SUPPORTING INFORMATION

Quantitative assessments of the distinct contributions of polypeptide backbone amides versus sidechain groups to chain expansion via chemical denaturation

Alex S. Holehouse[†], Kanchan Garai[†,‡], Nicholas Lyle[†], Andreas Vitalis[¶], and Rohit V. Pappu[*]

[†]Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, One Brookings Drive, Campus Box 1097, St. Louis, MO 63130, USA

[‡]TIFR Centre for Interdisciplinary Sciences, 21 Brundavan Colony, Narsingi, Hyderabad, 500075, India

[¶]Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-5807, Zurich, Switzerland

*Corresponding author: pappu@wustl.edu

Obtaining statistically equivalent data sets for the internal scaling profiles: The internal scaling profiles were calculated as outlined in the main manuscript. There is an intrinsic asymmetry in the volume of data associated with pairs at different separations. As an example, a sequence of length 17 (15 residues + 2 capping residues) has 16 pairs of residues separated with a linear sequence separation of one (|i–j|=1) and only one residue pair separated by 16 residues. To account for the disparity in data volume, we employ the following approach to create statistically equivalent datasets for each pair.

1. For each pair of residues at each sequence separation a high-resolution distance distribution is generated (bin widths of 0.05 Å). There are 16 distributions per replica for |i–j|=1 and two distributions per replica for |i–j|=15
2. From the distance distribution for each pair, we selected 2,500 distance values and 16 × 2500 values were generated from the distributions for |i–j|=1 while 2 × 2500 distributions were generated for |i–j|=15.
3. The steps 1 and 2 were repeated for each replica. This creates 20 × 16 × 2500 (800,000 values) and 20 × 2 × 2500 (100,000 values) for sequence separations |i–j|=1 and |i–j|=15, respectively.
4. From these data sets, we randomly selected 50,000 data points without replacement for each |i–j| separation between 1 and 15 to create 15 sets of data of equivalent size

This approach ensures that we use equivalent numbers of points from the distance distributions for all sequence separations thereby creating statistically equivalent datasets.

Relative occupancies of GdmCl around peptides: The values for π_{XY} were generated as described in the main text. In all of the plots shown in Figure SI-1, X refers to the nitrogen atoms or the central carbon atom of the Gdm$^+$ ion. The reference radial distribution function from the bulk solution pertains to the Gdm$^+$ / Cl$^-$ cation-anion pair such that equation (5) from the main text is rewritten as:

$$\pi_{\text{XY}} = \frac{\int_0^{4\text{Å}} g\left(r_{\text{XY}}\right) r_{\text{XY}}^2 \, dr_{\text{XY}}}{\int_0^{4\text{Å}} g_{\text{GdmCl}}\left(r_{\text{Gdm}^+\text{Cl}^-}\right) r_{\text{Gdm}^+\text{Cl}^-}^2 \, dr_{\text{Gdm}^+\text{Cl}^-}} \qquad \text{(S1)}$$

The resultant relative occupancy values are shown in Figure SI-1.

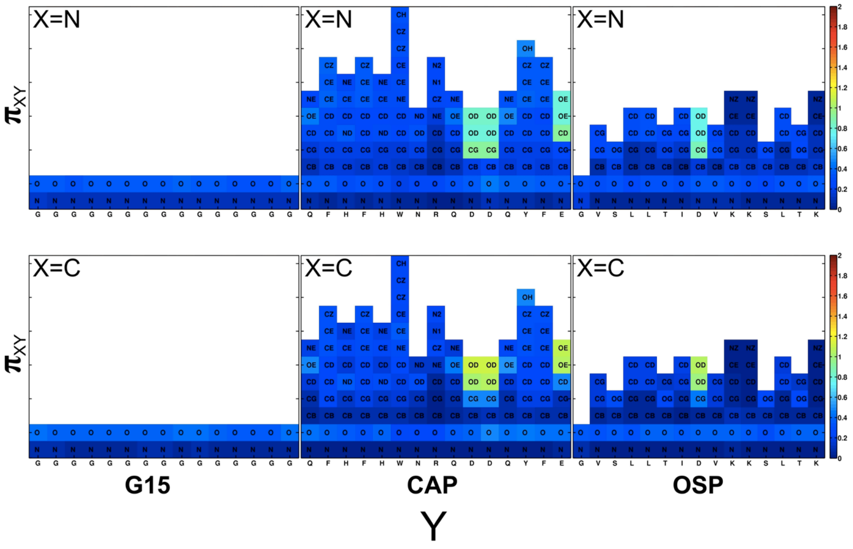

Figure SI-1: Relative occupancies of the Gdm$^+$ nitrogen atoms (top row) and central carbon atom around different backbone and sidechain sites of polyglycine, CAP, and OSP.

Analysis of amino acid compositional biases using a reduced alphabet: IDPs are deficient in bulky hydrophobic residues that make up the stable hydrophobic cores of folded proteins. Given this observation, amino acids are categorized as order promoting and disorder promoting residues.[1,2] The residues in the set {TAGRDHQKSEP} are designated as disorder promoting or **D**, while those in the set {WFYIMLVNC} are considered to be order promoting or **O**. We analyzed how the ratio disorder-to-order promoting residues, referred to hereafter as the D:O ratio, and distributions of D versus O-type residues vary within five sets of proteins. In all cases we considered only sequences of length greater than 20. The five datasets are as follows:

1. **Human proteome:** We extracted 220,202 distinct sequences corresponding to the complete and reviewed human proteome. These were obtained from the UniProt[3] database.

2. **D2P2 derived disordered regions:** We used the D2P2 database [4] of disorder predictions to scan the entire human proteome for regions predicted by five or more predictors to be disordered. This yielded 129,874 fragments, which reduces to 29,844 upon the length restriction of 20 residues or more.
3. **DISPROT[5]:** We obtained 693 sequences from the curated database of intrinsically disordered proteins. These sequences correspond to *bona fide* disordered proteins or regions.
4. **PDBSELECT25[6]:** This database is an inventory of non-redundant sequences that share less than 25% sequence homology and have structural models deposited in the protein data bank (PDB). In all, we analyzed 3,119 sequences from this dataset.
5. **Full set of human structural sequences:** We also analyzed the full set of sequences associated with human proteins obtained from the PDB. We first determined the set of unique human proteins in the PDB. For proteins with more than one structure associated with them, a single specific structure was selected at random. The sequences from the PDB files were collected. This ensures that we only collected sequences associated with well-structured proteins instead of collecting full sequences from proteins containing both folded and potentially disordered regions. In all, this dataset comprised of 8,924 sequences.

Table SI-1 summarizes the statistics for the sequences within each dataset. **Figure SI-2** shows the five histograms, each quantifying the number density of sequences with specific D:O ratios.

Table SI-1: The table below provides a summary of the statistics associated with sequences in each dataset.

Dataset	Number of sequences	Mean fraction disordered	Standard deviation of fraction disordered
Human proteome	20,202	0.63	0.073
D2P2 disordered	29,844	0.79	0.064
DISPROT	693	0.67	0.074
PDBSELECT	3,119	0.62	0.064
All human PDB	8,924	0.62	0.05

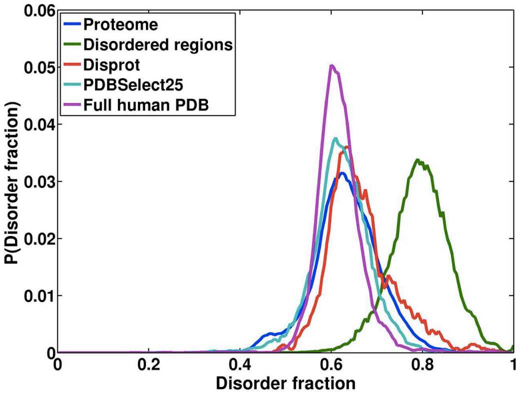

Figure SI-2: Number densities of the fraction of disorder promoting residues within sequences drawn from each of the five datasets.

The data shown in **Figure SI-2** highlight the similarities in the fraction of disorder promoting residues that are calculated using distributions for very different datasets of foldable proteins. In contrast, regions that are predicted to be disordered show a much higher fraction of disorder promoting residues. While sequences in DISPROT do show an increased fraction of disorder promoting residues, we suggest this increase is modest because a significant fraction of the sequences in the DISPROT database contain structured domains as well as disordered regions.

Internal scaling plots with error bars: In the interest of completeness, **Figures SI-3-5** show internal scaling plots for each of the three peptides with error bars.

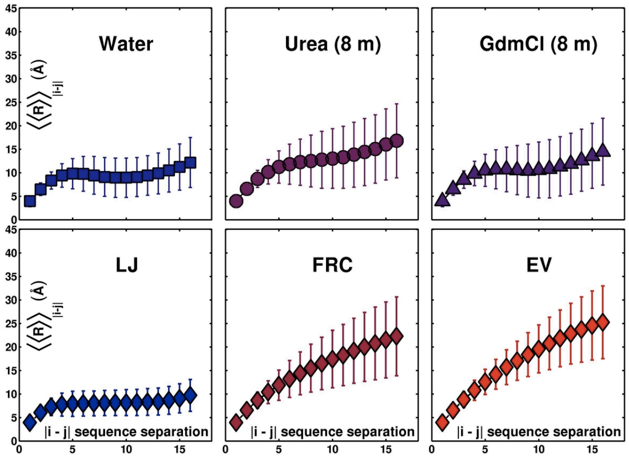

Figure SI-3: Internal scaling profiles for G_{15} in each of the three milieus (top row) and in the three reference ensembles (bottom row).

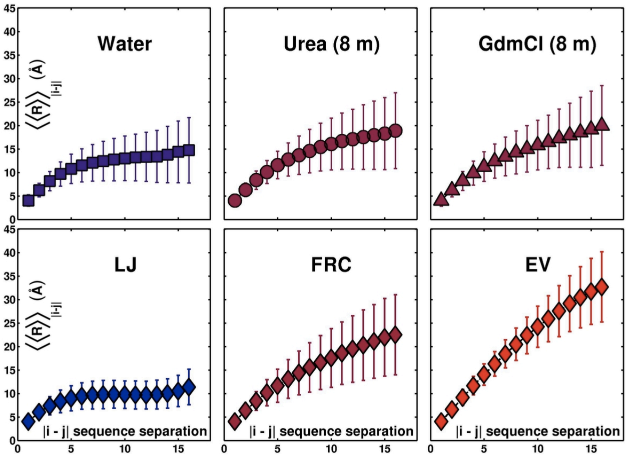

Figure SI-4: Internal scaling profiles calculated over the backbone atoms of CAP in each of the three milieus (top row) and in the three reference ensembles (bottom row).

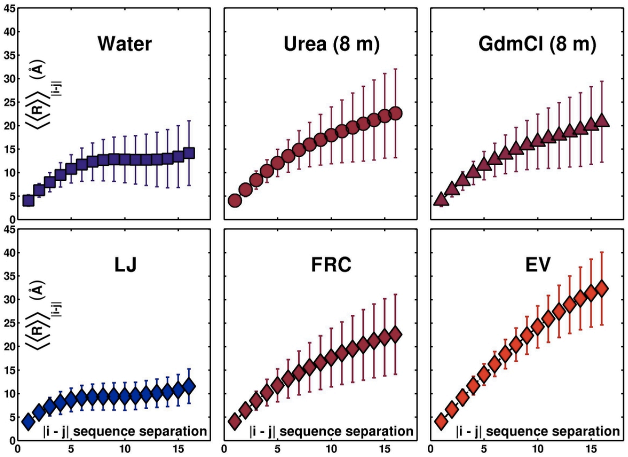

Figure SI-5: Internal scaling profiles calculated over the backbone atoms of OSP in each of the three milieus (top row) and in the three reference ensembles (bottom row).

References

(1) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. R.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C. H.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, M.; Garner, E. C.; Obradovic, Z. *J. Molec. Graph. Model.* **2001**, *19*, 26.

(2) Campen, A.; Williams, R. M.; Brown, C. J.; Meng, J.; Uversky, V. N.; Dunker, A. K. *Protein Pept. Lett.* **2008**, *15*, 956.

(3) Apweiler, R.; Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alam-Faruque, Y.; Alpi, E.; Antunes, R.; Arganiska, J.; Casanova, E. B.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Chan, W. M.; Chavali, G.; Cibrian-Uhalte, E.; Da Silva, A.; De Giorgi, M.; Fazzini, F.; Gane, P.; Castro, L. G.; Garmiri, P.; Hatton-Ellis, E.; Hieta, R.; Huntley, R.; Legge, D.; Liu, W.; Luo, J.; MacDougall, A.; Mutowo, P.; Nightingale, A.; Orchard, S.; Pichler, K.; Poggioli, D.; Pundir, S.; Pureza, L.; Qi, G.; Rosanoff, S.; Sawford, T.; Shypitsyna, A.; Turner, E.; Volynkin, V.; Wardell, T.; Watkins, X.; Zellner, H.; Corbett, M.;

Donnelly, M.; Van Rensburg, P.; Goujon, M.; McWilliam, H.; Lopez, R.; Xenarios, I.; Bougueleret, L.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Binz, P.-A.; Blatter, M.-C.; Boeckmann, B.; Bolleman, J.; Boutet, E.; Breuza, L.; Casal-Casas, C.; de Castro, E.; Cerutti, L.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; James, J.; Jungo, F.; Keller, G.; Lara, V.; Lemercier, P.; Lew, J.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T.; Paesano, S. *Nucl. Acid. Res.* **2014**, *42*, D191.

(4) Oates, M. E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztányi,S., Uversky,V.N., Obradovic,Z., Kurgan,L., Dunker,A.K., Gough,J. *Nucl. Acid. Res.* **2013**, *41*, D508.

(5) Sickmeier, M.; Hamilton, J. A.; LeGall, T.; Vacic, V.; Cortese, M. S.; Tantos, A.; Szabo, B.; Tompa, P.; Chen, J.; Uversky, V. N.; Obradovic, Z.; Dunker, A. K. *Nucl. Acid. Res.* **2007**, *35*, D786.

(6) Griep, S.; Hobohm, U. *Nucl. Acid. Res.* **2010**, *38*, D318.