Biochemistry

pubs.acs.org/biochemistry Article

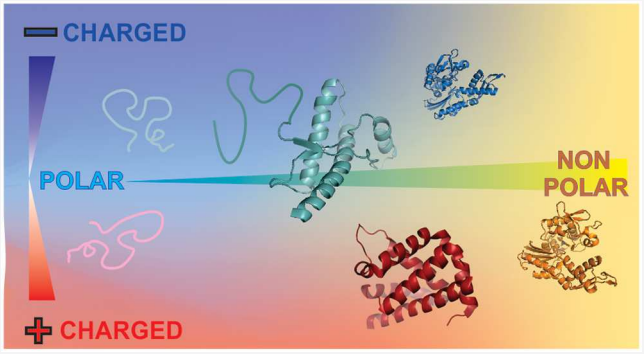
Net Charge and Nonpolar Content Guide the Identification of Folded and Prion Proteins

Susanna K. Yaeger-Weiss,[‖] Theodore S. Jennaro,[‖] Miranda Mecha, Jenna H. Becker, Hanming Yang, Gordon L. W. Winkler, and Silvia Cavagnero*

Cite This: *Biochemistry* 2020, 59, 1881−1895 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The degree of hydrophobicity and net charge per residue are physical properties that enable the discrimination of folded from intrinsically disordered proteins (IDPs) solely on the basis of amino acid sequence. Here, we improve upon the existing classification of proteins and IDPs based on the parameters mentioned above by adopting the scale of nonpolar content of Rose et al. and by taking amino acid side-chain acidity and basicity into account. The resulting algorithm, denoted here as net charge nonpolar or NECNOP, enables the facile prediction of the folded and disordered status of proteins under physiologically relevant conditions with >95% accuracy, based on amino-acid sequence alone. The NECNOP approach displays a much-enhanced performance for proteins with >140 residues, suggesting that small proteins are more likely to have irregular charge and hydrophobicity features. NECNOP analysis of the entire *Escherichia coli* proteome identifies specific net charge and nonpolar regions peculiar to soluble, integral membrane, and non-integral membrane proteins. Surprisingly, protein net charge and hydrophobicity are found to converge to specific values as chain length increases, across the *E. coli* proteome. In addition, NECNOP plots enable the straightforward identification of protein sequences corresponding to prion proteins and promise to serve as a powerful predictive tool for the design of large proteins. In summary, NECNOP plots are a straightforward approach that improves our understanding of the relation between the amino acid sequence and three-dimensional structure of proteins as a function of molecular mass.



The physical properties of intrinsically disordered and folded proteins are important for fully understanding the sequence−structure paradigm.[1−3] The polar and nonpolar characteristics of individual residues, for instance, are well-known contributors to protein foldability. Specifically, the degree of nonpolar content is proportional to the extent of the hydrophobic effect, which plays a significant role in folding through the energetically favorable burial of nonpolar surface away from the aqueous solvent.[4−8] From the early days of protein chemistry, it was noted that the first X-ray crystal structure of a folded protein, myoglobin, has most of its nonpolar side chains buried in the core.[9] This concept was later shown to apply to the majority of nonpolar side chains of all proteins.[6,7,10−12] The importance of nonpolar side chains in protein folding is pictorially illustrated in Figure 1a.

Net charge is another defining characteristic of proteins. As schematically illustrated in Figure 1b, electrostatic interactions between two charged groups can be attractive or repulsive.[13] Attractive electrostatic forces between two moieties of opposite charge typically result in salt bridges, also known as ion pairs.[14] The role of salt bridges in protein thermodynamic stability is overall moderate.[15,16] While the Coulombic interaction between charges is favorable in ion pairs, this effect is countered by the thermodynamically unfavorable charge desolvation that takes place upon salt-bridge formation.[15] The thermodynamic balance between these opposing effects ultimately governs the net thermodynamic effect of salt bridges. The latter is highly dependent upon the protein environment, which primarily leads to variations in the degree of surface solvation.[16] In case salt bridges have a net stabilizing effect, it was estimated that each of them contributes approximately 1−3 kcal/mol to protein stability.[17] Again, overall, salt bridges are believed to have a small, nondominant effect on protein thermodynamic stability.[17] Repulsive electrostatic forces increase the protein net charge and contribute to protein thermodynamic stability. These interactions are generally thermodynamically unfavorable, often resulting in unfolding.[17]

Received: December 24, 2019
Revised: April 28, 2020
Published: April 30, 2020

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

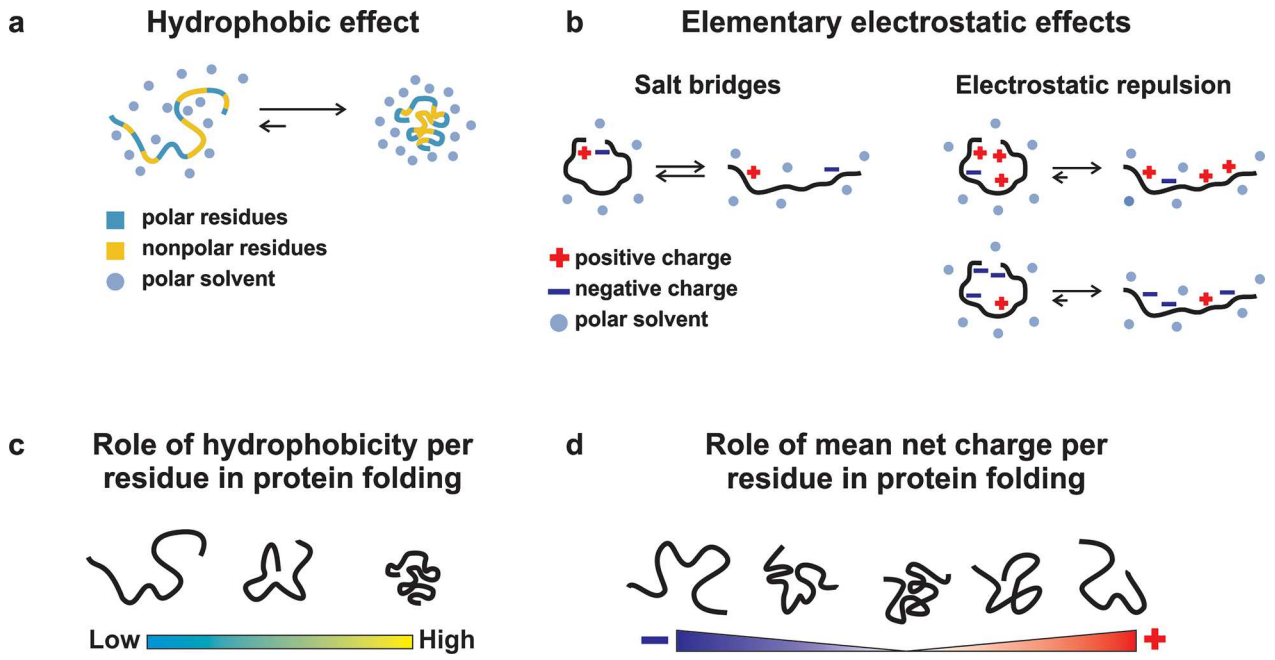


Figure 1. Schematic representation of the role of hydrophobic and electrostatic effects in protein folding. (a) Consistent with the hydrophobic effect, most nonpolar residues are typically buried from aqueous solvent upon folding. (b) Overall, salt bridges do not significantly stabilize mesophilic proteins while electrostatic repulsion can be strongly destabilizing. (c) Higher hydrophobicity and (d) moderate charge repulsion per residue are expected to favor protein folding.

In summary, among Coulombic effects, it is clear that electrostatic repulsion affects protein thermodynamic stability much more than ion pairs.

According to the criteria mentioned above, it is reasonable to make the highly qualitative conceptual prediction that if a protein is not sufficiently nonpolar or if it has a high net charge, it is expected to be thermodynamically unstable, hence unfolded. This concept is pictorially illustrated in panels c and d of Figure 1. Consistent with this prediction, Uversky, Gillespie, and Fink reported that two simple protein parameters, hydrophobicity and net charge, are sufficient to describe the folded versus disordered status of polypeptide and protein chains based on primary structure, i.e., amino acid sequence, alone.[18] Their method involves assessment of nonpolar content via the hydropathy scale by Kyte and Doolittle[15] and estimation of net charge via the absolute value of the difference between the number of positively (Arg and Lys) and negatively (Asp and Glu) charged residues in a protein. The resulting net charge−hydropathy plots show that intrinsically disordered proteins (IDPs) tend to have a lower nonpolar content per residue and a wider range of net charge per residue values than folded proteins.[18] More recently, Uversky and co-workers updated their original model by employing a computationally derived hydrophobicity scale denoted as IDP-hydropathy,[19] which leads to more accurate charge−hydropathy plots for IDPs.[19] In addition, Pappu and co-workers established a detailed classification of IDP conformational features upon taking the nonpolar content, net charge, and position of residues and residue clusters into account.[20] These advances targeted primarily amino acid sequences corresponding to IDPs and were extremely successful on that front.

On the other hand, a more reliable prediction of amino acid sequences corresponding to folded proteins (as opposed to IDPs) is highly desirable. First, accurate identification of amino acid sequences corresponding to folded proteins is expected to greatly benefit proteomics and structural genomics. Namely, correct identification of gene sequences compatible with folded proteins will acceleration the selection of open reading frames (ORFs) worthy of overexpression followed by structural analysis. Second, identification of amino acid compositions compatible with folded proteins is expected to be a valuable tool in protein design.

Here, we move a step forward toward the objectives mentioned above by developing an optimized algorithm to determine the mean net charge per residue (MNC) and mean nonpolar content per residue (MNPC) of proteins of any given amino acid sequence. Unlike previous studies, we took into account amino acid protonation/deprotonation and evaluated nonpolar amino acid content based on the scale by Rose et al.,[21] which quantifies the nonpolar nature of amino acids in proteins based on database analysis and first principles. Our method enables classification of folded proteins and IDPs with unprecedented accuracy for folded proteins of >140 amino residues. Extension to the analysis of the *Escherichia coli* proteome showed that the mean net charge per residue and the mean hydrophobicity per residue of proteins from this organism converge to specific values, as the chain length increases. Intriguingly, we also found that prion proteins, which are often partly folded and partly disordered in their non-infectious cellular form, lie along the discriminant line of NECNOP plots.

METHODS

Determination of the Mean Net Charge and Mean Nonpolar Content per Residue. At room temperature and at a given pH, the net charge per residue (Z) of a molecule bearing multiple independent ionizable functional groups (e.g., a protein), each with its own pK_a, is

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

$$Z = \sum_i N_i \frac{10^{pK_{ai}}}{10^{pH} + 10^{pK_{ai}}} - \sum_j N_j \frac{10^{pH}}{10^{pH} + 10^{pK_{aj}}} \quad (1)$$

where N is the number of residues and i and j denote positively and negatively charged groups, respectively. Equation 1 is based on the Henderson−Hasselbalch relation. In the case of a protein, this equation applies to the side-chain pK_a of the positively (Lys, Arg, and His) and negatively (Glu, Asp, Tyr, and Cys) charged amino acid side chains.[22] The mean net charge per residue (MNC) parameter was computed assuming neutral pH, upon dividing Z (see eq 1) by the total number of residues.

Note that the pK_a values of Tyr and Cys are 10.07 and 8.18, respectively. Therefore, at neutral pH a single Tyr and Cys contribute −0.00085 and −0.062 to the total net charge, respectively. While these contributions may be regarded as insignificant for proteins bearing only a very small fraction of Tyr and Cys, the role of Tyr and Cys deprotonation is expected to be non-negligible for proteins carrying a large percent of these residues.

For the sake of simplicity, MNC calculations assumed that all Cys residues are reduced. Analysis of a subset of proteins with known Cys disulfides (or lack thereof) revealed that this approximation is appropriate, given that the MNC values of the tested proteins do not significantly change in the absence and presence of disulfide bridges (data not shown). Mean nonpolar content per residue (MNPC) values were determined according to Rose et al.[21] with a five-residue sliding window. Values for each residue were summed, and the resulting number was divided by the total number of amino acids minus 4, to account for the sliding window. Both MNC and MNPC values were computed via a Python script (see below). We separately tested the effect of variable-size window sizes and found that a five-residue window yields optimal results (see also the Supporting Information).

Generation of NECNOP Plots. Five different databases of known folded proteins and IDPs were employed in this work. Database definitions are provided in the Supporting Information and Table S1. In addition, the specific proteins belonging to databases 1−3 are listed in Tables S2−S4. Databases 4 and 5 comprise a large number of proteins (1147 and 4305, respectively) whose identity can be retrieved directly from the Protein Data Bank (PDB) and from the UniProt repositories (see the Supporting Information).

MNC and MNPC values were computed, and NECNOP plots were generated with a custom-generated Python 2.7 script taking advantage of the Biopython tool set. The linear discriminant analysis function of MATLAB (The MathWorks Inc., Natick, MA, ver. 2016a), in combination with procedures by Guo et al.,[23] led to the establishment of the optimal discriminant line separating folded proteins from IDPs

$$|\text{MNC}| = 12.0698 \times \text{MNPC} - 8.4815 \quad (2)$$

where |MNC| denotes the absolute value of the MNC parameter. Two discriminant lines were plotted, corresponding to |MNC| and −|MNC|, to facilitate the classification of folded and intrinsically disordered proteins irrespective of MNC sign (see the Supporting Information).

Error bars in Tables 1 and 2 were estimated as follows. The percent of correctly predicted proteins (PCP), verified against databases of known folded proteins (databases 1 and 3) and IDPs (database 2), is defined as

Table 1. Summary of NECNOP Performance, Focusing on the Ability of This Algorithm to Correctly Predict IDPs[a] and Folded[b] Proteins Starting from Databases of Proteins with Known Structure or a Lack Thereof

protein type	no. of correctly predicted proteins	no. of incorrectly predicted proteins	percent of correctly predicted proteins[c] (PCP) (%)
folded (any size)	217	14	93.9 ± 3.1
folded with <140 residues	91	14	86.7 ± 6.5
folded with >140 residues	126	0	100 ± 0
IDP (any size)	41	5	89.1 ± 9.0
IDP with <140 residues	21	2	91.3 ± 5.9
IDP with >140 residues	20	3	86.9 ± 7.0

[a]Database 2 (see Table S3) was used to generate the data in this table. [b]Database 3 (see Table S4) was used to generate the data in this table. [c]The estimated error on PCP values is reported as a 95% confidence interval (see Methods).

Table 2. Summary of NECNOP Performance for the Prediction of Folded Protein Identity Employing a Large Database of Known Folded Proteins[a]

protein type	no. of correctly predicted proteins	no. of incorrectly predicted proteins	percent of correctly predicted proteins[b] (PCP) (%)
folded (any size)	1092	55	95.2 ± 1.2
folded with <140 residues	264	45	85.4 ± 3.9
folded with >140 residues	828	10	98.8 ± 0.7

[a]Database 4 (see Table S5 for outliers in Figure 5) was used to generate the data in this table. [b]The estimated error on PCP values is reported as the 95% confidence interval (see Methods).

$$\text{PCP} = 100 \times \frac{\text{no. of correctly predicted proteins}}{\text{total no. of proteins}} \quad (3)$$

where the parameters denoted as no. of correctly predicted proteins and total no. of proteins refer to the chosen category, i.e., either folded protein or IDP. We then assessed the accuracy of PCP by first estimating the standard error (SE) based on a binomial distribution of the fraction of correctly predicted proteins (p = PCP/100) according to[24]

$$\text{SE} = 100 \times \sqrt{\frac{p(1-p)}{n}} \quad (4)$$

where n is the total number of proteins known to belong to a given category (i.e., folded or IDP), followed by evaluation of the 95% confidence interval, which was estimated from the SE and the two-tailed Student's t distribution as described previously[24] and found to be equal to ±1.96 × SE. The accuracy of PCP was the estimated ±95% confidence interval.

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

All of the predictions listed this study imply ambient temperature and atmospheric pressure, given that they rely on databases of electrostatic and nonpolar properties of proteins assessed under these conditions.

Determination of Solvent-Accessible Protein Surface Area. Surface Racer (version 5.0)[25] was used to determine the solvent-accessible surface area (SASA) of 52 single-domain proteins of variable size. Proteins were selected using a random-number generator from databases 1−3 (see Tables S2−S4). To ensure the proper representation of large chain lengths, several proteins with >1000 residues were included in this set. A spherical solvent probe with a 1.4 Å radius and van der Waals atomic radii according to Chothia[26] were employed.

RESULTS AND DISCUSSION

NECNOP: A Refined Tool for the Identification of Folded and Disordered Protein Sequences Based Solely on Amino Acid Composition. Identification of protein structure corresponding to given amino acid sequences is the hallmark of protein structure prediction.[27−32] A much less specific, yet extremely useful, goal is to predict whether the amino acid sequence of a known protein of unknown structure is expected to be folded or intrinsically disordered, under ambient conditions.

Here, we target the latter topic and hypothesize that a sufficiently high hydrophobicity as well as a sufficiently low electrostatic repulsion promotes the folded status (relative to IDP) of a protein with a known amino acid sequence, under ambient conditions. This concept is pictorially illustrated in Figure 1. The hypothesis described above is supported by a class of plots by Uversky and co-workers[18,19] featuring net charge per residue versus mean net hydrophobicity per residue. We started by improving upon the algorithms underlying the latter plots as outlined below and took advantage of the resulting knowledge to unveil novel protein properties.

First, we employed the Henderson−Hasselbalch equation, combined with known pK_a information about amino acid side chains, to generate realistic mean net charge values of proteins of known amino acid sequence at neutral pH and ambient temperature and pressure. Next, we took advantage of the scale by Rose and co-workers to quantitatively assess the degree of nonpolar content from protein sequence.[21] The scale by Rose was used, instead of the Kyte−Dolittle hydropathy[33] employed by Uversky and co-workers, because the Rose scale is derived from a uniform set of rigorously defined principles, i.e., solvent-exposed surface area across the folded proteome, which more accurately define the amino acid degree of nonpolar nature.[21] We generated a Python script to perform the calculations and denoted the resulting two-dimensional views of net charge and nonpolar content per residue as NECNOP (net charge − nonpolar) plots.

The NECNOP tool performs very well and leads to the correct prediction of 93.9 ± 3.1% of the folded proteins and 86.7 ± 6.5% of the IDPs (Table 1 and Figure 2), out of databases comprising a total of 277 proteins. As shown in Figure 2, most folded proteins fall to the right of the discriminant line and are characterized by an MNPC larger than those of most IDPs, while the majority of IDPs falls to the left of the discriminant line. These results are consistent with the expectations, discussed above and illustrated in Figure 1, that (i) more nonpolar proteins are more likely to be folded and (ii) IDPs can accommodate a somewhat wider range of net charge than folded proteins.

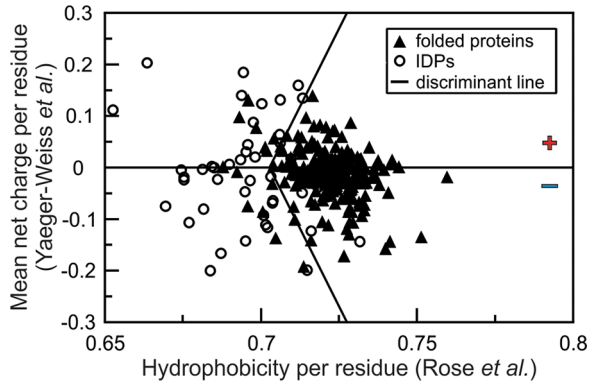


Figure 2. NECNOP plot illustrating the mean net charge per residue (MNC) and nonpolar content per residue (MNPC) of single-domain globular folded proteins and unfolded or intrinsically disordered proteins (IDPs). The black discriminant line (dl) separates regions pertaining to folded (right of the dl) and unfolded proteins/IDPs (left of the dl).

To provide a direct comparison with the previous literature on charge versus hydropathy plots, we computed the balanced-accuracy parameter of NECNOP plots, defined according to Huang et al.[19] This parameter accounts for the overall success in the prediction of both folded proteins and IDPs. The balanced-accuracy value of NECNOP plots is 92%. This value compares favorably with the 79% balanced accuracy achieved via net charge−hydropathy plots,[18] and with the more recent value of 90% obtained via the IDP-hydropathy scale by Huang et al.[19]

While the overall 2% improvement over that of Huang et al. is somewhat moderate, the most significant advantage of the NECNOP plots lies in the analysis of midsize to large (i.e., >140 residues) folded proteins, as highlighted in the next section.

Additional Comparisons with Charge−Hydropathy Plots by Uversky and Co-workers. To further compare our method with the original predictions by Uversky et al.,[18] we generated a Python script to determine MNC and MNPC values according to the criteria established by these authors.[18]

MNCs were determined by assigning each residue with an overall positively charged side chain with a charge of +1 (Asp and Glu). Similarly, we assigned each residue with an overall negatively charged side chain with a charge of −1 (Arg and Lys). We then divided the total net charge by the number of residues. MNPCs were evaluated with a five-residue window via the Kyte−Doolittle scale (normalized over the range of 0−1). We then divided the resulting score by the total number of residues minus 4, to account for the window size. The script carrying out the calculations described above was successfully validated by verifying that a few proteins listed in the work by Uversky et al.[18] yielded the published MNC and MNPC values.

We then compared the performance of Uversky's[18] and NECNOP methods for a protein set similar to that used by Uversky et al.[18] We found that Uversky's method predicts disordered proteins slightly better than the NECNOP approach, with 43 correctly predicted IDPs of 46 via Uversky's method (93.5% success rate) versus 41 correctly predicted IDPs via the NECNOP approach (89.1% success rate).

On the other hand, the NECNOP procedure is significantly better at predicting folded proteins. Namely, the NECNOP strategy correctly predicted 217 of 231 folded proteins,

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

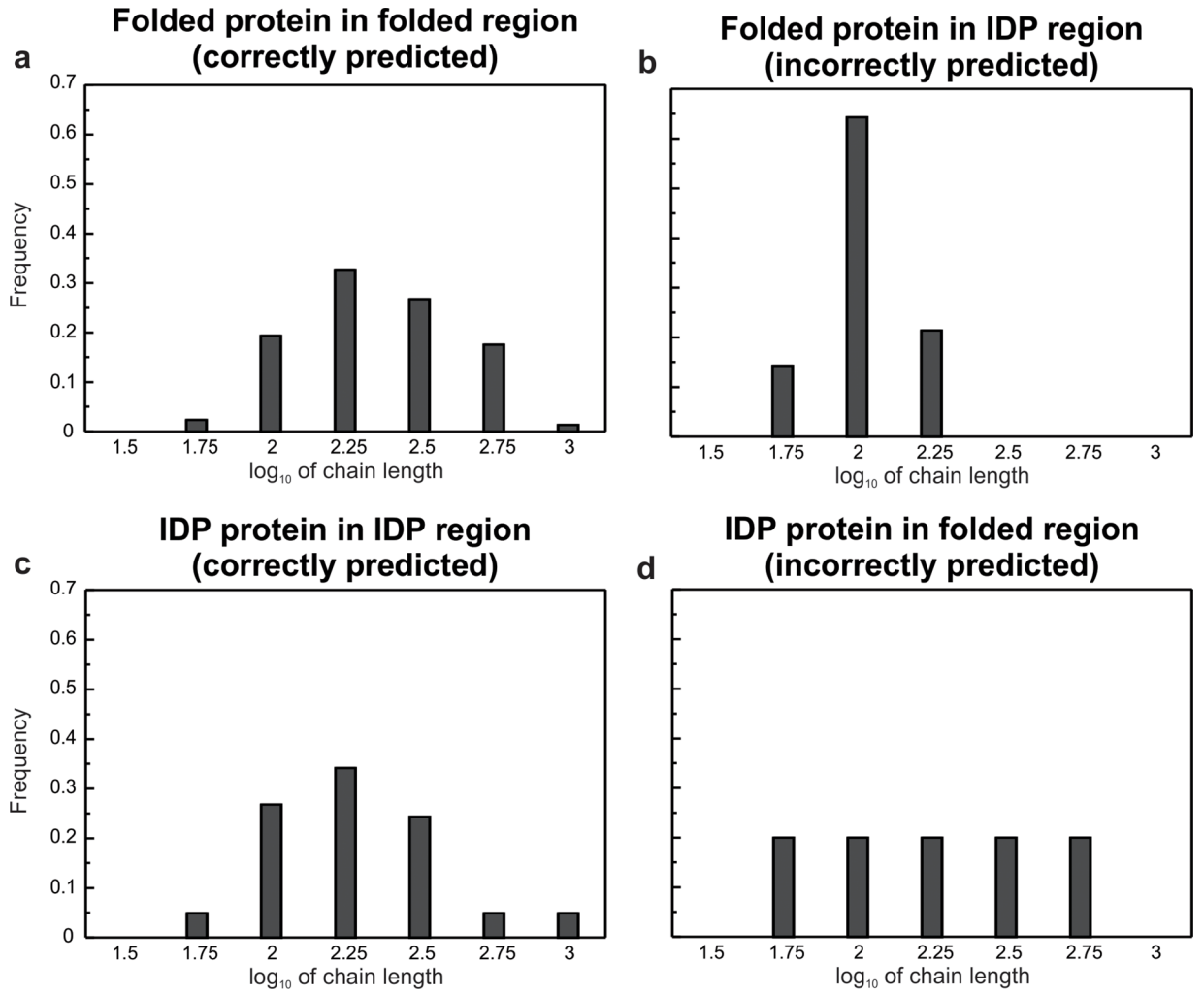


Figure 3. Histogram illustrating the chain-length distribution of single-domain globular proteins and IDPs of Fure 2. The frequency vs logarithm of chain length is plotted for (a) correctly predicted folded proteins, i.e., proteins whose MNC and MNPC fall within the folded-protein region; (b) incorrectly predicted folded proteins whose MNC and MNPC fall within the predicted disordered-protein region; and IDPs (c) correctly or (d) incorrectly predicted to fall within the IDP region.

corresponding to a 93.9% success rate, while Uversky's method correctly predicted only 181 of the 231 folded-protein sequences, resulting in a success rate of 78.4%.

The Prediction Accuracy of Folded Proteins Increases Dramatically for Chain Lengths of >140 Residues. Next, we focused on the correct prediction of folded proteins, given its potential impact on proteome analysis and protein structure prediction. As shown in panels a and b of Figure 3, which is based on the data in Figure 2, we noticed that the incorrectly predicted folded proteins tend to be shorter than the correctly predicted folded proteins. In contrast, correctly and incorrectly predicted IDPs are distributed in a more size-independent fashion (Figure 3c,d). Note that the distribution in Figure 3d is perfectly flat because a total of only five IDPs (of progressively increasing size) were incorrectly predicted.

The distributions in Figure 3 suggest that the net charge and nonpolar content criteria exploited in NECNOP plots are less reliable for small folded proteins.

Indeed, a replotting of the data of Figure 2 including only folded proteins with >140 residues leads to a large increase in the percent of correctly predicted folded proteins, from 93.9 ± 3.1% to 100 ± 0.0%. This result is detailed in Figure 4 and Table 1. On the other hand, no improvement was achieved in the ability to predict IDP status, when the >140-residue cutoff was applied. In summary, our NECNOP algorithm is

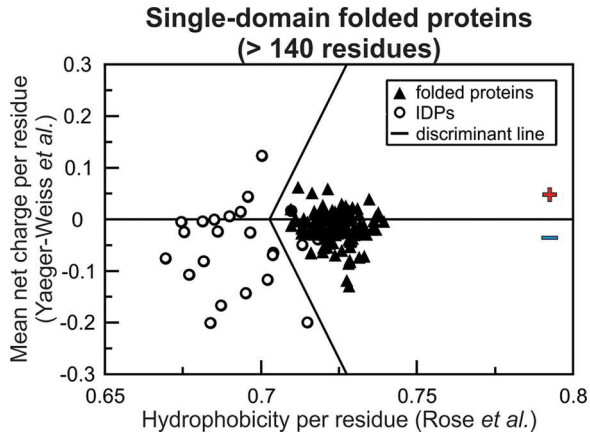


Figure 4. NECNOP plot illustrating the mean net charge and nonpolar content per residue of known folded, globular, single-domain proteins with >140 residues. The plot was generated starting from database 3 (see Table S4), which comprises 231 proteins from a variety of organisms.

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

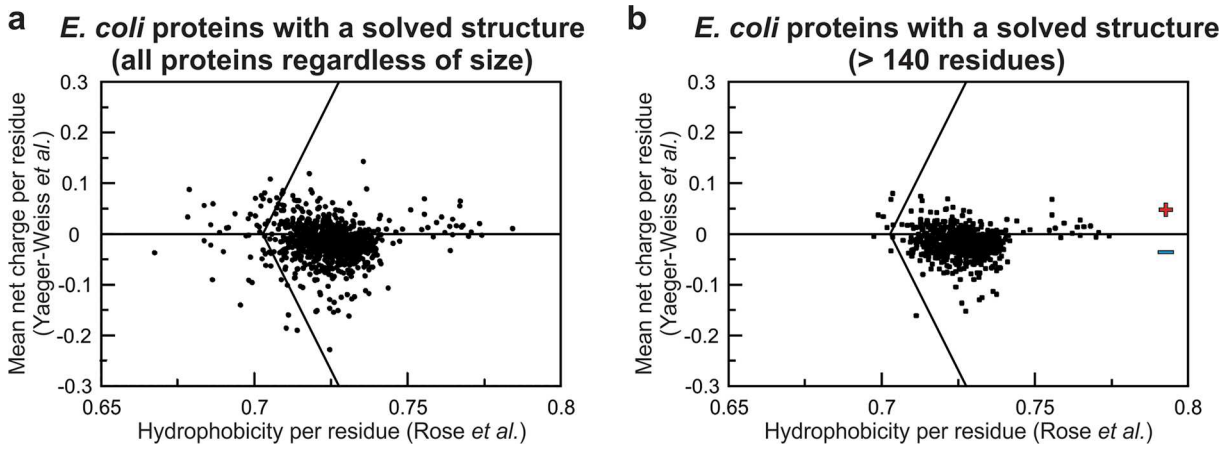


Figure 5. NECNOP plot illustrating the mean net charge per residue (MNC) and mean nonpolar content per residue (MNPC) of (a) all *E. coli* proteins with a determined structure and (b) *E. coli* proteins with a determined structure carrying >140 residues. Plots were generated starting from database 4 (see the Supporting Information), which comprises a large number (i.e., 1147) of proteins from *E. coli* that are known to be folded. Table 3 summarizes the general features of the proteins in this figure (see Table S5 for outliers).

particularly effective at predicting the folded status of proteins with >140 residues.

To more thoroughly validate the significance of the latter result and increase accuracy, we generated NECNOP plots for a larger protein database comprising 1147 folded proteins of known structure (see the Supporting Information for the definition of database 4) from *E. coli*. As shown in Figure 5a and Table 2, the folded status of proteins of all sizes is correctly predicted with 95.2 ± 1.2% accuracy.

Remarkably, the prediction accuracy for folded proteins improves to 98.8 ± 0.7%, when the analysis is restricted to proteins with >140 residues (Figure 5b and Table 2). Given the large size of the database used to generate the latter set of data, we regard 98.8 ± 0.7% as a more reliable estimate of the performance of the NECNOP algorithm for the prediction of folded proteins with >140 residues.

Table S5 describes the 10 proteins that are incorrectly predicted to lie in the IDP region of the plot in Figure 5b. No particular trends were identified. Hence, we conclude that these outlier proteins represent a true reflection of the limitations of the NECNOP analysis.

In summary, data in Figure 5 and Table 2 demonstrate that the folded state of midsize to large proteins (>140 residues) is reliably predicted by the NECNOP method (Table 3).

Table 3. Summary of General Characteristics of Proteins Plotted in Figure 5

	total no. of proteins	total no. of folded proteins	total no. of unfolded proteins
all proteins	1147	1092	55
proteins with >140 residues	838	828	10

Analysis of the *E. coli* Proteome. The NECNOP approach was then applied to the entire proteome of *E. coli* (K12 strain). We focused on *E. coli* proteins with >140 residues, given the higher reliability of NECNOP for this class of proteins. The resulting plots are shown in Figure 6. Panel a shows that most *E. coli* proteins fall within the folded region, with a densely populated section comprising proteins with a high nonpolar content (>0.7). The high representation of proteins on the right-hand side of the discriminant line suggests that the *E. coli* proteome contains very few or no IDPs of >140 residues. Indeed, the eight proteins lying in the IDP region of Figure 6a are fewer than the statistically expected number of folded proteins incorrectly assigned to the IDP region [1.2% proteins (see Table 2)]. Hence, our data predict that there are no IDPs with >140 residues in *E. coli*. This result is qualitatively consistent with the reported expectation that 5% of the *E. coli* proteome consists of disordered proteins.[34]

Interestingly, the NECNOP distribution of folded proteins, on the right of the discriminant line, is far from uniform and has a distinctly bimodal profile (Figure 6a). Region I (0.71 < MNPC < 0.75) includes relatively nonpolar proteins, while region II (0.75 < MNPC < 0.79) comprises proteins with a higher nonpolar content.

Additional insights are deduced from analysis of individual subclasses of the *E. coli* proteome defined according to the UniProt database. The soluble subclass of the proteome (Figure 6b) falls mostly within region I. The majority of these proteins has a negative mean net charge per residue. This result is consistent with the fact that most soluble *E. coli* proteins have a low isoelectric point and few positively charged amino acids and are enriched with aspartic and glutamic acid.[35,36]

As shown in panels c of Figure 6 and its enlarged version (Figure 6d), integral membrane proteins span regions I and II. A comparison with the full proteome (Figure 6a) shows that the fraction of proteins with MNPC values of >0.75 consists mostly of integral membrane proteins. Thus, while region I includes soluble proteins (panel b), non-integral membrane proteins (panel e), uncharacterized proteins (panel f), and some integral membrane proteins, the more highly nonpolar region II is nearly exclusively populated by integral membrane proteins.

Panel d of Figure 6 highlights the interesting finding that MNC increases linearly with MNPC, in integral membrane proteins. The origin of this trend is not clear at this juncture and is likely nontrivial. One potential rationalization is that a higher fraction of positive charges may be necessary for the proper membrane insertion and orientation of highly hydrophobic integral membrane proteins. Interestingly, integral membrane proteins with a higher hydrophobicity per residue, i.e., a larger MNPC, tend to be smaller in size and relatively richer in positively charged residues (see text below).

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

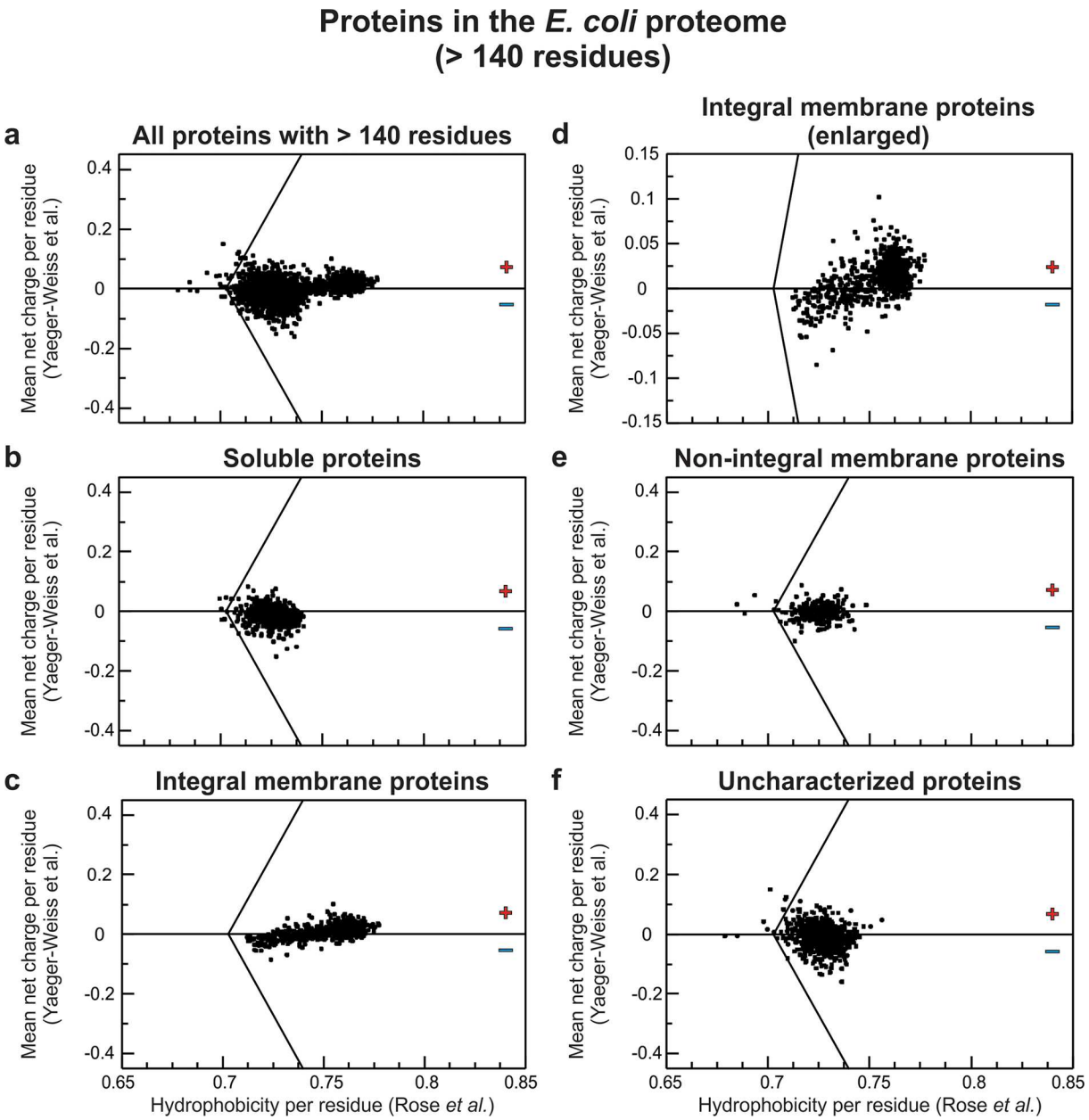


Figure 6. NECNOP plots showing the MNC and MNPC values of all *E. coli* proteins with >140 residues: (a) all proteins, (b) water-soluble proteins, (c) integral membrane proteins, (d) integral membrane proteins (enlarged), (e) non-integral membrane proteins, and (f) uncharacterized proteins. All protein categories are defined according to the UniProt database. All proteins in this figure belong to a subset of database 5 (see Supporting Information).

Therefore, the observed trends suggest that smaller integral membrane proteins are more hydrophobic and more positively charged.

Panel e shows that non-integral membrane proteins exhibit trends in mean net charge and nonpolar content per residue similar to those of soluble proteins. These results are not surprising as non-integral membrane proteins experience an environment largely dominated by bulk solution properties.

The remaining proteins, plotted in panel f, are defined as uncharacterized in the UniProt database. These proteins have a fairly wide distribution of mean net charge and a moderate nonpolar content.

Convergence of *E. coli*-Protein MNC at a High Molecular Mass. Figure 7 illustrates the dependence of MNC on protein chain length. Panel a shows that, across the *E. coli* proteome, MNC dramatically converges toward a small net negative value.

Proteins with fewer than ∼140 residues display a large variability in their mean net charge per residue with no bias toward either positive or negative values (Figure 7a). In contrast, larger proteins are characterized by a remarkably narrower distribution, up to a convergence point of approximately −0.05 MNC. This value is reached at chain lengths equal or larger than ∼1000 residues.

Partitioning of the *E. coli* proteome according to protein type [soluble, integral, and non-integral membrane proteins (panels b−d, respectively, of Figure 7)] shows that the observed convergence at large protein chain lengths is common to all protein types.

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

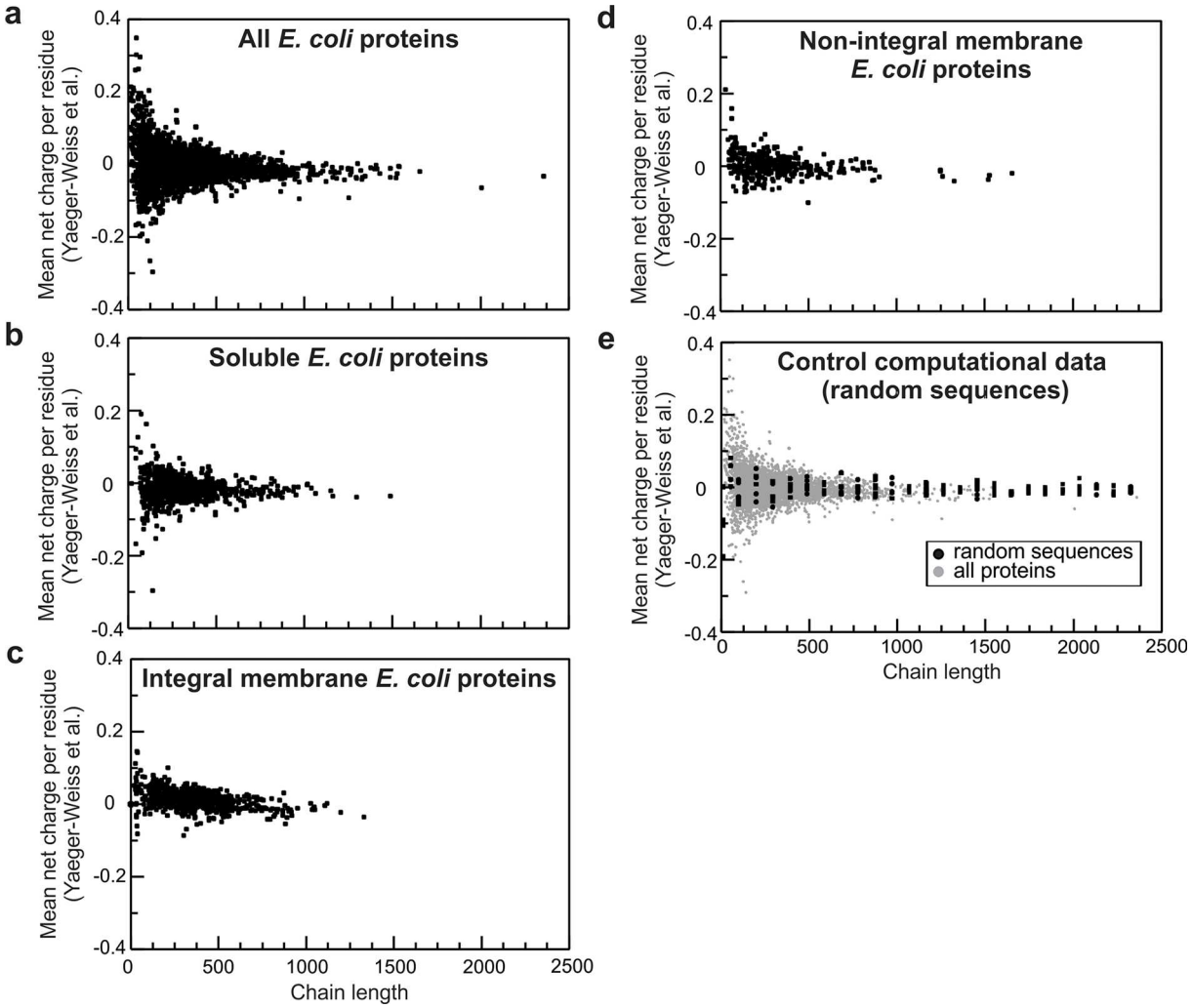


Figure 7. Plot of MNC as a function of chain length across the *E. coli* proteome (in the Supporting Information, see database 5): (a) all proteins, (b) water-soluble proteins, (c) integral membrane proteins, (d) non-integral membrane proteins, and (e) uncharacterized proteins. All protein categories are defined according to the UniProt database. Panel e includes computationally generated control data (black dots) consisting of random sequences of variable chain length, preserving the known frequency of occurrence of amino acids in proteins.

Control Experiments. Next, we carried out control computations to explore whether the observed convergence might be merely a result of statistical arguments. Toward this end, random protein sequences with lengths between 10 and 2400 residues were generated and constrained to fulfill the known amino acid frequency in proteins.[37] MNC values were calculated as a function of chain length. The results are shown in Figure 7e. As the number of residues increases, the fractional abundance of each amino acid in the full-length protein more accurately approaches the average proportion of that amino acid in Nature, leading to a narrower distribution. Panel e of Figure 7 shows that the shorter control sequences exhibit a slightly wider distribution in net charge compared to the longer control sequences (black dots in Figure 7e), yet the change in distribution width at short and long chains is considerably smaller than in the case of the experimental values based on existing proteins (see the gray dots in Figure 7e), especially in the case of very small proteins ($\lesssim$140 residues).

We conclude that the observed MNC convergence of large proteins is only partially explained by simple statistical arguments. Therefore, the observed convergence is a novel property of bacterial proteins identified by our NECNOP plots. These results highlight the large variability in MNC values for small proteins.

The control and experimental MNC values at long protein chain lengths match remarkably well, as shown in Figure 7e. The actual convergence value is entirely predictable on the basis of the amino acid abundance in proteins, consistent with the nature of the control computations (Figure 7e). Indeed, the plot of Figure 7e shows that the convergence value corresponds to the average MNC for proteins of all sizes. Given that this value was deduced from data on proteins from a variety of organisms, we conclude that the MNC convergence may be a general property of proteins; i.e., it may not be solely restricted to the bacterial realm.

Why Does Protein MNC Converge to Slightly Negative Values? Given the findings described above, it is natural to wonder why a slightly negative MNC convergence value is observed for proteins of high molecular mass. Recent studies showed that positively charged proteins translationally diffuse 100 times more slowly than negatively charged proteins in the bacterial cytoplasm, due primarily to nonspecific interactions with the negatively charged ribosomes.[38] Thus, we propose that the demand for effective intracellular

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

translational diffusion, particularly for large proteins that intrinsically diffuse more slowly, justifies the observed convergence to slightly net negative MNC values at high molecular masses.

What Is the Origin of the Observed Protein MNC Convergence? Additional Considerations. The shape of the MNC plots of Figure 7 can be further justified on the basis of simple geometrical arguments and protein surface properties. As shown in Figure 8, the ratio of surface area to volume is

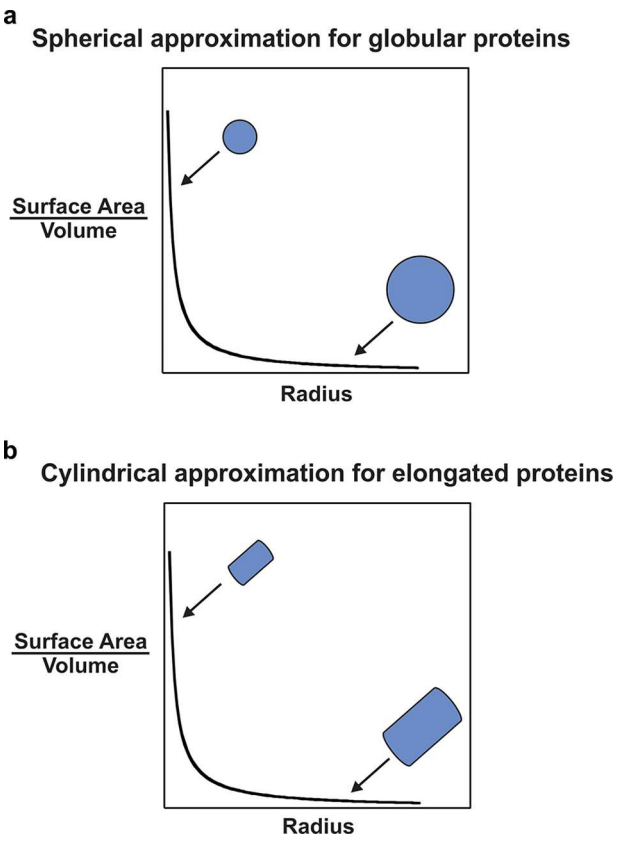


Figure 8. Expected protein surface area to volume ratio for molecular shapes resembling either (a) a sphere or (b) a cylinder.

expected to decrease as the protein radius increases, regardless of whether proteins have a globular [approximately spherical (Figure 8a)] or elongated [approximately cylindrical (Figure 8b)] shape. In other words, geometrical arguments show that larger folded proteins are expected to be more effective at burying surface area than smaller proteins. This prediction is consistent with the experimental finding, based on X-ray crystal structure data, that a greater fraction of surface area is buried in structured proteins with higher masses.[26] In summary, larger folded proteins are expected to have a smaller solvent-exposed area relative to their volume.

Let us consider this conclusion and, in addition, some fundamental surface properties of proteins. It is reasonable to propose that the ratio of the polar to nonpolar surface of folded proteins (necessary to grant solubility) is roughly size-independent. This assumption is corroborated by the surface−property calculations of panels a and b of Figure S1. These calculations were carried out with Surface Racer and were based on known values for single-domain folded proteins. Now, let us also assume, for the sake of simplicity, that the majority of a protein's net charge is contributed by surface residues.[39,40]

As shown in Figure 9, additional calculations reveal that the net charged ASA/total ASA ratio of the folded state is widely

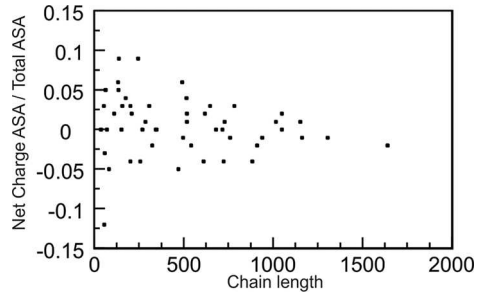


Figure 9. Plot of the ratio of net charged solvent-accessible surface area to total solvent-accessible surface area (net charged ASA/total ASA) as a function of chain length, for 52 single-domain proteins from databases 1−3 (see Methods). Surface areas were computed with Surface Racer.

distributed at short chain lengths and converges toward a slightly negative value for larger proteins. This trend is qualitatively similar to the MNC trends observed in Figure 7a.

Therefore, we propose that the wide distribution in MNC at short chain lengths and the narrowing observed at larger lengths result from the fact that, in larger proteins, a smaller fraction of residues contributing to the net charge is needed due to the decreased surface area/volume ratio. Intriguingly, our results imply that smaller proteins can accommodate a wider range of net charge values on their surface. This effect may relate to the size dependence of protein translational diffusion in the cell.[41,42] However, the actual origin of this phenomenon is still unclear at this juncture, and its understanding needs to await further investigations.

Convergence of *E. coli*-Protein MNPC at High Molecular Masses. Figure 10a shows that the distribution of mean nonpolar content per residue (MNPC) as a function of protein chains length is bimodal. The first cluster, characterized by a lower value for MNPC (centered at MNPC ∼ 0.72), is populated primarily by soluble proteins and by non-integral membrane proteins, with some representation by integral membrane proteins. The second cluster, centered at a higher value of MNPC (∼0.76), is populated almost exclusively by integral membrane proteins.

Soluble proteins and non-integral membrane proteins are likely to be less hydrophobic than integral membrane proteins, consistent with their existence in the cytoplasm in water-soluble form.

Panels b−d of Figure 10 show that each protein subclass populates a characteristic range of MNPC values. Soluble proteins (panel b) have hydrophobicity ranging from ∼0.68 to 0.75, converging around 0.72 as the chain length increases. Integral membrane proteins (panel c) span a wider MNPC range, from 0.71 to ∼0.80, with a higher population density at large MNPC values. Overall, integral membrane proteins also converge to an MNPC of ∼0.72, at large chain lengths. Non-integral membrane proteins behave like soluble proteins, consistent with the fact that these proteins have only a small fraction of their length embedded in the membrane.

Figure 10e shows an overlay of the plot for all *E. coli* proteins (gray dots) and control sequences (black dots). The same random sequences were employed for the controls in Figures

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

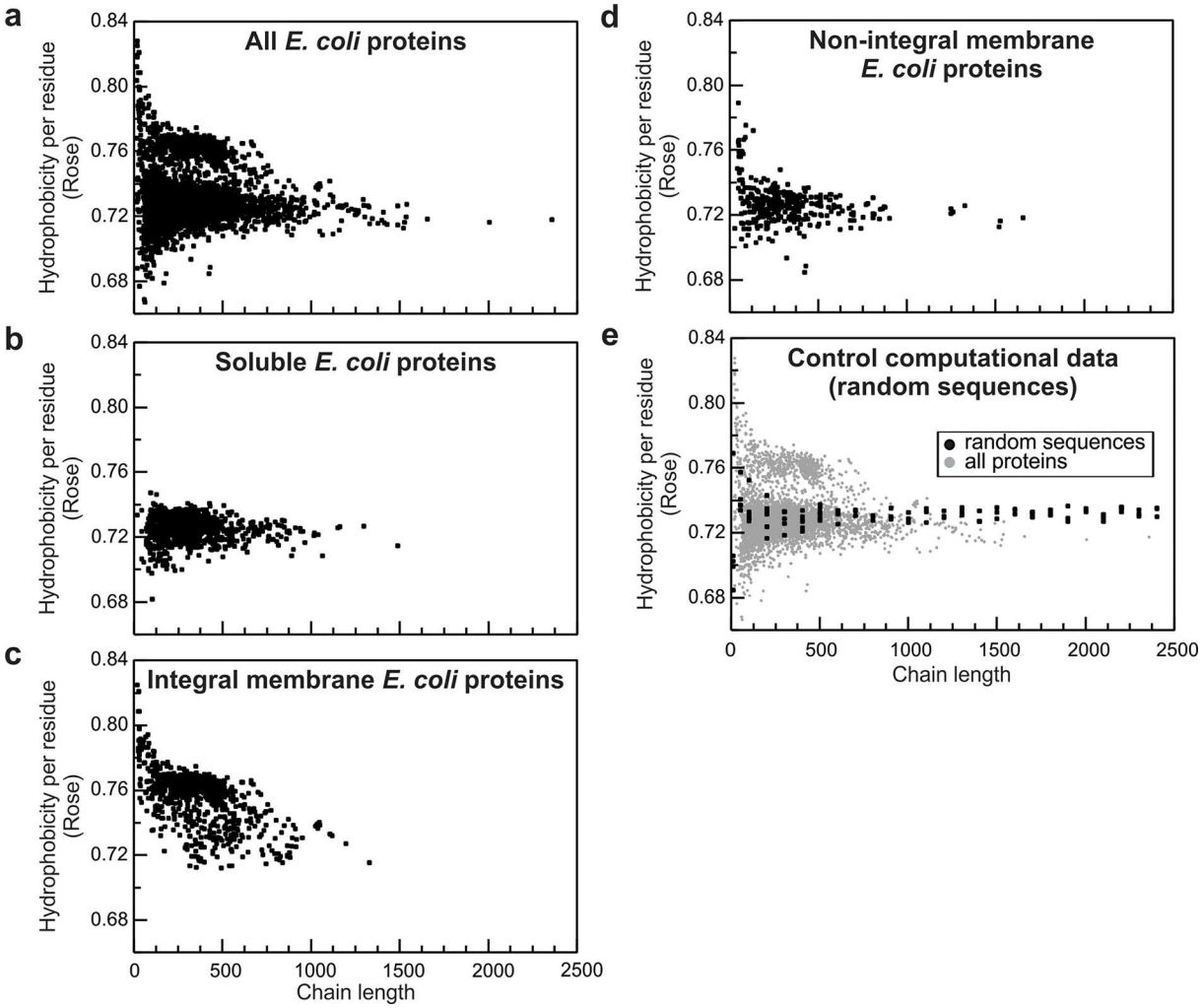


Figure 10. Plot of MNPC as a function of chain length across the entire *E. coli* proteome (in the Supporting Information, see database 5): (a) all proteins, (b) water-soluble proteins, (c) integral membrane proteins, (d) non-integral membrane proteins, and (e) uncharacterized proteins. All protein categories were defined according to the UniProt database. Panel e includes computationally generated control data (black dots) consisting of variable-length random sequences carrying the known frequency of occurrence of amino acids in proteins.

7e and 10e. Interestingly, the control plots of Figure 10e show a degree of convergence very similar to that of the experimental data for soluble and non-integral membrane proteins. Hence, we conclude that the MNPC convergence observed for this class of proteins at a high molecular mass is simply due to the more complete statistical averaging of MNPC.

An integral membrane protein can be modeled as a series of repeating units consisting of nonpolar sections that are embedded in the membrane and polar sections that are exposed (Figure 11a). The number of these units increases with protein length. Therefore, consistent with the data of Figure 10c, the average MNPC is expected to decrease as the chain length increases because the average ratio of nonpolar to polar residues in all units approaches a mean value. Figure 10c shows that the integral membrane proteins with the highest MNPC per residue values are fewer than 85 residues in length. On the basis of the width of the *E. coli* membrane and the length of α helices, we calculated that approximately 42 residues span the width of the membrane. Thus, proteins with fewer than 84 residues are likely to pass through the membrane only twice. This result indicates that short integral membrane proteins may be significantly more hydrophobic, because they have a higher fraction of residues embedded within the membrane compared to the sections of protein that extend out of the membrane (Figure 11b).

In summary, the observed MNPC trends are accounted for by a combination of statistical arguments (dominant effect for soluble and non-integral membrane proteins) and the expected decrease in nonpolar content per residue at high molecular masses (dominant effect for integral membrane proteins).

NECNOP Plots Are of General Significance. NECNOP plots shed light on the intrinsic nature of protein structure and how mean net charge and hydrophobicity modulate it. Despite the fact that this work focuses solely on the *E. coli* proteome, the NECNOP plots can be generated for proteomes of any organism. Hence, NECNOP plots are expected to be generally useful for enabling predictions on proteins that are uncharacterized or have undetermined structures.

NECNOP Plots Are Valuable Tools for the Prediction of Prions. In addition to the binary classification of proteins as folded or intrinsically disordered outlined above, NECNOP plots have additional useful applications. Here, we highlight NECNOP plots in the context of prion discovery.

Prions are proteins whose typically monomeric native cellular state, often denoted PrP^C, has two distinct domains,

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

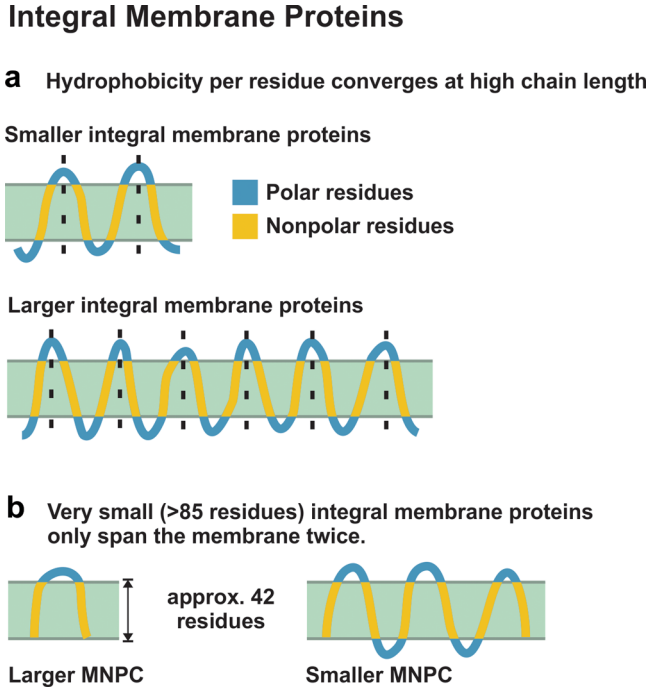


Figure 11. Cartoons qualitatively illustrating the fact that (a) as integral membrane proteins increase in length, the ratio of the number of nonpolar to polar residues in each repeat unit is expected to converge to an average value, assuming that the overall composition is not chain length-dependent. The cartoons in panel b show that short integral membrane proteins are expected to be more hydrophobic than long integral membrane proteins of similar composition, due to the higher fraction of nonpolar membrane-embedded residues.

one folded and one disordered.[43,44] Both domains are usually of comparable length.[43,44] Misfolded prion protein isoforms, known as PrP^{Sc}, are deemed infectious in the sense that they induce misfolding and aggregation of PrP^{C}, causing a class of neurodegenerative diseases known as transmissible spongiform encephalopathies.[45−47] The structure of PrP^{Sc} and the mechanism of conversion of PrP^{C} to PrP^{Sc} are not fully understood, to date.[48] While the existence of mammalian and yeast prions has been known for a long time, the potential presence of prions in bacteria has been proposed only recently,[49,50] Two bacterial prion-like proteins have been experimentally identified,[51−53] and a biological assay for identifying bacterial prions has been developed.[54]

Given its importance for basic science and the development of antibacterial strategies, computational methods for the prediction of prions have been recently developed. These methods are based on the identification of regions enriched with glutamine and asparagine and devoid of prolines and charged residues, which are characteristic of yeast prions.[50,55] This approach led to the identification of several prion candidates in a variety of organisms.[49,56]

NECNOP plots of known prion amino acid sequences, shown in Figure 12, display considerable potential for the prediction of novel representatives of this class of proteins. Due to the presence of folded and intrinsically disordered IDP-like domains of comparable size in prions, NECNOP plots of PrP^{C} from chickens, mice, humans, and yeast happen to fall either on top of or extremely close to the NECNOP discriminant line (Figure 12). In addition, most of these prions display nearly identical values of the MNC and MNPC

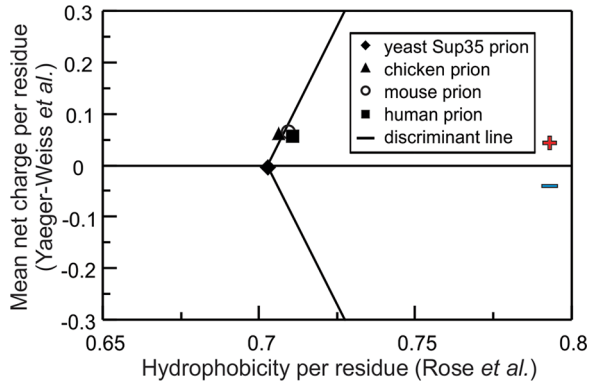


Figure 12. NECNOP plots of mammalian and yeast prion proteins. Prion proteins are readily identified as they overlap with, or fall extremely close to, the discriminant line. In addition, all prion proteins share very similar MNC and MNPC values, except for the yeast prion, which falls on the discriminant line but bears a smaller MNC and MNPC than prions from other organisms. Prion-protein sequences were obtained from the UniProt database (see details in Supporting Information).

parameters, suggesting that the overall nonpolar and electrostatic characteristics of prions may be nonrandom.

With these facts in mind, it is interesting to note that any proteins with more than one intrinsically disordered region (IDR) of total length comparable to the total length of folded regions are also likely to lie along the discriminant line of NECNOP plots. Hence, our predictions are not necessarily restricted to the identification of prion proteins alone. Folded proteins with several IDRs totaling to a length comparable to that of the folded regions are also likely to fall on (or close to) the NECNOP discriminant line.

In summary, the finding described above shows that the NECNOP method can be employed as an aid in the identification of novel prions in bacteria or other organisms. A combination of NECNOP plots and the other prion-specific computational and experimental tools listed above are a promising avenue for future more refined investigations.

Implications of NECNOP Plots for *De Novo* Protein Design. Naturally occurring proteins comprise only a small fraction of the vast number of possible sequences and structures.[57] Consequently, there is a great deal of interest in designing non-naturally occurring protein sequences, potentially able to perform *ad hoc* functions, both computationally and experimentally.[58−61]

De novo computational protein design has been successful. For instance, this approach has enabled the generation of small (<50 residue) protein sequences that can bind specific therapeutic targets.[62,63] However, one of the greatest challenges associated with this field has been the time and computing power required for the design of proteins of a large size.[57,64,65]

The NECNOP algorithm has the potential to become a useful tool in protein design. Namely, we envisage NECNOP to be helpful during all stages of protein design, to narrow down the number of amino acid compositions (>140 residues) that are compatible with a folded status under ambient conditions. NECNOP may be particularly useful in the design of large proteins with >140 residues. Proteins fulfilling this size constraint were shown here to fall within a well-defined and accurately characterized range of MNC and MNPC values.

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

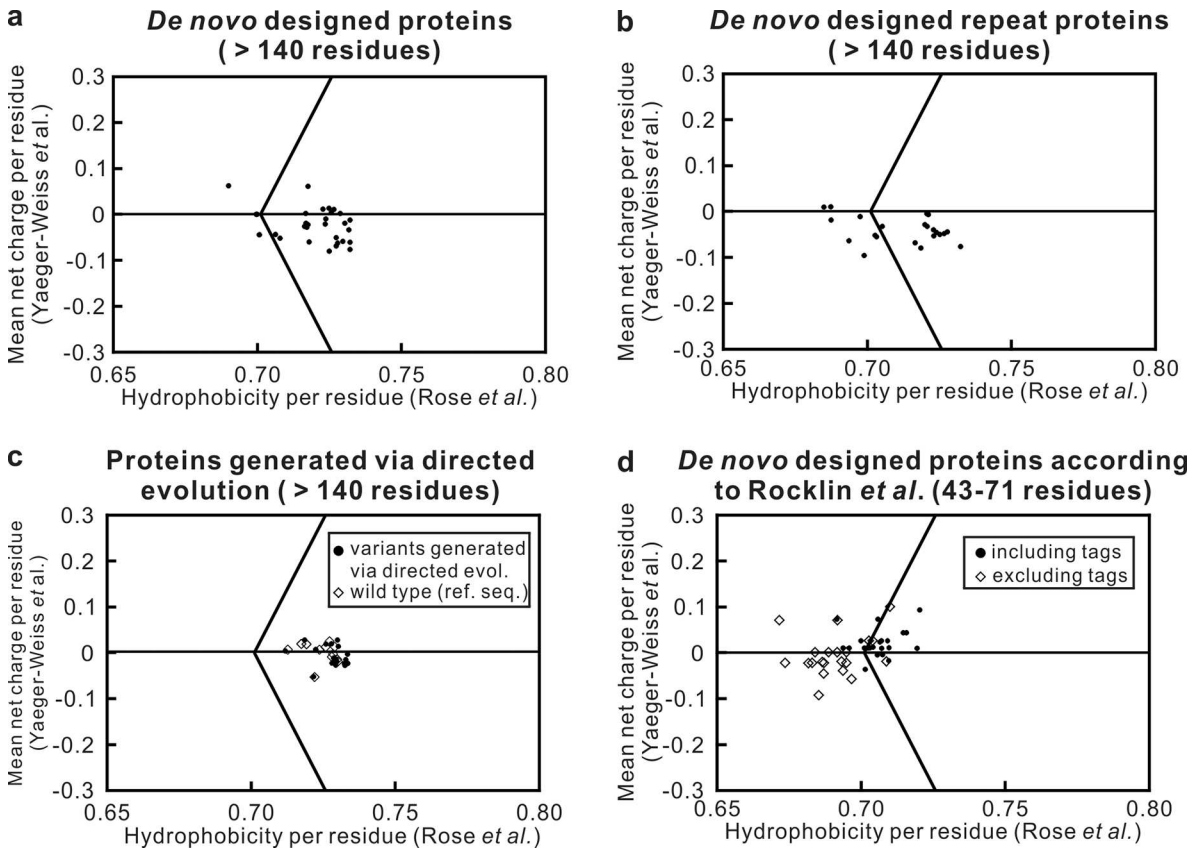


Figure 13. NECNOP plots of a variety of *de novo*-designed proteins as well as proteins generated via directed evolution. All proteins were experimentally prepared, and their three-dimensional structures were determined and deposited in the Protein Data Bank (PDB): (a) *de novo*-designed proteins with >140 residues with a well-folded three-dimensional structure, excluding proteins with large stretches of identical amino acid sequences; (b) *de novo*-designed proteins with >140 residues with a well-folded three-dimensional structure and bearing large stretches of identical amino acid sequences, denoted here as repeat proteins; (c) proteins generated via directed evolution and reference wild-type proteins; and (d) proteins that were *de novo*-designed as well as experimentally generated and characterized by far-ultraviolet circular dichroism and thermal or chemical denaturation by Rocklin et al.[63] and found to have a folded structure. Note that the proteins in panels a–c were identified via an exhaustive PDB search focusing on structures with a single amino acid chain and bearing no ligands or cofactors. See Tables S6–S9 for additional information about each of the proteins in this figure.

To directly test the potential of NECNOP plots to serve as tools in protein design, we applied the NECNOP algorithm to a set of well-characterized *de novo*-designed proteins. We initially focused on successfully *de novo*-designed sequences of >140 residues that were experimentally prepared and purified, and whose three-dimensional structures were independently determined and deposited in the Protein Data Bank (PDB) (see Tables S6 and S7). The results are shown in panels a and b of Figure 13. Interestingly, most of the *de novo*-designed proteins in Figure 13a (which have no stretches of identical amino acid sequence) lie in the proper folded-protein region of the NECNOP plot. On the other hand, a significant number of *de novo*-designed repeat proteins, which bear stretches of identical amino acid sequence and are shown in Figure 13b, lie in the IDP region. This result suggests that indeed *de novo*-designed protein sequences may be further optimized, and NECNOP plots have the potential to serve as a valuable tool to facilitate and streamline this process.

Next, we generated NECNOP plots of novel protein sequences that were generated via directed evolution,[58,59,61] as opposed to *de novo* approaches. The results are shown in Figure 13c (see also Table S8). Interestingly 100% of the directed-evolution-produced proteins fall well within the NECNOP folded-protein region. Given that the directed-evolution approach typically yields sequences fairly close to those of the corresponding reference wild-type protein, with only a small percent of the overall residues being mutated, the features of the NECNOP plot in Figure 13c are perhaps not entirely surprising. On the other hand, this plot highlights the fact that directed evolution is overall a safe approach, in that the proteins generated via this technique tend to retain the global, overall physical properties of the parent wild-type protein sequence.

Finally, as shown in Figure 13d and Table S9, we generated a NECNOP plot for a set of small (43–71 residues) *de novo*-designed proteins. These proteins, which were designed by Rocklin and co-workers and were experimentally shown to have a folded state in solution,[63] fall mostly within the IDP region of the NECNOP plot, in the absence of the N-terminal histidine tag (Figure 13d and Table S9). Interestingly, addition of the N-terminal tag increases the net charge of all of the proteins designed by Rocklin et al. and also augments their nonpolar nature. As a result, most proteins bearing the tag lie in the folded-protein region of the NECNOP plot (Figure 13d and Table S9). This result highlights the fact that small *de novo*-designed proteins may lie in unexpected regions of the NECNOP plot, relative to the majority of proteins found in Nature. These data also clearly show that introduction of a

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881–1895

simple unstructured tag can have dramatic effects, in this case clearly advantageous, on the overall physical properties of a protein.

We conclude that NECNOP plots are a promising tool for protein design, especially—but not only—in the case of amino acid sequences with >140 residues.

CONCLUSIONS

In summary, we developed the NECNOP method to accurately discriminate folded proteins from IDPs, based on amino acid sequence alone. NECNOP plots are particularly effective for the prediction of folded proteins whose chain length is >140 amino acids.

Analysis of the *E. coli* proteome shows that the folded region is densely populated at relatively low MNC (±0.1) and high MNPC (>0.7) values. Consistent with the literature, the soluble *E. coli* proteome is folded and mostly populated by negatively charged proteins. Non-integral membrane proteins exhibit trends similar to those of soluble proteins. Integral membrane proteins may or may not be more hydrophobic than soluble proteins and are characterized by increasingly positive MNC as the chain length increases.

In total, this work improves the classification of folded proteins based on mean nonpolar and mean net charge content per residue. In addition, it identifies peculiar characteristics of the amino acid sequence of very large proteins and prions. It is hoped that the criteria established here will ultimately serve to improve the *a priori* evaluation of protein structural characteristics based on amino acid sequence alone and contribute to the enhancement of the available toolkit for the *de novo* design of new proteins.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.9b01114.

Supporting Methods detailing protein database definitions and window-size comparisons, supporting tables listing protein characteristics (e.g., PDB IDs and other basic properties), and plots of solvent-accessible surface areas as a function of protein chain length (PDF)

AUTHOR INFORMATION

Corresponding Author

Silvia Cavagnero − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States;* orcid.org/0000-0002-4290-2331; Phone: 608-262-5430; Email: cavagnero@chem.wisc.edu

Authors

Susanna K. Yaeger-Weiss − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Theodore S. Jennaro − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Miranda Mecha − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Jenna H. Becker − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Hanming Yang − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Gordon L. W. Winkler − *Department of Chemistry, University of Wisconsin—Madison, Madison, Wisconsin 53706, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.9b01114

Author Contributions

‖S.K.Y.-W. and T.S.J. contributed equally to this work.

Funding

This work was funded by National Science Foundation (NSF) Grants MCB 1616459 and CBET 1912259 (to S.C.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful to Justin Dang for critical reading of the manuscript.

REFERENCES

(1) Bright, J. N., Woolf, T. B., and Hoh, J. H. (2001) Predicting properties of intrinsically unstructured proteins. *Prog. Biophys. Mol. Biol. 76*, 131−173.

(2) Cordes, M. H. J., Davidson, A. R., and Sauer, R. T. (1996) Sequence space, folding and protein design. *Curr. Opin. Struct. Biol. 6*, 3−10.

(3) Shenoy, S. R., and Jayaram, B. (2010) Proteins: sequence to structure and function - current status. *Curr. Protein Pept. Sci. 11*, 498−514.

(4) Kauzmann, W. (1954) in *The mechanism of enzyme action* (McElroy, W. D., and Glass, B., Eds.) pp 70−110, Johns Hopkins Press, Baltimore.

(5) Widom, B., Bhimalapuram, P., and Koga, K. (2003) The hydrophobic effect. *Phys. Chem. Chem. Phys. 5*, 3085−3093.

(6) Southall, N. T., Dill, K. A., and Haymet, A. D. J. (2002) A view of the hydrophobic effect. *J. Phys. Chem. B 106*, 521−533.

(7) Kauzmann, W. (1957) The physical chemistry of proteins. *Annu. Rev. Phys. Chem. 8*, 413−438.

(8) Clarke, S. (1981) The hydrophobic effect: formation of micelles and biological membranes. *J. Chem. Educ. 58*, A246.

(9) Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature 181*, 662−666.

(10) Chothia, C. (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature 248*, 338−339.

(11) Richards, F. M. (1963) Structure of proteins. *Annu. Rev. Biochem. 32*, 269−300.

(12) Kauzmann, W. (1987) Protein stabilization - thermodynamics of unfolding. *Nature 325*, 763−764.

(13) Creighton, T. E. (1993) *Proteins: structures and molecular properties*, 2nd ed., W. H. Freeman, New York.

(14) Kumar, S., and Nussinov, R. (2002) Close-range electrostatic interactions in proteins. *ChemBioChem 3*, 604−617.

(15) Bosshard, H. R., Marti, D. N., and Jelesarov, I. (2004) Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J. Mol. Recognit. 17*, 1−16.

(16) Pluharova, E., Marsalek, O., Schmidt, B., and Jungwirth, P. (2012) Peptide salt bridge stability: From gas phase via micro-hydration to bulk water simulations. *J. Chem. Phys. 137*, 185101.

(17) Dill, K. A. (1990) Dominant forces in protein folding. *Biochemistry 29*, 7133−7155.

(18) Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct., Funct., Genet. 41*, 415−427.

(19) Huang, F., Oldfield, C. J., Xue, B., Hsu, W. L., Meng, J. W., Liu, X. W., Shen, L., Romero, P., Uversky, V. N., and Dunker, A. K. (2014) Improving protein order-disorder classification using charge-hydropathy plots. *BMC Bioinf. 15*, S4.
(20) Das, R. K., and Pappu, R. V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A. 110*, 13392−13397.
(21) Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science 229*, 834−838.
(22) Nelson, D. N., and Cox, M. M. (2013) *Lehninger principles of biochemistry*, 6th ed., Freeman and Company, New York.
(23) Guo, Y. Q., Hastie, T., and Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics 8*, 86−100.
(24) Wallis, S. (2013) Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Joural of Quantitative Linguistics 20*, 178−208.
(25) Tsodikov, O. V., Record, M. T., and Sergeev, Y. V. (2002) Novel computer program for fast exact calculation of accessible an molecular surface areas and average surface curvatures. *J. Comput. Chem. 23*, 600−609.
(26) Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol. 105*, 1−12.
(27) Zhang, Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol. 18*, 342−348.
(28) Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004) Protein structure prediction using Rosetta. In *Numerical Computer Methods, Part D* (Brand, L., and Johnson, M. L., Eds.) p 66, Elsevier Academic Press Inc., San Diego.
(29) Kandathil, S. M., Greener, J. G., and Jones, D. T. (2019) Recent developments in deep learning applied to protein structure prediction. *Proteins: Struct., Funct., Genet. 87*, 1179−1189.
(30) Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019) Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins: Struct., Funct., Genet. 87*, 1011−1020.
(31) Kuhlman, B., and Bradley, P. (2019) Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol. 20*, 681−697.
(32) Wardah, W., Khan, M. G. M., Sharma, A., and Rashid, M. A. (2019) Protein secondary structure prediction using neural networks and deep learning: A review. *Comput. Biol. Chem. 81*, 1−8.
(33) Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105−132.
(34) Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005) Comparing and combining predictors of mostly disordered proteins. *Biochemistry 44*, 1989−2000.
(35) Niwa, T., Ying, B. W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. U. S. A. 106*, 4201−4206.
(36) Chan, P., Curtis, R. A., and Warwicker, J. (2013) Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci. Rep. 3*, 3333.
(37) UniProtKB/Swiss-Prot protein knowledgebase release 2013_04 statistics (2013) UniProt.
(38) Schavemaker, P. E., Smigiel, W. M., and Poolman, B. (2017) Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. *eLife 6*, e30084.
(39) Pey, A. L., Rodriguez-Larrea, D., Gavira, J. A., Garcia-Moreno, B., and Sanchez-Ruiz, J. M. (2010) Modulation of buried ionizable groups in proteins with engineered surface charge. *J. Am. Chem. Soc. 132*, 1218−1219.
(40) Nielsen, J. E., Gunner, M., and García-Moreno E., B. (2011) The pKa cooperative: A collaborative effort to advance structure-based calculations of pKa values and electrostatic effects in proteins. *Proteins: Struct., Funct., Genet. 79*, 3249−3259.
(41) Kapanidis, A. N., Uphoff, S., and Stracy, M. (2018) Understanding Protein Mobility in Bacteria by Tracking Single Molecules. *J. Mol. Biol. 430*, 4443−4455.
(42) Wang, Y., Li, C., and Pielak, G. J. (2010) Effect of Proteins on Protein Diffusion. *J. Am. Chem. Soc. 132*, 9392−9397.
(43) Baral, P. K., Yin, J., Aguzzi, A., and James, M. N. G. (2019) Transition of the prion protein from a structured cellular form (PrP (c)) to the infectious scrapie agent (PrPSc). *Protein Sci. 28*, 2055−2063.
(44) Vazquez-Fernandez, E., Young, H. S., Requena, J. R., and Wille, H. (2017) The Structure of Mammalian Prions and Their Aggregates. In *Early Stage Protein Misfolding and Amyloid Aggregation* (Sandal, M., Ed.) pp 277−301, Elsevier Academic Press Inc., San Diego.
(45) Prusiner, S. B. (1998) Prions. *Proc. Natl. Acad. Sci. U. S. A. 95*, 13363−13383.
(46) Baskakov, I. V., and Breydo, L. (2007) Converting the prion protein: What makes the protein infectious. *Biochim. Biophys. Acta, Mol. Basis Dis. 1772*, 692−703.
(47) Collinge, J., and Clarke, A. R. (2007) A general model of prion strains and their pathogenicity. *Science 318*, 930−936.
(48) Collinge, J. (2016) Mammalian prions and their wider relevance in neurodegenerative diseases. *Nature 539*, 217−226.
(49) Iglesias, V., de Groot, N. S., and Ventura, S. (2015) Computational analysis of candidate prion-like proteins in bacteria and their role. *Front. Microbiol. 6*, 1123.
(50) Espinosa Angarica, V., Ventura, S., and Sancho, J. (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains. *BMC Genomics 14*, 316.
(51) Pallares, I., and Ventura, S. (2017) The transcription terminator rho: A first bacterial prion. *Trends Microbiol. 25*, 434−437.
(52) Shahnawaz, M., Park, K. W., Mukherjee, A., Diaz-Espinoza, R., and Soto, C. (2017) Prion-like characteristics of the bacterial protein Microcin E492. *Sci. Rep. 7*, 45720.
(53) Yuan, A. H., and Hochschild, A. (2017) A bacterial global regulator forms a prion. *Science 355*, 198−201.
(54) Fleming, E., Yuan, A. H., Heller, D. M., and Hochschild, A. (2019) A bacteria-based genetic assay detects prion formation. *Proc. Natl. Acad. Sci. U. S. A. 116*, 4605−4610.
(55) Lancaster, A. K., Nutter-Upham, A., Lindquist, S., and King, O. D. (2014) PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics 30*, 2501−2502.
(56) Espinosa Angarica, V., Angulo, A., Giner, A., Losilla, G., Ventura, S., and Sancho, J. (2014) PrionScan: an online database of predicted prion domains in complete proteomes. *BMC Genomics 15*, 102.
(57) Huang, P. S., Boyken, S. E., and Baker, D. (2016) The coming of age of de novo protein design. *Nature 537*, 320−327.
(58) Yang, K. K., Wu, Z., and Arnold, F. H. (2019) Machine-learning-guided directed evolution for protein engineering. *Nat. Methods 16*, 687−694.
(59) Zeymer, C., and Hilvert, D. (2018) Directed Evolution of Protein Catalysts. In *Annual Review of Biochemistry* (Kornberg, R. D., Ed.) Vol. *87*, pp 131−157, Annual Reviews, Palo Alto, CA.
(60) Renata, H., Wang, Z. J., and Arnold, F. H. (2015) Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution. *Angew. Chem., Int. Ed. 54*, 3351−3367.
(61) Chowdhury, R., and Maranas, C. D. (2020) From directed evolution to computational enzyme engineering-A review. *AIChE J. 66*, e16847.
(62) Chevalier, A., Silva, D. A., Rocklin, G. J., Hicks, D. R., Vergara, R., Murapa, P., Bernard, S. M., Zhang, L., Lam, K. H., Yao, G. R., Bahl, C. D., Miyashita, S. I., Goreshnik, I., Fuller, J. T., Koday, M. T., Jenkins, C. M., Colvin, T., Carter, L., Bohn, A., Bryan, C. M., Fernandez-Velasco, D. A., Stewart, L., Dong, M., Huang, X. H., Jin, R.

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895

S., Wilson, I. A., Fuller, D. H., and Baker, D. (2017) Massively parallel de novo protein design for targeted therapeutics. *Nature 550*, 74−79.
(63) Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C. H., and Baker, D. (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science 357*, 168−175.
(64) Goldenzweig, A., and Fleishman, S. (2018) Principles of protein stability and their application in computational design. *Annu. Rev. Biochem. 87*, 105−129.
(65) Khoury, G. A., Smadbeck, J., Kieslich, C. A., and Floudas, C. A. (2014) Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol. 32*, 99−109.

https://dx.doi.org/10.1021/acs.biochem.9b01114
Biochemistry 2020, 59, 1881−1895