# Supporting Information for: "Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins"

Wenwei Zheng[1a], Gregory Dignon[2], Matthew Brown [1], Young C. Kim [3] and Jeetain Mittal [2b]

[1] College of Integrative Sciences and Arts, Arizona State University, Mesa, AZ 85212, USA

[2] Department of Chemical and Biomolecular Engineering, Bethlehem, PA 18015, USA

[3] Center for Materials Physics and Technology, Naval Research Laboratory, Washington, DC 20375, United States

[a]Electronic mail: wenweizheng@asu.edu

[b]Electronic mail: jeetain@lehigh.edu

## 1. Supplementary Methods

**Molecular simulation methods.** The simulations of sequences in the database (Fig. S1) were conducted using the HOOMD-Blue v2.1.5.[1] For each sequence, the simulation were run for 500 ns using a Langevin thermostat with a friction coefficient of 1 $ps^{-1}$, and the first 100 ns was dumped as equilibration. A time step of 10 fs and a temperature of 300 K were used for all the simulations. The ionic strength is set to be 100 mM and characterized using Debye-Hückle electrostatic screening [2]. There are three types of interactions, including bonded, electrostatic and short-range pairwise interaction term characterized by amino acid hydropathy.[3] Bonded interactions are modelled by a harmonic potential with a spring constant of 10 kJ/Å$^2$ and a bond length of 3.8 Å. Electrostatic interactions are modeled using a Coulombic term with Debye-Hückel [2] electrostatic screening,

$$E_{ij}(r) = \frac{q_i q_j}{4\pi D r} \exp(-r/\kappa), \tag{S1}$$

in which $\kappa$ is the Debye screening length and $D = 80$, the dielectric constant of the solvent. The short-range pairwise interaction is modelled using Ashbaugh-Hatch functional form[4],

$$\Phi(r) = \begin{cases} \Phi_{LJ} + (1-\lambda)\epsilon, & \text{if } r \leq 2^{1/6}\sigma \\ \lambda\Phi_{LJ}, & \text{otherwise} \end{cases} \tag{S2}$$

in which $\Phi_{LJ}$ is the standard Lennard-Jones potential

$$\Phi_{LJ} = 4\epsilon\Big[\Big(\frac{\sigma}{r}\Big)^{12} - \Big(\frac{\sigma}{r}\Big)^{6}\Big]. \tag{S3}$$

The $\lambda$ value in the pairwise interaction term is the arithmetic average of the $\lambda$ values of the two amino acids (Table S1). We refer the readers to the literature of our HPS model for the details of the coarse-grained model.[5]

## References


[1] Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comput. Phys.* **2008**, *227*, 5342–5359.
[2] Debye, P.; Hückel, E. *Physikalische Zeitschrift* **1923**, *24*, 185–206.
[3] Kapcha, L. H.; Rossky, P. J. *J. Mol. Biol.* **2014**, *426*, 484–498.
[4] Ashbaugh, H. S.; Hatch, H. W. *J. Am. Chem. Soc.* **2008**, *130*, 9536–9542.
[5] Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. *PLoS Comput. Biol.* **2018**, *14*, e1005941.
[6] Das, R. K.; Pappu, R. V. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392–13397.
[7] Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. *Science* **2017**, *358*, 238–241.
[8] Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155–16160.
[9] Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schuler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Best, R. B.; Schuler, B. *J. Am. Chem. Soc.* **2016**, *138*, 11714–11726.
[10] Fuertes, G.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E6342–E6351.

[11] Martin, E. W.; Holehouse, A. S.; Peran, I.; Farag, M.; Incicco, J. J.; Bremer, A.; Grace, C. R.; Soranno, A.; Pappu, R. V.; Mittag, T. *Science* **2020**, *367*, 694–699.
[12] Rieloff, E.; Skepö, M. *J. Chem. Theory Comput.* **2020**, *16*, 1924–1935.
[13] Cragnell, C.; Durand, D.; Cabane, B.; Skepö, M. *Proteins* **2016**, *84*, 777–791.

## 2. Supplementary Figures

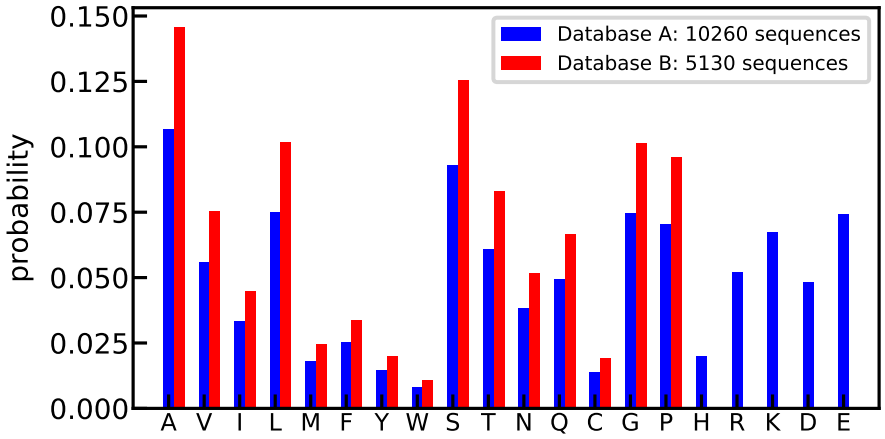


Figure S1. Amino acid probabilities for the randomly generated sequences. Database A contains 10260 charged random sequences (blue), in which there are 60 sequences for each of the chain length from 30 to 200, and database B contains 5130 random sequences without charged amino acids (red), in which there are 30 sequences for each of the chain length.

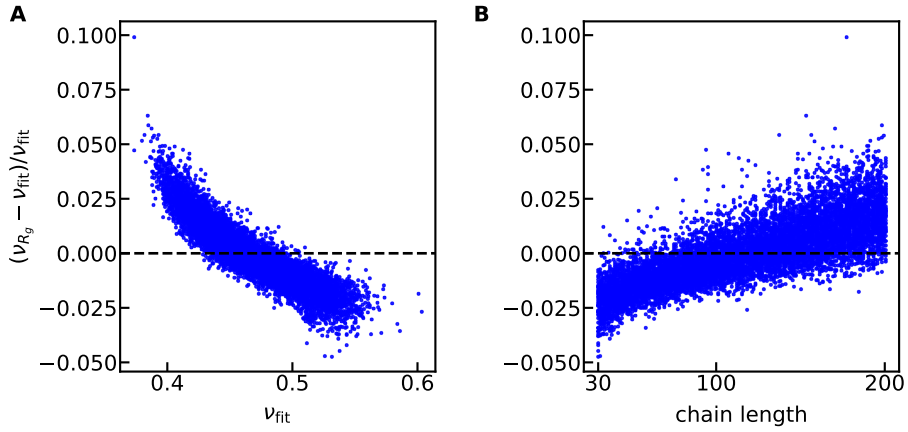


Figure S2. Relative difference between $\nu_{\mathrm{R_g}}$ and $\nu_{\mathrm{fit}}$ as a function of $\nu_{\mathrm{fit}}$ (A) and chain length (B).

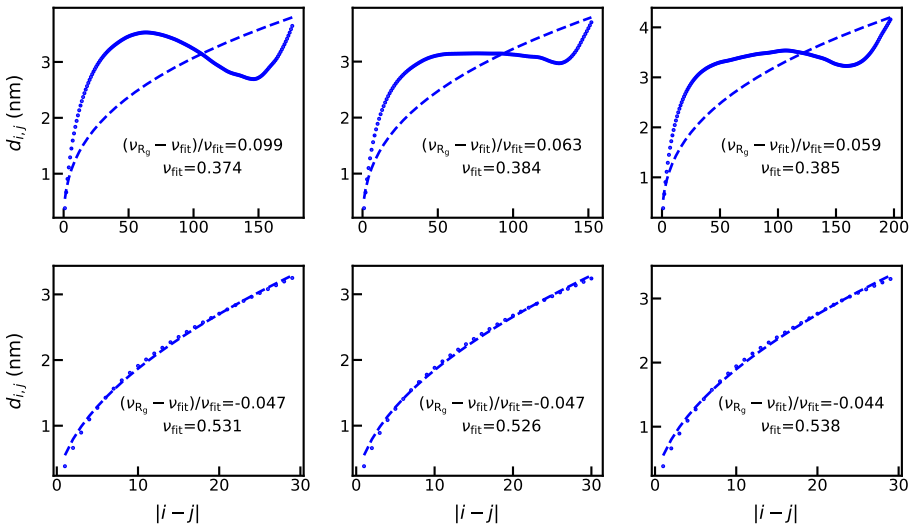


FIGURE S3. Intrachain distances as a function of the sequence separations for the three cases with most deviations between $\nu_{R_g}$ and $\nu_{\rm fit}$: $\nu_{R_g} > \nu_{\rm fit}$ (top row) and $\nu_{R_g} < \nu_{\rm fit}$ (bottom row). Dotted lines come from simulations and dashed lines show the exponential fitting curves.

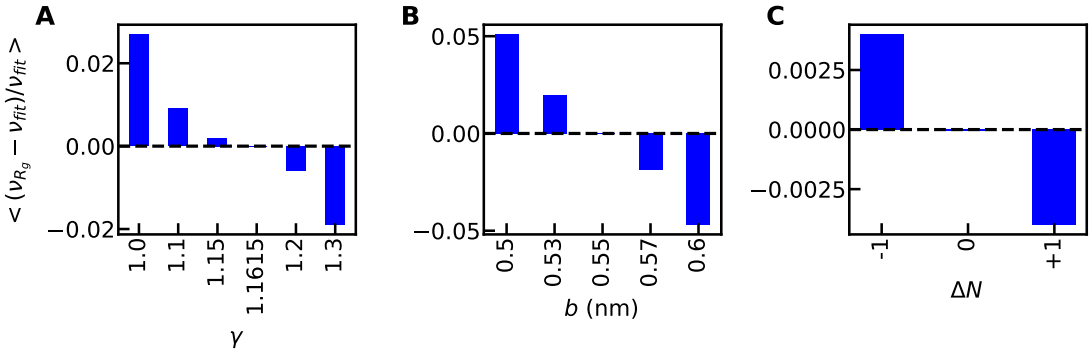


FIGURE S4. Average relative difference between $\nu_{\rm R_g}$ and $\nu_{\rm fit}$ as a function of $\gamma$ (A), $b$ (B) and variaton of $N$ (C) in Eq. 2 of the main text. $\Delta N = 0$ suggests that $N$ is defined as the number of peptide bonds and $\Delta N = 1$ suggests $N$ is defined as the chain length.

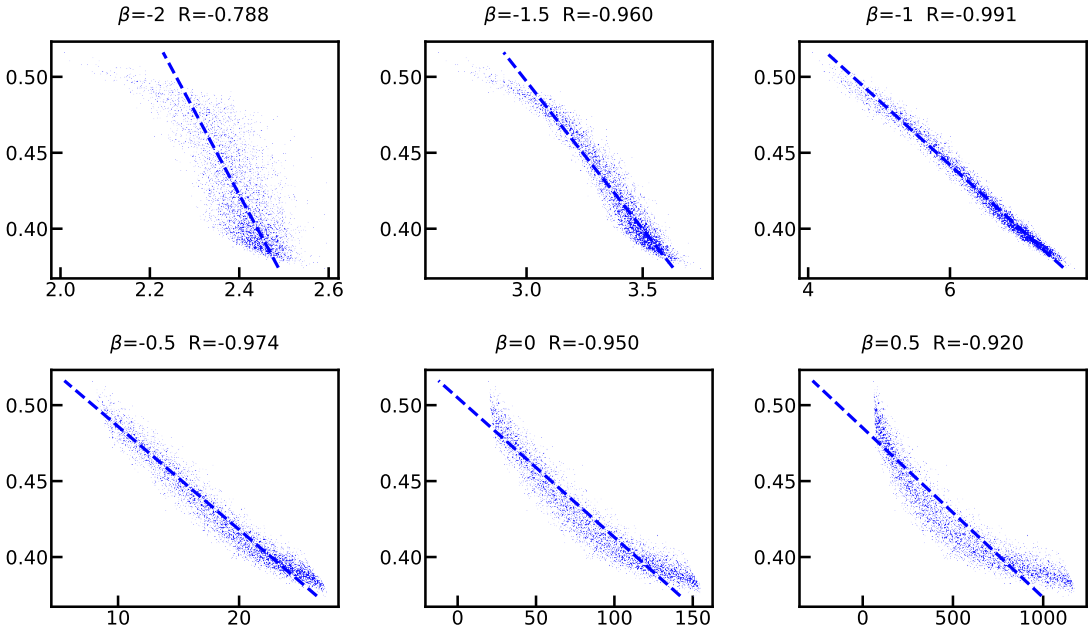


FIGURE S5. Scanning $\beta$ in defining $SHD$ (Eq. 4 of the main text). $R$ is the Pearson correlation coefficient between the two variables.

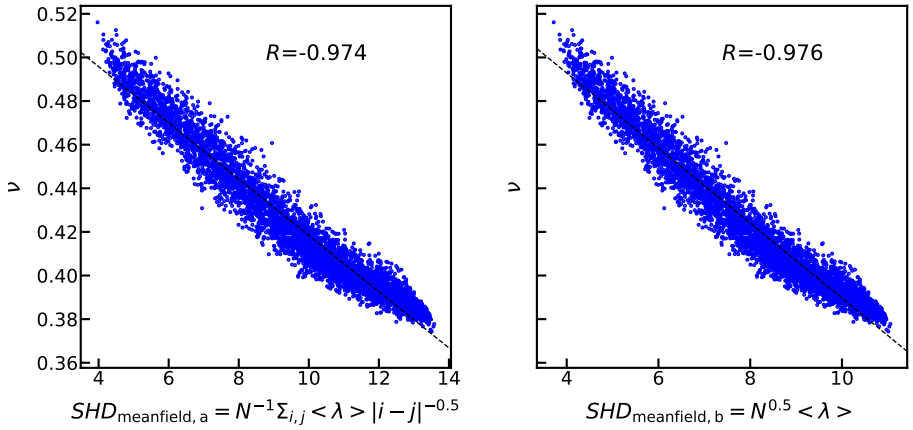


FIGURE S6. Mean field definitions of $SHD$ using average hydropathy instead of residue specific hydropathy.

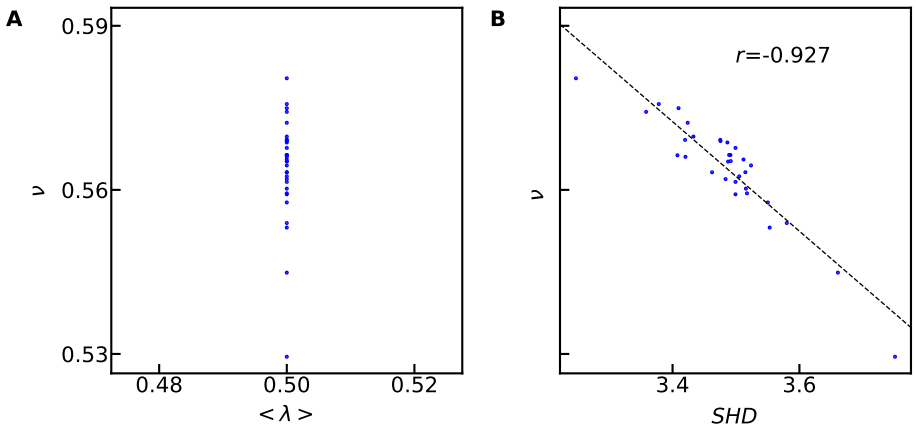


FIGURE S7. Comparison between using the mean hydrophathy $< \lambda >$ (A) and the hydropathy patterning parameter $SHD$ (B) to capture the scaling exponents of binary sequences (Table S2). The dash lines show the linear fitting between $SHD$ and $\nu$ and the legends show the Pearson correlation coefficients.

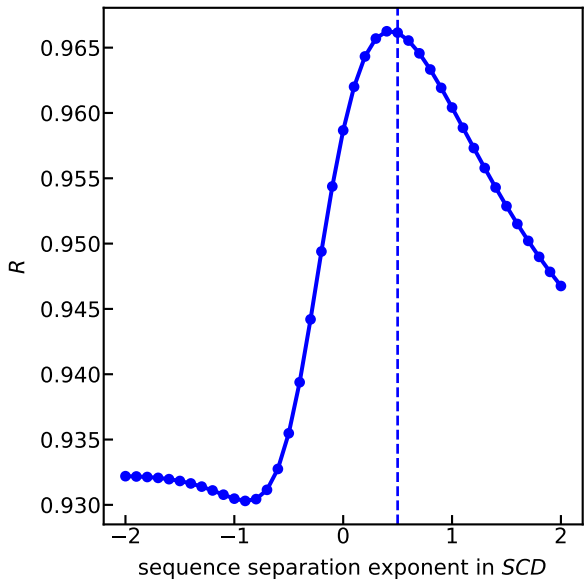


FIGURE S8. Test of the sequence separation exponent in $SCD$. We use random sequence database A (Fig. S1) for the test. For each scanned exponent in $SCD$, $SHD$ is always calculated using an optimal -1 exponent whereas $SCD$ is calculated using a different exponent shown in the x-axis. A multilinear regression is then applied to find the best parameters for using $SCD$ and $SHD$ to fit $\nu$. The Pearson correlation coefficient between the modeled and reference $\nu$ is shown.

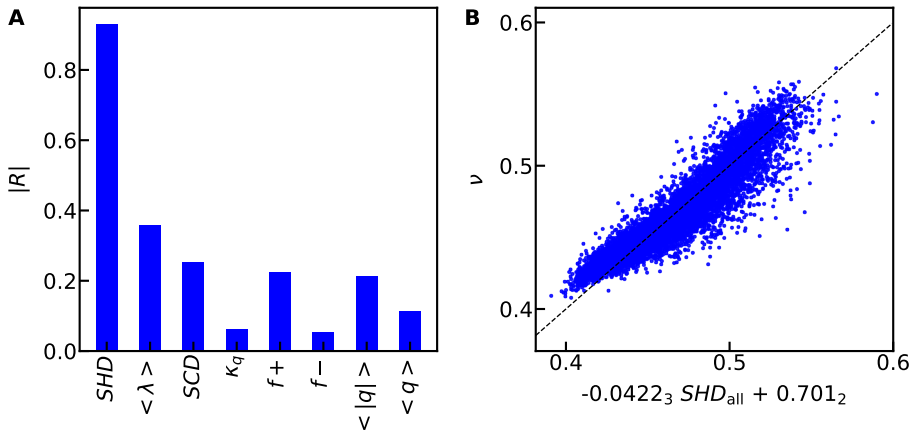


FIGURE S9. Capturing the scaling exponents using one sequence descriptor. A) Pearson correlation coefficients between the sequence descriptors and simulated scaling exponents. B) The best one sequence-descriptor model with the linear equation shown in labels of x-axis.

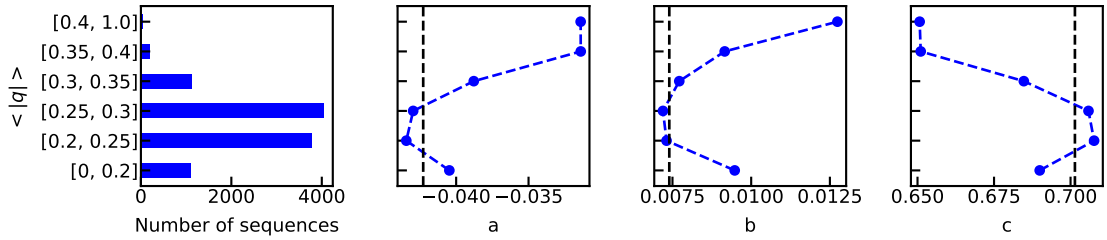


FIGURE S10. Left: Number of sequences of the database containing charged sequences (Fig. S1) at different range of $< |q| >$ values. Right: The three fitting free parameters in the multi-linear equations of $SHD$ and $SCD$ (Eq. 5 of the main text) when using sequences at different range of $< |q| >$ values. The black dashed lines show the fitting parameters using all the sequences.

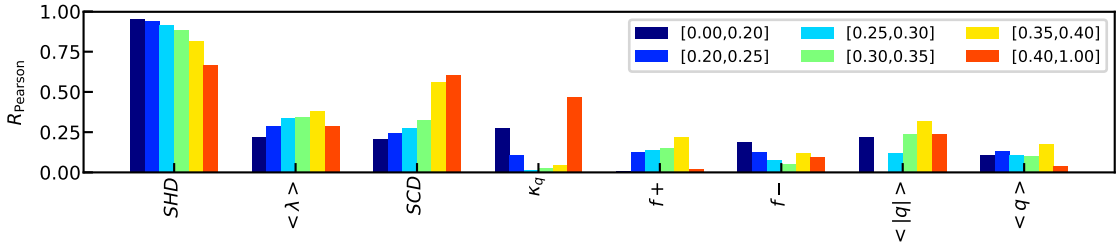


FIGURE S11. Pearson correlation coefficients between the sequence descriptors and $\nu$ for sequences with different range of $< |q| >$ shown in the legend.

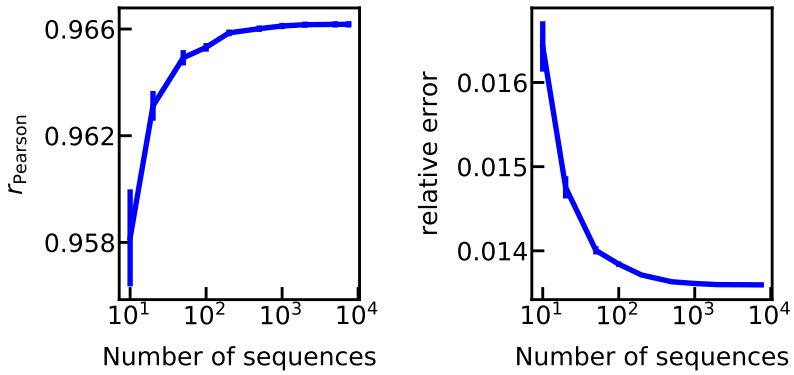


FIGURE S12. Pearson correlation coefficient (left) and relative errors (right) for assessing the model, when increasing the number of sequences. The error bars are estimated from randomly splitting the database for the training and test sets 100 times.

## 3. Supplementary Tables

Table S1. The amino acid hydropathy scales [3]. It has been rescaled within the range of 0 to 1.

| Amino acid | Hydropathy |
|---|---|
| ALA | 0.730 |
| ARG | 0.000 |
| ASN | 0.432 |
| ASP | 0.378 |
| CYS | 0.595 |
| GLN | 0.514 |
| GLU | 0.459 |
| GLY | 0.649 |
| HIS | 0.514 |
| ILE | 0.973 |
| LEU | 0.973 |
| LYS | 0.514 |
| MET | 0.838 |
| PHE | 1.000 |
| PRO | 1.000 |
| SER | 0.595 |
| THR | 0.676 |
| TRP | 0.946 |
| TYR | 0.865 |
| VAL | 0.892 |

TABLE S2. Compactness of 32 binary hydrophobic sequences. The binary sequences consist of two possible units, with a hydropathy of 0 and 1, respectively. So the sequences are written in numbers for illustration. The $R_g$ is obtained from simulations using HPS model.

| Sequences shown in numbers | $R_g$ (nm) | $\nu$ | $SHD$ |
|---|---|---|---|
| 10101010101010101010101010101010101010101010101010 | 2.040 | 0.568 | 3.499 |
| 11100011100011100011100011100011100011100011100010 | 2.024 | 0.565 | 3.488 |
| 01000100110011010101011000110101010001101011001111 | 2.049 | 0.569 | 3.475 |
| 01010011010011100101010001100011011001100011011101 | 2.048 | 0.569 | 3.476 |
| 01011010001111010000110101010110011000011011010101 | 2.046 | 0.569 | 3.486 |
| 11100100110110010010111000101100111000101111000010 | 2.032 | 0.566 | 3.491 |
| 11110000111100001111000011110000111100001111000010 | 2.004 | 0.562 | 3.484 |
| 00001111000011110000111100001111000011110000111101 | 2.011 | 0.563 | 3.515 |
| 11001110101011111001001001000110101000100001011101 | 2.049 | 0.569 | 3.420 |
| 10000001100011111000111000100110101101001001101111 | 2.011 | 0.563 | 3.462 |
| 10100000111001011110111100000101110110011010001100 | 1.977 | 0.558 | 3.550 |
| 10011111101001111010100101101001000100111010000100 | 2.019 | 0.564 | 3.524 |
| 01000101001000111000111010001100100100111111101101 | 2.032 | 0.566 | 3.489 |
| 10010110111100000110100100001000001111110110101011 | 2.070 | 0.572 | 3.424 |
| 00100100010010011101010010000101001111111101100111 | 2.024 | 0.565 | 3.492 |
| 10101100011000010010110010101001111111110100100001 | 2.007 | 0.562 | 3.505 |
| 10100000101000010100100101110110101001100111111111 | 2.029 | 0.566 | 3.420 |
| 01100111111101100000100010011100011000111111000010 | 2.026 | 0.566 | 3.512 |
| 11111000001111100000111110000011111000001111100000 | 2.000 | 0.561 | 3.499 |
| 11011111101110110011101001001011001000000000000111 | 2.031 | 0.566 | 3.408 |
| 11111111101000001011000000100100001001111110111000 | 2.087 | 0.575 | 3.409 |
| 01111011011000010110100000000000010011111111010111 | 2.092 | 0.576 | 3.378 |
| 11111011111111111101101000001000000010100001001100 | 2.053 | 0.570 | 3.433 |
| 11101111101111111111110001100000100000001000000000 | 1.993 | 0.560 | 3.516 |
| 11111111111011110110110100000000000000000011001101 | 2.083 | 0.574 | 3.358 |
| 01111111011011111111101111011000000000000000000001 | 1.987 | 0.559 | 3.517 |
| 00100010011111111111111111111011000000000000000100 | 1.899 | 0.545 | 3.661 |
| 10000000000000000000001111111111111111111011111010 | 1.948 | 0.553 | 3.553 |
| 01111011111111111111111111100000000000000000000000 | 1.953 | 0.554 | 3.580 |
| 11111111111111111111111100000000000000000000000000 | 1.986 | 0.559 | 3.499 |
| 00000000000001111111111111111111111111000000000000 | 1.809 | 0.529 | 3.751 |
| 11111111111100000000000000000000000000111111111111 | 2.123 | 0.580 | 3.248 |

TABLE S3. List of sequence descriptors we have tested.

| Sequences descriptors | Symbol |
|---|---|
| Charge properties | |
| Net charges per residue | $< q >$ |
| Fraction of charged residues | $< \|q\| >$ |
| Fraction of positively charged residues | $f+$ |
| Fraction of negatively charged residues | $f+$ |
| Sequence charge decoration | $SCD = N^{-1}\Sigma_i\Sigma_{j,j>i}(q_i q_j)\|j-i\|^{1/2}$ |
| $\kappa$ | see reference[6] |
| Hydrophathy properties | |
| Average hydrophathy | $< \lambda >$ |
| Sequence hydrophathy decoration | $SHD = N^{-1}\Sigma_i\Sigma_{j,j>i}(\lambda_i + \lambda_j)\|j-i\|^{-1}$ |

TABLE S4. List of sequences with experimentally determined $R_g$. [a)] Since both FRET and SAXS measurements were provided in these two publications and were shown to be consistent, we only reanalyze the FRET data set here with $R_0$=5.4 nm and dye length correction of 9 residues. [b)] Molecular form factor method [7]. [c)] Eq. 1 in the main text is used to calculate $\nu$ from reported experimental Rg values.

| Name | Chain length | $R_g$ (nm) | $\nu$ |
|---|---|---|---|
| Hofmann et al [8], FRET, SAW-$\nu$ | | | |
| CSP | 71 | 2.226 (0.102) | 0.535 (0.013) |
| R15 | 114 | 2.612 (0.087) | 0.511 (0.008) |
| hCyp | 167 | 2.534 (0.078) | 0.458 (0.007) |
| IN | 60 | 2.221 (0.063) | 0.562 (0.009) |
| ProT$\alpha$-N | 55 | 2.549 (0.076) | 0.619 (0.009) |
| ProT$\alpha$-C | 55 | 2.998 (0.134) | 0.668 (0.013) |
| Borgia et al [9], FRET and SAXS, SAW-$\nu$ [a] | | | |
| ACTR | 79 | 2.510 (0.044) | 0.552 (0.005) |
| R17d | 118 | 2.817 (0.056) | 0.525 (0.005) |
| Fuertes et al [10], FRET and SAXS, SAW-$\nu$ [a] | | | |
| N49 | 37 | 1.411 (0.072) | 0.494 (0.018) |
| NLS | 45 | 1.669 (0.062) | 0.518 (0.012) |
| NUS | 81 | 2.439 (0.075) | 0.538 (0.008) |
| IBB | 98 | 2.544 (0.076) | 0.523 (0.008) |
| NUL | 113 | 2.680 (0.081) | 0.517 (0.008) |
| Riback et al [7], SAXS, MFF [b] | | | |
| pNT | 334 | 5.11 | 0.542 |
| fHua | 144 | 3.34 | 0.543 |
| RNasea | 124 | 3.36 | 0.545 |
| Martin et al [11], SAXS, MFF [b] | | | |
| hnRNPA1 | 137 | 27.2 (0.2) | 0.45 (0.004) |
| hnRNPA1 Aro- | 137 | 28.9 (0.1) | 0.48 (0.003) |
| hnRNPA1 Aro– | 137 | 30.1 (0.1) | 0.51 (0.005) |
| hnRNPA1 Aro+ | 137 | 24.4 (0.6) | 0.41 (0.02) |
| Rieloff et al [12] , SAXS, pair distance distribution [c] | | | |
| SN15n | 15 | 0.99 (0.01) | 0.561 (0.006) |
| Cragnell et al [13] , SAXS, pair distance distribution [c] | | | |
| His5 | 24 | 1.38 (0.004) | 0.584 (0.001) |