations were analyzed using the weighted histogram analysis method (WHAM)[38–40] as implemented by Grossfield[41] to calculate the desired PMFs. Figure 1 shows the $R_g$ distributions obtained for $G_{15}$ in water and illustrates the overlap between adjacent windows. This overlap is necessary because it improves the accuracy and convergence properties of WHAM.

In the interest of computational efficiency, we used smaller box sizes containing fewer solvent molecules for smaller target $R_g^0$ values, and conversely larger box sizes and increased numbers of solvent molecules for larger target $R_g^0$ values. In all cases, the box sizes were large enough to ensure that the minimum distance of approach between periodic images did not fall below the 14 Å threshold used as cut offs for long-range Coulomb interactions. A similar strategy was used recently by Athawale et al.[42] to construct PMFs for different types of hydrophobic polymers in water.[42] Tables 1–3 provide a detailed inventory of the number of solvent molecules and average box sizes for each of the restrained simulations for $G_{15}$ and $(GS)_8$ in both water and 8 M urea, respectively.

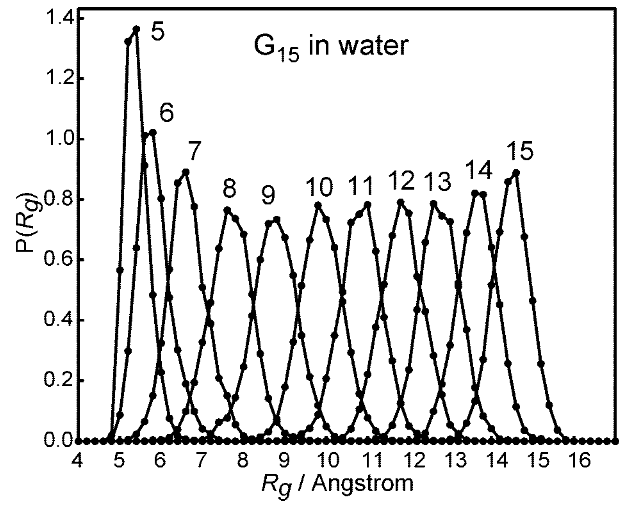


***Figure 1.*** $R_g$ distributions for G15 in water with 11 different harmonic restraints. Each curve is labeled with the target $R_g$ (in Angstroms), $R_g^0$, used to restrain the peptide chain.

**MRMD Simulations.** To analyze properties other than PMFs as a function of $R_g$, we carried out multiple, independent molecular dynamics simulations for $G_{15}$ in both water and 8 M urea. The simulation parameters for treatment of nonbonded interactions, choice of integrator, thermostat, manostat, and the use of constraints are as described above. Other details were altered as follows: The peptide, $G_{15}$, was soaked in a bath of either 6000 TIP3P water molecules or 5535 water and 1230 urea molecules. The average box sizes for $G_{15}$ in water and 8 M urea were 57 and 63 Å, respectively. Boxes for individual simulations were prepared by soaking a random, self-avoiding peptide conformation, followed by adding or deleting water molecules such that we ended up with the same number of water molecules for all replicas. In each case, steepest-descent minimization to remove steric clashes was followed by an equilibration run of 10 ns in the isothermal−isobaric ensemble ($T$ = 298 K, $P$ = 1 atm). We used the final configuration of the latter as the starting point for production runs. We carried out 10 independent simulations for $G_{15}$ in both water and 8 M urea, respectively, and the total simulation time in each of the 20 simulations was 100 ns, which includes the 10 ns of equilibration. Therefore, the cumulative simulation time, including equilibration, for $G_{15}$ in each environment was 1 μs.

**Simulations for $G_1$.** We also performed molecular dynamics simulations for a capped glycine residue, *N*-acetyl−glycine-*N*-methylamide, or $G_1$. The simulations parameters were as described for $G_{15}$. For $G_1$ in water, we used 887 water molecules and for $G_1$ in 8 M urea, we used 540 water molecules and 120 urea molecules. For both systems, the average box size was ca. 30 Å.

**Monte Carlo Simulations for Describing Reference Conformational Equilibria.** Conformational equilibria of polymers in generic good and poor solvents can be simulated using implicit solvent models.[43,44] Reference conformational equilibria for chains in good solvents can be obtained using a purely repulsive potential of the form:[37,45]

$$U_{\rm EV} = 4\sum_{i}\sum_{j<i}\varepsilon_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \tag{1}$$

Equation 1 represents the excluded volume (EV) limit, in which the only interactions are steric repulsions. Conformations generated in the EV limit represent conformations accessible in athermal/ideal good solvents. Chains in generic poor solvents can be modeled using a Lennard-Jones potential, which promotes the formation of

(34) Mountain, R. D.; Thirumalai, D. *J. Am. Chem. Soc.* **2003**, *125*, 1950–1957.

(35) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(36) Leach, A. R. *Molecular Modelling: Principles and Applications*, 2nd ed.; Prentice Hall: Edinburgh Gate, 2001.

(37) Vitalis, A.; Wang, X.; Pappu, R. V. *Biophys. J.* **2007**, *93*, 1923–1937.

(38) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195–1198.

(39) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(40) Roux, B. *Comput. Phys. Commun.* **1995**, *91*, 275.

(41) Grossfield, A. http://dasher.wustl.edu/alan/wham/index.html, 2003.

(42) Athawale, M. V.; Goel, G.; Ghosh, T.; Truskett, T. M.; Garde, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 733–738.

(43) Reddy, G.; Yethiraj, A. *Macromolecules* **2006**, *39*, 8536–8542.

(44) Steinhauser, M. O. *J. Chem. Phys.* **2005**, *122*, 094901.

(45) Tran, H. T.; Pappu, R. V. *Biophys. J.* **2006**, *91*, 1868–1886.

***Table 1.*** Number of Water Molecules and Average Box-Length for Restrained Simulations of $G_{15}$ and $(GS)_8$ in Water

| $R_g^0$ (Å) | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of water molecules | 2700 | 2700 | 2900 | 3300 | 3800 | 4300 | 5100 | 6000 | 7000 | 8200 | 9400 |
| Average box-length (Å) | 44 | 44 | 45 | 47 | 49 | 51 | 54 | 57 | 60 | 63 | 66 |

***Table 2.*** Number of Water Molecules, Number of Urea Molecules, and Average Box-Length for Restrained Simulations of $G_{15}$ in 8 M Urea

| $R_g^0$ (Å) | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of water molecules | 1872 | 1899 | 2007 | 2286 | 2583 | 2907 | 3483 | 4329 | 4761 | 5535 | 6372 |
| No. of urea molecules | 416 | 422 | 446 | 508 | 574 | 646 | 774 | 962 | 1058 | 1230 | 1416 |
| Average box-length (Å) | 44 | 45 | 45 | 47 | 49 | 51 | 54 | 58 | 60 | 63 | 66 |

***Table 3.*** Number of Water Molecules, Number of Urea Molecules, and Average Box-Length for Restrained Simulations of $(GS)_8$ in 8 M Urea

| $R_g^0$ (Å) | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of water molecules | 1863 | 1863 | 2016 | 2259 | 2592 | 2871 | 3483 | 4293 | 4734 | 5517 | 6381 |
| No. of urea molecules | 414 | 414 | 448 | 502 | 576 | 638 | 774 | 954 | 1052 | 1226 | 1418 |
| Average box-length (Å) | 44 | 44 | 45 | 47 | 49 | 51 | 54 | 57 | 60 | 63 | 66 |

nonspecific globules. The potential function used to simulate reference conformational equilibria in generic poor solvents was [43,44]

$$U_{\mathrm{LJ}}=4\sum_{i}\sum_{j<i}\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (2)$$

In eqs 1 and 2, $r_{ij}$ denotes the distances between nonbonded atoms $i$ and $j$; $\sigma_{ij}$ are contact distances, and $\varepsilon_{ij}$ are the Lennard-Jones dispersion parameters. The parameters for $\sigma_{ij}$ and $\varepsilon_{ij}$ are those used in previous work.[37,46] Metropolis Monte Carlo simulations were performed as described previously[45,46] to obtain reference ensembles for chains in generic good and poor solvents. These simulations were carried out for G15 and $(GS)_8$, respectively. Ensembles from simulations of $G_{15}$ and $(GS)_8$ in water and in 8 M urea were compared to two sets of reference ensembles using methods derived from polymer theory as described in previous work [37]

## 3. Results

**Justification for the Choices of Chain Lengths Studied.** Separation of length scales is an important hallmark of polymer solutions, and the concept of "blobs" is of particular importance.[21,47] A blob is the length scale beyond which the balance of chain−chain, chain−solvent, and solvent−solvent interactions is at least of order $k_BT$, where $k_B$ is the Boltzmann constant and $T$ is the temperature. For chains longer than blob lengths, solvent quality dictates the types of conformations i.e., the average spatial arrangement of blobs around each other. In a good solvent, the balance of interactions between blobs is net repulsive and chains swell to accommodate favorable contacts with the surrounding solvent. In poor solvents, the balance of interactions between blobs is net attractive and an ensemble of compact, globular conformations is preferred. In contrast, within blob-sized chain segments, the balance of interactions is smaller than $k_BT$, and concepts of solvent quality do not apply for describing conformations within blobs. [21]

We selected chain lengths for our simulations to ensure that the chains were long enough to allow us to discern preferences for collapsed versus swollen conformations and make adjudications regarding solvent quality. This requires the presence of multiple blobs within each chain. In homopolymers and block copolymers, the number of residues within a blob varies with sequence composition and can be estimated by calculating the length scale over which spatial correlations in the chain decay.[45] Let $\mathbf{l}_i$ be the vector of length $l$ from the peptide unit nitrogen of residue $i$ to the carbonyl carbon on the same residue; $\mathbf{l}_j$ is defined in a similar manner for residue $j$. The ensemble-averaged cosine of the angle $\Theta_{ij}$ between the vectors of residue $i$ and $j$ is $\langle|\cos \Theta_{ij}|\rangle = \langle|\mathbf{l}_i\cdot\mathbf{l}_j/l|\rangle$.

An estimate of the blob length for specific sequence constructs is obtained by quantifying the sequence separation, $|i - j|$, at which $\langle|\cos\Theta_{ij}|\rangle$ decays to $1/e$ in the excluded volume limit (i.e., using ensembles generated from Monte Carlo simulations based on the excluded volume potential shown in equation 1). For polyglycine and glycine-serine block copolypeptides, we estimated the blob length to be 2−3 residues as shown in Figure 2. Therefore, the sequences used in our molecular dynamics simulations have ca. 5 blobs for both sequence constructs, $G_{15}$ and $(GS)_8$, respectively.

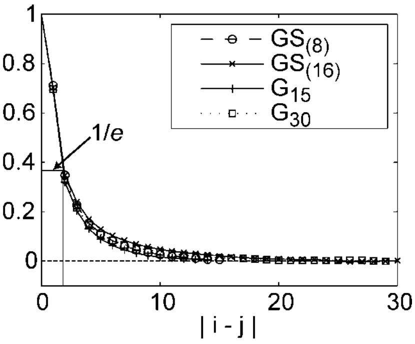


***Figure 2.*** Decay of spatial correlations, $\langle\cos \Theta_{ij}\rangle$, as a function of sequence separation $|i - j|$, for four peptide chains in the EV limit.

**Conformational Equilibria for $(GS)_8$ in Water versus 8 M Urea.** Möglich et al.[12] used time-dependent Förster resonance energy transfer experiments to show that a 32-residue glycine-serine block copolypeptide forms collapsed structures in water and expands significantly in 8 M GdnCl, which is a strong denaturant for polypeptides akin to 8 M urea. We expect that the polymeric properties, namely, the preference for collapsed versus swollen states, of $(GS)_{16}$ studied by Möglich et al. will be equivalent to the polymeric properties of $(GS)_8$. Figure 3 shows PMFs, $W(R_g)$, for $(GS)_8$ in water versus 8 M urea. The

[46] Tran, H. T.; Wang, X.; Pappu, R. V. *Biochemistry* **2005**, *44*, 11369–11380.

[47] Grosberg, A. Y.; Khokhlov, A. R. *Statistical Physics of Macromolecules*; AIP Press: New York, 1994.

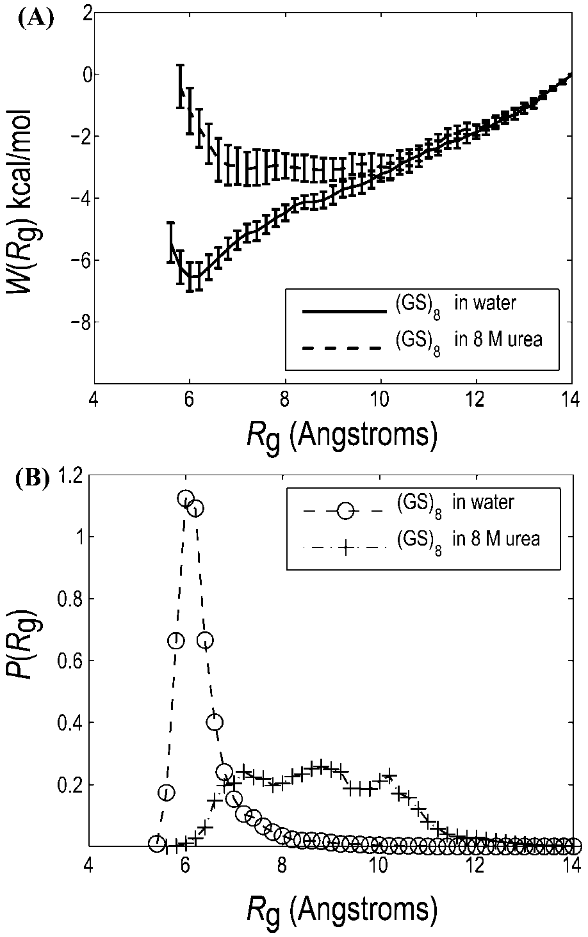


***Figure 3.*** (A) PMFs, $W(R_g)$, for $(GS)_8$ in water versus 8 M urea. Error bars are standard errors and were calculated using block-averaging with 10 blocks per simulation window. PMFs were arbitrarily shifted to be zero at a reference value of $R_g = 14$ Å. (B) Corresponding probability densities for $(GS)_8$ were calculated as $P(R_g) = A_0 \exp[-W(R_g)/RT]$, where $R = 1.98 \times 10^{-3}$ kcal/mol-K is the molar gas constant. $A_0$ was calculated such that $\sum P(R_g)\Delta R_g = 1$ where $\Delta R_g = 0.2$ Å.

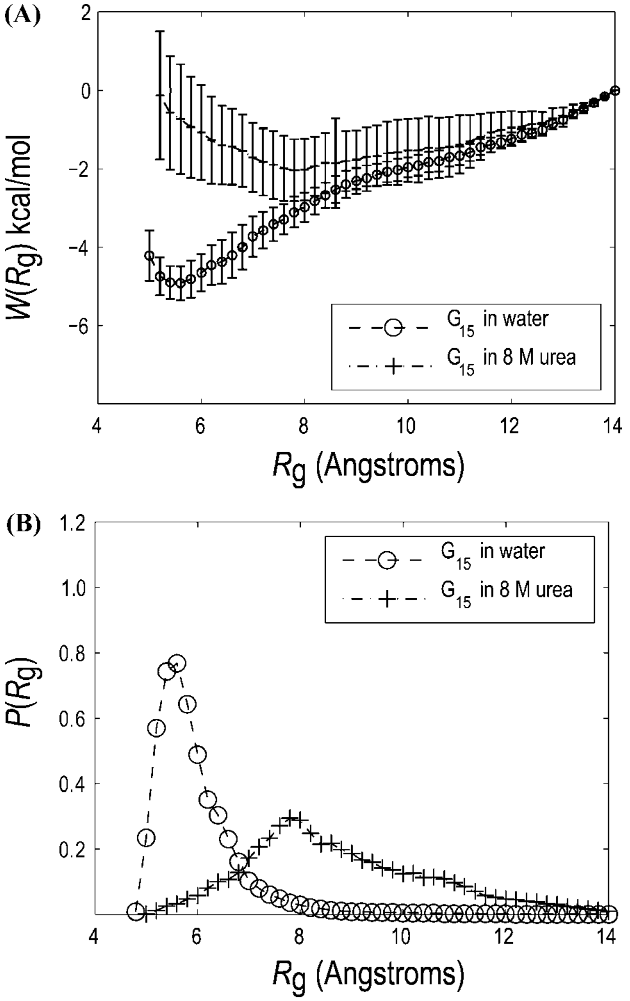


***Figure 4.*** (A) PMFs, $W(R_g)$, for $G_{15}$ in water and in 8 M urea, respectively. Error bars are standard errors, which we calculated using block-averaging techniques. We arbitrarily shifted all PMFs to be zero at a reference value of $R_g = 14$ Å. (B) Corresponding probability distributions for $G_{15}$, which were calculated as described in Figure 3.

calculated PMFs indicate a pronounced bias for collapse (small values of $R_g$) of $(GS)_8$ in water. Conversely, the PMF for $(GS)_8$ in 8 M urea is indicative of broad conformational equilibrium between a range of $R_g$ values. The two PMFs shown in Figure 3 are in agreement with expectations from the experiments of Möglich et al.[12] The remainder of our analysis is focused on assessing if the preference for distinct conformational ensembles for sequences such as glycine-serine copolypeptides in water versus 8 M urea is reproduced by generic polypeptide backbones.

**Polypeptide Backbones Show a Preference for Collapsed Conformations.** We performed umbrella sampling simulations for $G_{15}$ to determine the conformational preferences of "backbone-only" chains in both water and 8 M urea. Figure 4 shows the PMFs and probability distributions obtained for $G_{15}$ in water and in 8 M urea. $G_{15}$ prefers collapsed conformations in water and samples a variety of expanded conformations in 8 M urea, a behavior that is similar to that of $(GS)_8$. PMFs calculated for $G_{15}$ in the EV and limit, using the potential in eq 1, are qualitatively similar to the PMF in 8 M urea (data not shown) and the PMF in water bears qualitative resemblance to the PMF calculated in the nonspecific globule limit, using the potential in eq 2 (data not shown). We know that in the EV and nonspecific globule limits, chains mimic the global characteristics of polypeptides in generic good versus poor solvents, respectively. Therefore, the data suggest that water is a poor solvent for $G_{15}$, whereas 8 M urea is a good solvent for the backbone-only $G_{15}$ chain.

As shown in previous work,[37] we can calculate a range of quantities in accordance with polymer theories to assess the validity of our suggestion that water is a poor solvent for polypeptide backbones. Prior to carrying out these tests, we sought to rule out the possibility that the observed preference for collapsed states for $G_{15}$ in water is an artifact of our choice of the OPLS-AA/L forcefield, which is known to have problems recapitulating the experimentally measured free energy of hydration for *N*-methylacetamide.[48]

To address concerns that the collapse we observe is due to anomalies of the OPLS-AA/L force field, we performed simulations of $G_{15}$ and $(GS)_8$ in aqueous solutions using the GROMOS 53A6[26] force field and the SPC water model.[27] The initial configurations were generated using the EV model. These were soaked and equilibrated for 10 ns, after which we collected data for 100 ns for each peptide sequence. Figure 5 shows that the distributions of $R_g$ calculated using the two force fields for

[48] Udier-Blagovic, M.; De, Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322–1332.

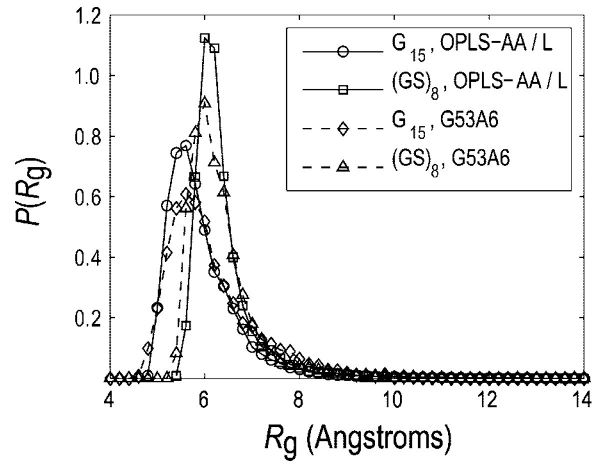


***Figure 5.*** Probability distribution of radius of gyration, $P(R_g)$, for various chains in aqueous solution, using the OPLS-AA/L force field (computed using PMFs in Figures 3 and 4) and the GROMOS 53A6 force field.

$G_{15}$ and $(GS)_8$ in water are qualitatively similar vis-à-vis locations of peaks and the range of values sampled. However, the $R_g$ distribution calculated with the GROMOS force field indicates broader conformational equilibrium and hence greater fluctuations around collapsed conformations. Thus, we conclude that the *qualitative* preference for collapse of $G_{15}$ in water is not an artifact of the force field we use. However, the *quantitative* preference i.e., the stability of collapsed states and magnitude of fluctuations depends on details of the force field. Of course, differences in the quality of conformational sampling are also important factors in comparing the two sets of distributions. Given a lack of consensus regarding the correct force field to use, it would be prohibitively expensive to carry out simulations with a range of force fields. Instead, we assume that simulations with different force fields are likely to show qualitative similarities. With this assumption in hand, we continue with the remainder of our analysis for simulations of $G_{15}$ based on the OPLS-AA/L force field and the TIP3P water model.

**Ensembles in Aqueous Solutions and 8 M Urea are Distinct.** We analyzed the magnitudes of fluctuations in shape and size, the scaling of internal distances, and solvent accessible surface area to show that global characteristics for $G_{15}$ ensembles in water are clearly distinct from those in 8 M urea. For a specific conformation of a polymer, one can calculate eigenvalues of the gyration tensor. These eigenvalues are in turn used to compute $R_g$, which characterizes the polymer size/density and a parameter $\delta^*$ referred to as asphericity which characterizes the polymer shape.[44,45,49] For a perfect sphere, $\delta^* = 0$, and for a perfect rod, $\delta^* = 1$; for intermediate values, the chain assumes ellipsoidal shapes. Therefore, $\delta^*$ quantifies the degree to which chain shape deviates from that of a perfect sphere.

Figure 6 shows two-dimensional histograms as a function of $R_g$ and $\delta^*$ for $G_{15}$ in water, 8 M urea, the EV limit, and nonspecific globule limit, respectively. Conformations with low asphericity and low $R_g$ are favored for $G_{15}$ in water and in the nonspecific globule limit. In stark contrast, $G_{15}$ in 8 M urea and in the EV limit prefer conformations with larger $R_g$ and asphericity values. The ensembles sampled in water and in the nonspecific globule limit overlap minimally with the ensembles sampled in 8 M urea and the EV limit. From simulations of the reference poor solvent ($G_{15}$ reference globule) and good solvent ($G_{15}$ EV limit) chains, we see that the dividing value of $R_g$ between the two ensembles is located at $R_g = 7$ Å (Figure 6). There is significant overlap between the conformational ensembles for $G_{15}$ in water and the reference globule. Similarly, the overlap of the distribution in 8 M urea is significant with the EV limit. Comparison of the probabilities of $R_g$ greater or less than 7 Å for $G_{15}$ shown in Table 4 underscores the distinct nature of ensembles sampled in water/the nonspecific globule limit versus 8 M urea/the EV limit. The fluctuations in size and shape for $G_{15}$ in water and in the nonspecific globule limit are considerably smaller than fluctuations for $G_{15}$ in 8 M urea and the EV limit.

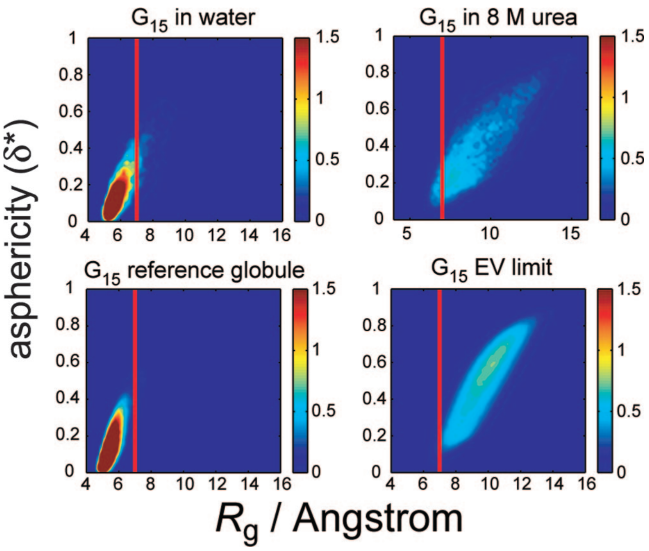


***Figure 6.*** Two-dimensional histograms in $R_g$ and $\delta^*$ for G15 in water, in 8 M urea, the reference nonspecific globule limit, and the EV limit, respectively. Vertical lines are drawn at $R_g = 7$ Å which separates the two reference ensembles. The bin size used was ($\mu R_g = 0.2$ Å) × ($\mu\delta^* = 0.02$).

***Table 4.*** Probability for $G_{15}$ to Adopt Compact ($R_g$ <7 Å) or Expanded ($R_g \geq 7$ Å) Conformations under Different Conditions

| | probability of sampling conformations with $R_g$ < 7 Å | probability of sampling conformations with $R_g \geq$ 7 Å |
|---|---|---|
| $G_{15}$ in 8 M urea | 0.06 | 0.94 |
| $G_{15}$ in water | 0.83 | 0.17 |
| $G_{15}$ (reference globule) | 0.97 | 0.03 |
| $G_{15}$ (EV limit) | 0.01 | 0.99 |

Solvent accessible surface area (SASA) is a commonly used measure to quantify the extent of interaction between solute and solvent. Figure 7 shows two-dimensional histograms as a function of $R_g$ and SASA for $G_{15}$ in different environments. In the EV limit, we note that a wide range of sterically allowed conformations (characterized by $R_g$) have roughly equivalent probabilities and SASA values. This suggests that in good solvents, there is minimal correlation between chain conformation and SASA. Essentially a single, average SASA value (ca. 1400 Å$^2$) characterizes all realizable, albeit disparate conformations in 8 M urea/the EV limit. In a good solvent, the chain is fully accessible on the length scale of the solvent molecule (the probe radius used to calculate SASA) and this is true irrespective of chain conformation. Figure 7 shows that the SASA distribution in 8 M urea shows minimal correlation with chain conformation, as observed for the EV limit. Conversely, in the

[49] Dima, R. I.; Thirumalai, D. *J. Phys. Chem. B* **2004**, *108*, 6564–6570.

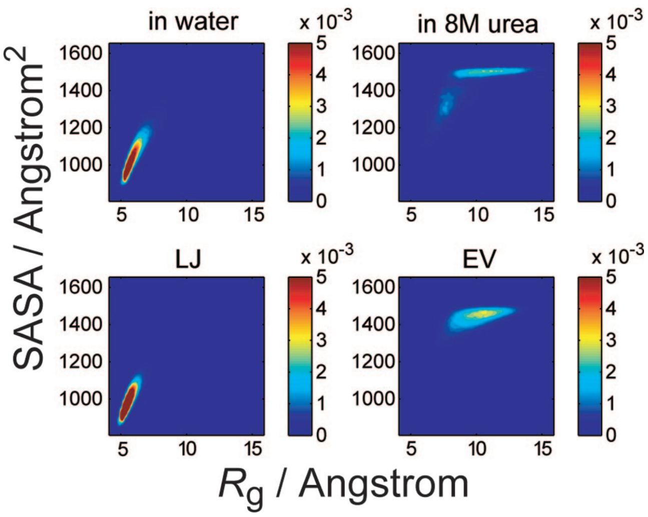


*Figure 7.* Two-dimensional histograms for $G_{15}$ as a function of Rg and SASA for G15 in water, 8 M urea, the reference nonspecific globule limit (LJ), and the EV limit, respectively. The calculated average SASA values are as follows: 1085 Å$^2$ in water; 1401 Å$^2$ in 8 M urea; 1427 Å$^2$ in the EV limit; and 998 Å$^2$ in the nonspecific globule (LJ) limit. We used a probe radius of 1.4 Å to calculate SASA values. The bin size used was($\Delta R_g$ = 0.2 Å) × ($\Delta S$ ASA = 10 Å$^2$).

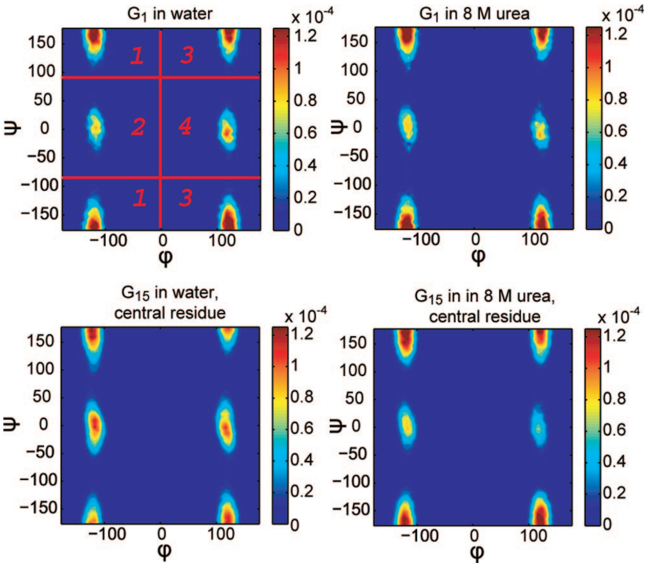


*Figure 8.* (Top) Conformational probability densities in Ramachandran space for $G_1$ in water and 8 M urea. (Bottom) Conformational probability densities in Ramachandran space for the central glycine residue of $G_{15}$ in water and in 8 M urea. We used regions 1−4 for analysis of conformational propensities in Table 5. The bin size used for contour plots was 6° × 6°.

*Table 5.* Conformational Probabilities, in Percentages, for Regions of Ramachandran Space As Defined in Figure 8, for $G_1$ and the Central Residue in $G_{15}$[a]

| region | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $G_1$ in water | 35 ± 6 | 14 ± 3 | 36 ± 6 | 15 ± 3 |
| $G_1$ in 8 M urea | 39 ± 5 | 14 ± 2 | 33 ± 5 | 13 ± 2 |
| $G_{15}$ in water | 30 ± 6 | 23 ± 6 | 26 ± 9 | 22 ± 6 |
| $G_{15}$ in 8 M urea | 37 ± 7 | 15 ± 8 | 36 ± 10 | 12 ± 3 |

[a] Total simulation time for $G_1$ in each environment was 200 ns. The error bars denote standard errors from block-averaging.

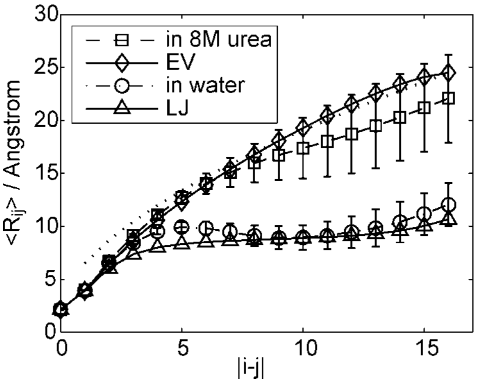


*Figure 9.* Scaling of intrachain distances $\langle R_{ij}\rangle$for $G_{15}$ as a function of sequence separation. Data are shown for $G_{15}$ in water, in 8 M urea, in the nonspecific globule limit (LJ), and in the EV limit, respectively. The theoretical good solvent scaling law is indicated by the dotted curve. Error bars, which measure standard error, were calculated using block-averaging.

nonspecific globule limit there is clear, positive correlation between chain conformation ($R_g$) and SASA and the magnitude of the average SASA is smaller, i.e., ca. 1000 Å$^2$. For $G_{15}$ in water, we find similar features of positive correlation between $R_g$ and SASA and smaller, average solvent accessibility (ca. 1100 Å$^2$). The pattern of solvent accessibility, as probed using a sphere of radius 1.4 Å, is similar for $G_{15}$ in water and the reference globule. These patterns are distinct from those for $G_{15}$ in 8 M urea and the EV limit.

**Distinct Preferences for $G_{15}$ in Water versus 8 M Urea Originate on Length Scales beyond the Blob Size.** We introduced the concept of "blobs", which we estimated to be ca. 2−3 residues. By definition, conformational equilibria within blob-sized segments should show weak or no dependence on solvent type, whereas the conformation of blobs with respect to each other must depend on solvent type. Figure 8 and Table 5 show comparative analysis of backbone dihedral angle propensities for $G_1$ in water and 8 M urea, respectively. There are no statistically significant differences in dihedral angle propensities for $G_1$ in the two solvent environments. This observation is consistent with our estimate of the blob size being greater than one residue. Alternatively, comparison of backbone dihedral angle propensities for the central residue in $G_{15}$ between water and 8 M urea shows statistically significant changes. These results suggest that there is a modulation of local conformational propensities in the context of longer chains, which happens when interactions with the solvent promote chain collapse. Therefore, conformational propensities for isolated dipeptides in water are only partially predictive of propensities in the context of longer chains in water.

One might speculate that the preference of $G_{15}$ for collapsed conformations in water is due to the intrinsic flexibility of the glycine residue. However, lack of overlap between conformational distributions for $G_{15}$ in water and distributions in the EV limit (Figures 6 and 7) indicates that intrinsic steric flexibility alone cannot lead to collapse. The presence of the ternary component, namely, 8 M urea, promotes large-scale conformational fluctuations in longer chains, thereby facilitating sampling of the full spectrum of sterically realizable conformations. In the absence of urea, $G_{15}$ is restricted to a manifold of compact conformations.

**Scaling of Internal Distances and Adjudication of Solvent Quality.** Figure 9 shows the scaling of ensemble averaged internal distances $\langle R_{ij}\rangle$ between residues *i* and *j* as a function of sequence spacing $|i - j|$ for $G_{15}$ in water, 8 M urea, the

*Table 6.* Average Numbers of Donor−Acceptor Contacts for $G_{15}$ in Water Quantified As a Function of a Range of Values for $r_c$

| $r_c$(Å) | N−OW | N−O | O−OW | total around N | total around O |
|---|---|---|---|---|---|
| 3.0 | 0.39 | 0.12 | 1.08 | 0.51 | 1.20 |
| 3.1 | 0.53 | 0.15 | 1.30 | 0.68 | 1.45 |
| 3.2 | 0.66 | 0.19 | 1.51 | 0.85 | 1.70 |
| 3.3 | 0.80 | 0.22 | 1.71 | 1.02 | 1.93 |
| 3.4 | 0.95 | 0.25 | 1.93 | 1.20 | 2.18 |
| 3.5 | 1.12 | 0.29 | 2.16 | 1.41 | 2.45 |

nonspecific globule limit (LJ), and the EV limit, respectively. For $|i - j|$ greater than the blob size, theory predicts that $\langle R_{ij} \rangle \approx |i - j|^{0.59}$ for chains in good solvents.[37,50] From Figure 9 we see that the scaling of internal distances follows theoretical predictions for $G_{15}$ in 8 M urea/the EV limit.

For chains in a poor solvent, theory predicts that, for $|i - j|$ greater than the blob size, $\langle R_{ij} \rangle$ should reach a plateau value, which corresponds to the average density of the globule that results from collapse.[37,50] This feature is apparent in Figure 9 for $G_{15}$ in both water and the nonspecific globule limit. The minor differences between plateau values for $\langle R_{ij} \rangle$ suggest that the average density of globules in water is slightly smaller than densities in the nonspecific globule limit.

The distinctive scaling of internal distances with sequence separation provides a rigorous tool for assessing the solvent quality of a specific milieu for a given polymer.[37,50] On the basis of the results shown in Figure 9, we conclude that water is a poor solvent whereas 8 M urea is a good solvent for $G_{15}$. The preference for two distinct ensembles as a function of solvent quality is realizable for a backbone-only chain. The implication is that the preference for compact geometries observed previously for polyglutamine and glycine-serine block copolymers in aqueous environments originates, at least partially, in the polypeptide backbone.

**Analysis of Solvation Characteristics of $G_{15}$.** Möglich et al.[12] suggested that chain collapse in glycine-serine block copolypeptides is driven primarily by intramolecular hydrogen bonding. We examined the distributions of intramolecular and chain-solvent hydrogen bond donor−acceptor contacts to develop a preliminary assessment of driving forces that lead to solvent-mediated differences in global conformational preferences for $G_{15}$, which we have noted is a backbone-only polypeptide. Possible donor−acceptor pairs for analysis of intramolecular and chain-solvent contacts are: Amide nitrogen and carbonyl oxygen (N−O); Water oxygen and carbonyl oxygen (OW-O); Urea nitrogen and carbonyl oxygen ($N_{urea}$−O); Amide nitrogen and water oxygen (N-OW); Amide nitrogen and urea oxygen (N−$O_{urea}$). For each donor−acceptor pair, there is a contact when the site−site distance is less than a specified cutoff, $r_c$. The value of $r_c$ can be somewhat arbitrary, so we use a range of cutoff values $r_c$ = (3.0, 3.1, 3.2, 3.3, 3.4, 3.5 Å). We selected this range to ensure that the estimates of donor−acceptor contacts bracket the stoichiometric expectations, namely 1 donor−acceptor contact per backbone amide nitrogen, and 2 such contacts per carbonyl oxygen atom.

Tables 6 and 7 show statistics for intramolecular and solvent−solute donor−acceptor contacts found using different values for the contact parameter $r_c$ for $G_{15}$ in water and in 8 M urea, respectively. The average number of intrabackbone donor−acceptor contacts is negligible for $G_{15}$ in 8 M urea, and only slightly larger for $G_{15}$ in water. To interpret these average values, we analyzed the distribution of intramolecular and chain-solvent contacts in both environments for $r_c$ = 3.3 Å.

*Table 7.* Average Numbers of Donor−Acceptor Contacts for $G_{15}$ in 8 M Urea Quantified As a Function of a Range of Values for $r_c$

| $r_c$(Å) | N−OW | N−$O_{urea}$ | N−O/O−N | O−OW | O−$N_{urea}$ | total N | total O |
|---|---|---|---|---|---|---|---|
| 3.0 | 0.17 | 0.25 | 0.02 | 0.58 | 0.69 | 0.44 | 1.29 |
| 3.1 | 0.24 | 0.34 | 0.02 | 0.70 | 0.92 | 0.60 | 1.64 |
| 3.2 | 0.29 | 0.42 | 0.03 | 0.81 | 1.13 | 0.74 | 1.97 |
| 3.3 | 0.36 | 0.50 | 0.03 | 0.92 | 1.31 | 0.89 | 2.26 |
| 3.4 | 0.43 | 0.58 | 0.04 | 1.03 | 1.48 | 1.05 | 2.55 |
| 3.5 | 0.51 | 0.66 | 0.04 | 1.15 | 1.65 | 1.21 | 2.84 |

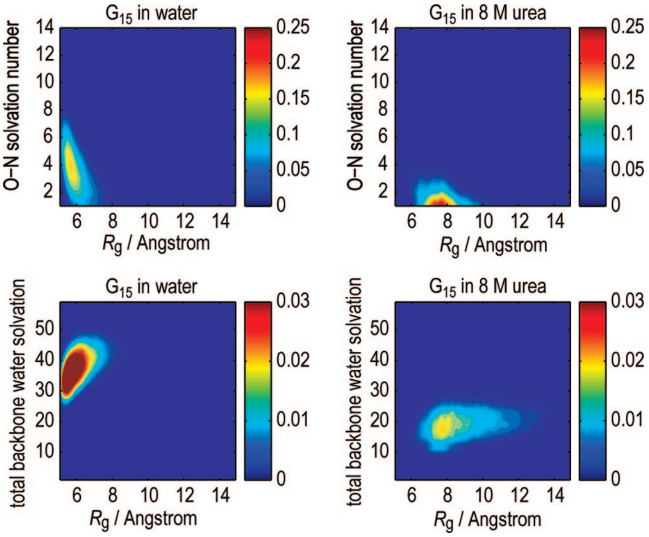


*Figure 10.* (Top) Two-dimensional histograms showing the conformational dependencies of donor−acceptor contacts. The abscissa denotes conformation $R_g$ and the ordinate is the number of intramolecular donor−acceptor contacts for $G_{15}$ in water (top left) and in 8 M urea (top right). (Bottom) Similar two-dimensional histograms, with the difference being that the ordinate is the total number of donor−acceptor contacts between the chain and water for $G_{15}$ in water (bottom left) and in 8 M urea (bottom right). In all plots, the donor−acceptor distance cutoff was 3.3 Å. The bin size for all contour plots was ($\Delta R_g$ = 0.2 Å) × (Δ(donor−acceptor contacts) = 1).

We first assessed how the number of intramolecular donor−acceptor contacts varies as a function of conformation ($R_g$) for $G_{15}$ in water and 8 M urea. For $G_{15}$ in water, the number of intramolecular donor−acceptor contacts shows negatively correlated fluctuations with $R_g$, i.e., this number increases as $R_g$ decreases and vice versa. No such correlations are evident for $G_{15}$ in 8 M urea where the number of intramolecular donor−acceptor contacts is uniformly low. In water, as the chain expands, intramolecular contacts are replaced with chain-solvent donor−acceptor contacts, and this positive correlation is illustrated in Figure 10. Collapse does not imply the preference for a single conformation. Instead, spontaneous fluctuations lead to a heterogeneous distribution of intramolecular and chain-solvent contacts. Figure 10 also shows that while swollen, aspherical conformations are populated in 8 M urea, there is a pronounced diminution in the number of donor−acceptor contacts between the peptide and water. This suggests that in the presence of 8 M urea, water is excluded from the vicinity of the polypeptide.

Further analysis of the distributions of intramolecular and chain-solvent donor−acceptor contacts is shown in Figure 11. Panel A in Figure 11 shows a narrow Poisson-like distribution

[50] Imbert, J. B.; Lesne, A.; Victor, J. M. *Phys. Rev. E* **1997**, *56*, 5630–5647.

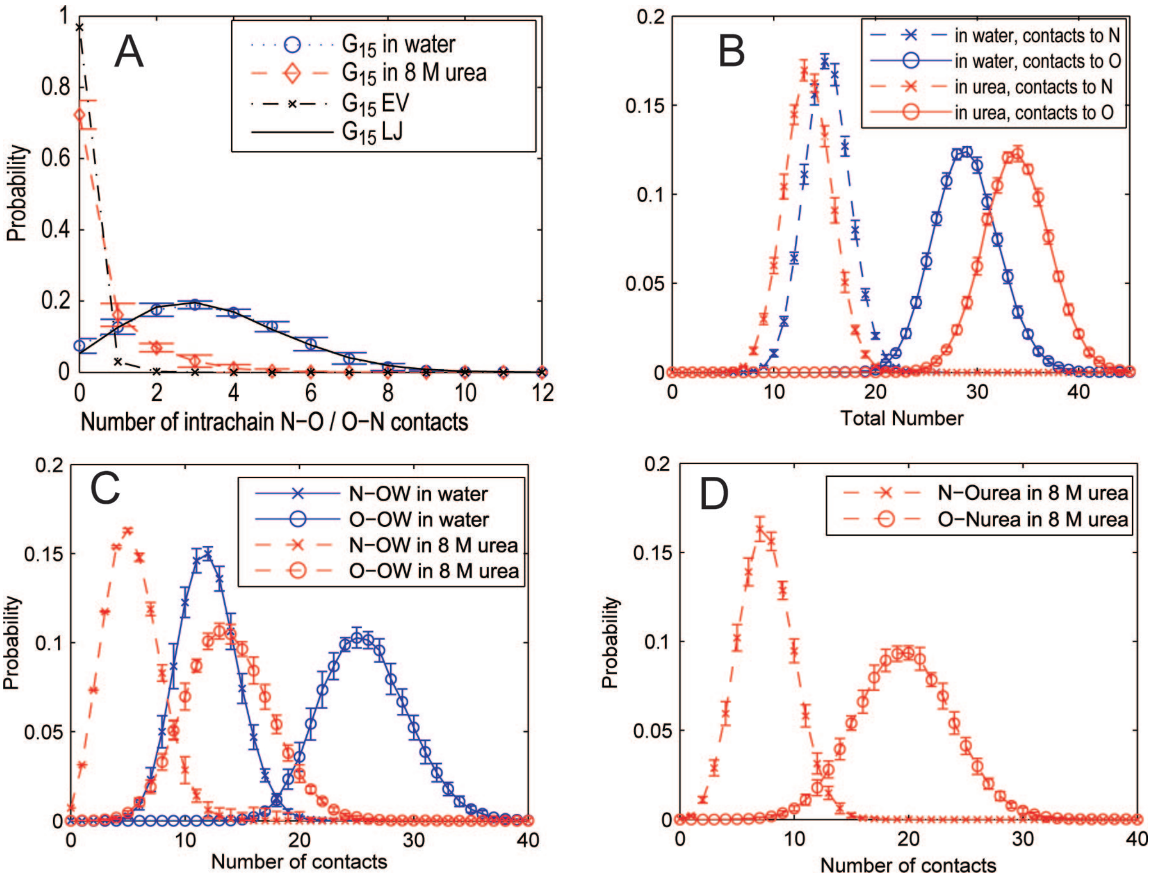


***Figure 11.*** Probability distributions for different donor−acceptor contacts: (A) distribution of intrachain donor−acceptor contacts for $G_{15}$ in water, in 8 M urea, the reference nonspecific globule (LJ) and the EV limit; (B) distribution of the total number of intramolecular and chain-solvent donor−acceptor contacts; (C) backbone-water contacts for $G_{15}$ in water and 8 M urea; (D) distribution of backbone-urea contacts. The contact cutoff distance used was 3.3 Å.

for the number of intramolecular donor−acceptor contacts in 8 M urea as opposed to a broad distribution for $G_{15}$ in water. On average, surface, and internal hydration, as well as backbone hydrogen bonding is characteristic for globules of $G_{15}$ in water.

Panel A in Figure 11 also compares distributions of intramolecular donor−acceptor contacts for conformations in the two reference ensembles to the distributions obtained for $G_{15}$ ensembles in water and 8 M urea. The distribution calculated for the reference globule matches that for $G_{15}$ in water and similarly, the distribution for the EV limit is consistent with that of $G_{15}$ in 8 M urea. The former result is surprising because the potential function used to generate the reference conformational ensembles in the nonspecific globule limit does not have a specific hydrogen bonding term. From polymer theory,[47] we know that collapse of generic polymers in poor solvents is nonspecific and characterized as a random walk on a compact manifold. The fact that distributions of intramolecular donor−acceptor contacts for $G_{15}$ in water match those calculated using a model with no specificity suggests that collapse in water is unlikely to be driven solely by specific intramolecular hydrogen bonding, and collapse does not have to promote the formation of a specific structure. The data in Panel A do not necessarily mean that van der Waals interactions drive the collapse of $G_{15}$ in water. In explicit solvent, several factors can contribute toward driving the collapse of $G_{15}$ in water.[51] The analysis of intramolecular and chain-solvent donor−acceptor contacts does not provide access to all of the information required to assess if the driving forces for collapse are primarily entropic versus enthalpic in nature. To make this assessment, we require knowledge of the temperature dependence of PMFs, $W(R_g)$, and the distributions of intramolecular as well as chain-solvent contacts, which is work in progress.

The distribution of intramolecular donor−acceptor contacts (Figure 11A) may be a consequence of inaccuracies in the OPLS-AA/L forcefield. Morozov et al.[52] have compared hydrogen bonded geometries and binding energies for formamide dimers in the gas phase calculated using quantum mechanics to those obtained using three molecular mechanics forcefields including OPLS-AA. The molecular mechanics formamide binding energies show qualitative agreement with the two quantum calculations, especially when the dimerization energies are compared as a function of proton-acceptor distance, the bending angle at the acceptor, and the bending angle at the proton. (Disagreements between results from two distinct quantum mechanical calculations are in the same range as disagreements between the molecular mechanics and quantum mechanics). However, one major shortcoming identified by Morozov et al.[52] was that molecular mechanics forcefields favor planar arrangements of the acceptor moieties in the donor plane leading to a pronounced, apparently erroneous preference for linear hydrogen bonds. Morozov et al. also investigated the hydrogen bond geometries found in high resolution protein

[51] Ben-Naim, A. *Solvation Thermodynamics*, 1st ed.; Plenum Press: New York, NY, 1987; p 246.

[52] Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6946–6951.

structures and noted that the deviations of molecular mechanics predictions from quantum mechanical estimates are more likely to lead to errors in predictions for hydrogen bonding geometries of polar sidechains. Conversely, they suggest that there ought to be fewer problems in recapitulating main-chain hydrogen bond geometries using standard molecular mechanics forcefields. Additional support for the latter point comes from the work of Pappu et al.[53] who showed that global minima for polyalanine are canonical regular hydrogen bonded structures in the gas phase. These calculations were performed with the OPLS forcefield, which was the precursor to the OPLS-AA/L forcefield used in this work. Taken together, the results of Morozov et al. and Pappu et al. suggest that intramolecular backbone hydrogen bonds are modeled adequately using molecular mechanics forcefields and therefore our observation that polyglycine prefers a heterogeneous ensemble of collapsed conformations is likely to be reasonable for *dilute, aqueous solutions* of this system of molecules.

**Preferential Solvation of Backbones by Urea.** The data in Tables 6 and 7 indicate the presence of preferential interactions between the backbone of $G_{15}$ and urea. Panels B–D in Figure 11 show the distributions for different types of chain-solvent donor–acceptor contacts for $G_{15}$ in water and 8 M urea, respectively. Panel B shows the distributions of the total number (sum of intramolecular and chain-solvent) of donor–acceptor contacts involving the polypeptide backbone. This analysis verifies that the stoichiometric value (with 16 peptide units, there are 16 donors and 32 acceptor lone pairs that need to be satisfied) for the number of donor–acceptor contacts is maintained on average. The distribution of donor–acceptor contacts between water and the peptide backbone sites shifts to smaller values for $G_{15}$ in 8 M urea as shown in Panel C of Figure 11. This indicates that water is excluded from the vicinity of the backbone in lieu of favorable direct contacts between urea and the backbone. Evidence for the latter is presented in Panel D of Figure 11, which shows the distribution of urea-backbone contacts. The distribution of backbone donor and water acceptor (N–OW) contacts in 8 M urea (Panel C) is similar to the distribution of backbone donor to urea acceptor contacts (N–$O_{urea}$), also in 8 M urea (Panel D). Conversely, the distribution of backbone acceptor to urea donor contacts (O–$N_{urea}$) shows a pronounced shift (Panel D) toward larger values vis-à-vis backbone acceptor to water donor contacts (O–OW) in 8 M urea (Panel C). This clearly indicates that in 8 M urea, the preferential interactions involve exclusion of water from the vicinity of the backbone and favorable direct interactions between urea and the backbone, which arise primarily from contacts between backbone acceptors and hydrogen bond donors of urea molecules. Similar conclusions were reached by O'Brien et al. [54] based on their simulation work and by Auton et al. [55] who used their group based transfer free energy model to provide an "anatomy" of changes associated with urea denaturation.

The mechanism by which urea denatures proteins is a subject of intense debate. One proposal is that urea solvates hydrophobic groups in the unfolded state, thus weakening the hydrophobic effect.[56] In the water structuring hypothesis, the effect of cosolutes, such as urea, on protein stability is attributed to the ability of the cosolute to increase or decrease "water structure".[57,58] Conversely, the direct/preferential interaction model posits that the effect of urea is due to the direct interaction of urea with polar or charged moieties on the protein,[59] in contrast with stabilizing cosolutes which are preferentially excluded from the vicinity of the protein.[59,60]

Recent studies have cast doubt on the water structuring and hydrophobic weakening theories. Batchelor et al.[61] classified a variety of cosolutes as water "structure-making" or "structure-breaking" using measurements based on pressure perturbation calorimetry. They were unable to correlate structure-making or structure-breaking cosolutes with their role as stabilizers or destabilizers of proteins. Simulations show that increasing urea concentration does not have an effect on the association of hydrophobic model solutes. [54] This is consistent with our data, which imply that urea interacts preferentially with polypeptide backbones, and clearly promotes chain swelling in backbone-only constructs.

## 4. Conclusions and Discussion

**Summary.** Experimental studies of archetypal, polar IDPs show that these sequences prefer ensembles of compact structures in aqueous milieus,[11–13] which is surprising because there are no hydrophobic residues in these archetypal IDPs. One might argue that collapse of polar tracts in aqueous milieus is a consequence of side chain-mediated interactions. Eberhardt and Raines[62] showed that 25 M formamide is equivalent to 55 M water as a solvent for peptide units, which are secondary amides. Formamide is a primary amide and mimics polar moieties in the sidechains of glutamine and asparagine. In fact, Wang et al.[63] proposed that polyglutamine favors an ensemble of collapsed structures in water because this increases the effective concentration of side chain primary amides around the polypeptide backbone. However, in this work, we find that a 15-residue polyglycine peptide $G_{15}$, which is devoid of sidechains, prefers an ensemble of collapsed structures in water. Conversely, in 8 M urea $G_{15}$ shows preference for a heterogeneous ensemble of swollen conformers. Therefore, by extrapolation we propose that the experimentally observed preferences for archetypal, polar IDPs in aqueous milieus must originate, at least partially, in the conformational preferences of generic polypeptide backbones in water.

**Water is a Poor Solvent for Backbone-Only Polyglycine Chains.** Our assessment of order parameters such as the scaling of internal distances and comparison of conformational equilibria for $G_{15}$ in water and 8 M urea to the EV and nonspecific globule limits allow us to conclude that water is a poor solvent for polyglycine, whereas 8 M urea is a good solvent for this system. In a poor solvent, polypeptides form a homogeneous solution of collapsed globules providing the solution is sufficiently dilute (with concentrations in the nanomolar range or below).[11,13] For

[53] Pappu, R. V.; Hart, R. K.; Ponder, J. W. *J. Phys. Chem. B.* **1998**, *102*, 9725–9742.

[54] O'Brien, E. P.; Dima, R. I.; Brooks, B.; Thirumalai, D. *J. Am. Chem. Soc.* **2007**, *129*, 7346–7353.

[55] Auton, M.; Holthauzen, L. M. F.; Bolen, D. W. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15317–15322.

[56] Alonso, D. O. V.; Dill, K. A. *Biochemistry* **1991**, *30*, 5974–5985.

[57] Vanzi, F.; Madan, B.; Sharp, K. *J. Am. Chem. Soc.* **1998**, *120*, 10748–10753.

[58] Zou, Q; Bennion, B. J; Daggett, V; Murphy, K. P *J. Am. Chem. Soc.* **2002**, *124*, 1192–1202.

[59] Schellman, J. A. *Biophys. J.* **2003**, *85*, 108–125.

[60] Arakawa, T.; Timasheff, S. N. *Biophys. J.* **1985**, *47*, 411–414.

[61] Batchelor, J. D; Olteanu, A; Tripathy, A; Pielak, G. J *J. Am. Chem. Soc.* **2004**, *126*, 1958–1961.

[62] Eberhardt, E. S.; Raines, R. T *J. Am. Chem. Soc.* **1994**, *116*, 2149–2150.

[63] Wang, X.; Vitalis, A.; Wyczalkowski, M. A.; Pappu, R. V. *Proteins* **2006**, *63*, 297–311.

dilute polymer solutions in poor solvents quantities such as $R_g$ and hydrodynamic size $R_h$ scale with chain length as $N^{0.33}$. The preference for globules can be inferred using fluorescence correlation spectroscopy (FCS),[11,13] which requires small concentrations (1−100 nM) of labeled species. The preference for collapsed structures does not imply the existence of a unique folded conformation. For archetypal IDPs, the conformational ensemble is heterogeneous because there is no unique way to partition residues in the chain between the interior and the surface of a globule. Therefore, conformations of equivalent compactness have equivalent stability. This situation is especially true for homopolymeric sequences. In direct contrast, for sequences that fold into well-defined three-dimensional structures a single family of self-similar globular conformations is preferred over all other globular conformations.[16,18]

**Comparisons with Findings from Other Experiments.** In FCS[11,13] experiments, there are very few molecules in the observation volume. Indeed, these experiments can also be performed in the single molecule limit. Data from FCS and related fluorescence-based experiments allow us to probe intrinsic polymeric properties as influenced by the balance of intrachain and chain-solvent interactions without interference from interchain interactions. Recent data from such experiments provided the necessary motivations for our current work.

We now place our results in the context of earlier experimental investigations into structural preferences of high molecular weight polymers of polyglycine, which are insoluble in aqueous solutions and form two types of extended structures in the solid state.[64–68] Intermolecular (as opposed to intramolecular) interactions stabilize structures adopted by polyglycine in the solid state.[67,68] Furthermore, the types of structures formed depend on the method of preparation. Air-dried polyglycine cast from dichloroacetic acid on a mercury surface forms oriented films of $\beta$-glycine (polyglycine I).[65] In contrast, when polyglycine in dichloroacetic acid solution is precipitated by dilution with water, the chains take on mostly polyglycine II structures in the precipitate.[65] In polyglycine II all backbone ($\phi$,$\psi$)-angles are expected to be equal to (−150°, 150°) making each chain molecule rather extended.[68] At issue is the relevance of the solid-state polyglycine I and polyglycine II structures for the observations reported in this work.

It is conceivable that polyglycine preferentially adopts extended conformations in dilute aqueous solutions. This would have to be true despite the absence of stabilizing intermolecular interactions that are available in aggregates. Hence, there is no *a priori* reason to expect that an individual polyglycine molecule in dilute aqueous solution should adopt the extended conformations inferred from the X-ray diffraction data for precipitates. Crick and Rich have made this point rather emphatically in their work on the structure of polyglycine II.[67]

A more complete explanation for the insolubility of polyglycine is available from the theory of flexible polymers in poor solvents,[21] which encompasses the full spectrum of phase behavior observed thus far for polyglycine, and accommodates our observations as well as experimental data on precipitates.[64–68] In poor solvents, there is a strong driving force for separation into polymer rich and solvent rich phases.[21] Dilute solutions of flexible polymers in poor solvents will collapse to form isolated globules. Collapse is a manifestation of intramolecular phase separation. As concentration increases, aggregates that are rich in polymer and deficient in solvent become the preferred thermodynamic state. These aggregates, which can be marginally soluble or become part of the so-called sediment[21] are stabilized and characterized by significant intermolecular interactions. The concept of a blob (introduced earlier) is central to understanding the balance between chain-chain and chain-solvent interactions. As reviewed in recent work,[69] pairs of spatially adjacent blob-sized segments within a globule realize attractive contacts. Conversely, a blob-sized segment that is exposed to solvent is deprived of the favorable intrapolymer interactions and the free energy penalty for exposing blob-sized segments to solvent leads to the surface tension effect. At high concentrations, in the regime where the phase separated state is thermodynamically favored, chains are surrounded by other chains, and pairs of blobs between chains now have access to attractive, intermolecular, interblob interactions. These interactions are akin to intramolecular, interblob contacts. Consequently, the driving force that confines individual chains to globules is counterbalanced, the surface tension is lowered, and $R_g$ scales with chain length as $N^{0.5}$ in precipitates and soluble aggregates.[21] Therefore, in aggregates, flexible polymers in poor solvents are considerably more extended than in dilute solutions.

Recently, Ohnishi et al.[70] studied peptide constructs of the form Ac-YES-Gly$_n$-ATD using nuclear magnetic resonance, fiber diffraction, and small-angle X-ray scattering; *n*, the number of glycine residues in the constructs was 1, 2, 6, and 9. Peptide concentrations were in the millimolar range and this is 6 orders of magnitude larger than concentrations used in FCS experiments.[11,13] Under these conditions, one should expect significant intermolecular interactions, akin to that of polyglycine in the precipitate, and the authors note that they encounter significant problems due to oligomerization and insolubility. Indeed, the X-ray fiber diffraction data of Ohnishi et al. are congruent with earlier data on polyglycine. Ohnishi et al. also found that constructs with greater than nine glycine residues form insoluble aggregates. At these high peptide concentrations, constructs with six glycine residues have average $R_g$ values of 9.1 Å, indicating a preference for relatively extended conformations. Ohnishi et al. assumed that the conformational ensemble for their peptide constructs do not vary significantly with concentration. Consequently, they extrapolated scattering profiles measured in the 1.5−7.5 mM range to the ultradilute regime and concluded that the “marked tendency” of their peptide constructs to be insoluble “is consistent with the elongated ensemble-averaged structure of polyglycine in solution”.

Separation of length scales is an important hallmark of flexible polymers. We reviewed the concept of “blobs” and estimated the size of a blob to be ca. 2−3 residues (see start of Results section). The length of a chain needs to be approximately an order of magnitude longer than a blob for the balance between chain-chain and chain-solvent interactions to “encode” a preference for collapsed versus swollen states. Shorter chains as well as blob-sized segments within longer chains will prefer ensembles of extended structures. Data shown in Figure 9 underscore this point. This figure shows that the internal

[64] Meyer, K. H.; Go, Y. *Helv. Chim. Acta* **1934**, *17*, 1488–1492.

[65] Elliott, A.; Malcolm, B. R. *Trans. Farad. Soc.* **1954**, *50*, 1011–1011.

[66] Bamford, C. H.; Brown, L.; Cant, E. M.; Elliott, A.; Hanby, W. E.; Malcolm, B. R. *Nature* **1955**, *176*, 396–397.

[67] Crick, F. H. C.; Rich, A. *Nature* **1955**, *176*, 780–781.

[68] Ramachandran, G. N.; Sasisekharan, V.; Ramakrishnan, C. *Biochim. Biophys. Acta* **1966**, *112*, 168–170.

[69] Pappu, R. V.; Wang, X.; Vitalis, A.; Crick, S. L. *Arch. Biochem. Biophys.* **2008**, *469*, 132–141.

[70] Ohnishi, S; Kamikubo, H; Onitsuka, M; Kataoka, M; Shortle, D. *J. Am. Chem. Soc.* **2006**, *128*, 16338–16344.

distances for segments of length 4–5 residues are similar to the EV limit irrespective of solvent. In the EV limit, intrachain interactions are purely repulsive. Therefore, blob-sized segments (4–5 residues long) with $G_{15}$ prefer extended structures in both water and 8 M urea. We propose that the observed preference for locally extended conformations in $G_{15}$ is consistent with the data of Ohnishi et al. Conversely, the spatial arrangements between blob-sized segments differ for $G_{15}$ in water versus 8 M urea, leading to a preference for global collapse in water and swollen states in 8 M urea.

We have interpreted the insolubility of high polymers of polyglycine (chain lengths that are at least an order of magnitude longer than the size of a blob) to be supportive of our finding that water is a poor solvent for glycine-rich polypeptides. In this view, the balance between unfavorable interactions of blob-sized segments with solvent and favorable interblob interactions makes water a poor solvent. Consequently, in dilute solutions, polyglycine should form isolated globules, providing it is long enough to do so. Conversely, as concentration increases, linear aggregates stabilized by intermolecular hydrogen bonds, become thermodynamically preferred. An alternative interpretation is that all glycine-rich systems, irrespective of chain length, are insoluble only because of intermolecular hydrogen bonding. In this view, favorable enthalpic interactions between the backbone and solvent are traded for more favorable intermolecular hydrogen bonds, and even in dilute solutions, irrespective of chain length glycine-rich systems adopt extended conformations.

We need more data to adjudicate between the two views for the insolubility of polyglycine. For example, we need to carry out simulations that include more than one molecule. These simulations must be designed to assess the interplay between intramolecular collapse and intermolecular interactions, and hence we will need multiple, independent simulations, each characterized by different chain lengths and number of molecules. Such simulations will be feasible with novel simulation methodologies that use implicit solvation models.[71] Additionally, we will need novel FCS-based methods to study polyglycine systems at low (nanomolar) concentrations and provide clear adjudication regarding the competition between collapse and intermolecular interactions in polyglycine systems as a function of chain length.

**Driving forces.** We have shown that water is a poor solvent for polyglycine, which is a poly secondary-amide. However, the transfer free energy for N-methylacetamide (NMA) from the gas phase into water is −10 kcal/mol at 298 K. Hence, extrapolations from the transfer free energy model do not explain the behavior of the longer polyamides, a feature that was first noted by Roseman[72] when he tried to explain water/octanol partition coefficients of *N*-acetylamino acid derivatives. In recent work, Avbelj and Baldwin[73] have also proposed that solvation of peptide groups within longer polypeptide chains should be dependent on conformation and cannot be accurately inferred by extrapolations from studies of the hydration of the model compound alone.

Our preliminary analysis suggests that one cannot invoke intramolecular hydrogen bonding[12,74] as the sole driving force to rationalize our findings. Instead, we need additional insights. Clues emerge from the subtle balance between enthalpy and entropy, which is apparent even at the level of free energies of hydration for model compound amides. The favorable free energy of hydration ($\Delta G_{\text{hydration}} \approx -10$ kcal/mol) at 25 °C for NMA is the result of an intricate balance between highly favorable enthalpy ($\Delta H_{\text{hydration}} \approx -20$ kcal/mol) and negative entropy ($-T\Delta S_{\text{hydration}} \approx 10$ kcal/mol).[75] The large negative entropy offsets at least half the favorable enthalpy. Graziano has proposed that this "negentropic" term derives mainly from the excluded volume penalty associated with creation of a solute-sized cavity in water[76] Our working hypothesis is that the negentropic term becomes increasingly unfavorable for hydration of long, intrinsically flexible chains. Flexibility creates the problem that the number of conceivable conformations that can maximize the interface with solvent will increase exponentially with chain length, and the work done to create solute-sized cavities for expanded conformations will be significant. We postulate that the free energy penalties associated with cavitation for this heterogeneous ensemble of swollen conformations increases in a nontrivial manner with chain length. Consequently, longer chains collapse to minimize the entropic penalties of solvent organization around swollen, loosely packed conformations. The result is a heterogeneous ensemble of compact conformations characterized by different degrees of internal hydrogen bonding and hydration. In our hypothesis, intrinsic chain flexibility dictates the length scale for collapse. Dill,[77] in his influential analysis of "additivity principles," suggested that nonadditivities in entropic terms can arise in aqueous solutions because the degrees coupling between solvent and chain degrees of freedom are likely to vary significantly with chain conformation. Partial support for our hypothesis comes from the high solubility observed for proline-rich sequences,[78,79] and the prevalence of proline residues in IDPs,[5] which indicates that the presence of semirigid proline-rich tracts promotes mixing with water on all length scales. Our hypothesis can also be tested through quantitative, albeit challenging studies of the differences in entropic and enthalpic contribution to the free energies of hydration for extended versus collapsed conformations of polyglycine.

**Implications for other IDPs.** The archetypal sequences studied thus far are reasonable models for IDPs because of the absence of hydrophobic residues. IDP sequences can also have high net charge and are akin to polyelectrolytes with different degrees of charge asymmetry.[80] If aqueous milieus are poor solvents for polypeptide backbones, then IDP sequences with low net charge should also prefer heterogeneous ensembles of collapsed conformations under physiological conditions. Conversely, in sequences that have high net charge, long-range electrostatic repulsions between sidechains will compete with the drive of the backbone to form collapsed structures. Dobrynin et al.[81,82] predict that such sequences should prefer ensembles of elongated "necklace-globule" structures, and the preference for these

[71] Vitalis, A.; Pappu, R. V. *J. Comput. Chem.* **2008**, in press.

[72] Roseman, M. A *J. Mol. Biol.* **1988**, *200*, 513–522.

[73] Avbelj, F.; Baldwin, R. L. *Proteins: Struct. Funct. Bioinf.* **2006**, *63*, 283–289.

[74] Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16623–16633.

[75] Makhatadze, G. I.; Lopez, M. M.; Privalov, P. L. *Biophys. Chem.* **1997**, *64*, 93–101.

[76] Graziano, G *J. Phys. Soc. Jpn.* **2000**, *69*, 3720–3725.

[77] Dill, K. A *J. Biol. Chem.* **1997**, *272*, 701–704.

[78] Mattice, W. L.; Mandelkern, L. *Macromolecules* **1971**, *4*, 271–274.

[79] Kitamura, M.; Kakinoki, S.; Hirano, Y.; Oka, M. *Polym. Bull.* **2005**, *54*, 303–310.

[80] Bright, J. N.; Woolf, T. B.; Hoh, J. H. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 131–173.

[81] Dobrynin, A. V.; Rubinstein, M.; Obukhov, S. P. *Macromolecules* **1996**, *29*, 2974–2979.

[82] Dobrynin, A. V.; Rubinstein, M. *Prog. Polym. Sci.* **2005**, *30*, 1049–1118.

extended geometries is expected to vary with salt concentration and pH. These ideas suggest that charge characteristics of sidechains in IDPs can modulate the intrinsic preferences of polypeptide backbones, thereby generating a rich variety of conformational possibilities from globules, as shown in this work, to ensembles of necklace globules as predicted by polymer theory.[81,82]

**Acknowledgment.** We are grateful to Andreas Vitalis, Alan Chen, Xiaoling Wang, Matthew Wyczalkowski, and Professor D. Wayne Bolen for helpful discussions. The National Science Foundation supported this work through grants MCB 0416766 and MCB 0718924.

**Supporting Information Available:** Complete reference 3. This material is available free of charge via the Internet at http://pubs.acs.org.

JA710446S