HHS Public Access

Author manuscript

Annu Rev Biophys. Author manuscript; available in PMC 2022 June 13.

Published in final edited form as:

Annu Rev Biophys. 2022 May 09; 51: 355–376. doi:10.1146/annurev-biophys-120221-095357.

Rules of Physical Mathematics Govern Intrinsically Disordered Proteins

Kingshuk Ghosh[1,2], **Jonathan Huihui**[1], **Michael Phillips**[1], **Austin Haider**[2]

[1]Department of Physics and Astronomy, University of Denver, Denver, Colorado, USA

[2]Molecular and Cellular Biophysics Program, University of Denver, Denver, Colorado, USA

Abstract

In stark contrast to foldable proteins with a unique folded state, intrinsically disordered proteins and regions (IDPs) persist in perpetually disordered ensembles. Yet an IDP ensemble has conformational features—even when averaged—that are specific to its sequence. In fact, subtle changes in an IDP sequence can modulate its conformational features and its function. Recent advances in theoretical physics reveal a set of elegant mathematical expressions that describe the intricate relationships among IDP sequences, their ensemble conformations, and the regulation of their biological functions. These equations also describe the molecular properties of IDP sequences that predict similarities and dissimilarities in their functions and facilitate classification of sequences by function, an unmet challenge to traditional bioinformatics. These physical sequence-patterning metrics offer a promising new avenue for advancing synthetic biology at a time when multiple novel functional modes mediated by IDPs are emerging.

Keywords

heteropolymer; polyampholyte; proteome; disorder; function; liquid–liquid phase separation

1. INTRODUCTION

Intrinsically disordered proteins and regions (collectively termed IDPs in this review, except in Section 4.3), do not fold into unique folded structures. However, the absence of unique structures does not mean that IDP conformations are featureless (92) or that IDPs lack functions (29). In fact, they can have conformational signatures specific to their sequences, and these features may be responsible for their specific functions. The defining features can range from such broad and simple observables as radius and scaling exponent (109) to detailed interresidue distance profiles (19, 46, 63), structure factors (5, 63, 64, 77), and other measures of ensemble properties (17, 36, 53, 59). These specific features must be encoded in the sequence, but how do we unlock that code? The answer to the question

kghosh@du.edu .

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

The Annual Review of Biophysics is online at biophys.annualreviews.org

of how—at first—seems just as complicated as the sequence-structure puzzle of folded proteins that perplexed protein biology for decades. Yet we can turn the disorder of IDPs to our advantage. First, the high degree of disorder allows averaging over different degrees of freedom with few constraints, facilitating models that are analytically tractable. Second, electrostatics plays a prominent role in determining IDP conformations (19, 20, 43, 44, 61, 62, 67, 93). Thanks to decades of advances in theoretical polymer physics (70), the electrostatics can now be incorporated into analytically tractable models. The same analytical framework also helps us gain insight into IDP functions. The formalism relating an IDP sequence to its conformational features—embedded in an analytical framework based on physicochemical models—is termed the physical mathematics (PM) of IDPs.

PM can provide fundamental insights into issues of IDP biophysics, from deciphering the rules of IDP regulation and formulating principles for sequence design to detecting evolutionary trends. First, principles derived from closed-form mathematical expressions can decouple the complex interplay between biological and chemical regulation (Figure 1). Biological regulators (BRs), for the purposes of this review, include posttranslational modifications (PTMs), alternate splicing, and mutations that can change the composition and placement of amino acids in IDP sequences. The coupling of sequence changes (due to BRs) to environmental conditions [chemical regulators (CRs)] such as salinity, temperature, pH, and crowding can be complicated. However, mathematical formulas that describe sequences and their responses to CRs can reveal these complex relationships. Second, understanding the intertwined effects of BRs and CRs can help in designing novel sequences and tuning solution conditions to favor desired conformations. Third, PM provides metrics that can help classify functionally similar IDPs. Functionally similar IDPs can have low sequence homology (51), rendering functional classification challenging using traditional bioinformatic tools.

Mathematical relationships involving sequence, rather than composition alone, provide foundational insights that are not otherwise possible. For example, to fully appreciate the combination of BRs and CRs, we need to analyze numerous IDP sequences under diverse solution conditions. This causes a combinatorial explosion that is difficult to handle using computational tools. Likewise, understanding IDP functions often requires analyzing multiple long sequences (lengths greater than 500 amino acids; see 6) across diverse species, which is well beyond the capacities of current all-atom simulations. Coarse-grained simulations of IDPs (16, 21–25, 40, 71, 79, 89, 98) are being developed to address such challenges, particularly to model the emerging role of IDPs in forming biomolecular condensates via liquid–liquid phase separation. However, these simulations—although highly insightful and necessary to benchmark analytical theory—are not yet capable of describing functions of IDPs other than formation of condensates. Even coarse-grained simulations are not always feasible in the face of the combinatorial explosion involved in simulating IDP sequences under diverse solution conditions, nor can they simulate the large collections of long proteins that are typically needed to model evolutionary trends. Newly developed deep-learning tools (2, 49) built for predicting protein structures cannot predict ensembles and thus are not suitable for modeling IDPs. In fact, AlphaFold—not surprisingly—tends to yield very low-confidence structures when applied to IDPs (80).

Simple physics-based polymer models have revealed general principles in protein science (12–15, 34, 35, 50, 66, 81, 84, 88, 94, 95, 108). The analytical tractability of these models relies on two major simplifying assumptions. First, sequence complexity is reduced either by a homopolymer assumption (all amino acids are identical) or by adopting models with at most two different types of amino acids, for example, hydrophobic and polar. The second assumption—even in models with both hydrophobic and polar amino acids, or models with different flavors of monomers—often ignores the exact sequence positioning of the amino acids. Neglecting exact positioning is equivalent to averaging over multiple different sequences, assuming that the disorder is annealed (28, 38, 41, 83). Building physical models amenable to analytical treatments while respecting the exact placement of amino acids (sequence patterning) is challenging, despite its importance. The new era of PM of IDPs addresses this challenge (57) and is the focus of this review.

2. SEQUENCE-BASED METRICS CAN DESCRIBE THE CONFORMATIONS OF AN INTRINSICALLY DISORDERED PROTEIN AND REGION SEQUENCE

2.1. Brief Background on Homopolymer Theory and Applications to Intrinsically Disordered Proteins and Regions

We first define a few terms. The ensemble average radius of gyration, R_g, defined as $R_g = \sqrt{\langle r_g^2 \rangle}$ (where $\langle .. \rangle$ denotes the ensemble average and r_g refers to the radius of gyration for a given conformation), is typically used to describe the overall size of a polymer. Similarly, another useful metric for size is ensemble average end-to-end distance, R_{ee}, defined as $R_{ee} = \sqrt{\langle r_{ee}^2 \rangle}$, where r_{ee} is the stochastic value of the end-to-end distance for a given conformation. In polymer theory, for a homopolymer without any interactions (also termed a Gaussian chain), R_g and R_{ee} are related: $\langle r_g^2 \rangle = \langle r_{ee}^2 \rangle / 6 = Nbl/6$, where N is the number of monomers, b is the bond length, and l is the Kuhn length. The Kuhn length is a measure of the correlation in the direction of connecting bonds between different monomers (or amino acids, in the case of IDPs). For a protein, typical values are $b = 3.8$ Å and $l = 8$ Å (43, 111). Kuhn length can also vary between different amino acids. However, a uniform value of Kuhn length is a reasonable approximation for typical protein sequences. The scaling of R_g is the hallmark of Gaussian chain behavior and is generalized as $R_g \propto N^{\nu}$, where $\nu = 1/2$ recovers the Gaussian chain reference (27).

Any polymer with a dimension such as R_g less than the corresponding dimension of the Gaussian chain is considered to be collapsed. In common parlance, $\nu \approx 1/3$ is also termed a globule. A polymer is considered to be expanded when the dimension is greater than that of the Gaussian chain. IDP sequences tend to be enriched in charged amino acids (compared with the sequences of foldable proteins). Consistent with this statistical observation, early works found that charge composition can provide rules of thumb for distinguishing the globule and expanded states (61) and can influence several measures of IDP sizes, including R_g, R_{ee}, and ν (43, 62).

2.2. Intrinsically Disordered Protein and Region Conformations Depend on Charge Patterning

Charge composition gives the number of charges, but it gives no information about the placement of the charges in the sequence. Consider the two sequences shown in Figure 2. They have the same composition, or number of positive and negative charges, but the charges are distributed in different sequence orders, called patternings or decorations. Srivastava & Muthukumar (97) demonstrated that polymers with the same charge composition but different charge patternings can differ significantly in their sizes. More recently, Das & Pappu (19) revisited the role of charge decoration by simulating 30 sequences, each having 25 glutamic acids (with −1 charge each) and 25 lysines (with +1 charge each) distributed in different orders. They found that sequences with well-mixed or alternating positive and negative charges (similar to the top sequence in Figure 2) tend to have greater dimensions compared with sequences where positive and negative charges are segregated in blocks (similar to the bottom sequence in Figure 2). They defined an empirical charge-segregation metric to quantify this intuitive expectation (19). Thirumalai and colleagues (5) have also performed coarse-grained simulations to highlight the observation that charge composition alone is not sufficient to describe the subtle features of IDP conformations. The effects of varying charge patterning while keeping the same composition have also been observed in IDP functions (6, 72, 90). Intriguingly, the fact that sequence patterning alters conformation of the denatured state (of foldable proteins) has also been shown to be critical for function (10).

2.2.1. The sequence charge decoration metric can describe the global dimensions of intrinsically disordered proteins and regions.—

Recent advance in heteropolymer theory provide an analytical framework for determining the ensemble average end-to-end distance R_{ee} of a heteropolymer as a function of its sequence of charged monomers (amino acids, glutamic acid, aspartic acid, lysine, and arginine for proteins). The theory builds on a coarse-grained energy function with four essential ingredients ($I1$, $I2$, $I3$, and $I4$): $I1$ is the connectivity of monomers in the polymer; $I2$ is the two-body short-range interaction, which can be attractive or repulsive; $I3$ is the three-body short-range repulsive interaction; and $I4$ is the long-range electrostatic interaction among charged monomers (Figure 3). The three-body repulsive interaction is needed to avoid polymer collapse when the two-body interaction and electrostatics are highly attractive. The detailed form of the Hamiltonian (H) can be found in References 31 and 46.

In this derivation, the effective $\left\langle r_{ee}^2 \right\rangle$ is $\left\langle r_{ee}^2 \right\rangle = Nbl_r$, where l_r is the renormalized Kuhn length, which is different from the bare Kuhn length l. The details of sequence specificity are effectively captured by l_r. This technique, originally developed by Edwards & Singh (30), provides analytical tractability. It has been used in work on homopolymers, including polyelectrolytes (PEs) (32, 39, 68). The ratio of the two Kuhn lengths is defined as a dimensionless variable: $x = l_r/l$. For the Gaussian chain reference state, x is equal to unity. x will deviate from unity due to the composition and patterning of amino acids. The ranges of x provide different regimes of IDP conformation, such as coil-like ($x \approx 1$), globule ($x \ll 1$), and expanded ($x \gg 1$). An analytical expression for $F(x)$, the free energy as a function of x,

can be written by explicitly incorporating the sequence patterning as described by Equations 1 and 2:

$$\beta F(x) = \frac{3}{2}(x - \ln x) + \left(\frac{3}{2\pi}\right)^{3/2}\frac{\Omega}{x^{3/2}} + \omega_3\left(\frac{3}{2\pi}\right)^{3}\frac{B}{2x^3} + \frac{l_b}{l}\frac{Q}{x^{1/2}}\sqrt{\frac{6}{\pi}}, \quad 1.$$

where Ω, Q, and B contain details of the sequence and are given by

$$\Omega = \frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1}\omega_{m,n}(m-n)^{-1/2},$$
$$Q = \frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1}q_m q_n (m-n)^{1/2}, \quad 2.$$
$$B = \frac{1}{N}\sum_{p=3}^{N}\sum_{m=2}^{p-1}\sum_{n=1}^{m-1}\frac{(p-n)}{[(p-m)(m-n)]^{3/2}},$$

where l_b is the Bjerrum length, assumed to be l_b = 7.2 Å (298/T), and T is the absolute temperature. The first term defining the free energy F is the entropy respecting chain connectivity ($I1$ in Figure 2). The terms for Ω, B, and Q represent the contributions of interactions $I2$, $I3$, and $I4$, respectively (Figure 2). For a two-body residue pair (m, n, representing residue numbers), the specific short-range (or excluded volume) interaction parameter is given by $\omega_{m,n}$, while ω_3 is the three-body repulsive interaction parameter, which is assumed to be independent of amino acid type.

For a given sequence, Q is calculated from the sum in Equation 2 by assigning a negative charge (q = −1) to glutamic and aspartic acids and a positive charge (q = 1) to lysines and arginines. Histidines can also be assigned a charge of 0.5 if desired. Thus, Q explicitly accounts for sequence charge decoration (SCD). Two sequences with the same charge composition but different charge patterning will have different values of SCD. Reference 85 provides a detailed derivation of the SCD metric. Sequences with highly segregated positive and negative charges tend to have lower SCD values, and thus smaller sizes, compared with sequences with well-dispersed positive and negative charges. For example, the alternating and blocky sequences in Figure 2 have SCD values of −0.45 and −2.02, respectively. SCD captures the overall size variation of the sequences reported by Das & Pappu (19) when compared with the dimensions (R_g) generated by all-atom Monte Carlo (55) (Figure 4a) and coarse-grained molecular dynamics simulations (65). Equations 1 and 2 provide a formalism to directly estimate the size of a chain if the two-body interaction parameters $\omega_{m,n}$ and three-body repulsive parameter ω_3 values are known. Given a sequence with a high proportion of charged residues, as in the Das & Pappu sequences, it is reasonable to replace the two-body short-range interaction with a constant term (i.e., $\omega_{m,n} \approx \omega_2$). Figure 4b shows end-to-end distance as a function of mean-field ω_2 for a fixed value of ω_3 = 0.1 (for this typical choice of ω_3, see Reference 31). The difference in charge patterning (captured by SCD) is manifest in the chain dimensions between two sequences having the same charge composition (25 glutamic acids and 25 lysines) but different patterning (Figure 2). The variation in ω_2 can be attributed to changes in temperature or local solution condition inside

the cell that can happen due to weak nonspecific interactions (103). Figure 4 also shows that changing ω_2 can cause a chain to undergo a sharp transition in conformation.

2.2.2. Heuristic derivation of the sequence charge decoration metric.

—The functional form of the SCD metric can be appreciated with a scaling argument. The variational calculation stipulates a quantity I, defined as $I = \left\langle r_{ee}^2(H_r - H_t)\right\rangle - \left\langle r_{ee}^2\right\rangle\left\langle(H_r - H_t)\right\rangle = 0$(30, 85), where H_r is the Hamiltonian renormalized with the Gaussian form having effective bond length l_r, and H_t is the total form. The electrostatic contribution of the total Hamiltonian, defined as H_t^{el}, can be written as $H_t^{el} = \sum_{m,n} q_m q_n / |R_{m,n}|$ (ignoring constants). We also note the decomposition $r_{ee}^2 = r_{n,m}^2 + r_{m,n}^2 + r_{n,1}^2$. With these two relationships, the relevant term in I from the electrostatics becomes

$$\sum_{m,n} q_m q_n \left\langle r_{m,n}^2 \frac{1}{|r_{m,n}|}\right\rangle \propto \sum_{m,n} q_m q_n \langle |r_{m,n}|\rangle \propto \sum_{m,n} q_m q_n (m-n)^{1/2}, \quad 3.$$

where the last equality uses Gaussian chain (random walk) statistics, $[\langle |r_{m,n}^2|\rangle]^{1/2} = (m - n)^{1/2}$, neglecting all of the prefactors. For rigorous derivations, readers should consult Reference 85. SCD captures long-range correlations in the sequence, unlike the charge patterning metric κ (not to be confused with the inverse Debye length introduced below) introduced by Das & Pappu (19). For this reason, the two charge patterning metrics—although broadly correlated (85)—can differ in their ability to capture trends such as R_g variance in sequences having the same charge composition but different patterning (55).

2.3. Noncharge Patterning Can Also Influence the Sizes of Intrinsically Disordered Proteins and Regions

Equations 1 and 2 also provide a framework for modeling the sequence specificity of noncharged amino acids by choosing interaction parameters $\omega_{m,n}$ (between any residue pair m and n) without assuming the constant ω_2 described above. Zheng, Mittal, and colleagues (110) recently adopted a normalized hydrophobicity score λ_i for each amino acid i to estimate $\omega_{m,n} = \lambda_m + \lambda_n$. The new metric, which they called sequence hydropathy decoration (SHD), has been combined with SCD to predict scaling exponents ν and R_g for multiple sequences and benchmarked against coarse-grained simulations. The correlations between theoretically predicted and simulated values of ν were optimized to modify Ω and define SHD as

$$\mathrm{SHD} = \frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1} \omega_{m,n}(m-n)^{\beta}, \quad 4.$$

with $\beta = -1$, in contrast to $\beta = -1/2$ originally derived in Equation 2 (85). Zheng, Mittal, and colleagues' definition of SHD (with $\beta = -1$ instead of $\beta = -1/2$), along with their parameterization scheme of $\omega_{m,n} = \lambda_m + \lambda_n$, yielded a sequence-dependent equation to predict ν and R_g as

$$\nu = -0.0423\mathrm{SHD} + 0.0074\mathrm{SCD} + 0.701;\ R_g = \left[\frac{\gamma(\gamma+1)}{2(\gamma+2\nu)(\gamma+2\nu+1)}\right]^{1/2}(bl)^{1/2}N^{\nu}, \quad 5.$$

where $\gamma \approx 1.1615$ (37), $b = 3.8$ Å, and $l = 8$ Å. Reference 110 provides more details and comparisons with simulation data. Other noncharge patterning metrics—such as local hydrophobic clustering (HpC) (10) and local asymmetry in aromatic residues (Ω_{aro}) (64)—have also been used to design sequences. Future studies are needed to test and improve the choice of the $\omega_{m,n}$ parameter set by using the variational formalism given by Equations 1 and 2. There are several advantages of using this general formalism: (*a*) First-principle patterning metrics describe local and long-range correlations in sequence, unlike other intuitive local metrics such as HpC (10), κ (19), and Ω_{aro} (64); (*b*) chain dimensions and size can be directly predicted and compared (similar to Figure 2; see 85) against experiment and/or simulation beyond just seeking the size–metric correlation shown in Figure 4a; and (*c*) both electrostatic and nonelectrostatic interactions can be coupled in one framework, avoiding ad hoc fitting parameters to weight SHD and SCD.

2.4. Sequence Decoration Matrices Can Provide Detailed Conformational Features

Global metrics of chain dimensions such as R_{ee} and R_g and/or the scaling exponent ν are typically used to describe protein sizes (36, 43, 77). Internal distance profiles, defined as $\left\langle (r_i - r_j)^2 \right\rangle = \left\langle r_{i,j}^2 \right\rangle$ between any two amino acids i and j can provide conformational features beyond R_{ee}, R_g, and ν. For homopolymers, such profiles are mostly redundant because homopolymers display fractal-like behavior with an approximate scaling relation $\left\langle (r_i - r_j)^2 \right\rangle \propto |i-j|^{2\nu}$. In contrast, heteropolymers such as IDPs display strong heterogeneity in their internal distance profiles (19, 63) and may exhibit more nuanced features.

Recent progress in PM provides a framework for computing these ensemble average distances—besides global dimensions—as functions of sequence. The internal distance profiles are calculated as $\left\langle (r_i - r_j)^2 \right\rangle = |i-j| b l_r(i,j)$, where $l_r(i,j)$ is the residue pair (i, j)-specific renormalized Kuhn length, analogous to the calculation of $\left\langle r_{ee}^2 \right\rangle$ given above. These distances can be written in terms of a nondimensionalized quantity, $x_{i,j} = l_r(i,j)/l$, with a corresponding free energy function, $F(x_{i,j})$ (46):

$$\beta F(x_{i,j}) = \frac{3}{2}(x_{i,j} - \ln x_{i,j}) + \left(\frac{3}{2\pi}\right)^{3/2}\frac{\mathrm{SHDM}_{i,j}}{x_{i,j}^{3/2}} + \left(\frac{3}{2\pi}\right)^{3}\frac{\omega_3 T_{i,j}}{2(i-j)x_{i,j}^{3}} + \frac{l_b}{l}\sqrt{\frac{6}{\pi}}\frac{\mathrm{SCDM}_{i,j}}{x_{i,j}^{1/2}}. \quad 6.$$

Equation 6 is a generalization of Equation 1 with four terms that account for the physical contributions of $I1$, $I2$, $I3$, and $I4$ shown in Figure 3. The charge patterning is encoded by a specific metric $\mathrm{SCDM}_{i,j}$ for a given residue pair i, j, for which the internal distance is to be computed. Consequently, the general formalism yields an SCD matrix (SCDM) with elements $\mathrm{SCDM}_{i,j}$ given by

$$\mathrm{SCDM}_{i,j} = \frac{1}{(i-j)}\left[\sum_{m=j}^{i}\sum_{n=1}^{j-1} q_m q_n \frac{(m-j)^2}{(m-n)^{3/2}} + \sum_{m=j+1}^{i}\sum_{n=j}^{m-1} q_m q_n (m-n)^{1/2} + \sum_{m=i+1}^{N}\sum_{n=1}^{j-1} q_m q_n \frac{(i-j)^2}{(m-n)^{3/2}} + \sum_{m=i+1}^{N}\sum_{n=j}^{i} q_m q_n \frac{(i-n)^2}{(m-n)^{3/2}}\right]. \quad 7.$$

It is instructive to note that SCD, defined above, is just one element of the SCDM, and is only applicable to describing R_{ee}. Notably, we have $\mathrm{SCDM}_{i=N,j=1} = \mathrm{SCD}$ (assuming $N \approx N-1$, which is reasonable for large N).

The noncharged patterning can be similarly generalized by creating an SHD matrix (SHDM) defined as

$$\mathrm{SHDM}_{i,j} = \frac{1}{(i-j)}\left[\sum_{m=j}^{i}\sum_{n=1}^{j-1} \omega_{m,n} \frac{(m-j)^2}{(m-n)^{5/2}} + \sum_{m=j+1}^{i}\sum_{n=j}^{m-1} \omega_{m,n} (m-n)^{-1/2} + \sum_{m=i+1}^{N}\sum_{n=1}^{j-1} \omega_{m,n} \frac{(i-j)^2}{(m-n)^{5/2}} + \sum_{m=i+1}^{N}\sum_{n=j}^{i} \omega_{m,n} \frac{(i-n)^2}{(m-n)^{5/2}}\right]. \quad 8.$$

The derivation of a high-dimensional SCDM and SHDM also shows the power of variational calculation that is not limited to computing a scalar metric such as SCD; a single charge patterning metric (κ, defined in 19; not to be confused with the inverse Debye length); or SHD, local hydropathy cluster (HpC) (10), and Ω_{aro} (64).

The detailed expression for the three-body repulsion denoted by $T_{i,j}$, omitted here for brevity, can be found in Reference 46. This formalism provides a framework for quantitatively estimating the effects of electrostatics for highly charged sequences by assuming $\omega_{m,n} \approx \omega_2$. Within this approximation, the role of charge patterning on local dimensions is exclusively modeled by the SCDM. For a given value of ω_2, ω_3 and a pair of residues *i, j*, the free energy equation, Equation 6, is minimized to determine $x_{i,j,\mathrm{min}}$. The ensemble average distance ($R_{i,j}$) between these two residues *i, j* is calculated using $R_{i,j}^2 = \langle r_{i,j}^2 \rangle = |i-j| b l x_{i,j,\mathrm{min}}$.

Figure 5 shows distance maps under zero-salt conditions for two different IDP sequences with specific values of ω_2 and ω_3. The best estimates for ω_2 and ω_3 were obtained by matching the values of $\langle (r_i - r_j)^2 \rangle$ generated in an all-atom simulation. The details of the protocol for determining ω_2 and ω_3 are given in Reference 46. Reasonable agreement between analytical theory and the all-atom simulation, noted in Figure 5, shows that the SCDM can reveal the prominent features of distance maps from the placement of charges in the sequence. It is interesting to note that different residue pairs—within the same chain—can exhibit different conformations, even as different as expanded and collapsed (Figure 5; 46, figure 4). Approaches rooted in homopolymer models ignore variations in these local features. A recent mathematical model supported by coarse-grained simulation has also shown independence of the distribution of R_{ee} and R_g for IDPs (96), in stark contrast to homopolymer predictions and highlighting limits of homopolymer-centric models.

Maps of the SCDM can be generated from sequence charge information even when ω_3 and ω_2 values are not known. These graphs can provide critical information about the sets of amino acid residue pairs that experience repulsive (or attractive) electrostatic interactions, which expand (or compact) the chain. Figure 5c,d shows the electrostatic contributions to the intrachain interaction topology (quantified by the SCDM) for two IDPs (for which the distance maps are shown in Figure 5a,b). The most repulsive regions correspond to the most expanded regions in the distance maps. This feature of SCDM maps has proven useful for functional classifications (see Section 4.3).

3. INTRINSICALLY DISORDERED PROTEIN AND REGION CONFORMATIONS ARE SENSITIVE TO BIOLOGICAL AND CHEMICAL REGULATORS

3.1. Physical Mathematics Can Identify Phosphorylation Hot Spots

PTMs such as phosphorylation can alter IDP sequences by adding negatively charged phosphate groups to neutral amino acids such as serine and threonine (3). How do these modifications alter IDP conformations? Equation 6 provides a framework for predicting PTM-induced changes in SCDMs and consequent changes in internal distance profiles. Figure 6 illustrates phosphorylation-induced conformational changes for the wild-type (WT) IDP P0A8H9 and its two variants S2T15 and S54S56. The first variant, S2T15, is modified by adding negative charges (to mimic the effect of phosphorylation) at amino acids 2 and 15 in the WT sequence. The second variant, S54S56, has negative charges at residues 54 and 56 in the WT sequence. Differences in distance maps between the WT and the variants show phosphorylation can significantly alter distance profiles. Furthermore, it is evident that the choice of phosphorylation site matters. Both the S54S56 and S2T15 variants have the same charge composition, with near identical patterning except at two amino acids. These differences—due to the long-range nature of electrostatics—are enough to significantly alter the SCDMs, resulting in vastly different distance maps. Figure 6a,b shows changes in the distance maps that arise from the changes in the SCDMs. These distance maps largely agree with the results from the all-atom simulations. There are also noticeable disagreements that could be due to finer details of the all-atom simulation that the coarse-grained energy function (H) ignores.

The results of phosphorylating different residues in the WT IDP P0A8H9 also show that IDPs propagate phosphorylation signals to distant parts of the sequence. For example, modifications at residue numbers 2 and 15 (Figure 6a) can induce changes far away. PM provides a formalism to search through numerous combinations of phosphorylation sites to identify those hot spots that can induce drastic changes locally and/or far from the site of the modification.

3.2. Intrinsically Disordered Protein and Region Conformations Are Sensitive to Salt

The dimensions of charged polymers depend on salt concentration (an example of a CR). Consider a PE consisting of charged monomers of only one type—either positive or negative. For a PE-like IDP, the electrostatics is purely repulsive; consequently, added

salt will screen the repulsions, and the chain will become more compact. However, IDPs are typically polyampholytes with both positive and negative charges present in the same sequence. Would polyampholytic IDPs contract or expand with salt? The salt dependence can be modeled by a screened Coulomb potential within a Debye-Huckle formalism with an analytical solution. First, we present the salt dependence of the end-to-end distance (R_{ee}) of an IDP. The electrostatics contribution (Q in Equation 2) is modified to Q' as a function of salt. Specifically, Q' is expressed in terms of κ, the inverse Debye length is defined as $\kappa^2 = 8\pi l_b c_s$, and c_s is the salt concentration (for the exact expression of Q', see 45, equation 3).

To predict whether an IDP will expand or collapse with the addition of salt, Q' can be expanded in the limit of zero salt, i.e., small κl, as

$$Q' \approx \frac{1}{2}\sqrt{\frac{6\pi}{x}}\frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1} q_m q_n \sqrt{m-n} - (\kappa l)\frac{\pi}{2}\frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1} q_m q_n (m-n) + \ldots \mathscr{H}.\mathscr{O}., \quad 9.$$

where $\mathscr{H}.\mathscr{O}.$ are higher-order terms in κl. The first term can be identified as the SCD (ignoring constants) expected for the end-to-end distance. The second term yields a new sequence charge patterning metric ($SCD_{low\ salt}$) (see Reference 45, equations 4 and 5) defined as

$$SCD_{low\ salt} = \frac{1}{N}\sum_{m=2}^{N}\sum_{n=1}^{m-1} q_m q_n (m-n). \quad 10.$$

In the vicinity of zero salt, if $SCD_{low\ salt}$ is positive (negative), the ensemble average end-to-end distance will shrink (expand) upon the addition of salt. An immediate consequence is that two IDPs with identical charge compositions but different patternings can have different responses to salt due to the different signs of $SCD_{low\ salt}$. Reference 45 provides examples of this. The trend predicted by the $SCD_{low\ salt}$ metric agrees with the single-molecule Förster resonance energy transfer measurements of Schuler and colleagues (67) on a limited set of four proteins: CspTm, integrase, and the N-terminal and C-terminal ends of prothymosin-α. However, the salt dependence of charged–polar and polar–polar interactions can cause deviations from the simple trend based on the sign of $SCD_{low\ salt}$ alone. Another important physical contribution, which we do not discuss, is the salting-out effect—as proposed by Zheng and colleagues (104)—that can modulate salt-dependent shape changes in IDP sequences.

The discussion above shows the role of $SCD_{low\ salt}$ in determining sequence-specific, salt-dependent responses in end-to-end distances. The strong variation noted among specific residue pair distance profiles suggests that the trends observed for end-to-end distances may not be representative of the entire chain. Beyond just the end-to-end distance, the salt dependence of internal distances is also revealed by the salt-dependent (i.e., κl) SCDM, defined as $SCDM_{i,j}(\kappa l) = Q'_{i,j}/(i-j)$; salt-dependent $Q'_{i,j}$ is defined in Reference 46 (equations 4 and 5).

Expanding near zero salt, an SCDM matrix of low-salt metrics defined as $\mathrm{SCDM}_{\text{low salt}}$ (analogous to $\mathrm{SCD}_{\text{low salt}}$, defined above for the end-to-end distance) is given by

$$\mathrm{SCDM}_{\text{low salt},i,j} = \frac{1}{(i-j)}\left[\sum_{m=j}^{i}\sum_{n=1}^{j-1} q_m q_n \frac{(m-j)^2}{(m-n)} + \sum_{m=j+1}^{i}\sum_{n=j}^{m-1} q_m q_n (m-n) + \sum_{m=i+1}^{N}\sum_{n=1}^{j-1} q_m q_n \frac{(i-j)^2}{(m-n)} + \sum_{m=i+1}^{N}\sum_{n=j}^{i} q_m q_n \frac{(i-n)^2}{(m-n)}\right]. \quad 11.$$

Positive (negative) values of $\mathrm{SCDM}_{\text{low salt},i,j}$ would cause ensemble average $\langle r_{i,j}^2 \rangle$ to shrink (expand) with the addition of salt. An intriguing outcome of this is that some sequences may have both positive and negative elements in their $\mathrm{SCDM}_{\text{low salt}}$, implying drastically different responses to salt at different distances. Figure 7 shows an example of an $\mathrm{SCDM}_{\text{low salt}}$ map for protein P0A8H9 with positive values in some regions expected to shrink—in contrast to other regions expected to expand—upon addition of salt near the zero-salt regime. Differential salt response between different pairs of amino acid residues (in the context of the full protein) have been experimentally observed by Schuler and colleagues (67). Hofmann and colleagues (102) also reported different salt responses for different segments of the full protein. However, we emphasize that theoretical predictions based on $\mathrm{SCDM}_{\text{low salt}}$ are purely within the Debye-Hückle approximation. Deviations from these predictions may also occur for additional driving forces, such as polar-charge and salting-out effects not included in this review. As expected, $i = N$, $j = 1$, and $N \approx (N-1)$ yield $\mathrm{SCD}_{\text{low salt}}$ in Equation 10.

4. PHYSICAL MATHEMATICS MODELS FOR INTRINSICALLY DISORDERED PROTEIN AND REGION FUNCTIONS

PM is increasingly being used in modeling the functional features of IDPs. IDPs participate in many biological processes, from signaling and cellular differentiation to formation of membraneless organelles. Many of these functions rely on interactions among multiple copies of the same IDP or between IDPs and other macromolecules in disordered modes. Many of these modes of interaction are amenable to approximate models with closed-form solutions.

4.1. Analytical Models Predict Sequence-Dependent Phase Separation

Solutions of IDPs, either by themselves or in association with other macromolecules, can phase separate into protein-rich and dilute liquid phases. This phenomenon, known as liquid–liquid phase separation (LLPS), is emerging as a key mechanism in the formation of membraneless organelles and in numerous biological functions (4, 8, 42, 48, 52, 82, 91). In many such examples, IDPs in a homogeneous phase will phase separate below a critical temperature T_c. Below T_c, for a range of bulk protein densities, the solution will coexist in both the protein-rich and dilute phases, yielding a phase coexistence curve. Experiments are beginning to measure sequence-dependent T_cs and phase coexistence curves by modulating the amino acid composition or by shuffling the amino acid patterning at a fixed composition (11, 64, 72, 87, 100). In a seminal work, Chan and colleagues

(56) developed a theoretical formalism using the random phase approximation (RPA; detailed below) to determine T_c as a function of sequence charge patterning. The sequence dependence of RPA theory explained remarkably well experimentally observed differences in phase-separation propensities, and their salt dependence, between two sequences: WT Ddx4 protein and its charge-scrambled (CS) version (64, 72). The CS version, created by keeping the same composition as in the WT but dispersing charges more evenly, did not phase separate, in contrast to the WT sequence. The theory of Chan and colleagues was the first of its kind to successfully explain the sequence dependence of an IDP's propensity for phase separation. Figure 8 shows an application of the same theory to contrasting the phase separation propensity of the two different phosphorylated versions of P0A8H9 discussed above. The theory predicts that the phase separation propensity of the S54S56 variant is higher than that of the S2T15 variant, indicating the critical role of phosphorylation hot spots in determining LLPS. Both modifications have same charge composition but different patterning.

4.1.1. Simple explanation of random phase approximation and its improvement.—

Models of phase separation must account for two key elements: the presence of multiple chains interacting with one another and the connectivity of each individual chain. The RPA formalism introduces field variables $\{\phi\}$ corresponding to the density of monomers and describing the presence of multiple chains. These variables are not directly constrained by chain connectivity. Variables $\{R\}$, describing the single-chain conformational ensemble, in contrast, must respect the connectivity of the monomers. The central aim of RPA is to account for an effective field (a function of $\{\phi\}$) experienced by a single chain due to the presence of other chains. Consequently, the distribution of conformations $\{R\}$ depends on $\{\phi\}$, and $\{\phi\}$ itself is a function of $\{R\}$. For analytical tractability, RPA makes a simplifying approximation by neglecting higher-order fluctuations in $\{\phi\}$, which is reasonable when densities are not too low. As part of this simplification, any explicit dependence of $\{R\}$ on $\{\phi\}$ is ignored, and contributions from $\{\phi\}$ are kept only to the second order. This simplification allows chain connectivity to be accounted for within the simplest framework of a Gaussian chain. The simplified RPA predicts that the T_c of a PE (an IDP with only one type of charge) will increase with the length of the PE, contradicting earlier theoretical results, which had predicted that T_c saturates with increased chain length (70). A modified RPA called renormalized Gaussian RPA (rG-RPA) adopts a self-consistent formalism in which correlations in $\{\phi\}$ are calculated using a single-chain conformational ensemble $\{R\}$, which in turn depends on $\{\phi\}$ (54). This modification incorporates single-chain conformations correctly in building models for multichain interactions. rG-RPA recovers the correct limiting results for PE phase behavior. Furthermore, rG-RPA is able to explain differences in phase-separation propensities in WT and CS forms of Ddx4, capturing the sequence effect. A unified analytical framework such as rG-RPA demonstrates the power of first-principle PM models that can describe both polyampholyte and PE phase behavior (54). It is also important to note that the sequence-specific RPA formalism—despite the assumptions of weak fluctuations—is different from purely mean-field Flory-Huggins theory (for the interaction term), which completely neglects spatial dependence and fails to capture sequence specificity, providing another example of the importance of PM.

4.1.2. Unresolved issues.—There are still many caveats to rG-RPA. First, the theory assumes a uniform dielectric constant, even in the presence of salt. Recent work from the Chan group (23, 101) has addressed this issue and has shown that the approximation may cause modest deviations but not major changes. Second, extensions of the rG-RPA models are needed to describe the combinations of charge and noncharge patterning in chains that have charges, π, and hydrophobic interactions (23, 64, 72). Coarse-grained simulations are being developed with different interaction parameters to explain experimental data (23, 64, 99). Insights gained from these simulations may help to further advance first-principles analytical theory using rG-RPA. It is also important to develop PM models to estimate the effect of neglecting fluctuations beyond the second order, which are expected to be important in systems at low densities. At present, strong fluctuations are modeled by numerical solutions of stochastic differential equations using a self-consistent field theory formalism (18, 65). However, the quantitative, sequence-specific effects of these corrections are still unclear. For calculating the phase diagrams of PEs, in analytical calculations where higher-order terms to infinite order were approximately added within a closure relation (69), corrections beyond RPA were found to be insignificant compared with other terms (70). Improved models of the actual temperature dependence of the interaction parameters are also needed to address the lower critical solution temperatures, where proteins phase separate upon increases in temperature (26). Analytical models are needed to delineate the competition between liquid and other phases (such as solids and gels) that dictate the properties of condensates (7, 86).

4.1.3. Single-chain and many-chain physics can be related.—The success of rG-RPA also highlights the importance of single-chain conformations in predicting multichain phase behavior. The apparent interdependence between single-chain and multichain physics, inspired by earlier homopolymer results, has been elucidated in recent work. First, Lin & Chan (55) showed that the T_cs predicted using RPA [for Das & Pappu (19) sequences] strongly correlate with the SCD values for the single chain. Consistent with their observation, S54S56 has lower SCD and a higher propensity to phase separate (Figure 8) compared with S2T15 even though both sequences have same charge composition. Mittal and colleagues (24) performed coarse-grained simulations to establish the relationships between the temperature dependence of single-chain conformations and phase diagrams. Pappu and colleagues (108) have quantitatively shown how the temperature dependence of single-chain conformations can be used to predict solution phase diagrams. More recently, Lindorff-Larsen and colleagues (99) optimized a coarse-grained model against single-chain biophysical properties to predict multichain phase separation. Rana and colleagues (76) made an intriguing observation that normalized SCD can approximately determine competition between phase separation and aggregation in coarse-grained models of disordered proteins. These and other experimental studies (64, 78) highlight the importance of studying single-chain behavior to advance our understanding of solution behavior.

4.2. A Joint Sequence Charge Decoration Metric Describes Complexation Between Two Chains

Complexation is a frequent mode of protein–protein interaction. Much of the dimeric complexation of proteins leads to the formation of ordered structures. Schuler and colleagues (9) demonstrated a novel interaction mechanism in which two dissimilar IDPs can form a fully disordered complex. Their coarse-grained simulations show that simple electrostatics modeled by Debye-Hückle theory is sufficient to capture experimentally observed conformational features. Chan and colleagues (1) provided an analytical theory to quantify the binding constant between identical or dissimilar (*A* and *B*) pairs of IDPs. Intriguingly, the theory finds a metric, joint SCD (jSCD), that correlates with the binding constant. jSCD is defined as

$$\mathrm{jSCD}(\{q_A, q_B\}) = -\frac{1}{2N_A N_B}\sum_{s,t=1}^{N_A}\sum_{l,m=1}^{N_B} q_s^A q_t^A q_l^B q_m^B [|s-t| + |l-m|]^{1/2}, \qquad 12.$$

where $\{q_A\}$ is the sequence of charged amino acids in chain *A*, and $\{q_B\}$ is that of chain *B*. Reference 1 provides the derivation of jSCD and its application to different pairs of heteropolymer sequences. Coacervation of polycation and polyanion chains—a limiting case of the example studied by Schuler and colleagues—is attracting a lot of interest in studies of synthetic polymers (60, 74). An interesting problem arises at the interface of complexation and phase separation. Recent work by Chan and colleagues (58) provided a mathematical model to couple the two processes and compare the results with experiments.

4.3. Charge Decoration Metrics Can Be Useful for Functional Annotation

Functionally similar intrinsically disordered regions (IDRs)—disordered segments between two structured regions—can have low sequence similarity detectable by traditional sequence alignment algorithms. Consequently, sequence and/or structure alignments, which are traditionally applied to predicting the functions of folded proteins, are not useful for predicting functionally similar IDRs. How then do we classify functionally similar IDRs? Is it possible to identify functional similarities among IDRs by finding similarities in their sequence decoration metrics, even when they are not similar in sequence alignments? Large-scale bioinformatic analyses have found underlying molecular features that are hidden in sequences but can be used to discern functionally similar proteins (105). Given that the SCDM can provide a high-dimensional representation of the charged patterning of a protein, it is natural to test the ability of SCDMs to classify IDPs (47, 73).

Further motivation for using SCDMs in discerning functional classes comes from two observations: (*a*) SCDMs have the ability to describe subtle features of conformations, and (*b*) conformational features can be similar among homologous IDPs (75). Indeed, high-dimensional features embedded in the SCDM can be used to distinguish functional and nonfunctional proteins within the Ste50 protein family (47). Ste50 is an IDR between two folded domains (107). It is critical to cell growth and shape. Moses and colleagues (107) classified five IDRs in functional and nonfunctional categories based on cell growth, shape, and basal protein expression. The five IDRs are the WT Ste50 from *Saccharomyces cerevisiae*; the doubly phosphorylated version of that WT; the WT Ste50 from a different

organism, *Lachancea kluyveri*; and two other, unrelated IDRs, RAD26 and Pex5. The SCDM-based classification scheme is consistent with the experimental classification scheme (see Reference 47 for more).

Briefly, the classification algorithm proceeds in a few steps (47): (*a*) An IDR's SCDM is first binarized to capture the repulsive or attractive nature of the electrostatics interactions, (*b*) binarized SCDMs (bSCDMs) are properly resized to compare proteins of different sizes, and (*c*) these bSCDMs are subsequently decomposed using principal component analysis to keep only the essential features and avoid any spurious information that may cause an artifact. The same algorithm was also reasonably successful in functionally classifying two other protein families: (*a*) protein family PSC, which is critical in chromatin remodeling (6), and (*b*) the RAM region of the Notch receptor (90). Each of these three proteins is highly charged, which justifies the use of the SCDM for functional classification while neglecting the contribution of other interactions (embedded in the SHDM). However, for IDPs with fewer charges, SHDM metrics in combination with the SCDM may be important for functional predictions. Recent work from Moses and colleagues (105, 106) revealed multiple molecular features—different from the patterning metrics noted above—that can be used to rescue and predict function.

ACKNOWLEDGMENTS

We acknowledge support from National Institutes of Health grant R01GM138901. We are indebted to H.S. Chan, M. Gruebele, J.D. Forman-Kay, B. Schuler, D. Thirumalai, and W. Zheng for critical reading of the manuscript. We thank T. Firman, W. Kimbro, Y.H. Lin, M. Muthukumar, R.V. Pappu, S. Pathak, V Prabhu, A. Sorano, L. Sawle, S. Vaiana, T. Sosnick, and members of the Protein Folding Consortium (Research Coordination Network supported by National Science Foundation award number 1516959) for many insightful discussions and/or collaboration over many years. We also acknowledge Sarina Bromberg for help with figures and manuscript edits.

LITERATURE CITED

1. Amin AN, Lin YH, Das S, Chan HS. 2020. Analytical theory for sequence-specific binary fuzzy complexes of charged intrinsically disordered proteins. J. Phys. Chem. B 124:6709–20 [PubMed: 32639157]
2. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. 2021.Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–76 [PubMed: 34282049]
3. Bah A, Forman-Kay JD. 2016. Modulation of intrinsically disordered protein function by post-translational modifications. J. Biol. Chem 291:6696–705 [PubMed: 26851279]
4. Banani SF, Lee HO, Hyman AA, Rosen MK. 2017. Biomolecular condensates: organizers of cellular biochemistry. Nat. Rev. Mol. Cell Biol 18:285–98 [PubMed: 28225081]
5. Baul U, Chakraborty D, Mugnai ML, Straub JE, Thirumalai D. 2019. Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. J. Phys. Chem. B 123(16):3462–74 [PubMed: 30913885]
6. Beh LY, Colwell LJ, Francis NJ. 2012. A core subunit of polycomb repressive complex 1 is broadly conserved in function but not primary sequence. PNAS 109(18):E1063–71 [PubMed: 22517748]
7. Bhandari K, Cotten MA, Kim J, Rosen MK, Schmit JD. 2021. Structure-function properties in disordered condensates. J. Phys. Chem. B 125(1):467–76 [PubMed: 33395293]
8. Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, et al. 2018. Protein phase separation: a new phase in cell biology. Trends Cell Biol 28(6):420–35 [PubMed: 29602697]
9. Borgia A, Borgia MB, Bugge K, Kissling VM, Heidarsson PO, et al. 2018. Extreme disorder in an ultrahigh-affinity protein complex. Nature 555:61–66 [PubMed: 29466338]

10. Bowman M, Riback J, Rodriguez A, Guo H, Li J,et al. .2020. Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. PNAS 117(38):23356–64 [PubMed: 32879005]
11. Brady JP, Farber PJ, Sekhar A, Lin YH, Huang R, et al. 2017. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. PNAS 114(39):E8194–203 [PubMed: 28894006]
12. Camacho CJ, Schanke T 1997. From collapse to freezing in random heteropolymers. Europhys. Lett 37:603–8
13. Camacho CJ, Thirumalai D. 1993. Minimum energy compact structures of random sequences of heteropolymers. Phys. Rev. Lett 71:2505–8 [PubMed: 10054697]
14. Chakraborty AK. 2001. Disordered heteropolymers: models for biomimetic polymers and polymers with frustrating quenched disorder. Phys. Rep 342:1–61
15. Chan HS, Dill KA. 1994. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys 100:9238–57
16. Choi JM, Dar F, Pappu RV 2019. Lassi: a lattice model for simulating phase transitions of multivalent proteins. PLOS Comput. Biol 15:e1007028 [PubMed: 31634364]
17. Cohan M, Ruff K, Pappu R. 2019. Information theoretic measures for quantifying sequence-ensemble relationships of intrinsically disordered proteins. Protein Eng. Des. Sel 32(4):191–202 [PubMed: 31375817]
18. Danielson S, McCarty J, Shea JE, Delaney K, Fredrickson G.2019. Molecular design of self-coacervation phenomena in block polyampholytes. PNAS 116(17):8224–32 [PubMed: 30948640]
19. Das RK, Pappu RV. 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. PNAS 110(33):13392–97 [PubMed: 23901099]
20. Das RK, Ruff KM, Pappu RV. 2015. Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr. Opin. Struct. Biol 32:102–12 [PubMed: 25863585]
21. Das S, Amin AN, Lin YH, Chan HS. 2018. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. Phys. Chem. Chem. Phys 20:28558–74 [PubMed: 30397688]
22. Das S, Elsen A, Lin YH, Chan HS. 2018. A lattice model of charge-pattern-dependent polyampholyte phase separation. J. Phys. Chem. B 122(21):5418–31 [PubMed: 29397728]
23. Das S, Lin YH, Vernon RM, Forman-Kay JD, Chan HS. 2020. Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. PNAS 117(46):28795–805 [PubMed: 33139563]
24. Dignon GL, Zheng W, Best RB, Kim YC,Mittal J. 2018. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. PNAS 115(40):9929–34 [PubMed: 30217894]
25. Dignon GL, Zheng W, Kim YC, Best RB, Mittal J. 2018. Sequence determinants of protein phase behavior from a coarse-grained model. PLOS Comput. Biol 14(1):e1005941 [PubMed: 29364893]
26. Dignon GL, Zheng W, Kim YC, Mittal J. 2019. Temperature-controlled liquid-liquid phase separation of disordered proteins. ACS Cent. Sci 5(5):821–30 [PubMed: 31139718]
27. Dill KA, Bromberg S. 2002. Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology Oxford, UK: Garland. 1st ed.
28. Dobrynin AV, Colby RH, Rubinstein M. 2004. Polyampholytes. J. Polym. Sci. B 42:3513–38
29. Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol 6(3):197–208 [PubMed: 15738986]
30. Edwards SF, Singh P 1979. Size of a polymer molecule in solution. Part 1: Excluded volume problem. J. Chem. Soc. Faraday Trans. 2 75:1020–29
31. Firman T, Ghosh K. 2018. Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins. J. Chem. Phys 148(12):123305 [PubMed: 29604827]
32. Ghosh K, Carri GA, Muthukumar M. 2001. Configurational properties of a single semiflexible polyelectrolyte. J. Chem. Phys 115:4367–75

33. Ghosh K, de Graff AMR, Sawle L, Dill KA. 2016. Role of proteome physical chemistry in cell behavior. J. Phys. Chem. B 120(36):9549–63 [PubMed: 27513457]
34. Ghosh K, Dill KA. 2009. Theory for protein folding cooperativity: helix bundles. J. Am. Chem. Soc 131(6):2306–12 [PubMed: 19170581]
35. Ghosh K, Ozkan SB, Dill KA. 2007. The ultimate speed limit to protein folding is conformational searching. J. Am. Chem. Soc 129:11920–27 [PubMed: 17824609]
36. Gomes GNW, Krzeminski M, Namini A, Martin EW, Mittag T, et al. 2020. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. J. Am. Chem. Soc 142(37):15697–710 [PubMed: 32840111]
37. Guillou J, Zinn-Justin J. 1977. Critical exponents for the n-vector model in three dimensions from field theory. Phys. Rev. Lett 39:95–98
38. Gutin AM, Shakhnovich EI. 1994. Effect of a net charge on the conformation of polyampholytes. Phys. Rev. E 50:R3322–25
39. Ha BY, Thirumalai D. 1999. Conformations of a polyelectrolyte chain. Phys. Rev. A 46:R3012–15
40. Harmon TS, Holehouse AS, Rosen MK, Pappu RV 2017. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. eLife 6:e30294 [PubMed: 29091028]
41. Higgs P, Joanny J. 1991. Theory of polyampholyte solutions. J. Chem. Phys 94:1543–54
42. Hnisz D, Shrinivas K, Young RA, Chakraborty A, Sharp PA. 2017. A phase separation model for transcriptional control. Cell 169(1):13–23 [PubMed: 28340338]
43. Hofmann H, Soranno A, Borgia A, Gast K, Nettels D, Schuler B. 2012. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. PNAS 109(40):16155–60 [PubMed: 22984159]
44. Holehouse A, Das R, Ahad J, Richardson M, Pappu R. 2017. Cider: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. Biophys. J 112:16–21 [PubMed: 28076807]
45. Huihui J, Firman T, Ghosh K. 2018. Modulating charge patterning and ionic strength as a strategy to induce conformational changes in intrinsically disordered proteins. J. Chem. Phys 149:085101 [PubMed: 30193467]
46. Huihui J, Ghosh K. 2020. An analytical theory to describe sequence-specific inter-residue distance profiles for polyampholytes and intrinsically disordered proteins. J. Chem. Phys 52:161102
47. Huihui J, Ghosh K. 2021. Intrachain interaction topology can identify functionally similar intrinsically disordered proteins. Biophys. J 120(10):1860–68 [PubMed: 33865811]
48. Hyman AA, Weber CA,Julicher F. 2014. Liquid-liquid phase separation in biology. Annu. Rev. Cell. Dev. Biol 30:39–58 [PubMed: 25288112]
49. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–89 [PubMed: 34265844]
50. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, et al. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. PNAS 101:12491–96 [PubMed: 15314214]
51. Lange J, Wyrwicz LS, Vriend G. 2016. KMAD: knowledge-based multiple sequence alignment for intrinsically disordered proteins. Bioinformatics 32:932–36 [PubMed: 26568635]
52. Larson AG, Elnatam D, Keenen MM, Trnka MJ,Johnston JB, et al. 2017. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature 547:236–40 [PubMed: 28636604]
53. Lazar T,Martínez-Pérez E, Quaglia F, Hatos A, Chemes LB, et al. 2021. PED in 2021: a majorupdate of the protein ensemble database for intrinsically disordered proteins. Nucleic Acids Res 49:D404–11 [PubMed: 33305318]
54. Lin YH, Brady J, Chan HS, Ghosh K. 2020.A unified analytical theory of heteropolymers for sequence specific phase behaviors of polyelectrolytes and polyampholytes. J. Chem. Phys 152:045102 [PubMed: 32007034]
55. Lin YH, Chan HS. 2017. Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. Biophys. J 112(10):2043–46 [PubMed: 28483149]

56. Lin YH, Forman-Kay JD, Chan HS. 2016. Sequence-specific polyampholyte phase separation in membraneless organelles. Phys. Rev. Lett 117:178101 [PubMed: 27824447]
57. Lin YH, Forman-Kay JD, Chan HS. 2018. Theories for sequence-dependent phase behaviors of biomolecular condensates. Biochemistry 57:2499–508 [PubMed: 29509422]
58. Lin YH, Wu H, Jia B, Zhang M, Chan HS. 2022. Assembly of model postsynaptic densities involves interactions auxiliary to stoichiometric binding. Biophys. J 121:4–6 [PubMed: 34932937]
59. Lincoff J, Haghighatlari M, Krzeminski M, Teixeira JMC, Gomes GN, et al. 2020. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. Commun. Chem 3:74 [PubMed: 32775701]
60. Madinya JJ, Chang LW, Perry SL, Sing CE. 2020. Sequence-dependent self-coacervation in high charge-density polyampholytes. Mol. Syst. Des. Eng 5:632–44
61. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV 2010. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. PNAS 107:8183–88 [PubMed: 20404210]
62. Marsh JA, Forman-Kay JD. 2010. Sequence determinants of compaction in intrinsically disordered proteins. Biophys. J 98:2383–90 [PubMed: 20483348]
63. Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu RV, Mittag T. 2016. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. J. Am. Chem. Soc 138:15323–35 [PubMed: 27807972]
64. Martin EW, Holehouse AS, Peran I, Farag M, Incicco JJ, et al. 2020. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science 367:694–99 [PubMed: 32029630]
65. McCarty J, Delaney KT, Danielson SPO, Fredrickson GH, Shea JE. 2019. Complete phase diagram for liquid-liquid phase separation of intrinsically disordered proteins. J. Phys. Chem. Lett 10(8):1644–52 [PubMed: 30873835]
66. Miyazawa S,Jernigan R. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–52
67. Muller-Spath S, Soranno A, Hirschfeld V, Hofmann H, Ruegger S, et al. 2010. Charge interactions can dominate the dimensions of intrinsically disordered proteins. PNAS 107:14609–14 [PubMed: 20639465]
68. Muthukumar M. 1987.Adsorption of a polyelectrolyte chain to a charged surface.J. Chem. Phys 86:7230–35
69. Muthukumar M 1996. Double screening in polyelectrolyte solutions: limiting laws and crossover formulas. J. Chem. Phys 105:5183–99
70. Muthukumar M 2017. 50th anniversary perspective: a perspective on polyelectrolyte solutions. Macromolecules 50(24):9528–60 [PubMed: 29296029]
71. Nguemaha V, Zhou HX. 2018. Liquid-liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation. Sci. Rep 8:6728 [PubMed: 29712961]
72. Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, et al. 2015. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol. Cell 57(5):936–47 [PubMed: 25747659]
73. Ozkan SB. 2021. Can sequence-specific and dynamics-based metrics allow us to decipher the function in IDP sequences? Biophys.J 120(10):1857–59 [PubMed: 33951452]
74. Peng B, Muthukumar M. 2015.Modeling competitive substitutionin a polyelectrolyte complex. J. Chem. Phys 143:243133 [PubMed: 26723618]
75. Portz B, Lu F, Gibbs EB, Mayfield JE, Mehaffey MR, et al. 2017. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nat. Commun 8:15231 [PubMed: 28497792]
76. Rana U, Brangwynne CP, Panagiotopoulos AZ. 2021. Phase separation versus aggregation behavior for model disordered proteins. J. Chem. Phys 155:125101 [PubMed: 34598580]
77. Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, et al. 2017. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. Science 358(6360):238–41 [PubMed: 29026044]

78. Riback JA, Katanski C, Kear-Scott J, Pilipenko E, Rojek A, et al. 2017. Stress-triggered phase separation is an adaptive, evolutionary tuned response. Cell 168(6):1028–40 [PubMed: 28283059]
79. Ruff KM, Harmon TS, Pappu RV. 2015. Camelot: a machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. J. Chem. Phys 143(24):243123 [PubMed: 26723608]
80. Ruff KM, Pappu RV 2021.AlphaFold and implications for intrinsically disordered proteins. J. Mol. Biol 433(20):167208 [PubMed: 34418423]
81. Rustad M, Ghosh K. 2012. Why and how does native topology dictate the folding speed of a protein? J. Chem. Phys 137:205104 [PubMed: 23206039]
82. Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, et al. 2018. Coactivator condensation at super-enhancers links phase separation and gene control. Science 361:eaar3958 [PubMed: 29930091]
83. Samanta HS, Chakraborty D, Thirumalai D. 2018. Charge fluctuation effects on the shape of flexible polyampholytes with applications to intrinsically disordered proteins. J. Chem. Phys 149:163323 [PubMed: 30384718]
84. Samanta HS, Zhuravlev PI, Hinczewski M, Hori N, Chakrabarti S, Thirumalai D. 2017. Protein collapse is encoded in the folded state architecture. Soft Matter 13:3622–38 [PubMed: 28447708]
85. Sawle L, Ghosh K. 2015. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. J. Chem. Phys 143:085101 [PubMed: 26328871]
86. Schmit JD, Bouchard JJ, Martin EW, Mittag T. 2020. Protein network structure enables switching between liquid and gel states. J. Am. Chem. Soc 142:874–83 [PubMed: 31845799]
87. Schuster BS, Dignon GL, Tang WS, Kelley FM, Ranganath AK, et al. 2020. Identifying sequence perturbations to an intrinsically disordered protein that determines its phase-separation behavior. PNAS 117:11421–31 [PubMed: 32393642]
88. Shakhnovich EI. 1994. Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett 72:3907–10 [PubMed: 10056327]
89. Shea JE, Best RB, Mittal J. 2021. Physics-based computational and theoretical approaches to intrinsically disordered proteins. Curr. Opin. Struct. Biol 67:219–25 [PubMed: 33545530]
90. Sherry KP, Das RK, Pappu RV, Barrick D. 2017. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. PNAS 114(44):E9243–52 [PubMed: 29078291]
91. Shin Y, Brangwynne CP. 2017. Liquid phase condensation in cell physiology and disease. Science 357:eaaf4382 [PubMed: 28935776]
92. Showalter S 2014. Intrinsically disordered proteins: methods for structure and dynamics studies. eMagRes 3:181–90
93. Sizemore SM, Cope SM, Roy A, Ghirlanda G, Vaiana SM. 2015. Slow internal dynamics and charge expansion in the disordered protein CGRP: a comparison with amylin. Biophys. J 109(5):1038–48 [PubMed: 26331261]
94. Socci ND, Onuchic JN. 1994. Folding kinetics of proteinlike heteropolymers. J. Chem. Phys 101:1519–28
95. Socci ND, Onuchic JN. 1995. Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique. J. Chem. Phys 103:4732–44
96. Song J, Li J, Chan HS. 2021. Small-angle X-ray scattering signatures of conformational heterogeneity and homogeneity of disordered protein ensembles. J Phys. Chem. B 125:6451–78 [PubMed: 34115515]
97. Srivastava D, Muthukumar M. 1996. Sequence dependence of conformations of polyampholytes. Macromolecules 29:2324–26
98. Statt A, Casademunt H, Brangwynne CP, Panagiotopoulos AZ. 2020. Model for disordered proteins with strongly sequence-dependent liquid phase behavior. J. Chem. Phys 152:075101 [PubMed: 32087632]
99. Tesei G, Schulze TK, Crehuet R, Larsen-Lindorff K. 2021. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. PNAS 118(44):e2111696118 [PubMed: 34716273]

100. Wang J, Choi JM, Holehouse A, Lee HO, Zhang X, et al. 2018. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell 174:688–99 [PubMed: 29961577]
101. Wessen J, Pal T, Das S, Lin YH, Chan HS. 2021. A simple explicit-solvent model of polyampholyte phase behaviors and its ramifications for dielectric effects in biomolecular condensates. J. Phys. Chem. B 125(17):4337–58 [PubMed: 33890467]
102. Wiggers F, Wohl S, Dubovetskyi A, Rosenblum G, Zheng W, Hofmann H. 2021. Diffusion of a disordered protein on its folded ligand. PNAS 118(37):e2106690118 [PubMed: 34504002]
103. Wirth A, Gruebele M. 2013. Quinary protein structure and the consequence of crowding in living cells: leaving the test-tube behind. Bioessays 35(11):984–93 [PubMed: 23943406]
104. Wohl S, Jakubowski M, Zheng W 2021. Salt-dependent conformational changes of intrinsically disordered proteins. J. Phys. Chem. Lett 12(28):6684–91 [PubMed: 34259536]
105. Zarin T, Strome B, Nguyen Ba AN, Alberti S, Forman-Kay JD, Moses AM. 2019. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. eLife 8:46883
106. Zarin T, Strome B, Peng G, Pritisanac I, Forman-Kay JD, Moses AM. 2021. Identifying molecular features that are associated with biological function of intrinsically disordered regions. eLife 10:e60220 [PubMed: 33616531]
107. Zarin T, Tsai CN, Nguyen Ba AN, Moses AM. 2017. Selection maintains signaling function of a highly diverged intrinsically disordered region. PNAS 114(8):E1450–59 [PubMed: 28167781]
108. Zeng X, Holehouse AS, Chilkoti A, Mittag T, Pappu RV. 2020. Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. Biophys. J 119:402–18 [PubMed: 32619404]
109. Zerze GH, Best RB, Mittal J. 2015. Sequence- and temperature-dependent properties of unfolded and disordered proteins from atomistic simulations. J. Phys. Chem. B 119:14622–30 [PubMed: 26498157]
110. Zheng W, Dignon G, Brown M, Kim YC, Mittal J. 2020. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. J. Phys. Chem. Lett 11(9):3408–15 [PubMed: 32227994]
111. Zheng W, Zerze GH, Borgia A, Mittal J, Schuler B, Best RB. 2018. Inferring properties of disordered chains from FRET transfer efficiencies. J. Chem. Phys 148:123329 [PubMed: 29604882]

SUMMARY POINTS

1. The PM of IDPs describes rules of regulation that control IDP conformations.
2. Models for describing single-chain conformations inform the multichain physics of phase separation.
3. The physical mathematical relationships of IDPs provide insights into biological functions.
4. Physical mathematical equations—balancing accuracy and efficiency—overcome combinatorial challenges that may impede design strategies for new polymers or proteins.

FUTURE ISSUES

1. Further improvement in electrostatic modeling is needed to include the roles of degree of ionization (of charged moieties) and of charge–dipole and dipole–dipole interactions. Furthermore, to improve the predictability of the conformational properties of single chains, transferable nonelectrostatics interaction parameters that can be combined with electrostatic models are needed.
2. Models for sequence-dependent phase separations are needed to account for noncharge interactions, strong fluctuations at low density, and lower critical solution temperatures. Existing theoretical platforms can be used to learn about the competition between LLPS and other physical processes such as gelation, aggregation, and complexation for which analytical models exist.
3. Efforts to build functional classification algorithms for IDPs are limited by the small numbers of data on IDP functions now available for suitable comparisons. More experimental measurements with designed variants are needed for building algorithms beyond the SCDM and for developing new patterning metrics (such as SHDMs that describe noncharge amino acids), as well as including molecular features beyond patterning metrics.

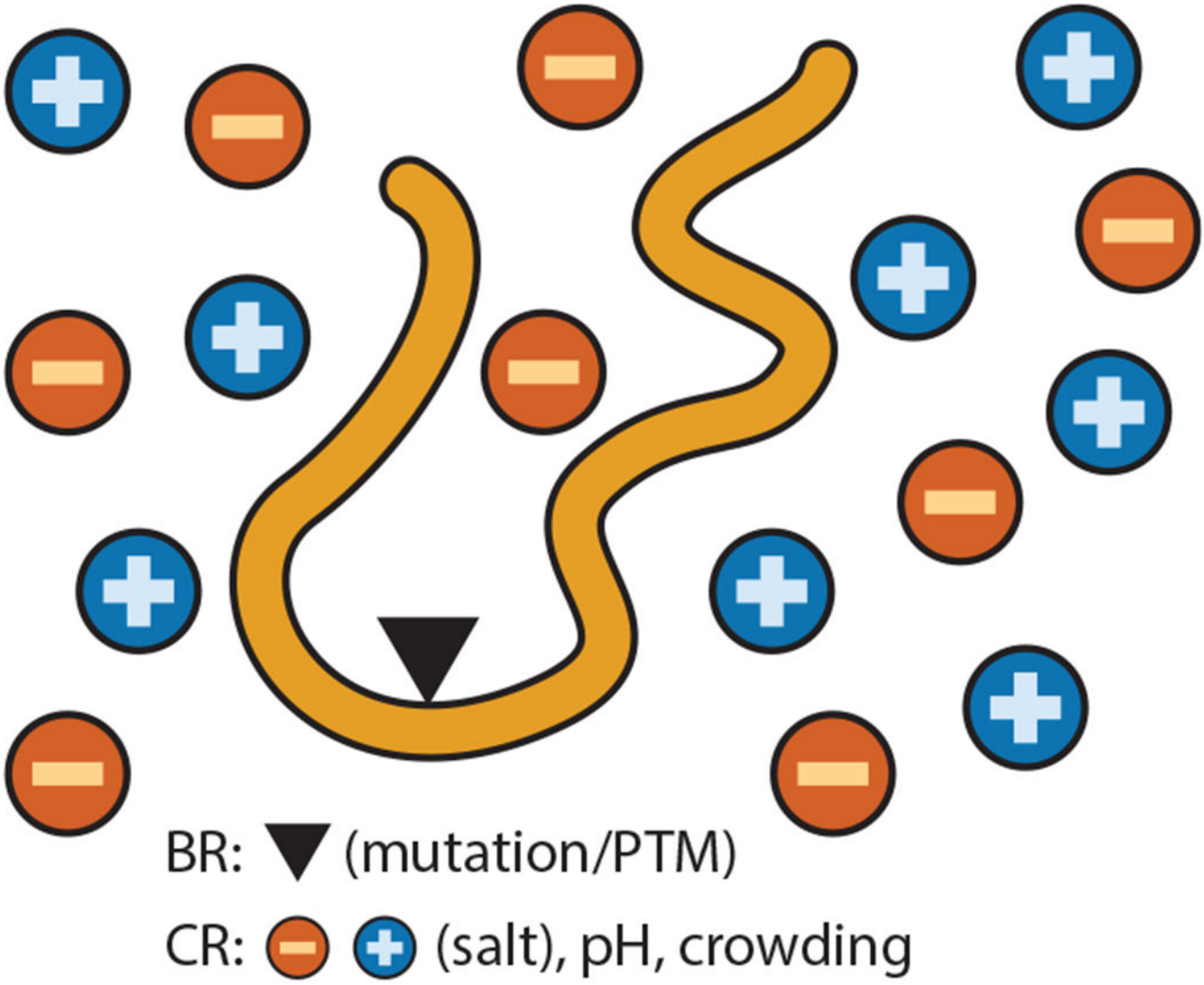

Figure 1.

Interplay between BRs (e.g., mutations or PTMs that directly change the sequence) and CRs (e.g., changes in solution conditions) can be complex and lead to a combinatorial explosion, requiring a PM-based approach. Abbreviations: BR, biological regulator; CR, chemical regulator; PM, physical mathematics; PTM, posttranslational modification.

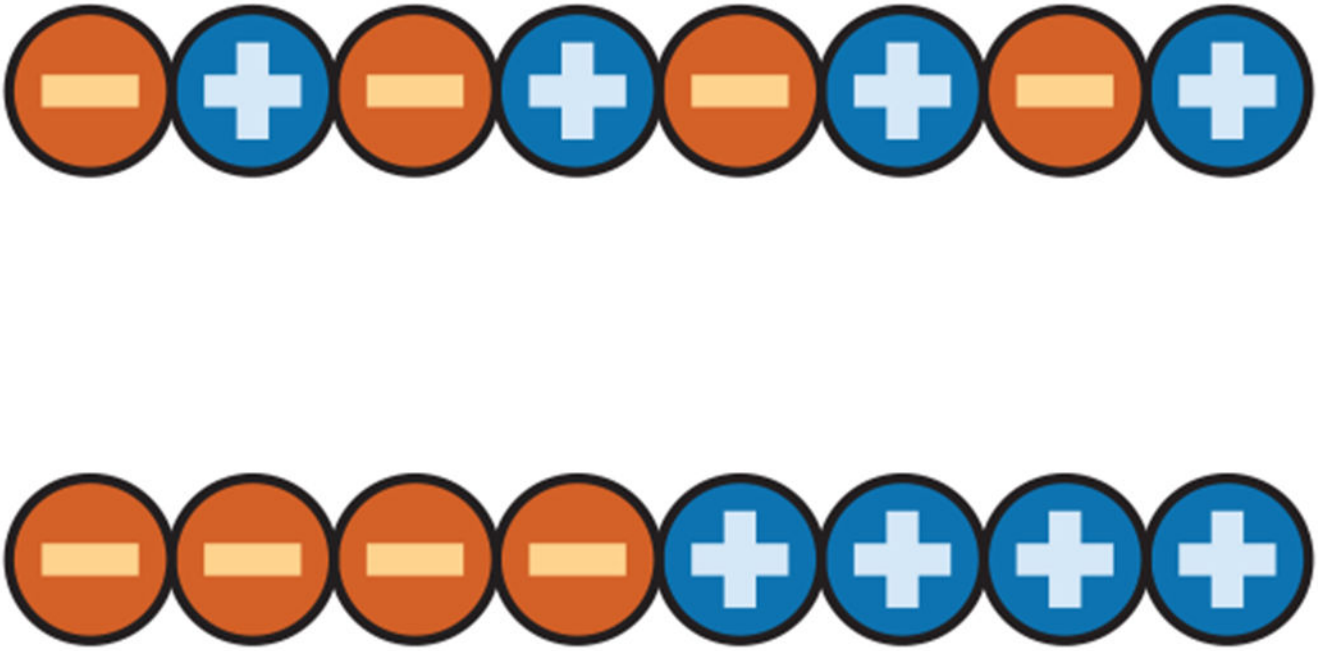

Figure 2.

The SCD metric (the same as Q in Equation 2) is a measure of the patterning of positive and negative charges. Two sequences (*top*, *bottom*) can have the same number of positive and negative charges (composition) but differences in patterning, reflected in their differing SCDs. Blocky sequences (*bottom*) tend to have lower SCDs compared with sequences where charges are more dispersed (*top*). Abbreviation: SCD, sequence charge decoration. Figure adapted with permission from Reference 33.

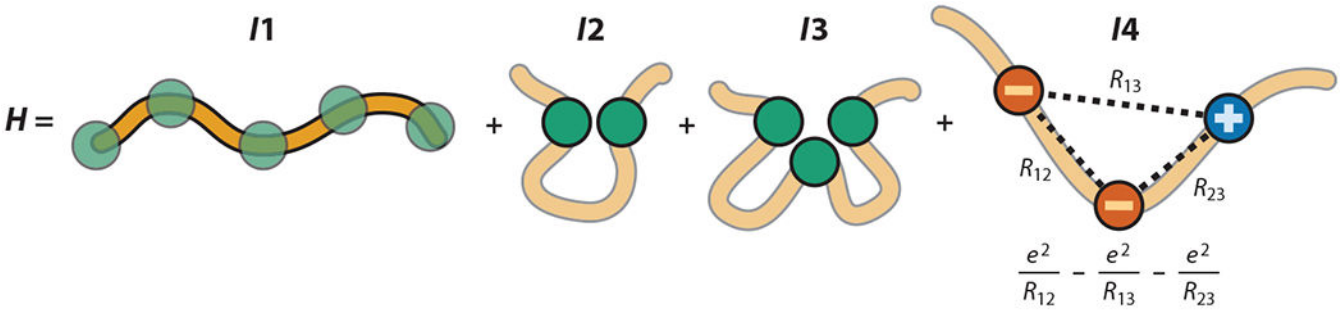

Figure 3.

The energy function (H) of an analytical framework for determining the R_{ee} of a sequence that includes charged monomers accounts for chain connectivity ($I1$), two-body ($I2$) and three-body ($I3$) short-range interactions, and long-range electrostatic interactions ($I4$). The exact functional form of H can be found in References 31 and 46.

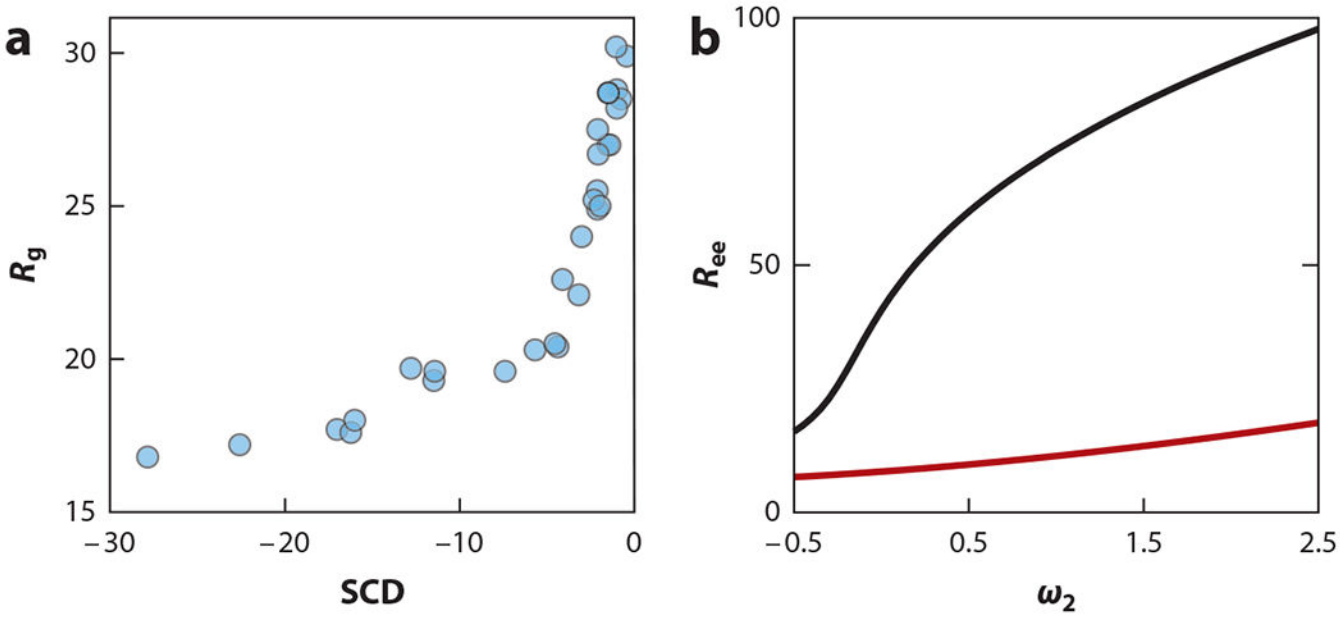

Figure 4.

(*a*) SCD captures variations in size (R_g in angstroms) obtained from all-atom Monte Carlo simulations for 30 sequences, each having 25 Es and 25 Ks arranged in different orders. (*b*) Equations 1 and 2 can be used to predict the size ($R_{ee} = \langle r_{ee}^2 \rangle^{1/2}$ in angstroms) dependence on ω_2 (proxy for temperature or solution conditions) for a typical choice of $\omega_3 = 0.1$, $b = 3.8$ Å, and $l = 8$ Å, $l_b = 7.2$ Å. The difference between the black curve (representing a sequence with alternating Es and Ks) and the red curve (a sequence with a block of 25 Es followed by 25 Ks) highlights the effect of charge patterning revealed by the SCD embedded in the Equation 2. Abbreviations: E, glutamic acid; K, lysine; SCD, sequence charge decoration.

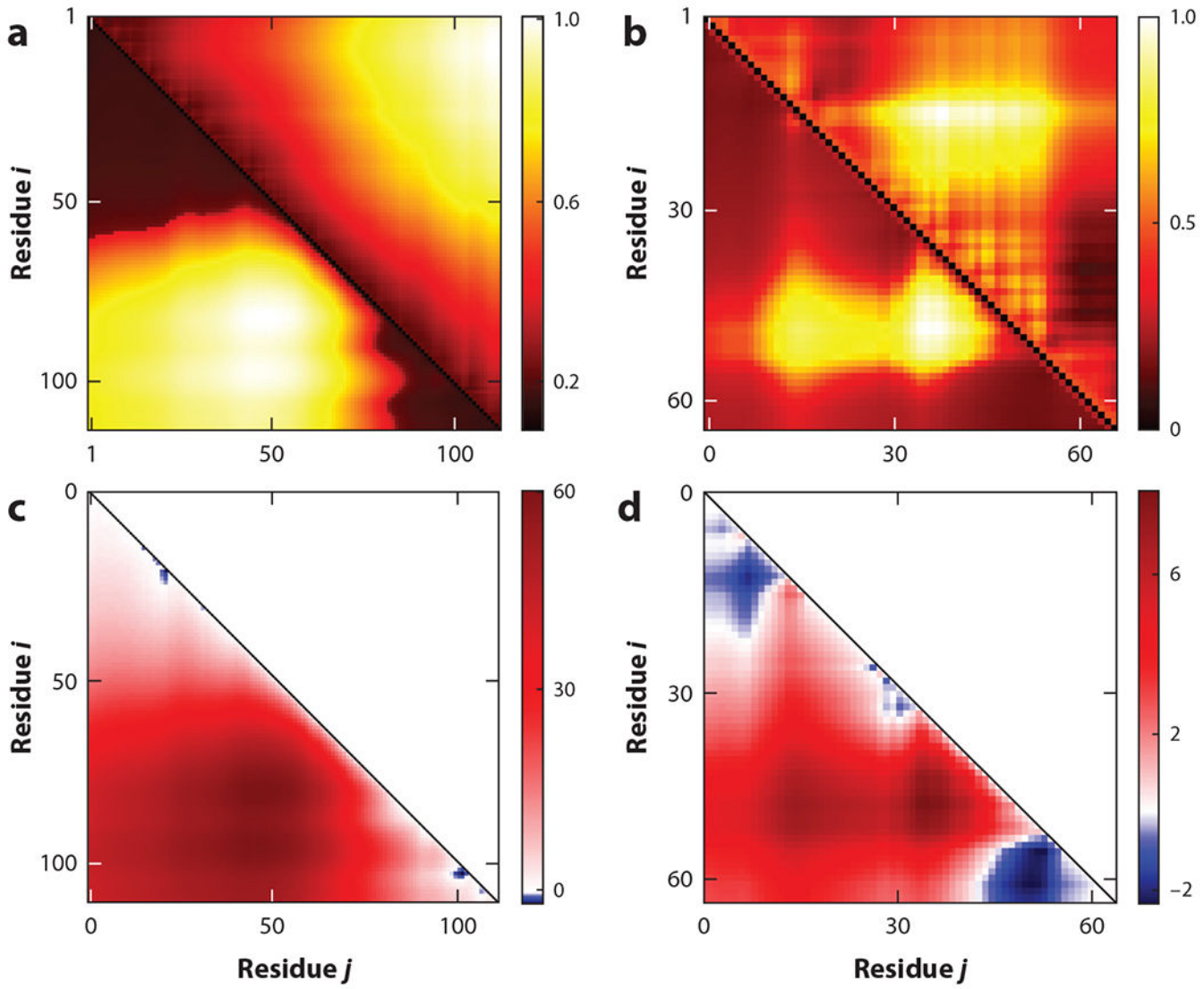

Figure 5.

Equation 6 provides distance maps quantified by $x_{i,j}$ values (*lower triangles*) for two intrinsically disordered protein and region sequences: (*a*) prothymosin-α and (*b*) DP00877. Color coding denotes the different values of $x_{i,j}$ (normalized between 0 and 1) for residue pairs *i*, *j* (*x* and *y* axes). The noncharge parameters ω_3 and ω_2 were obtained to best match the all-atom simulation data of $x_{i,j}$ shown in the upper triangle (for details, see Reference 46). (*c*, *d*) The most expanded regions (*bright yellow*) in distance maps in panels *a* and *b* correspond to the bright red (repulsive) regions in the respective sequence charge decoration matrix maps shown for (*c*) prothymosin-α and (*d*) DP00877. Figure adapted from Reference 46 with permission from AIP Publishing.

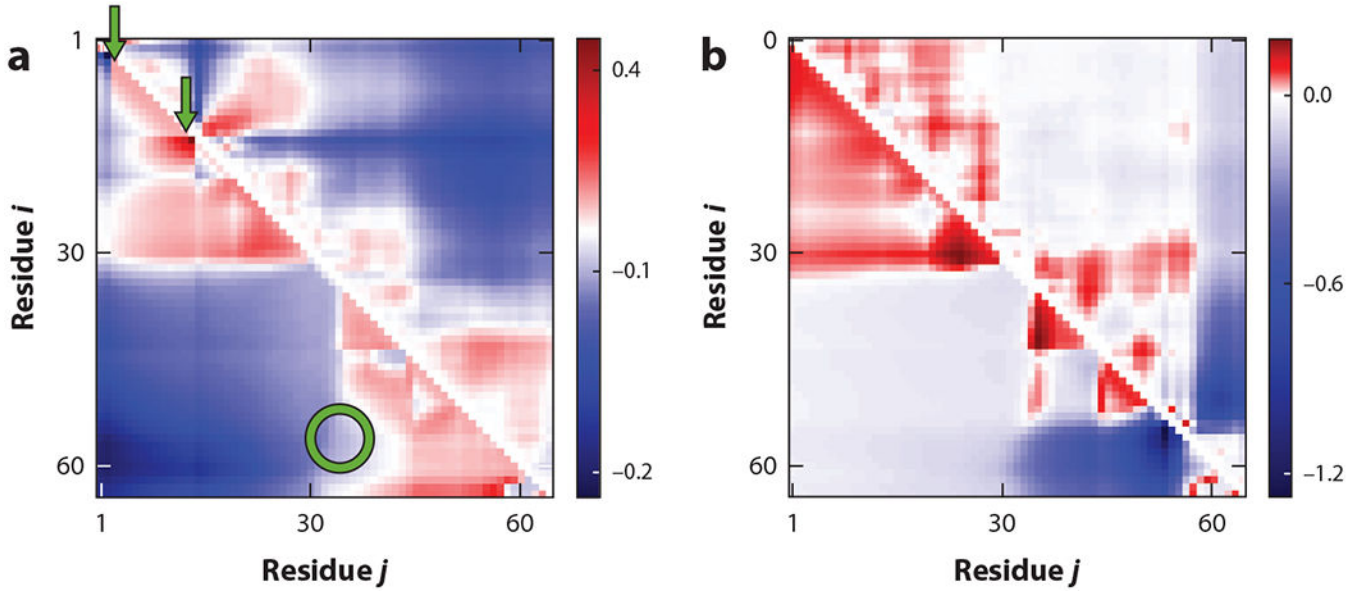

Figure 6.

Intrinsically disordered proteins and regions have hot spots for phosphorylation. Distance profiles, given by sequence-specific $x_{i,j}$ (relating to $\langle r_{i,j}^2 \rangle$) for amino acid pairs i, j, for two different phosphorylated forms, (*a*) S2T15 and (*b*) S54S56, of the wild-type protein P0A8H9 show that specific phosphorylation sites induce drastic conformational changes at all scales. Positive (*red*) and negative (*blue*) differences are evident in these heat maps. Theoretical results (*lower triangles*) exhibit trends similar to the all-atom simulation results (*upper triangles*). The arrows point to sequence sites of phosphorylation that can generate changes in distances among sites that are remote in the sequence (*circle*). Figure adapted from Reference 46 with permission from AIP Publishing.

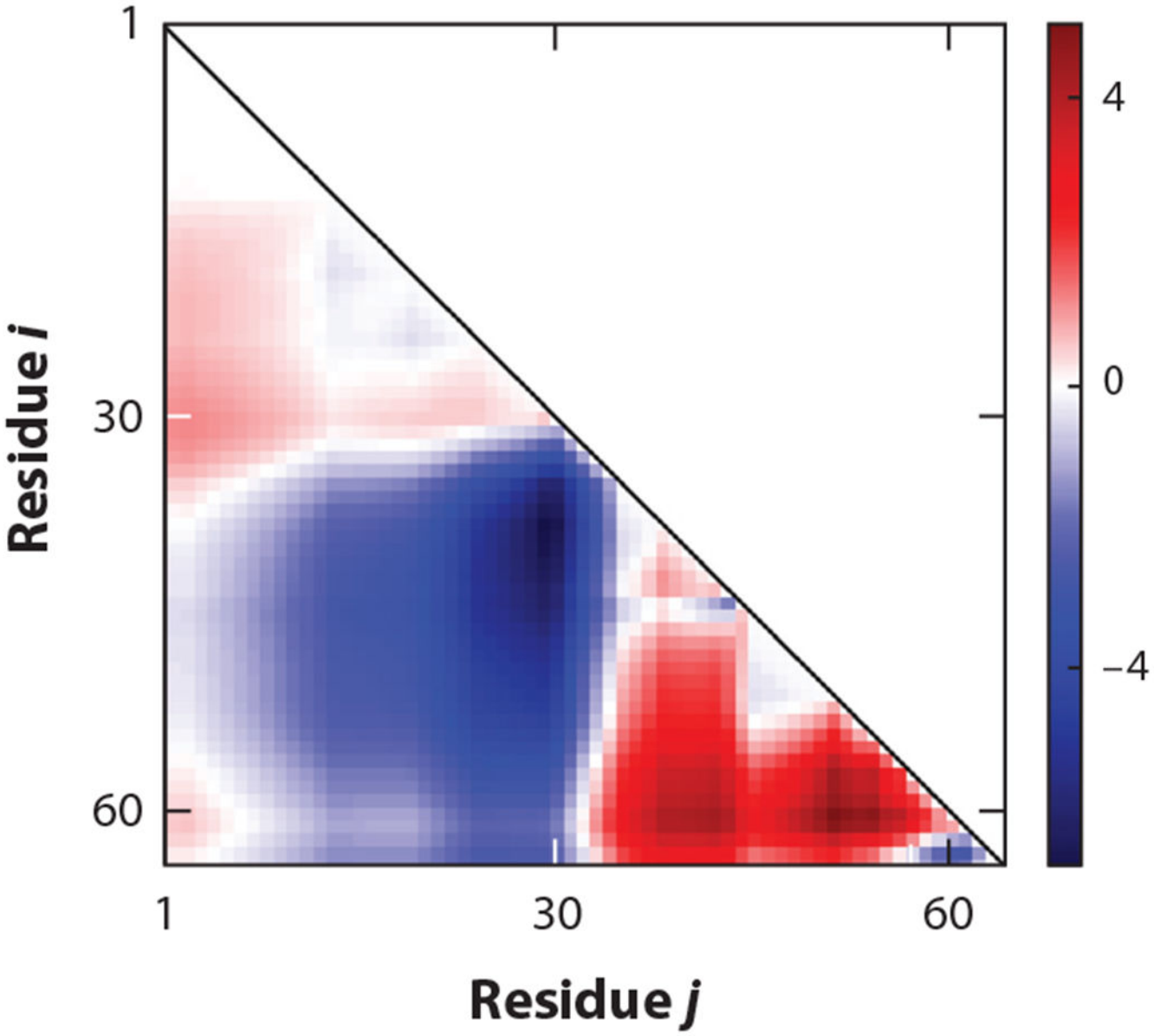

Figure 7.

Residue pair–specific distances can have varying salt responses. Elements of $SCDM_{low\ salt}$ (calculated using Equation 11) corresponding to different residue pairs (in the x and y axes) are shown for protein P0A8H9. The red regions denote positive values of the matrix elements, implying that the addition of salt near the zero-salt limit will shrink the distances among the respective residue pairs, while the blue regions denote pairs of amino acids for which the distances will expand upon addition of salt. Abbreviation: SCDM, sequence charge decoration matrix.

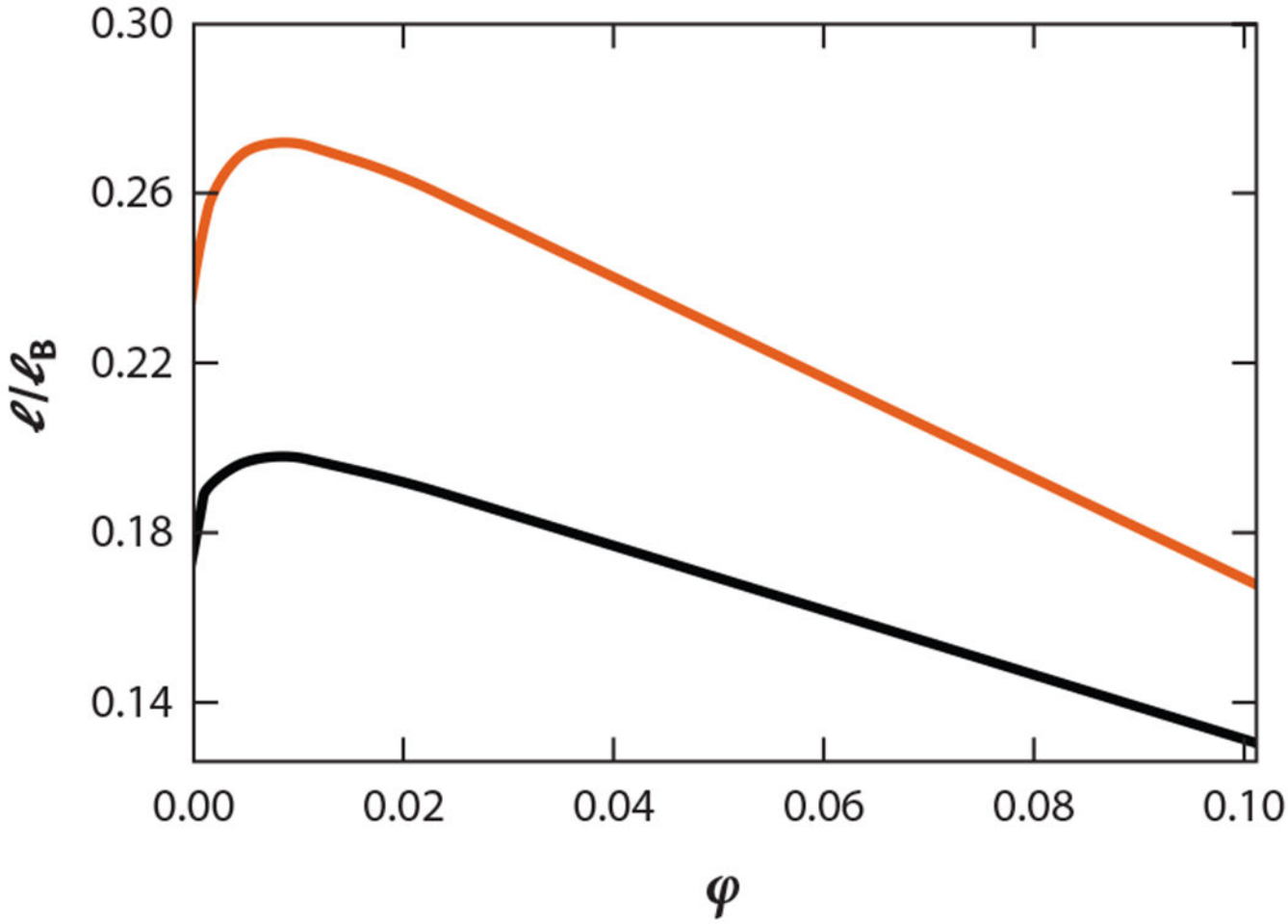

Figure 8.

Phosphorylation hot spots can influence the propensity to phase separate. Phase diagrams (density ϕ in the x axis and dimensionless temperature l/l_b in the y axis) of two different phosphorylated versions (S54S56 in orange and S2T15 in black) of the wild-type protein P0A8H9 have been computed using the theory of Chan and colleagues (56). This theory predicts that two sequences with the same charge composition but with different patterning can have different critical points and propensities to phase separate. The effect of phosphorylation has been modeled by introducing a minus charge at the site of phosphorylation.