Supplementary Information:

Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers

Felipe García Quiroz\textsuperscript{1,2} and Ashutosh Chilkoti\textsuperscript{1}

\textsuperscript{1}Department of Biomedical Engineering, Duke University, Durham, North Carolina, 27708-0281, USA

\textsuperscript{2}Present address: Howard Hughes Medical Institute, Robin Chemers Neustein Laboratory of Mammalian Cell Biology and Development, The Rockefeller University, New York, New York 10065, USA.

1. Supplementary Methods

Materials

Restriction endonucleases, calf intestinal alkaline phosphatase (CIP) and T4 DNA ligase (400 U/ul) were purchased from New England Biolabs. PfuUltra\textsuperscript{TM} II Fusion HS DNA polymerase (Pfu) was purchased from Stratagene; Go Taq Green Master Mix\textsuperscript{TM} was purchased from Promega; Circligase\textsuperscript{TM} ssDNA ligase was purchased from Epicentre Biotechnologies, and T4 DNA ligase (1 Weiss U/µl) was purchased from Fermentas. Plasmid DNA was purified using Qiaprep\textsuperscript{TM} spin miniprep kits and PCR products were purified using the Qiaquick\textsuperscript{TM} PCR purification kit from QIAGEN. Double stranded DNA was visualized by electrophoresis on agarose gels (EMD chemicals) prestained with SYBR\textsuperscript{®} Safe (Invitrogen). BL21(DE3) and EB5α E. coli strains were purchased from EdgeBio. All cultures were grown in Terrific Broth (TB) from BioExpress, and expression was induced with isopropyl β-D-thiogalactopyranoside (IPTG) from GoldBio. Custom oligonucleotides were synthesized by IDT (Integrated DNA Technologies) and proteins were visualized with Precast SDS–PAGE gels (BioRad).

Compositional analysis of natural Pro and Gly-rich proteins
We implemented custom-made scripts for analysis of proteomes and protein sequences using MATLAB R2013a. The full names and sequence identification numbers (gi) for the NCBI (National Center for Biotechnology Information) record of each of the prototypical Pro- and Gly-rich proteins studied herein are reported in Table S1. The amino acid sequence of these proteins was retrieved from the NCBI and saved as .TXT files for analysis. Full proteomes were downloaded as FASTA files from UniProt (http://www.uniprot.org/) and proteins selected from these large protein datasets are directly presented according to their UniProt identifier. Script 1 (Supplementary Methods) quantifies the fraction of P-X\textsubscript{n}-G motifs (where X is any residue but Pro and Gly) for each n value with respect to the total number of motifs (i.e., n=0-4). Using a similar strategy to the one shown in Script 1 to store the residue positions of all Pro (vector ‘pstarts’) and Gly (vector ‘gends’) residues participating in P-X\textsubscript{n}-G motifs (for n=0-4), we also calculated the average distance between the Gly residue in motif i-1 and the Pro residue in motif i (Script 2), as well as the amino acid composition of these inter-residue segments. For proteome-wide analysis, Script 3 scanned a proteome to generate a list of the subset of proteins with a given Pro and Gly content as well as other characteristics of interest (e.g., length, removal of elastins and collagens, etc). This information, stored as text file, is then processed by a variant of Script 1 that analyses the composition of all proteins at once (Script 4).

For the compositional analysis of resilins and tropoelastins, Script 5 (Supplementary Methods) calculated the percent abundance of all amino acids in these proteins excluding Pro and Gly. We excluded these two residues because they are highly and similarly abundant in these proteins and because their fundamental role is to contribute to the unstructured character of these proteins and not to their hydropathy balance. Script 5 also calculated the hydropathy index for each amino acid according to the Kyte-Dolittle scale of hydropathy 1.

**Proteome-wide search for proteins with features of UCST IDPPs**
To identify proteins with potential UCST phase behavior, using a minor variant of Script 3 we scanned the human proteome for proteins that fulfill a set of 5 sequence parameters that control UCST phase behavior. The minimal threshold values for these parameters are those derived from the compositional analysis of IDPPs with validated UCST phase behavior and from the compositional analysis of resilins and LC domains. These critical sequence parameters are: i) protein length, ii) charge content (R + K + D +E), iii) percent of positively charged residues (R + K) with respect to all (R +K + D+ E) charged residues—as a measure of the zwitterionic character of the protein—, iv) aromatic content (Y+H+W+F) and v) percent of Arg residues with respect to all (R+ K) positively charged residues. In the Supplementary Text we provide a detailed description for the minimal threshold values that we chose for our proteome-level screen.

**Genetically encoded synthesis of protein polymers**

**Overlap-extension rolling circle amplification (OERCA).** To enable the rapid screening of a large number of protein polymers in our exploration of amino acid sequence determinants of IDPPs with phase behavior, we took advantage of our recently developed method for the parallel synthesis of peptide polymers, OERCA. The large majority of polymers reported in this manuscript were synthesized using this methodology (Table S2 and Table S4) and following the instructions described in reference 2. Briefly, the oligomer sequences shown in Table S2 were circularized using CircLigase I and un-circularized DNA was removed by restriction digestion with Exonuclease I. The purified circular DNA was subjected to 40 cycles of OERCA in the presence of the corresponding forward and reverse primers as listed in Table S2 and Table S4. The degree of extension was assessed by DNA electrophoresis and when necessary (i.e., when visible products were≤ 400 bp in length) constructs were further extended with 10 cycles of OE-PCR in the absence of primers 2. These OERCA products were blunt ligated (using T4 DNA ligase from Fermentas) into a modified pET25 vector that incorporates an N-terminal MSKGP sequence and a C-terminal HHHHHHYG tag to all protein polymers synthesized by OERCA. The ligation mixture was then transformed
into BL21 cells. The resulting colonies were screened by directional colony PCR (using T7R and the forward primer specific to each polymer) and positive clones were grown overnight, and their plasmids were purified using a plasmid miniprep kit (Qiagen), which were then further screened by restriction endonuclease digestion and DNA sequencing (Eton Bioscience Inc., NC, USA). This resulted in a library of genes encoding for protein polymers of varying length for each repeat unit in Table S2 and Table S4. Only clones with fully confirmed DNA sequences were retained for further study.

Table S2 and S4 only report the length of the forward and reverse primers for brevity. All primers were designed to flank the two ends of the DNA oligomer. Hence, the DNA sequence of these primers can be easily extracted from the 5’-end (forward primer) and 3’-end (reverse primer) of each oligomer. Note that the DNA sequence of the reverse primer is the reverse complement of the respective sequence in the oligomer.

**Recursive Directional Ligation by Plasmid reconstruction (PRe-RDL).** For the synthesis of the large genes encoding SM1 and SM2 (each composed of 144 repeats of some form of a hexapeptide motif), as well as genes encoding for additional IDPPs and copolymers therefrom, we used the plasmid reconstruction variant (PRe-RDL) of the recursive directional ligation (RDL) method as it offers a remarkably high cloning efficiency at each step of gene assembly 3. The oligomers used for the synthesis of these genes are reported in Table S3 and Table S5. The overall methodology was as originally described by McDaniel and collaborators 3. Briefly, the oligomers were annealed to form double stranded DNA cassettes with CC and GG overhangs that enable the concatenemerization of the cassettes and their eventual ligation (using T4 DNA Ligase from NEB) into a modified pET24 (JMD3) vector that incorporates a short trailer sequence encoding for the GWP peptide. This led to the construction of a small library of plasmids with 2-4 copies of the original cassettes, which were then assembled through multiple cycles of PRe-RDL to the desired length. The final genes were subjected to an additional cycle of Pre-RDL to incorporate a short leader sequence encoding for the MSKGP peptide and a trailer sequence encoding for a His-tag (HHHHHHYG). We did not include
a His-tag in diblock copolymers and in polymers used for zeta potential measurements (Supplemental Fig. 10a). The multiple cycles of gene assembly were typically carried out using EB5α cells, whereas the final, full-length gene in the corresponding plasmid was transformed into BL21 cells for expression.

**Expression and purification of IDPPs**

**Expression.** Starter cultures (5 mL starter culture per 1 L of E. coli cells) of Terrific broth (TB) media (BioExpress) supplemented with 100 μg/mL ampicillin (for all peptide polymers synthesized by OERCA) or 45 μg/mL Kanamycin (for peptide polymers prepared by PRe-RDL) were inoculated with transformed cells from DMSO stocks stored at −80 °C, and incubated overnight at 37 °C while shaking at 250 rpm. The starter cultures were then centrifuged at 3000 g for 7 min and resuspended in 1 mL of fresh TB medium. Expression cultures (4 L flasks containing 1 L of TB media with ampicillin or kanamycin as in the starter cultures) were inoculated with the resuspended starter culture and incubated at 37 °C with shaking at 200 rpm. After 8-9 h of growth, expression was induced by the addition of IPTG to a final concentration of 1 mM. Cells were harvested 24 h after inoculation, resuspended in PBS and sonicated for a total time of 1.5 min at ~80 W to induce cell lysis.

**Purification.** IDPPs were purified from cell lysates typically pre-treated with 10% PEI to remove contaminating DNA. Polymers that exhibited LCST were purified by conventional inverse transition cycling (ITC) —with minor modifications— as described elsewhere 4, whereas the purification of polymers that exhibited UCST was carried out with a new purification strategy that exploited the UCST behavior of the protein polymers. Because of the high yield expression of these UCST IDPPs in *E. coli*, they form an insoluble phase that can be readily collected from the insoluble fraction obtained after sonication and centrifugation. Briefly, cells were sonicated and centrifuged at 14000 rpm for 15 min at 4 °C. The clear supernatant and the viscous phase were discarded and the pellet was resuspended in 3-5 ml of PBS or PBS supplemented with 2-4 M urea. Because urea supplementation lowers the UCST cloud point of IDPPs, polymers with
very low cloud points are readily solubilized in PBS but supplementation with 4 M urea is preferred for most UCST IDPPs with physiologically relevant UCST behavior to eliminate the need to work at high—above the cloud point—temperatures throughout the purification process. These solutions were heated to 65 °C for 10 min and centrifuged at 15000 rpm for 15 min at room temperature. The supernatants were further purified by one or two rounds of UCST phase separation, here referred as UCST transition cycling, as follows. The samples supplemented with urea were diluted 0.5 to 1 times their volume in PBS and heated to 65 °C for 10 min, followed by centrifugation at room temperature for 15 min at 15000 rpm. The resulting supernatant was incubated on ice for 5 min, typically followed by 2 min at -20 °C, to trigger a phase transition that is visible by an increase in turbidity. The coacervate was then harvested by centrifugation at 14000 rpm for 3 min at -4 °C and the pellet was resuspended in PBS supplemented with 2-4 M Urea. This solution was then heated to 65 °C for 5 min and centrifuged at 15000 rpm for 5 min at room temperature to remove any contaminants. Although most IDPPs were pure at this stage, we often conducted an additional cycle of UCST transition cycling.

For the protein polymers with repeating Cys residues (SM1 and SM2), purification was carried out in 20 mM TCEP (pH 7.0) to prevent undesired disulfide bond formation throughout the purification process. These polymers were extensively dialyzed against water to remove the TCEP and were lyophilized for storage at -20 °C. Polymers of VPSALYGVG, VGPVG, VGPAVG, VRPVG, hEndo and mEndo were purified from the insoluble fraction obtained after sonication and centrifugation. Briefly, cells were sonicated and centrifuged at 14000 rpm for 15 min at 4 °C. The supernatant was discarded and the pellet was resuspended in 15 ml 6M GnCl2. These solutions were centrifuged at 14000 rpm for 30 min at 4 °C and the supernatants were extensively dialyzed against water at 4 °C. These IDPPs were then purified through regular ITC in PBS, with the exception of polymers of VPSALYGVG, which were purified in PBS with 8 M Urea.
Characterization of the phase transition behavior, zeta potential and secondary structure of IDPPs

To characterize the phase behavior of the synthesized protein polymers, the optical density of peptide polymer solutions (at the concentrations indicated in the manuscript and figures) was monitored at a wavelength of 350 nm as a function of temperature, with heating and cooling performed at a rate of 1 °C min⁻¹, on a Cary 300 UV-visible spectrophotometer equipped with a multicell thermoelectric temperature controller (Varian Instruments, Walnut Creek, CA). To avoid artifacts from the handling of aggregated polymers, LCST-exhibiting polymers were first heated from temperatures below the LCST toward high temperatures (75-80 °C), whereas UCST-exhibiting polymers were first cooled from temperatures above the UCST (65-80 °C) toward low temperatures (5-15 °C). The LCST and UCST cloud points were calculated as the temperature at the maximum dAb₃₅₀/dT (i.e., the first derivative of the turbidity profiles). For most of the UCST IDPPs, the turbidimetry experiments were conducted in pure PBS but small amounts of residual urea were usually present in the range of 0.01-0.05 M — because of the dilution of concentrated samples stored in PBS supplemented with 2-4 M urea. To study the concentration dependence of the UCST behavior without varying amounts of residual urea, the final concentration of urea was fixed at 0.2 M for all polymer concentrations in Fig. 4d.

The secondary structure displayed by these protein polymers was studied by circular dichroism (CD) using an Aviv Model 202 instrument and 1 mm quartz cells (Hellma) by scanning from 260 nm to 180 nm with 1 nm steps and a 3 second averaging time at various temperatures. Purified polymers were dialyzed overnight against Milli-Q water, protein purity was assessed by SDS-PAGE and the polypeptides were diluted to 5 µM in water. Raw CD data in millidegrees was first corrected by subtracting the corresponding CD signal from water blanks and transformed into Mean Residue Ellipticity (θ) as reported elsewhere⁶.
The zeta potential of IDPPs was measured at a polymer concentration of 50 µM in both water (pH 6-7) and PBS (pH 2.5, pH 7.3 and pH 12.5) using a Malvern Zetasizer NanoZS with folded capillary cells (DTS1060, Malvern Instruments). The zeta potential values were calculated from electrophoretic mobility data by the Zetasizer Software (v. 7.11) using the Smoluchowski model.

**Evaluation of the cell adhesiveness of RGD-based IDPPs**

The rationale for this experiment is that the cell adhesiveness of the polymers in solution is expected to delay cell adhesion to an underlying tissue culture-treated polystyrene substrate by binding of the UCST IDPPs to integrins on the cell surface. Because we hypothesized that cell adhesiveness depended on the UCST behavior of the polymer, we compared a RGD-containing IDPP with an UCST cloud point below 37 °C for the concentration range used in the experiment, with a RGD-containing IDPP that has a UCST cloud point above 37°C over the same concentration range. To test their effect on cell adhesion, 5x10⁵ PC3-luc-C6 cells per well were seeded on a 24-well plate and immediately after, while at room temperature, the indicated amounts of the filter-sterilized UCST IDPPs (or PBS supplemented with urea) were added to the media. Because there is a residual concentration of urea of up to 0.1 M for the cultures with the highest polymer concentration, control cultures with matching concentrations of urea were also prepared. The cultures were incubated for either 3 or 18 h at 37 °C in 5% CO2. The media was then removed and the number of adherent cells was measured using a conventional MTT assay. Briefly, media was replaced and MTT (5mg/ml in PBS) was added to a final concentration of 0.5 mg/ml before incubating the cultures for 1 h at 37 °C/5% CO2 while protected from light. Cultures were then washed with PBS. The formazan crystals resulting from the metabolization of the MTT salt were solubilized by adding 200 µl DMSO per well. A 50 µl aliquot per well was transferred to a 96-well plate to measure the absorbance at 560 nm using a Victor³ microplate reader (PerkinElmer).
2. Supplementary Figures

Supplementary Fig. 1. Pro and Gly content of all Pro- and Gly-rich IDPs analyzed in Fig. 1 of the main manuscript.
Supplementary Fig. 2. Analysis of the average number of residues (i.e., average distance) that separate consecutive P-X<sub>n</sub>-G motifs (n=0-4) among (a) prototypical Pro- and Gly-rich IDPs and (b) among the top 9 human proteins with the most P-X<sub>n</sub>-G motifs from Fig. 1b. The left panel in (a) shows the characteristic recurrence of P-X<sub>n</sub>-G motifs in elastin and resilin, as evidenced through a digitalization of their protein sequences into a P-X<sub>n</sub>-G signal in which residues—other than Gly—participating in a P-X<sub>n</sub>-G motif are assigned a value of 1, while all other residues are set to 0. Proteins with low (average) number of residues between motifs have domains with a high density of P-X<sub>n</sub>-G motifs in their sequence. Proteins in (a) and (b) are identified by their UniProt accession numbers.
Supplementary Fig. 3. (a) Pro-rich proteins (>10% Pro) are rare in the proteomes of multi-cellular eukaryotes and virtually absent in prokaryotes. They only represent a small percentage of the proteins in a major database of IDPs (DisProt). (b) Abundance of proteins highly enriched in Gly—we set > 16% as an arbitrary threshold—as a function of Pro content.

Supplementary Fig. 4. (a) Amino acid composition of the spacer regions that link P-X_n-G motifs (where X is any amino acid different from Pro and Gly and n=0-4) for proteins
in Fig. 1. The percent abundance is calculated relative to the total length of the spacer regions for each protein. Unlike the biased composition of the spacer regions in canonical Pro and Gly-rich IDPs (shown in black), these same linkers in proteins with recurrent P-X_n-G motifs from the human proteome (in red) or from Disprot (mostly corresponding to human proteins; in blue) make use of the entire palette of amino acid residues at frequencies that are in accordance with the observed frequencies in the overall human proteome (b).

Supplementary Fig. 5. (a) Abundance of individual P-X_n-G motifs among the top 9 proteins (>10% Pro, >10% Gly, >50 residues) with the most number of P-X_n-G motifs in the IDP repository. (b) Average number of residues that separate consecutive P-X_n-G motifs (for n=0-4) among proteins in (a). Proteins are identified by their UniProt accession numbers.
Supplementary Fig. 6. Peptide motifs to explore the sequence space of P-Xₙ-G-containing polymers where X is any amino acid and n varies from 0 to 4. We recombinantly synthesized polymers of various lengths for each of these 40 peptide motifs and included two additional Pro-devoid, Ala mutants. Motifs colored red have an average hydropathy index that lies within the hydropathy range of tropoelastins, while those colored blue have a hydropathy index reminiscent of resilins (inset of Fig. 1d).
Supplementary Fig. 7. Pro-devoid UCST IDPPs are highly overexpressed in E. coli and are readily purified by UCST transition cycling. Copper-stained PAGE exemplifies the high yield purification of (GRGDSAY)16 (lanes A1-A4) and of (GRGDSAY)32 (lanes B1-B3) from E. coli cell lysates. Lanes A1/B1 are the cell lysates; lane A2 is the insoluble fraction of the cell lysate before centrifugation; lane A3/B2 is the resulting soluble fraction after centrifugation; lanes A4/B3 show the purity of the samples after a first cycle of UCST transition cycling (see Methods for details).

Supplementary Fig. 8. Temperature-dependent turbidity data as a function of concentration for polymers of (RGDAPYQG)28. (a) Raw turbidimetry data and (b) the
corresponding UCST values calculated as the maximum in the dAb/dT vs. Temperature plots (i.e., the first derivative of the turbidity profiles in panel (a)).

Supplementary Fig. 9. Zeta potential data for a subset of IDPPs. (a) Full amino acid sequence for each of the studied IDPPs and the corresponding zeta potential values in water as extracted from data in (b). The color assigned to each sequence is used across all panels in this figure. (b) Zeta potential distribution in water for a canonical LCST-exhibiting IDPP devoid of charged residues (VPGVG-80; at pH 6.0) and for several UCST IDPPs—at pH 7.0—with two oppositely charged residues per repeat. (b) Zeta potential as a function of pH in PBS. To visualize overlapping data at pH 2.5 and pH 12.5, some data points were displaced 0.25 pH units along the abscissa—samples were analyzed in PBS at pH 2.5, pH 7.3 or pH 12.55. Two measurements are typically shown for each polymer at each pH. All data were acquired at a concentration of 50 µM and at 40 °C.
Supplementary Fig. 10. A Lys-based IDPP exhibits a UCST transition in water but not in PBS. By substituting Lys (K) for Arg (R) in GRGDSPYG, we obtained a highly hydrophilic zwitterionic IDPP that only exhibits UCST in water. A polymer composed of 20 repeats of the original, hydrophobic motif exhibited a UCST at ~ 62 °C in PBS at 50 µM (Fig. 2a). In water, zwitterionic IDPPs experience attractive electrostatic interactions between oppositely charged residues that provide a UCST driving force that is negligible for polymers in PBS. Turbidimetry data was obtained at 200 µM.
Supplementary Fig. 11. Polymers of the cell adhesion motif GRGDSP display UCST phase behavior and bioactivity. (a) Two IDPPs composed of 12 and 20 repeats of the octapeptide GRGDSPYG were studied over a range of concentrations. (b) Temperature-dependent turbidimetry shows that polymers with n=12 (coded in blue) are soluble at 37 °C, whereas polymers with n=20 (coded in green) are insoluble at this temperature. (c-d) Effect of these polymers on the cell adhesion of 0.5x10^5 PC-3 cells to 24-well tissue-culture plates. The cells were allowed to adhere in the presence of varying concentrations of the UCST IDPPs for either 3 hours (c) or 18 hours (d), and the cultures were washed with PBS before quantifying cell metabolic activity (indicated by a metabolic product that absorbs strongly at ~560 nm). Control cultures (0 µM) were prepared by adding PBS supplemented with the maximum amount of residual urea (~0.1 M) in these experiments. Residual urea is typically left over after dilution of UCST IDPPs from highly concentrated stocks (2-8 mM) in PBS supplemented with 2-4 M Urea. Additional controls tested for the effect of the residual urea on cell adhesion and showed that 100 mM urea had no effect on cell viability (data not shown).
Supplementary Fig. 12. Phase behavior of IDPPs built from matrikine motifs. We chose the Gly-X-X-Pro-Gly motif because specific peptide motifs that belong to this class are matrikines that bind the elastin receptor to exert a variety of biological functions.\(^7\) (a) We synthesized LCST IDPPs with a repeat unit composed of a recurring matrikine motif (SM1) or a scrambled motif that completely disrupts the Gly-X-X-Pro-Gly pattern (SM2), and wherein each repeat unit is enriched for hydrophobic residues and includes a single Cys residue to promote *in vivo* retention by formation of intermolecular disulfide crosslinks.\(^5\) (b) Turbidimetry data in PBS (pH 7.4) for SM1 and SM2.
Supplementary Fig. 13. Phase behavior of polymers built from the biologically active domain of endostatin. (a) Peptide sequences forming the bioactive domains of murine (PDB file: 1DY0) and human endostatin (PDB file: 1BNL), shown here in blue and red, respectively. These endostatin-like IDPPs display CD spectra characteristic of IDPs (b) and exhibit LCST transition behavior (c). CD spectra in (b) were measured in water at a polypeptide concentration of 5 µM and at 25 °C. (d) Temperature-dependent turbidimetry of mEndo1-6 at 5 µM in PBS (pH 6.4) and (e) corresponding — for the same sample — CD spectroscopy as a function of temperature. The phase transition of this IDPP was
accompanied by the characteristic decrease in the disorder of the polypeptide conformation, as evidenced by a greatly diminished random coil negative peak at ~197 nm.
Supplementary Fig. 14. The ability to encode phase behavior in protein polymers consisting of monomer units that have local secondary structure propensities, as in human and murine endostatin, may enable the development of a broader set of “smart”, drug-like protein-polymers derived from the growing list of polypeptide hormones. Here we show the primary sequence and predicted secondary structures of polypeptides composed of two repeat units of each hormone. The amino acid sequences of human and murine preptin, as well as the human polypeptide hormones obestatin, glicentin-related polypeptide (GRP), osteostatin, and ghrelin were modified with the C-terminal tripeptide PGG. Residues that are not part of the polypeptide hormone, typically added at the N or C-terminus of the repeat unit, are shown in orange. The UniprotKB/Swiss-Prot code for...
the pro-hormone corresponding to each peptide is shown under the name of each peptide. Secondary structures were predicted using the Jnet algorithm (http://www.compbio.dundee.ac.uk/www-jpred/index.html). H: alpha helix, E: β-sheet, and ‘−’: random coil.
Supplementary Fig. 15. The low complexity (LC) domains of RNA-binding proteins are compositionally similar to resilins. We studied the composition of the LC domains in three human RNA-binding proteins — hnRNPA1 [186-320], hnRNPA2 [181-341], TIA-1 [280-375] — that are involved in RNA granule formation and which appear to exhibit a UCST-type transition at high concentration and low temperatures (Ref 2 from the main manuscript). (a) Resilins and LC domains exhibit a similar hydropathy balance as judged by the overall distribution of Kyte-Dolittle hydropathy values. (b) LC domains are enriched for similar hydrogen bonding residues as resilins, namely R, N, S and Y. These domains are also biased towards Arg residues as they account for the majority (71% ± 0.2) of the positively charged residues. The occurrence of Lys residues at a higher frequency than in resilins, however, may contribute to their reduced aggregation propensity, as their phase transition is only observed at high concentrations.
**Supplementary Figure 16.** Human proteins that comply with the sequence heuristics that determine UCST phase behavior in IDPPs. We found 83 hits upon searching the human proteome for all proteins with >100 residues, wherein Y, W and F and H account for more than 10% of the residues (Aromatic>10%) and which contain a large fraction of charged residues (FCR>20%). For the charged residues, we filtered for proteins with a nearly zwitterionic character by restricting the Percent of Positively Charged residues (PPC) to 0.45-0.55. To account for the role of Arg, we also filtered for proteins wherein Arg accounts for most (R-K>=85%) of the positively charged residues. We did not apply a filter for Pro and Gly content, but calculated these values as an additional parameter to differentiate among the selected proteins. Notably, if relaxing the R-K criteria to 75%, the space of human proteins with UCST features would increase to 403 (data not shown).

### 3. Supplementary Tables and MATLAB scripts

**Supplementary Table 1.** Name and accession number of prototypical Pro and Gly-rich proteins and proteins with clustered P-X_n-G motifs analyzed in this manuscript.

| Protein                  | Species                  | Accession Number (gi) |
|--------------------------|--------------------------|-----------------------|
| Elastin                  | Homo sapiens             | 182021                |
| Elastin                  | Bos taurus               | 28461173              |
| Elastin                  | Mus musculus             | 31542606              |
| Elastin                  | Rattus norvegicus        | 55715827              |
| Elastin                  | Macaca mulatta           | 13182892              |
| Elastin a                | Danio rerio (Zebrafish)  | 121583675             |
| Alpha-1 Collagen Type I  | Homo sapiens             | 553615                |
| Dragline silk fibroin (Spidroin 2) | Nephila clavipes | 159714                |
| Flagelliform silk protein | Nephila clavipes         | 2833649               |
| High molecular weight    | Elymus alashanicus       | 84181091              |
| Protein                  | Species                    | Accession Number (gi) |
|-------------------------|----------------------------|-----------------------|
| gluten subunit          |                            |                       |
| Tropoelastin 1          | Xenopus Tropicalis (XT)    | 118403463             |
| Tropoelastin 2          | Xenopus Tropicalis (XT)    | 218664459             |
| Tropoelastin 2a         | Xenopus Laevis (XL)        | 296040424             |
| Tropoelastin 2b         | Xenopus Laevis (XL)        | 296040422             |
| Resilin Isoform B       | Drosophila melanogaster    | 45552671              |
| Resilin Isoform A       | Drosophila melanogaster    | 7302880               |
| Resilin Isoform A       | Haematobia irritans        | 351581713             |
| Resilin Isoform B       | Haematobia irritans        | 351581715             |
| Resilin Isoform A       | Ctenocephalides felis      | 351581709             |
| Resilin Isoform B       | Ctenocephalides felis      | 351581711             |

**Supplementary Table 2.** DNA sequences of genes encoding for protein polymers spanning P-X<sub>n</sub>-G motifs from n=0 to n=4. Both oligomers and primers were 5’-phosphorylated. DNA oligomers encoding for 4-5 copies of the peptide repeat unit (as shown) were further oligomerized by OERCA.

| Repeat unit | Oligomer sequence (5’-3’) | F. primer (bp) | R. primer (bp) |
|-------------|---------------------------|----------------|----------------|
| VPGVG       | GTACCTGGGGTAGGTGTGCCGGCGTCGGTGTCGCCG | 17             | 16             |
|             | GGCGTCGCGTGTCCCGGGCGTGGTGTCGCCGAGGCGTAG |               |                |
| AVPGV       | GTACCTGGGGTAGGTGTGCCGGCGTCGGTGTCGCCG | 17             | 16             |
|             | GGCGTCGCGTGTCCCGGGCGTGGTGTCGCCGAGGCGTAG |               |                |
| VAPGV       | GTACCTGGGGTAGGTGTGCCGGCGTCGGTGTCGCCG | 17             | 16             |
|             | GGCGTCGCGTGTCCCGGGCGTGGTGTCGCCGAGGCGTAG |               |                |
| Repeat unit | Oligomer sequence (5’-3’) | F. primer (bp) | R. primer (bp) |
|------------|---------------------------|---------------|---------------|
| G          | TGGCGCCTGGCGTGCCCTGGCGGCTCGCGCCGCCGTCGGCG |              |               |
|            | TTGCACCAGGTGTAGGT          |               |               |
| VPGAV G    | GTACCAAGGTGCTTGTTGGTTGGTGTGCCGGCGCCGCTGGGCGGCG | 16            | 17            |
|            | TGCCTGGCGCCGTGGGCCGCTCCCAGCGGCTCGGCCCGTGGGCGGCGGCG | 15            | 17            |
| VPGVA G    | GTACCAAGGTGCTTGTTGGTTGGTGTGCCGGCGGCGGTGCCCAGCGGCTCGGCCCGGCGGCG | 15            | 17            |
|            | TGCCTGGCGCCGTGGGCCGCTCCCAGCGGCTCGGCCCGGCGGCGGCG | 15            | 17            |
| APGV G     | GTACCAAGGTGCTTGTTGGTTGGTGTGCCGGCGGCGGTGCCCAGCGGCTCGGCCCGGCGGCG | 16            | 16            |
|            | GCCCGCCGCAGCTGGCCGGCGGCGCGGCGGTGCCCAGCGGCTCGGCCCGGCG | 17            | 17            |
| VPGVA CA   | GTACCAAGGTGCTTGTTGGTTGGTGTGCCGGCGGCGGTGCCCAGCGGCTCGGCCCGGCGGCG | 16            | 18            |
|            | ACTGGTGCCTGGCGGGCGGCAAGTCGACCAGGGGCGGCGGCG | 16            | 18            |
|            | TGCACCCGGCGTGGGCGCTCCCGCAGGTCGCGGCGGCGGCG | 15            | 15            |
| VHPGV G    | GTACCAAGGTGCTTGTTGGTTGGTGTGCCGGCGGCGGTGCCCAGCGGCTCGGCCCGGCGGCG | 19            | 19            |
|            | TGCACCCGGCGTGGGCGCTCCCGCAGGTCGCGGCGGCG | 19            | 19            |
| VPAGV G    | GTACCAAGGTGCTTGTTGGTTGGTGTGCCGGCGGCGGTGCCCAGCGGCTCGGCCCGGCGGCG | 16            | 17            |
|            | TGCACCCGGCGTGGGCGCTCCCGCAGGTCGCGGCGGCG | 16            | 17            |
| Repeat unit | Oligomer sequence (5’-3’) | F. primer (bp) | R. primer (bp) |
|-------------|---------------------------|---------------|---------------|
| VGPVG       | GTAGGGCCTGTAGGTGTGGGCCCGGTGTCGGTGGGCC | 15            | 16            |
|             | CCGGTCTCGGTGTGGCCCGGTGTCGGCCCGGTAG |               |               |
| VAPVG       | GTTGTCCAGTTGGTGTTGGGCCCGGTGTCGGCCGGC | 16            | 16            |
|             | CTGTGGGCGTGCAGGCGGTGAGCACCAGTAGG     |               |               |
| VRPGV       | GTACGTCCTGGAGTTGGCGTCGAGCACCAGTAGG   | 18            | 18            |
|             | CCGGTGGCGTCCGAGCCGGGTGAGTTCGCAGTAGG  |               |               |
| VGAPV       | GTAGGTGCTCCAGTTGGTGTTGGGCCCGGTGGGCG | 16            | 17            |
| G           | TGGCGCCTACGCGTGGCGTGCGCAAGGTCGCGCTCG |               |               |
| VPGVG       | GTACCAACGGGTGTGGTGCCAGACAGGCGTGCGG | 19            | 17            |
| G           | GTGCCCTAATGGCGTGCGCAGAGGTCGCGGTCCGC |               |               |
| VPGVG       | GTTCAAGGCTTAGGCTAGGTGAGTTGGCGTGTGGGCG | 19            | 18            |
| VPGVG       | C            |               |               |
| VPGVG       | GTTCACACGTTGGTGTCGCGCTCCGCGCTAGCAATGTGAG | 19            | 18            |
| VPGVG       | ACGTGCGCCGTCGCGCTAGCAATGTGAG | 19            | 18            |
| VPGVG       | GTGCCTAATGGCGTGCGCAAGGTCGCGGTCCGC |               |               |
| VPGVG       | TGGCGCCTACGCGTGGCGTGCGCAAGGTCGCGCTCG |               |               |
| VPGVG       | TAGGTGACAGCAGTAGGGGTGTCGCGGTGCGGCG | 18            | 18            |
| VPGVG       | GTTACCCAGCTGTAGGTCGACACCAGCCTGCGGTGGGCG | 19            | 18            |
| VPGVG       | TGAACCTCAGGCGTGGCGTCAGCACCAGCCTGCGGTGGGCG | 19            | 18            |
| Repeat unit | Oligomer sequence (5′-3′) | F. primer (bp) | R. primer (bp) |
|-------------|--------------------------|---------------|---------------|
| TPVAV       | AACCCAGTTGCTGTGTTGACACCGGGTGCGGCGGCA | 16            | 18            |
|             | CTCCTGTGGCGGTCGCTCGCCTGCGGAC |               |               |
|             | GCCAGTACGACGACGGTGGGGTTGGG |               |               |
| VPSTDY      | GAGACTACGCGCGGTCGCTCGCCTGCCGAGC | 17            | 17            |
|             | GGGGTGCCTGTACCACTTGGATGATTACGGTTGGG |               |               |
| VPSALY      | GTACCTTCTGCGCTGTGTTGAGGGTGTGCCCCTGGTACGG | 18            | 16            |
|             | CCCCTGTACGCGTCGCTCCTGCCCCTGGCACAGTCGCGGAC |               |               |
| VPSDD       | GTACCTTCTGCACACTATGCGGAGGGTGTGCCCAGCGG | 20            | 19            |
|             | ATGACTACGCGCGGTCGCTGGGCATGACTACG |               |               |
|             | GCACCAGGGTGTACCATCGGGATGGATACGGCCAGGG |               |               |
| VPSDD       | GTACCTTCTGACACTATGCGGAGGGTGTGCCCAGCGG | 20            | 19            |
|             | ATGACTACGCGCGGTCGCTGGGCATGACTACG |               |               |
|             | GCACCAGGGTGTACCATCGGGATGGATACGGCCAGGG |               |               |
| GRGDSP      | GGGGCCTGGTACCTTCTGACATGCGGACAGCCGTCGCTGTGTTGAGGGTGTG | 18            | 18            |
|             | CTTTCGCCATTACGTGACTGCGCACAGTCGCTGTGTTGAGGGTGTG |               |               |
| GRGDSP      | GGGGCGGTACCTTCTCCACATGCGGACAGCCGTCGCTGTGTTGAGGGTGTG | 18            | 16            |
|             | CCGCATGCGGTCGCTGGGTCGAGCCATGCGGTCGCTGTGTTGAGGGTGTG |               |               |
|             | ATTCCGCTCAC |               |               |
| GRGDSP      | GGGGCCTGGTACCTTCTCCACATGCGGACAGCCGTCGCTGTGTTGAGGGTGTG | 18            | 18            |
|             | CCGGTTGCTGCGGACAGCCATGCGGTCGCTGTGTTGAGGGTGTG |               |               |
|             | ATTCCGCTCAC |               |               |
| GRGDSP      | GGGGCGGTACCTTCTCCACATGCGGACAGCCGTCGCTGTGTTGAGGGTGTG | 18            | 19            |
|             | CCGGTTGCTGCGGACAGCCATGCGGTCGCTGTGTTGAGGGTGTG |               |               |
|             | ATTCCGCTCAC |               |               |
| Repeat unit | Oligomer sequence (5’-3’) | F. primer (bp) | R. primer (bp) |
|-------------|---------------------------|---------------|---------------|
| YG          | AGCCCGTATGGCGGTCGGCGACAGTGTCGATGGCTATGGCTTGTCGATGGACAGTCTATGCACTTCATCCAGTTGCTGGCGACAGCTATCCCAGGTCGATGGCGACAGTTATCCGGGTCGTGGTGA | 20 | 20 |
| RPLGY DS    | CGCCCATATTAGGTTATGACTCTCTCGTCCGGCGGCTGGGCTATG | 20 | 20 |
| GRGDS YP    | GGCCGCGGCTGACGCTTTATCCAGGTTCGTCGGCGACAGCTACAGCCGTCCGGGACAGTTATCCTGGGTCGTGGTGA | 21 | 19 |
| VPHSNR GG   | GTTCCACACAGCCGTAACGTTGTTGTCGGGCAGGGCTATG | 19 | 19 |
| RPAGY DS    | CGCCCACTGCGTTATGACCTCTCTGCTCCGGCAGGGCTATGACAGCCGTCCGGGACAGTTATCCGGGTCGTGGTGA | 19 | 19 |
| GRGDA PYQ   | GGTCGTGGCGATGCTCTCCATACCCAGGGCCCGCGGTGAC | 20 | 21 |
| GRPSDS YG   | GGGCGCTCATCTGACTCTACGCGGCGTCGCGCCGGAGC | 19 | 20 |
| VAPDG       | GTGACCCGCGTCGGCGCCGGGATGCGGCTAGACGACCAGAGCG | 20 | 18 |
| GRASDS YG   | GGGCGTGCATCCATCTCCCTACGCGCGGCTCGCGCAAGCG | 19 | 20 |
### Supplementary Table 3. DNA sequence information for the synthesis of genes encoding P-X₉₋₀-G containing peptide polymers via Pre-RDL. Oligomers were 5’-phosphorylated.

| Repeat unit | Forward Oligomer (5’-3’) | Reverse Oligomer (5’-3’) |
|-------------|---------------------------|-------------------------|
| LGAPVG      | GCTGGGTGCTCCAGTTGGTCTGGG | TACTGGTGCAACCGAGCCGACC |
|             | CGCCCCGCTGCGGGCTGCCGCC   | GCGGGCCCAGCCACAGGGGCC |
|             | TGTGGCACCTGGGCGGCGGTCGG  | GCCAGGCCCACCGGGGCGGCCC |
|             | CCTGGTGACCCAGTACGG       | GACCAACTGGAGCACCCAGCCC |
| Repeat unit | Forward Oligomer (5’-3’) | Reverse Oligomer (5’-3’) |
|-------------|--------------------------|-------------------------|
| GRGDSPYQ    | GCGCGGTGACTCTCCATACCAGGG | TTGGTAAGGCAGAATCACCACGAC |
|             | TCGTGGCCAGCAGCGCGGTATACGGG | CCTGATAAGGACTGTCGCCACGA |
|             | TCGTGGCGACAGTCCGTATCAGGG  | CCCTGATAAGGCTGTCGCCACG  |
|             | TCGTGGTGAATTCGCTTACCAAGG  | ACCCTGGATAGGAAGGATCACCAC |
| GRGNSPYG    | GCGCGGTAAAATCCATACAGGG    | ACCGTAAGGCGAATCACCACGAC |
|             | TCGTGGCAACAGCCGCTATTGTTTGT | CGGCATAGGCAGCTTGGCCACGA |
|             | TCGTGGCAACTGCGGTATTCCGCGG | CCACCATACGGGCTTGGCCACG |
|             | TCGTGGTAAATTCCGCTTACGGTG  | ACCGCCCATAGGAAGGATCACCAC |
| QYPSDGRG    | GCAGTACCCATCTGACGGTCGCGG | ACGACCATCCGAGGGGATTGAC |
|             | TCGATATCCGAGCGAGCGCGGTCGG | CACGGCCGTCATCGATACGTA |
|             | TCGATATCCGAGTGACGGGCCGTTG | CCACGCCCAGTCGCTGATACGTA |
|             | TCAAATACCCCTCGGATGCTGG   | ACCGCGACCAGTGGAGGTTACT |
|             |                          | GCCC |
| GRGDSPH     | CCGCGGTGACTCTCCACATGGTCG | GTAGCCGGAATCACCACGACCAT |
|             | TGGCGAAGCGGCCATGGTCGTGG  | GCGAATGTCGCCACCGACCATGC |
|             | CGACAGTCGGATGGTCGTGG    | GGGCCTGTCGCCACCGACCATGTG |
|             | TTCGCTTCACGG             | AGAGTCACCAGCCGCCC |
| GRGDSPY     | CCGCGGTGACTCTCCATACGGTGC | GTAAGCGAATCACCACGCCCC |
|             | TGGCGAAGCGGCCATGGCTGG   | ACGGAATGTCGCCACCGACCATAC |
|             | CGACAGTCGGTGATGGCGTGTA  | GGGCCTGTCGCCACCGACCATGTG |
|             | TTCGCTTACGG             | AGAGTCACCACCGCCC |
| GRGDSAY     | CCGCGGTGACTCTGCTTACGGTGC | GTAAGCGAATCACCACGCCCC |
|             | TGGCGAAGCGGCCTATGGTGGTCG | ACGGACTGTCGCCACCGACCATAC |
|             | CGACAGTCGGTGATGGCGTGTA  | GGGCCTGTCGCCACCGACCATGTG |
|             | TTCGCGATACGG             | AGAGTCACCACCGCCC |
**Supplementary Table 4.** Synthesis of genes encoding for bioactive peptide polymers by OERCA. Both oligomers and primers were 5’-phosphorylated.

| Repeat unit | Oligomer Sequence (5’-3’) | F. primer (bp) | R. primer (bp) |
|-------------|---------------------------|----------------|----------------|
| hEndo:      | GTCGCCGGGGTCATAGCCATCGTGA | 20             | 19             |
| VPGHSRDFQPVLHLVALNSPL SGGMRG | TCCAGCCTGTCTGCACCTGTTGCT |               |                |
| mEndo:      | CATACCCACCAGGAATTCCAGCCTG | 20             | 19             |
| HTHQDFQPVLHLVALNTPLSGG MRGIRPGG | TTCTGCACTCTGCTCTGAACACC |               |                |
| DPGYIGSRG   | GACCCAGGCTACATCGGTTCGCG  | 19             | 21             |

**Supplementary Table 5.** DNA sequence information for the synthesis of genes encoding bioactive peptide polymers via Pre-RDL. Oligomers were 5’-phosphorylated.

| Repeat unit | Forward Oligomer (5’-3’) | Reverse Oligomer (5’-3’) |
|-------------|--------------------------|--------------------------|
| AVPGVGAVG | GGCTGTACCAGGTTGTGCGC | TACACCTGGAACGCAGCCACG |
| VGVAVPGVGA | GCCGGCCGTGGGCGGCGGCTGGTTC | CCGGCACGGGCGGCACGGCCAC |
| VPGVGAVG | CTGCGGCGGCGTCCCGGGGCGTGTCG | GACCACGCAGCCAGGCACGC |
| GCVPVGVG | CGCCGTGCACGCCGTGGGGCTGCCTG | GCCCGCCCGCCACAGCCAGCCAC |
|             | TCCAGGTGTAGG             | ACACCTGGTACAGC           |
| Repeat unit | Forward Oligomer (5’-3’) | Reverse Oligomer (5’-3’) |
|-------------|--------------------------|--------------------------|
| VPAGVGPAG   | GGTACCAGCTGGGTGGTGTTGGTGCC | TACACCGCATGGAACGCCACGCCACGC |
| VGVPAGVGP   | GGCACCGGTGGCCGCTTGCTCCGCGG | CGGGCCCGGACCGGCCACGCCACGCCAC |
| AGVGPAGVGPVG | CGTGGGGCTCTGCGGCGGCGGTCCGG | GGGACGCCCACGCCCGCGACAGGGCA |
| GVGPCVGVG   | CGTGCCGCGGCCGCGGCTTGCCGTTCC | CGCCCAAGGGCGCGGCCGACACACA |
|             | ATGCGGTGTTAGG             | ACACCGCTGCTGTTACCCC       |
| VPHSRNGGNS  | CGTTCCGACAGCCGTAACCGTG | GTACGGAGAATCGCCGCGACCGC |
| GGRGDSPYG   | CAACTCTGGTGCGCGGCGGTGACAG | CGCTGTTACCGCCATTACCGGAA |
|             | CCGTATGTTGCTGCCCATCTCCG | TGGCGCACCCATATGGGCTGTC |
|             | TAATGGGCCGTAACAGCGCGGCTG | ACCCGCGACCCAGATTGCGCAC |
|             | CGGCGATTCTCTCGTACGG       | CTACGGGTCTGGCGAAGCGCC    |

MATLAB Script 1

%*************************** Start of MATLAB Script 1

% Definition of main variables
% pstarts: stores the residue position of each Pro participating in a P-Xn-G motif
% gends: stores the residue position of each Gly participating in a P-Xn-G motif

clear all
fid=fopen('ProteinSequence.txt');
x=fscanf(fid,'%s');

% Start of code A to identify the location of Pro and Gly residues forming P-Xn-G motifs
t=1;n=1;
for j=1:length(x)
    if t==1
if strcmp('P', x(j))
    t=2;
    pstarts(n)=j;
end
else
    if strcmp('P', x(j))
        pstarts(n)=j;
    end
    if strcmp('G', x(j))
        t=1;
        gends(n)=j;
        n=n+1;
    end
end
end

% End of code A

% Start of code B to calculate the fraction of P-Xn-G motifs
pg=0;pxg=0;pxxg=0;pxxxg=0;pxxxxg=0;
for j=1:length(gends)
    if pstarts(j)-gends(j)==-1
        pg=pg+1;
    end
    if pstarts(j)-gends(j)==-2
        pxg=pxg+1;
    end
    if pstarts(j)-gends(j)==-3
        pxxg=pxxg+1;
    end
    if pstarts(j)-gends(j)==-4
        pxxxg=pxxxg+1;
    end
end
if pstarts(j) - gends(j) == -4
  pxxxg = pxxxg + 1;
end
if pstarts(j) - gends(j) == -5
  pxxxxg = pxxxxg + 1;
end
end

% End of code B

% Start of code C to calculate the distance between P-Xn-G motifs (n=0-4)
l1 = 1; t1 = 1;
for i = 2:n-1
  if t1 == 1
    if pstarts(i-1) >= gends(i-1) - 5
      t1 = 2;
      p1 = i-1;
    end
    if pstarts(i) >= gends(i) - 5
      Dist(l1) = pstarts(i) - gends(p1) - 1;
      l1 = l1 + 1;
      p1 = i;
    end
  end
else
  if pstarts(i) >= gends(i) - 5
    Dist(l1) = pstarts(i) - gends(p1) - 1;
    l1 = l1 + 1;
    p1 = i;
  end
end
end
end

end

% End of code C to calculate the distance between P-Xn-G motifs (n=0-4)

% Variables that summarize all the relevant data
FractPXG=[pg pxg pxxg pxxg pxxg]/length(pstarts);
Distance=sum(Dist)/length(Dist);

%*************************** End of MATLAB Script

MATLAB Script 2

%*************************** Start of MATLAB Script

clear all
fid=fopen('P24592.txt');
x=fscanf(fid,'%s');

% Start of code A to identify the location of Pro and Gly residues forming P-Xn-G motifs

for j=1:length(x)
    if t==1
        if strcmp('P',x(j))
            t=2;
            pstarts(n)=j;
        end
    end
end

% End of code C to calculate the distance between P-Xn-G motifs (n=0-4)
end
else
    if strcmp('P', x(j))
        pstarts(n) = j;
    end
    if strcmp('G', x(j))
        t = 1;
        gends(n) = j;
        n = n + 1;
    end
end
end

% End of code A

% Start of code B to concatenate all linking residues

d = 1;
for j = 1:n-1
    if j == 1
        Div(d:pstarts(j)-1) = x(1:pstarts(j)-1);
        d = length(Div) + 1;
    else
        dif = pstarts(j) - gends(j-1);
        Div(d:d+dif-2) = x(gends(j-1)+1:pstarts(j)-1);
        d = length(Div) + 1;
    end
end
if length(x) - gends(end) > 0
    Div(d:d+length(x) - gends(end) - 1) = x(gends(end)+1:end);
end
clear x
x = Div;
m = 1;
A = 0; V = 0; L = 0; M = 0; F = 0; Y = 0; W = 0; S = 0; T = 0; C = 0; N = 0; Q = 0; K = 0; H = 0; R = 0; D = 0; E = 0; P = 0; G = 0;
for i = 1:length(x)
    if strcmp('A', x(i)); hi(m) = 1.8; A = A + 1; end
    if strcmp('V', x(i)); hi(m) = 4.2; V = V + 1; end
    if strcmp('L', x(i)); hi(m) = 3.8; L = L + 1; end
    if strcmp('I', x(i)); hi(m) = 4.5; I = I + 1; end
    if strcmp('M', x(i)); hi(m) = 1.9; M = M + 1; end
    if strcmp('F', x(i)); hi(m) = 2.8; F = F + 1; end
    if strcmp('Y', x(i)); hi(m) = -1.3; Y = Y + 1; end
    if strcmp('W', x(i)); hi(m) = -0.9; W = W + 1; end
    if strcmp('S', x(i)); hi(m) = -0.8; S = S + 1; end
    if strcmp('T', x(i)); hi(m) = -0.7; T = T + 1; end
    if strcmp('C', x(i)); hi(m) = 2.5; C = C + 1; end
    if strcmp('N', x(i)); hi(m) = -3.5; N = N + 1; end
    if strcmp('Q', x(i)); hi(m) = -3.5; Q = Q + 1; end
    if strcmp('K', x(i)); hi(m) = -3.9; K = K + 1; end
    if strcmp('H', x(i)); hi(m) = -3.2; H = H + 1; end
    if strcmp('R', x(i)); hi(m) = -4.5; R = R + 1; end
    if strcmp('D', x(i)); hi(m) = -3.5; D = D + 1; end
    if strcmp('E', x(i)); hi(m) = -3.5; E = E + 1; end
    if strcmp('P', x(i)); hi(m) = 1.6; P = P + 1; end
    if strcmp('G', x(i)); hi(m) = -0.4; G = G + 1; end
m=m+1;
end

pAA=[A R N D C Q E H I L K M F S T W Y V P G]'*100/m;
AvHi=sum(hi)/length(hi);

MATLAB Script 3

% Start of MATLAB Script

% Analysis of a full proteome with filtering for Collagens and Elastins

% Definition of main variables
% ProRichProt: Structure that stores the information of select Pro-rich proteins
% pAA: Matrix with the percent abundance of select residues for all proteins in the proteome

clear all

[name,Prot]=fastaread('ProteomeFile.fasta');
% Start of Code A to filter out collagen and elastin proteins
j=1;j2=1;col=0;ela=0;
while j2<=length(Prot)

F=0;Y=0;W=0;K=0;H=0;R=0;D=0;E=0;P=0;G=0;A=0;
name2=char(name(j2));
x=char(Prot(j2));

if strfind(name2,'Collagen')
    col=col+1;
    j2=j2+1;
else
    if strfind(name2,'Elastin')
        ela=ela+1;
        j2=j2+1;
    else
        for i=1:length(x)
            if strcmp('A',x(i));A=A+1;end
            if strcmp('F',x(i));F=F+1;end
            if strcmp('Y',x(i));Y=Y+1;end
            if strcmp('W',x(i));W=W+1;end
            if strcmp('K',x(i));K=K+1;end
            if strcmp('H',x(i));H=H+1;end
            if strcmp('R',x(i));R=R+1;end
            if strcmp('D',x(i));D=D+1;end
            if strcmp('E',x(i));E=E+1;end
            if strcmp('P',x(i));P=P+1;end
        end
    end
end
if strcmp('G',x(i));G=G+1;end

end

pAA(j,1:11)=[P G D E R K F W Y H A]'*100/length(x);
lProt(j)=length(x);
Fname(j)={name2};
j=j+1;
j2=j2+1;
end
end

dend
%End of Code A

%Start of Code B for Identification of protein subsets
m=1;
L=size(pAA);
TotalCharged=0;
for i=1:L(1)
    if lProt(i)>=50  %Restriction to proteins with >50 residues
        if pAA(i,1)>10 %Specifies threshold for Pro content
            if pAA(i,2)>10 %Species threshold for Gly content
                ProRichProt(m,1)={char(Fname(i))};
                ProRichProt(m,2)={pAA(i,1)}; %Pro content
                ProRichProt(m,3)={pAA(i,2)}; %Gly content
                ProRichProt(m,4)={pAA(i,3)+pAA(i,4)+pAA(i,5)+pAA(i,6)}; %FCR
                m=m+1;
            end
        end
    end
end
%End of Code B
end
end
end
end

end
%End of Code B

%************************************************************************ End of MATLAB Script
%************************************************************************

MATLAB Script 4

%************************************************************************ Start of MATLAB Script
%************************************************************************

%Analysis of P-Xn-G motifs in large protein datasets from proteome-wide analyses

clear all

%State of Code A to obtain sequence information for a select group of proteins
[name,Prot]=fastaread('Zebrafish_7955.fasta');
[subset,empty]=fastaread('Zebrafish10Pro10Gly.txt');

for i=1:length(name)
    name2=char(name(1,i));
    name3(i)={name2(1:20)};
end
for i = 1:length(subset)
    subset2 = char(subset(1,i));
    subset3(i) = {subset2(1:20)};
end

Position = find(ismember(name3, subset3));  % points to protein location in fasta file

% End of Code A

% Start of Code B to estimate the abundance and distance between P-Xn-G motifs
% for all proteins in the subgroup
sp = 1;
for cycle = 1:length(subset)

    x = char(Prot(Position(cycle)));
    rndProt(1,:) = x;

    t = 1; n = 1;
    for j = 1:length(x)
        if t == 1
            if strcmp('P', rndProt(1,j))
                t = 2;
                pstarts(n) = j;
            end
        else
            if strcmp('P', rndProt(1,j))
                pstarts(n) = j;
            end
            if strcmp('G', rndProt(1,j))
t=1;
gends(n)=j;
n=n+1;
end
end
end

Gpm2=0;Gpm1=0;Ggp2=0;Ggp1=0;t=0; p=0; pg=0;pxg=0;pxxxg=0;pxxxxg=0;
l0=1;l1=1;t0=1;t1=1;
if n>1
for j=1:length(gends)
    if pstarts(j)-gends(j)==-1
        pg=pg+1;
    end
    if pstarts(j)-gends(j)==-2
        pxg=pxg+1;
    end
    if pstarts(j)-gends(j)==-3
        pxxg=pxxg+1;
    end
    if pstarts(j)-gends(j)==-4
        pxxxg=pxxxg+1;
    end
    if pstarts(j)-gends(j)==-5
        pxxxxg=pxxxxxg+1;
    end
end

if j<=length(gends) & j>1
    if t1==1

if pstarts(j-1)>=gends(j-1)-5
    t1=2;
    p1=j-1;
if pstarts(j)>=gends(j)-5
    EhomoD1(l1)=pstarts(j)-gends(p1)-1;
    l1=l1+1;
    p1=j;
end
end
else

if pstarts(j)>=gends(j)-5
    EhomoD1(l1)=pstarts(j)-gends(p1)-1;
    l1=l1+1;
    p1=j;
end
end

EhomoD(j-1)=pstarts(j)-gends(j-1)-1;
end
end
end
end
end

%Subset of proteins with repeating P-Xn-G motifs
if pg+pxg+pxxg+pxxxg+pxxxxg>1
if length(EhomoD1)>15 %Specifies minimum number of P-Xn-G motifs
    FractPXG=[pg pxg pxxg pxxxg pxxxxg]/length(pstarts);
    Distance=[sum(EhomoD1)/length(EhomoD1),length(EhomoD1)+1];
    Special(sp)={name(Position(cycle))};

FractSpecial(sp,1:5)=FractPXG;
DistSpecial(sp,1:2)=Distance(1:2);
sp=sp+1;
end
end
%end of susbet

clear rndProt pstarts gends EhomoD1 EhomoD

end
%End of code B

%********************************************* End of MATLAB Script
%*********************************************

MATLAB Script 5

%********************************************* Start of MATLAB Script 5
%*********************************************

% Definition of main variables
% Capitalized variables are counters named after the corresponding single-letter amino acid code.
% pAA: percent abundance of all amino acids when excluding Pro and Gly.
% Hi: vector containing the Kyte-Dolittle hydropathy index for each amino acid.

clear all
fid=fopen('ProteinSequence.txt');
x=fscanf(fid,'%s');

m=1;
A=0;V=0;L=0;I=0;M=0;F=0;Y=0;W=0;S=0;T=0;C=0;N=0;Q=0;K=0;H=0;R=0;D=0;E=0;

for i=1:length(x)
if ~strcmp('P',x(i)) && ~strcmp('G',x(i))
    if strcmp('A',x(i));hi(m)=1.8;A=A+1;end
    if strcmp('V',x(i));hi(m)=4.2;V=V+1;end
    if strcmp('L',x(i));hi(m)=3.8;L=L+1;end
    if strcmp('I',x(i));hi(m)=4.5;I=I+1;end
    if strcmp('M',x(i));hi(m)=1.9;M=M+1;end
    if strcmp('F',x(i));hi(m)=2.8;F=F+1;end
    if strcmp('Y',x(i));hi(m)=1.3;Y=Y+1;end
    if strcmp('W',x(i));hi(m)=0.9;W=W+1;end
    if strcmp('S',x(i));hi(m)=0.8;S=S+1;end
    if strcmp('T',x(i));hi(m)=0.7;T=T+1;end
    if strcmp('C',x(i));hi(m)=2.5;C=C+1;end
    if strcmp('N',x(i));hi(m)=3.5;N=N+1;end
    if strcmp('Q',x(i));hi(m)=3.5;Q=Q+1;end
    if strcmp('K',x(i));hi(m)=3.9;K=K+1;end
    if strcmp('H',x(i));hi(m)=3.2;H=H+1;end
    if strcmp('R',x(i));hi(m)=4.5;R=R+1;end
    if strcmp('D',x(i));hi(m)=3.5;D=D+1;end
    if strcmp('E',x(i));hi(m)=3.5;E=E+1;end
    m=m+1;
end
end
4. Supplementary Text

IDPPs built from biologically active peptide repeats

To further investigate the concept of reprogramming bioactive peptides to exhibit LCST phase behavior, we identified a number of candidate peptide drugs. We looked for peptides that are bioactive and that are predicted to be predominantly disordered (Supplementary Fig. 14). For polymerization we chose the 25-27 amino acid long bioactive peptide domains of endostatin from humans and mice —hEndo and mEndo, respectively (Supplementary Fig. 13a)—, as they belong to a region of endostatin that is largely random coil yet retains the anti-angiogenic activity of the full protein, and because their average hydropathy lies within the range of our LCST motifs (Fig. 3c). Because there is only one P-X_{2}-G motif (P-L-S-G) in these peptides and guided by our design heuristics, we incorporated a new PG motif at the C-terminus of mEndo and at the N-terminus of hEndo prior to polymerization to reduce the average distance between P-X_{n}-G motifs in the protein polymer to ~10-12 residues. The resulting IDPPs are intrinsically disordered as seen by their CD spectra (Supplementary Fig. 13b) and exhibit LCST behavior (Supplementary Fig. 13c) that is accompanied by a decrease in the intensity of the random coil peak at ~197 nm (Supplementary Fig. 13d-e) as observed for other LCST exhibiting IDPPs such as ELPs. The synthesis of LCST polymers with a predominantly disordered repeat unit and yet with stretches of amino acids with a high propensity to fold, which seems a common feature in many bioactive peptides of interest (Supplementary Fig. 14), suggests that even sequences that have some secondary
structure and useful biological function can yield IDPPs that retain bioactivity and simultaneously display phase behavior. Assessing the bioactivity of these endostatin-like and other drug-like “smart” polymers, however, is beyond the scope of this paper as it is likely to require the controlled release of monomer peptide from the polymer, which is non-trivial. In principle, this can be achieved by the incorporation of protease-cleavable peptide motifs between repeat units to release bioactive, intact peptide drugs at the site of interest, as we recently demonstrated with an inert polymer of a glucagon-like peptide-1 analog fused to an ELP to effectively treat type II diabetes in mice.\textsuperscript{9}

**Bioactivity of RGD-based syntactomers**

We show preliminary evidence that the UCST behavior exhibited by these polymers may constitute a bioactivity switch, as we observed that polymers with a UCST below 37 °C, which are thus soluble under regular cell culture conditions (Supplementary Fig. 11b), reduce cell adhesion in a dose-dependent manner (Supplementary Fig. 11c), whereas polymers with a UCST above 37 °C failed to affect cell adhesion in a similar manner — despite having more repeats of the RGD motif, which has been reported to lead to stronger adhesion properties.\textsuperscript{10} The ability of this RGD-based syntactomer to reduce cell adhesion to tissue-culture treated polystyrene was no longer observed when the cells were allowed to adhere for 18 h, which demonstrates that the observed response after 3 h is not due to a cytotoxic effect but rather by delayed cell adhesion mediated by the UCST syntactomer binding to integrins on the cell surface (Supplementary Fig. 11d). The future utility of these RGD-containing IDPPs is further motivated by our ability to effectively decouple the bioactivity of the peptide and the phase behavior of the polymer, as seen for syntactomers composed of scrambled GRGDSP motifs, such as QYPSDGRG and GRDGSPYG (Fig. 2b). Moreover, these peptide polymers already exhibit UCSTs in a temperature window that is relevant for mild hyperthermia, as they are insoluble at 37 °C and soluble at 41-42 °C (data now shown) in cell culture media supplemented with serum and at 25 µM; a concentration at which we already observe a marked response for the short peptide polymer studied in Supplementary Fig. 11c.
Proteome-wide identification of proteins that may exhibit UCST phase behavior

To identify proteins with potential UCST phase behavior, we implemented our sequence heuristics for UCST behavior in a search algorithm implemented in MATLAB (a minor variant of Script 3) with a set of 5 sequence parameters that control UCST phase behavior. The minimal threshold values for these parameters are those derived from the compositional analysis of IDPPS with validated UCST phase behavior and from the compositional analysis of resilins and LC domains. These critical sequence parameters are: i) protein length, ii) charge content \((R + K + D + E)\), iii) percent of positively charged residues \((R + K)\) with respect to all \((R + K + D + E)\) charged residues as a measure of the zwitterionic character of the protein, iv) aromatic content \((Y+H+W+F)\) and v) percent of Arg residues with respect to all \((R + K)\) positively charged residues. Below we provide the basis for the minimal threshold values that we chose for our proteome-level screens:

*Protein length:* \(>100\) residues. Because molecular weight is a potent modulator of UCST phase behavior, we set this value to the length of the shortest UCST IDPPs in our library.

*Charge content:* \(>20\%\). Most of our UCST IDPPs have at least 20\% charge content, but some UCST IDPPs had as little as 15\% charged residues and resilins typically have between 12 and 15\% charged content. Because more than 50\% of the human genome is composed of proteins with at least 20\% charged residues (data now shown), we decided to use 20\% charged content as a rather conservative threshold to target proteins that are reminiscent to UCST IDPPs.

*Zwitterionic character:* While most UCST IDPPs that we studied are zwitterionic, with positively charged residues accounting for exactly 50\% of all charged residues, we reported several examples wherein cationic IDPPs exhibit UCST under physiologically relevant conditions. Because we are aware that highly cationic, Arg-based proteins can have compositional features of UCST IDPPs, we set a conservative threshold for zwitterionic character wherein the percent of positively charged residues (PPC) can vary by 5\% from perfect zwitterionic character (i.e., \(45<\text{PPC}<55\%\)).
**Aromatic content:** >10%. Most UCST IDPPs herein studied have about 12% aromatic content, but at least one UCST IDPP contains only 10% aromatic residues (Fig. 6a). Because many eukaryotic proteins with the potential to exhibit UCST phase behavior may do so through IDP domains that only comprise 30-50% of the overall protein length, as in the case of RNA binding proteins with LC domains discussed in Supplementary Fig. 15, we established 10% as a conservative threshold for a proteome-wide search that ignores compositional biases observed in protein domains (i.e., even a protein with a high aromatic content in its UCST domain may exhibit significantly lower aromatic content if averaged across the entire length of the protein).

**Arg enrichment:** We know that Arg-enrichment is a critical parameter, but our UCST library does not provide a systematic study into Lys doping of Arg-based IDPPs to establish a definitive threshold for this parameter. Based on the compositional analysis of resilins, we could set a threshold of >88%. However, because we also considered LC domains that appear to exhibit UCST phase behavior with ~70% Arg-enrichment (Supplementary Fig. 15), we decided to consider a very conservative threshold of >85%.

**References**

1. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* **157**, 105-132 (1982).
2. Amiram, M., Quiroz, F. G., Callahan, D. J. & Chilkoti, A. A highly parallel method for synthesizing DNA repeats enables the discovery of 'smart' protein polymers. *Nature Materials* **10**, 141-148 (2011).
3. McDaniel, J. R., MacKay, J. A., Quiroz, F. G. & Chilkoti, A. Recursive Directional Ligation by Plasmid Reconstruction Allows Rapid and Seamless Cloning of Oligomeric Genes. *Biomacromolecules* **11**, 944-952 (2010).
4. Hassounah, W., MacEwan, S. R. & Chilkoti, A. Fusions of Elastin-Like Polypeptides to Pharmaceutical Proteins. *Methods in Enzymology* **502**, 215 (2012).
5. Asai, D. *et al*. Protein polymer hydrogels by *in situ*, rapid and reversible self-gelation. *Biomaterials* (2012).
6. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nature protocols* **1**, 2876-2890 (2007).
Brassart, B. *et al.* Conformational dependence of collagenase (matrix metalloproteinase-1) up-regulation by elastin peptides in cultured fibroblasts. *Journal of Biological Chemistry* **276**, 5222 (2001).

Tjin Tham Sjin, R. M. *et al.* A 27-amino-acid synthetic peptide corresponding to the NH2-terminal zinc-binding domain of endostatin is responsible for its antitumor activity. *Cancer research* **65**, 3656 (2005).

Amiram, M., Luginbuhl, K. M., Li, X., Feinglos, M. N. & Chilkoti, A. Injectable protease-operated depots of glucagon-like peptide-1 provide extended and tunable glucose control. *Proceedings of the National Academy of Sciences* **110**, 2792-2797 (2013).

Lee, B. W. *et al.* Strongly Binding Cell-Adhesive Polypeptides of Programmable Valencies. *Angewandte Chemie International Edition* **49**, 1971-1975 (2010).