research articles Journal of proteome research

Charge and Hydrophobicity Patterning along the Sequence Predicts the Folding Mechanism and Aggregation of Proteins: A Computational Approach

Joseph P. Zbilut,[†] Alessandro Giuliani,[‡] Alfredo Colosimo,[§] Julie C. Mitchell,[||] Mauro Colafranceschi,[‡,§] Norbert Marwan,[#] Charles L. Webber, Jr.,[⊥] and Vladimir N. Uversky*[,▽,##]

Department of Molecular Biophysics and Physiology, Rush Medical College, Chicago, Illinois 60612, Environment and Health Department, Istituto Superiore di Sanitá, Rome, Italy, Department of Human Physiology and Pharmacology, University of Rome "La Sapienza", Rome, Italy, Departments of Mathematics and Biochemistry, University of Wisconsin-Madison, 480 Lincoln Drive, Madison, Wisconsin 53706-1388, Nonlinear Dynamics Group, Institute of Physics, University of Potsdam, Potsdam, Germany, Department of Physiology, Loyola University Medical Center, Maywood, Illinois 61059, Institute for Biological Instrumentation of the Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia, and Department of Biochemistry and Molecular Biology, School of Medicine, IUPUI, 635 Barnhill Drive, MS 4021, Indianapolis, Indiana 46202

Received July 13, 2004

The presence of partially folded intermediates along the folding funnel of proteins has been suggested to be a signature of potentially aggregating systems. Many studies have concluded that metastable, highly flexible intermediates are the basic elements of the aggregation process. In a previous paper, we demonstrated how the choice between aggregation and folding behavior was influenced by hydrophobicity distribution patterning along the sequence, as quantified by recurrence quantification analysis (RQA) of the Myiazawa−Jernigan coded primary structures. In the present paper, we tried to unify the "partially folded intermediate" and "hydrophobicity/charge" models of protein aggregation verifying the ability of an empirical relation, developed for rationalizing the effect of different mutations on aggregation propensity of acyl-phosphatase and based on the combination of hydrophobicity RQA and charge descriptors, to discriminate in a statistically significant way two different protein populations: (a) proteins that fold by a process passing by partially folded intermediates and (b) proteins that do not present partially folded intermediates.

Keywords: protein folding • protein aggregation • recurrence quantification analysis • charge/hydrophobicity patterning • partially folded intermediate

It is known that that protein association represents an essential problem in biomedicine and biotechnology. Particularly, a number of human disorders, including the several neurodegenerative diseases [such as Alzheimer's disease, Pick's disease, Parkinson's disease, diffuse Lewy bodies disease, Lewy bodies variant of Alzheimer's disease, dementia with Lewy bodies, multiple system atrophy, Huntington disease, Creutzfeld−Jacob disease, Gerstmann−Straussler−Schneiker syndrome, fatal familial insomnia, different transmissible encephalopathies (kuru, bovine spongiform encephalopathy, and scrapie)] originate from the deposition of filamentous protein aggregates, known as amyloid fibrils. In each of these pathological states, a specific protein or protein fragment changes from its natural soluble form into insoluble fibrils, which accumulate in a variety of organs and tissues.[1−6] Currently, approximately 20 different proteins (unrelated in terms of their sequences or structures) are known to be involved in the amyloidoses (extracellular deposits). In addition, a number of diseases are also associated with the appearance of intracellular proteinaceous deposits. Prior to fibrillation, amyloidogenic polypeptides may be rich in β-sheet, α-helix, β-helix, or contain both α-helices and β-sheets.[7] They may be globular proteins with rigid 3D-structure or belong to the class of natively unfolded (or intrinsically unstructured) proteins [the phenomenon of intrinsically unstructured proteins is reviewed in refs 7−17]. Despite these differences, the fibrils from different pathologies display many common properties including a core cross-β-

* To whom correspondence should be addressed. Department of Biochemistry and Molecular Biology, School of Medicine, IUPUI, 635 Barnhill Drive, MS 4023, Indianapolis, IN 46202; E-mail: uversky@hydrogen.ucsc.edu.
[†] Department of Molecular Biophysics and Physiology, Rush Medical College.
[‡] Environment and Health Department, Istituto Superiore di Sanitá.
[§] Department of Human Physiology and Pharmacology, University of Rome "La Sapienza".
[||] Departments of Mathematics and Biochemistry, University of Wisconsin-Madison.
[#] Nonlinear Dynamics Group, Institute of Physics, University of Potsdam.
[⊥] Department of Physiology, Loyola University Medical Center.
[▽] Institute for Biological Instrumentation of the Russian Academy of Sciences.
[##] Department of Biochemistry and Molecular Biology, School of Medicine.

10.1021/pr049883+ CCC: $27.50

Published on Web 10/02/2004

sheet structure in which continuous β-sheets are formed with β-strands running perpendicular to the long axis of the fibrils.[18] Importantly, there is an increasing belief that the ability to fibrillate is a generic property of a polypeptide chain, and all proteins are potentially able to form amyloid fibrils under appropriate conditions.[3,19−23]

In biotechnology, the formation of inclusion bodies is a major problem in the overexpression of recombinant proteins;[24−29] production and in vivo delivery of protein drugs is often complicated by association.[30] It is also known that protein refolding is often accompanied by transient association of partially folded species. Furthermore, the propensity to aggregate is considered as a general characteristic of the nonnative proteins in diluted solutions.[26,29,31−43] Despite all this, little is currently know about the precise molecular mechanism driving protein to choose the aggregation pathway.

Recently, we have proposed that some key features of protein hydrophobicity patterns analysed by a nonlinear signal processing technique, recurrence quantification analysis (RQA), might determine necessary conditions for aggregation.[44] A significant finding included a correspondence between short deterministic patches of hydrophobicity distribution along the sequence, what we term laminarity, with 3-D "unstructured" portions of acylphosphatase (AcP). It was shown that the "ruggedness" of the hydrophobicity as measured by the derivative of hydrophobic change, [what we term TREND, (T)] coincided with Dunker's "disorder".[13,14] Beyond this, a counterpoint was defined as the degree of laminarity [LAM (L)]. Specifically, in an analysis of the protein engineering experiments of Chiti and his associates[45] aimed at modeling the effect of different mutations on aggregation propensity of AcP, it was shown that aggregation sensitive zones vs folding sensitive zones were distinguished by the two complementary concepts of trend/laminarity.[44] The implication of this finding is that these areas may be inherently unstable, and somehow involved in the promotion of (at least) partial unfolding and aggregation. The degree at which these conditions exist probabilistically determines the propensity for aggregation. Consonant with the findings of Chiti et al.,[46−48] we conclude that, the charge acts on this frame modulating the repulsive/attractive electrostatic forces between nearby molecules: as a matter of fact the isoelectric point (when the charge of the system is neutralized) was demonstrated, by means of molecular dynamics simulation and experimental studies, to be both the most flexible and aggregation prone condition of the protein molecule. These observations form the basis for the "charge/hydrophobicity" model of protein aggregation.

Another model of protein aggregation is based on a well-known link between the propensity of a given protein to form partially folded intermediate(s) and the propensity to aggregate. It has been proposed that fibrillation can occur when the rigid native structure of a protein is destabilized, favoring partial unfolding and formation of a partially unfolded intermediate.[1−7,17,19,26,49,50] This hypothesis is based on the following observation: independently on the original structure of a given protein, all fibrils have a common cross-β structure, this means that considerable conformational rearrangement has to occur for this to happen. Such changes cannot take place in the tightly packed native protein, due to the constraints of the tertiary structure. Thus, the formation of some flexible, non-native partially unfolded conformation(s) is required.[1−7,17,19,26,49,50] This hypothesis represents a "partially folded intermediate" model of protein aggregation that on a purely qualitative point of view, has a general affinity with the hydrophobicity/charge hypothesis as for the crucial role exerted by "flexibility".

In this paper, we present an attempt to unify "charge/hydrophobicity" and "partially folded intermediate" models of protein aggregation. Obviously, the most straightforward way to perform this "marriage" is to validate the ability of the formula derived for the prediction of aggregation propensity of AcP to discriminate between two protein sets—proteins that are able *to adopt equilibrium partially folded conformation(s) and proteins, which have been shown to unfold without the formation of any equilibrium intermediate.* It is important to stress that the aggregation/folding choice is a stochastic process, and any protein system can in principle undergo the aggregation. This implies that we can only model the "relative probability" of each protein system to aggregate, consequently we can only expect, at best, a statistically significant relation between aggregation propensity and the presence of partially folded intermediates, and this is what we actually found in this occasion. The demonstration of the ability of the same empirical formula to model both the aggregation propensity of a specific system and the existence of partially folded intermediates along the folding funnel represents some noteworthy evidence of the overlapping of the two phenomena, aggregation and the formation of partially folded intermediates.

Material and Methods

Strategy of Analysis. Experimental data on the effect of different mutations on fibrillation of AcP can be found in refs 45 and 51. These authors demonstrated the presence of *mutational aggregation zones* along the AcP sequence corresponding to the sequences 16−31 and 87−98.[45,51] In fact, only mutations intervening in these portions of the sequence were capable of significantly influencing the aggregation behavior of the protein. To derive a quantitative model, which could be able to generalize these observations to different protein systems, the different sequences were coded by means of Miyazawa−Jernigan hydrophobicity of each single residue and the patterning of hydrophobicity along the chain was quantified by means of the RQA descriptors. These descriptors, solely based on the sequence information, allowed for a statistically significant model describing the effect of mutations on aggregation propensity. The power of the obtained model was improved by inserting the information of the net charge of the system. This gave rise to a synthetic formula in which RQA descriptors and charge interact in a nonlinear manner.

This formula, together with the raw RQA parameters was applied to a 143 proteins data set constituted by 104 protein systems having at least one partially folded intermediate and 39 systems that fold without intermediate states (see below). The obtained data set was analyzed by means of a canonical correlation analysis that allowed for a very significant discrimination between the two groups.

Recurrence Quantification Analysis (RQA). Eckmann introduced a tool that can visualize the recurrence of states $\boldsymbol{x}_i$ in a phase space.[52] Usually, a phase space does not have a dimension (two or three), which allows it to be pictured. Higher dimensional phase spaces can only be visualized by projection into the two- or three-dimensional sub-spaces. However, Eckmann's tool enables one to investigate the m-dimensional phase space trajectory through a two-dimensional representation of its recurrences. Such a recurrence of a state at time i at a different time j is pictured within a two-dimensional squared matrix with black and white dots, where black dots mark a

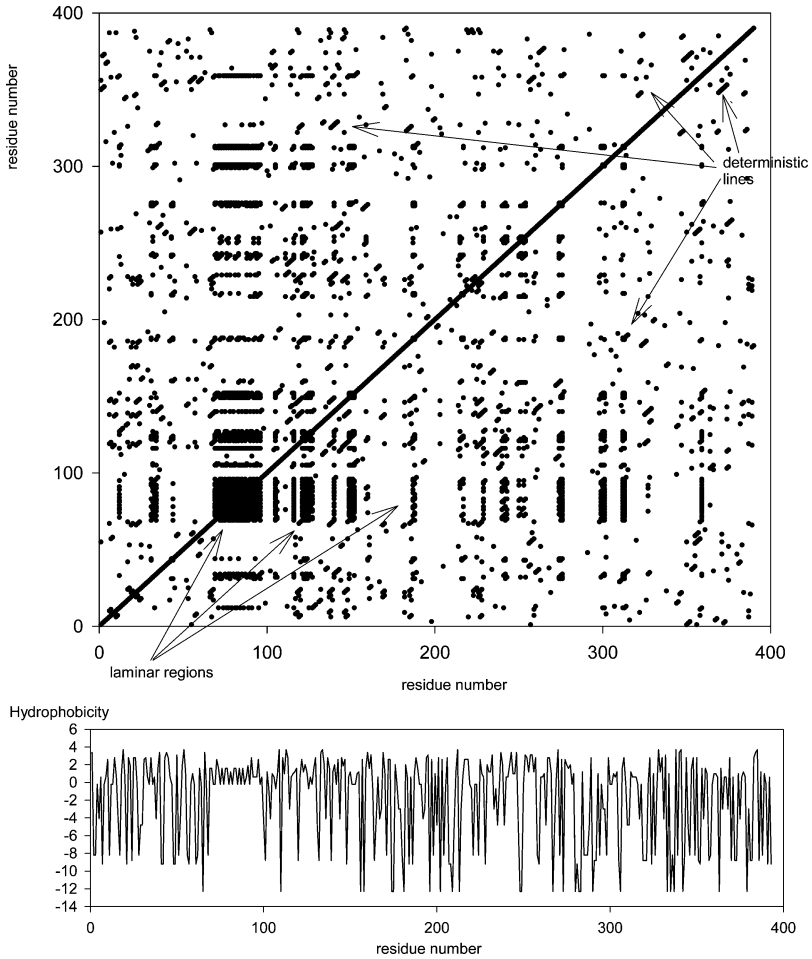

Figure 1. Recurrence Plot of P53 protein, together with the original sequence coded in terms of hydrophobicity of the different residues.

recurrence, and both axes are the ordered sequences. This representation is called *recurrence plot (RP)*. Such an RP can be mathematically expressed as

$$\mathbf{R}_{i,j} = \Theta\ (\epsilon - ||\ \mathbf{x}_i - \mathbf{x}_j||),\ \mathbf{x}_i \in \times c2^m,\ i,\ j{=}1,...,\ N \qquad (1)$$

where N is the number of considered states, ϵ is a threshold distance, $||\ ||$ a norm, and Θ the Heaviside function. The threshold distance, ϵ, determines if a given point is considered recurrent.

The initial purpose of RPs is the visual inspection of higher dimensional correlations. The view on RPs gives hints about the time evolution of these correlations. The advantage of RPs is that they can also be applied to rather short and even nonstationary data.

The closer inspection of the RPs reveals small-scale structures (the texture), which are *single dots*, *diagonal lines* as well as *vertical* and *horizontal lines* (the combination of vertical and horizontal lines obviously forms rectangular clusters of recurrence points). These structures are taken into account and quantified by the descriptors used in RQA.[53−56] In particular:

• *Single, isolated recurrence points* can occur if states are rare, if they do not persist or fluctuate heavily. However, they are not a unique sign of chance or noise. These dots are counted in the quantitative descriptor recurrence (REC).

• A *diagonal line* $\mathbf{R}_{i+k,j+k} = 1$ (for $k = 1...l$, where l is the length of the diagonal line) occurs when a segment of the numerical series runs parallel to another segment, i.e., the sequence visits the same region of the phase space at different intervals. The length of this diagonal line is determined by the duration of such similar local evolution of the segments. These diagonal structures are called deterministic lines and are counted by the descriptor determinism (DET), entropy (ENT) and maxline (MAXL).

• A *vertical (horizontal) line* $\mathbf{R}_{i,j+k} = 1$ (for $k = 1...v$, where v is the length of the vertical line) marks a length in which a state does not change or changes very slowly. It seems, that the state is trapped. These are called laminar regions and are counted in the descriptors laminarity (LAMIN) and traptime (TRAPT).

Figure 1 represents a typical RP of a protein (human P53) with the indication of the different features taken into consideration by quantitative descriptors that in turn are defined in Table 1.

The numerical series studied in this work are protein sequences coded by the hydrophobicity of the constituent residues. Discrete time and spatial series (like nonbranching polymers) are completely congruent mathematical objects, given they are both linear arrangements of discrete subsequent elements with a fixed and well-defined ordering.

Each protein sequence was coded by means of the Miyazawa−Jernigan[57] hydrophobicity scale (MJ) of amino acid residues, a choice dictated by our previous analysis of a 1141 random sample of protein sequences from the Swiss-Prot Database: (URL: ftp://ftp.ebi.ac.uk/pub/contrib/swissprot/testsets/signal). It has been emphasized that this scale corresponds to the first eigenvalue of the contact energy matrix as

Table 1. Definition of RQA Measures

measure	definition
recurrence, REC	percentage of recurrence points in an RP: $\mathrm{RR} = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j}$
determinism, DET	percentage of recurrence points which form diagonal lines: $\mathrm{DET} = \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\sum_{i,j}^{N} R_{i,j}}$ *P(l)* is the histogram of the lengths l of the diagonal lines.
laminarity, LAM	percentage of recurrence points which form vertical lines: $\mathrm{LAM} = \frac{\sum_{v=v_{\min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)}$ *P(v)* is the histogram of the lengths v of the vertical lines.
entropy, ENT	Shannon entropy of the distribution of the diagonal line lengths *p(l)*: $\mathrm{ENT} = -\sum_{l=l_{\min}}^{N} p(l)\ln p(l)$
trend, TREND	paling of the RP toward its edges: $\mathrm{TREND} = \frac{\sum_{i=1}^{N-2} [i-(N-2)](RR_i - \langle RR_i \rangle)}{\sum_{i=1}^{N-2} [i=(N-2)/2]^2}$
trapping time, TRAPT	average length of vertical lines: $\mathrm{TT} = \frac{\sum_{v=v_{\min}}^{N} vP(v)}{\sum_{v=v_{\min}}^{N} P(v)}$
longest diagonal line, MAXL	length of the longest diagonal line: MAXL $L_{\max} = \max(\{l_i; i = 1 \ldots N_l\})$

reported at the URL: http://us.expasy.org/tools/pscale/Hphob.Miyazawa.html. In that case, we demonstrated that the MJ was the code producing the largest separation in distance space for obtained patterns, as compared to a random assortment of amino acids.

The application of RQA implies the a priori setting of the working parameters embedding dimension, radius, and line (the minimum number of adjacent recurrent points to be considered as deterministic). On the basis of studies of the maximal information content of protein sequences as well as our previous analyses, the above parameters were set to the following: embedding dimension = 3; radius = 6 (first minimum of DET as determined by a plot of the radius from 0 to 100; see Figure 8, below), and line = 2.[58−61]

Canonical Correlation Analysis (CCA). The basic form of canonical correlation analysis is the classical linear model:

$$\mathbf{Y} = \mathbf{X}B + e$$

where **Y** is a matrix of dependent variables, **X** is a matrix of dependent variables, B are the regression coefficients, and e is the matrix of random errors. In the case of CCA, the algorithm find the linear combinations (one for **Y** and one for **X** variables) that maximize their mutual correlation coefficient, these linear combinations are called canonical variables. In our particular case, in which **Y** matrix is made by only one column variable (the dichotomous variable indicating the pertaining of each protein to the (a) and (b) groups), CCA is called canonical discriminant analysis (CDA) and the canonical variable corresponds to the linear combination of X variables (in our case RQA descriptors + charge + three formulas combining RQA and charge information + protein length) maximizing the discrimination between the two protein subsets. The normal-

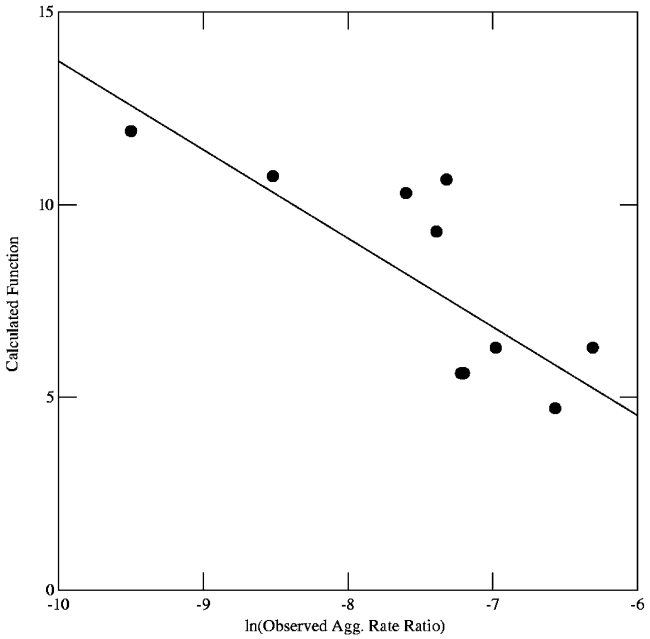

Figure 2. Calculated vs Observed aggregation rates based upon mutations performed by Chiti et al. (ref 45). The mutations were chosen to minimize hydrophobicity collapse and α-helix and β-sheet propensities. Note that the Calculated Function is nonlinear, and is plotted as a linear graph for convenience.

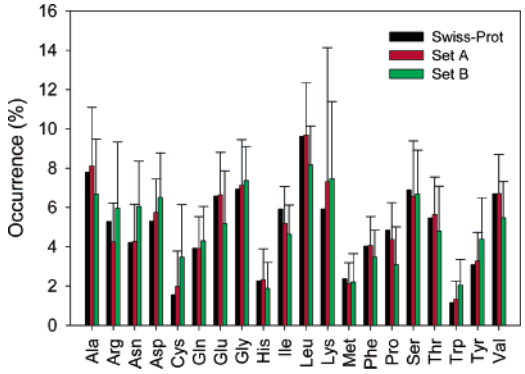

Figure 3. Comparison of the amino acid compositions of proteins from the set A (proteins that are able to form equilibrium intermediates, red bars), the set B (proteins shown to unfold without accumulation of partially folded conformations, green bars) and the averaged composition of a protein for the complete database Swiss-Prot (http://au.expasy.org/sprot/relnotes/relstat.html).

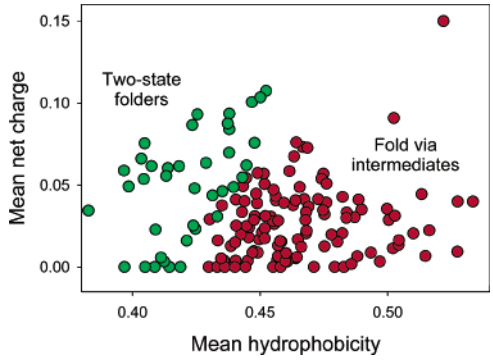

Figure 4. Comparison of mean net charge vs mean hydrophobicity (calculated according to Kyte and Doolittle approach, (ref 85)) for the set of 115 proteins able to form equilibrium intermediates (red symbols) and the set of 39 proteins shown to unfold without accumulation of partially folded conformations (green symbols).

ized regression coefficients indicate the relative importance of each single X variable in the discrimination task. The first canonical variable (CAN1) is then transformed into a dichotomous indicator variable (CLAPRE) having value = A if the protein has a value of CAN1 more similar to the average value of the A group (presence of partially folded intermediates) and a value = B if CAN1 is more similar to the average of the B group (no partially folded intermediates). The resemblance between the actual and predicted (CLAPRE) class was evaluated by means of a chi-square test.

Determination of the Nature of Folding Process. Literature data on the equilibrium unfolding of globular proteins (with unfolding induced by changes in pH, temperature or by increase in the concentration of strong denaturants, such as urea or guanidinium chloride) were analyzed. A test set of 154 globular proteins, *for which data are readily available,* was chosen based on this analysis. For the need of having a net charge other than zero in order to compute our 'aggregation formulas' the set was reduced to 143 proteins. This test set was subdivided into two subsets based on the published unfolding behavior. In the first subset, *104 globular proteins were included, each having been shown to adopt equilibrium partially folded conformation(s). The second subset contained 39 globular proteins, each of which has been shown to unfold without the formation of any intermediate state.*[62]

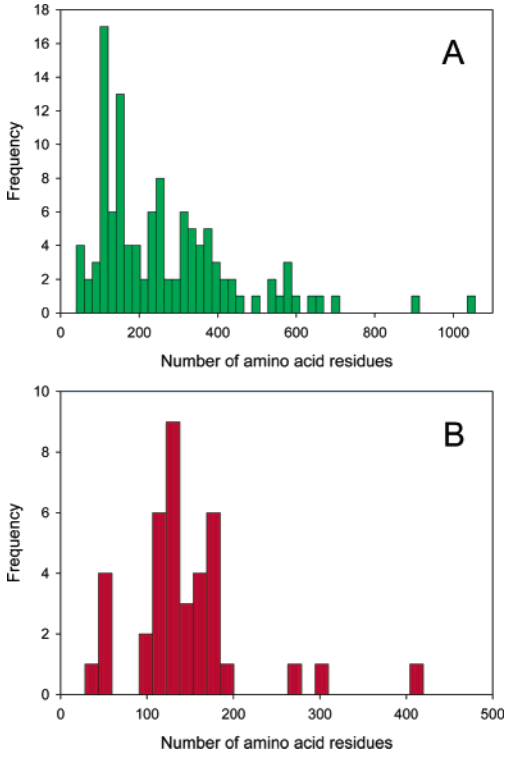

Figure 5. Comparison of the length distribution for the set of 115 proteins able to form equilibrium intermediates (**A**) and the set of 39 proteins shown to unfold without accumulation of partially folded conformations (**B**).

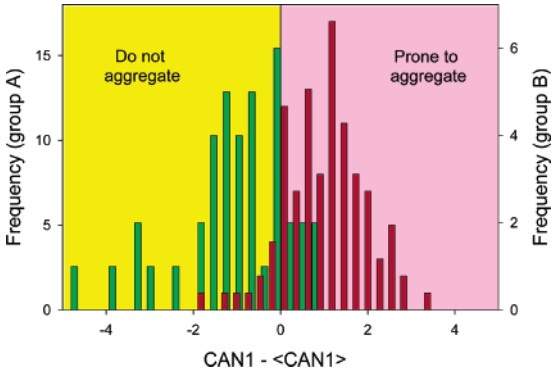

Figure 6. Peculiarities of distribution of charge/hydrophobicity patterning along protein sequences allow the prediction of protein folding mechanism and propensity to aggregate. Analysis revealed that proteins that fold without accumulation of intermediates do not generally aggregate. Green and red bars show two-state folders and proteins, whose folding is accompanied by the formation of partially folded intermediates, respectively. *X*-axis represents distances from the general-mean value of first canonical variable ($\langle CAN1 \rangle = -0.5095$) for the analyzed set of 143 proteins.

Results and Discussion

Prediction of the Effects of Mutations on Aggregation Propensity of AcP. As indicated by Chiti and his associates, an inverse relationship emerged between the aggregation rate of AcP mutant and the protein net charge.[45,51] To determine whether such a relationship was also exhibited by any recur-

rence variable, all 6 RQA variables were entered into a stepwise regression analysis for aggregation rate (Agg) changes between mutants (mut) and wild type (wt) AcP, expressed as $\ln(\nu_{mut}/\nu_{wt})$. TREND was shown to be significant, as well as its related variable LAM ($p = 0.045$, multiple $R = 0.617$). They, however, were somewhat collinear, with TREND explaining more variance. As a result, TREND was chosen to explore the charge/aggregation dependency.

When a general linear model, based only on charge, $|Q|$, and TREND, was applied to aggregation data, a statistically significant interaction term between TREND and $|Q|$ was found ($R = 0.682$), suggesting that a straightforward linear model was inappropriate.[63] Note that we have used a charge value based on a pH of 5.5. This was done in order to standardize the calculations and not make them dependent upon specific experimental conditions. Careful consideration of the relationship between TREND and LAM pointed to the fact that LAM is a modifier of TREND, namely, that the repetitive deterministic patches affect the overall TREND calculation. Thus, one possible formulation of these ideas could be

$$\text{Agg} = \text{Const} + a(|T|^{L*}|Q|) \quad (2)$$

On the other hand, LAM, in its turn, can be further specified by the Trapping Time (TT); i.e., the average length of the laminar segments. Thus, the relationship among RQA variables was formulated as $|T|^{(L^{TT})}$, with LAM being expressed as decimal fraction, and included in the following empirical formula, which conveys the idea of statistical singularity and is also mathematically singular, insofar as it does not contain continuous derivatives

$$\text{Agg} = \text{Const} + a(|T|^{(L^{TT})*}|Q|) \quad (3)$$

Notice that in expressions 2 and 3, a is an adjustable parameter and for both T and Q the absolute values have been taken, without loss of generality. Q is the absolute value of the net charge associated to the protein sequence and was calculated from the total number of positively (Arg, Lys) and negatively (Glu, Asp) charged residues at pH = 5.5 by standard pK_a values. Again, Q was found to be significantly related to the $|T|^{(L^{TT})}$-function via interaction ($R = 0.74$; $p = 0.001$; Figure 2), the reduction of net charge is thus stressed as a crucial factor in determining aggregation propensity. At the next step, the eq 3, which is used to calculate aggregation propensity, was further modified to normalize by total recurrence, *RR*, and to account for the peptide length, *N*, by invoking Coulomb's inverse radius law. This gives rise to a new formula

$$\text{Agg} = \text{Const} + a\left(\frac{|T|^{((L*0.01)^{TT})}}{RR}*\frac{|Q|}{N^2}\right) \quad (4)$$

The reason lamin, *L*, is multiplied by 0.01 is to make the exponent less than 1, thus making the function non-Lipschitz. In fact, this also singles out peptides with values of 0, which tend to be multidomain or large assembly proteins. On the basis of strictly formal arguments, one would assume that any net charge effects would be affected by Coulomb's inverse square law; i.e., the "net" electrostatic effects would not be linear, and are proportional to *1*/length^2. This is not to suggest that this relation is definitive since, as is well-known, molecular electrostatic forces are confounded by other factors, such as, e.g., van der Waals forces; nor to state that there are specific point charge effects. However, in this respect, we were guided by the experience of Plaxco[64] and Finkelstein[65] who revised their observation of contact order being important in protein folding to include protein size/length. This is to say that the "net" effects are screened by distance along the chain. Nonetheless, this may serve as a first approximation of length effect.

All three mentioned above equations (2, 3, and 4) are to be considered empirical relations, directly derived from the model fitting of the effect of mutations on the aggregation propensity of a peculiar system; i.e., AcP. The ability of the proposed formalism to model the discrimination between proteins giving rise to partially folded intermediates from proteins that do not present this kind of behavior is both a proof of the general value of the proposed quantitative model of aggregation tendency and of the involvement of partially folded intermediates in aggregation.

Discrimination of Proteins having Partially Folded Intermediates. Our previous analysis of literature data on equilibrium unfolding of globular proteins induced by changes in pH, temperature, or strong denaturants (urea or guanidinium chloride) revealed that unfolding in 115 proteins is accompanied by accumulation of equilibrium intermediate states of one sort or another. We combined these proteins in a set A. Another set, set B, comprises of 39 proteins, which were shown to unfold according to a simple two-state model; i.e., no equilibrium intermediate of any kind was formed during their unfolding. It is important to emphasize that all the proteins included in the testing sets where collected from literature based solely on the published mechanisms of their equilibrium unfolding, two-state or multi-state, assuming that these subsets (115 proteins in the subset A and 39 in the subset B) faithfully represent the protein universe. Another important point is that both groups include proteins of all major folds (all α, all β, α+β, and α/β), thus, potentially excluding biases related to the type of protein structure. Figure 3 compares the amino acid compositions of the proteins from sets A and B with the averaged composition calculated for the entire Swiss-Prot database (http://au.expasy.org/sprot/relnotes/relstat.html). It can be seen that there is a striking similarity between an averaged Swiss-Prot protein and a protein from set A in their amino acid compositions, whereas proteins from set B seems to be slightly depleted in Ala, Glu, Ile, Leu, Phe, Pro, Thr, and Val and are somehow enriched in Arg, Asn, Asp, Cys, Lys, Trp, and Tyr in comparison with the Swiss-Prot proteins. However, these differences are not of the greatest amplitude and all three distributions look very similar. This observation suggests that sets A and B do not generally have significant biases related to amino acid composition.

The full list of proteins from both groups is present in Table (see Supporting Information), where the proteins effectively used for the analysis are indicated. On the basis of these data, we have established that charge-to-hydrophobicity ratio of a polypeptide chain may represent a key determinant in discrimination proteins known to unfold via the equilibrium intermediates from proteins with two-state unfolding mechanism.[62] In fact, Figure 4 shows that proteins known to form equilibrium partially folded intermediates are specifically localized within a unique region of charge-hydrophobicity space. Thus, the competency of a protein to form equilibrium intermediate(s) may be determined by the bulk content of hydrophobic and charged amino acid residues rather than by the positioning of amino acids within the sequence.[62]

Figure 5 compares the sequence-length distributions for proteins from both groups and shows that proteins known to form equilibrium intermediates are generally larger than

proteins shown to unfold without accumulation of partially folded conformations, with the mean lengths of 270 ± 176 and 145 ± 66 amino acid residues, respectively. Interestingly, the proteins even in the subset A are somewhat shorter than the average sequence length in the Swiss-Prot database,[66,67] which is 368 amino acids (http://au.expasy.org/sprot/relnotes/relstat.html). Thus, long proteins seem to be more prone to form equilibrium partially folded intermediates than short polypeptides. Interestingly, this observation is in a good agreement with a general believe that there is a significant difference in the refolding kinetics of small and large proteins, with simple two-state kinetics being characteristic to smaller proteins and larger proteins having a multi-state folding kinetics (i.e., being characterized by accumulation of kinetic intermediates).[65,68−71] On the other hand, Figure 5 shows that length-distributions for both classes overlap considerably. Furthermore Table (see Supporting Information) shows that the unfolding of a short protein of 50 residues could be accompanied by the accumulation of an intermediate state, whereas a large protein of 412 residues could unfold according to the two-state model. This suggests that the chain length alone, being an important factor, cannot be used to discriminate proteins from the two classes. Although it would seem that the differences in the two groups could be accounted for by a relatively simple difference in the length of their amino acid sequence, analysis of covariance (see below) demonstrated that lengths were not significantly different when CHARGE and FORMULA3 were considered (ENT and MEAN were not significant covariates). This indicates that length of protein itself is not the main issue and other covariates should be considered.

The protein sequences reported in Table (see Supporting Information) were further analyzed by means of RQA descriptors: REC, DET, ENT, TREND, LAM, and TT. The RQA descriptors were supplemented by two simple sequence descriptors: LENGTH (number of residues) and MEAN (average hydrophobicity of the chain) and by the value of the net charge of the protein (CHARGE). The last three descriptors (FORMULA1, FORMULA2, and FORMULA3) correspond to the eqs 2, 3, and 4 described in the previous paragraph.

The entire set of proteins was subdivided into two groups called A and B corresponding to proteins that fold via partially unfolded intermediates and proteins that *unfold without the formation of any intermediate state*. The proteins coded with the above-mentioned descriptors were analyzed by CDA to find (if any) a significant separation between proteins from groups A and B. Before computing the canonical variables, the significance of the separation of the A and B groups for each single variable was assessed by means of univariate analysis of variance. The results of this analysis are reported in Table 2. Simply looking at the Table 2, and keeping in mind the meaning of *F* statistics as the ratio Between groups variance/Within group variance, we can order the relevance of each single element of our multidimensional description of proteins for the discrimination of A and B folding behaviors. It can be seen that average hydrophobicity (MEAN) and size (LENGHT, which can be considered as a proxy for the molecules size) have the greatest discrimination power as single variables.

The determinism of hydrophobicity patterning exerts a limited but nevertheless significant effect as denoted by ENT variable. On the other hand, charge and the interaction between hydrophobicity patterning and charge as described by formula1, formula2, and formula3 shows a strong effect on the separation between the two protein classes. Canonical discriminant analysis unifies the univariate information into a common frame, the general canonical correlation is equal to 0.62 ($p < 0.0001$), and the canonical structure (normalized regression coefficients) of the original variables on the canonical variate are reported in Table 3.

Table 2. Univariate Test Statistics[a]

variable	*F* statistic	p
length	**18.54**	**<0.0001**
mean	**29.23**	**<0.0001**
REC	0.36	0.5463
DET	1.95	0.1642
MAXL	0.10	0.7500
ENT	**3.75**	**0.05**
TREND	0.81	0.3688
LAMIN	0.72	0.3975
TRAPT	1.18	0.2795
charge	**17.73**	**<0.0001**
formula1	**27.12**	**< 0.0001**

[a] The significance of the univariate analysis of each single variable for the descrimination of groups A and B. The table reports the Univariate inferential statistics (one-way Analysis of Variance) relative to the discrimination power of the single variables used in the work. The statistically significant values are bolded. The *F* statistic has the usual meaning of the ratio between inter-group and intra-groups variance, while the *p* value is the probability of the observed result under the null hypothesis of no difference between the groups.

Table 3. Total Canonical Structure of the Original Variables[a]

variable	Can1
length	**0.536712**
mean	**0.616941**
REC	0.037816
DET	0.101336
MAXL	−0.012699
ENT	**0.242842**
TREND	0.040875
LAMIN	−0.057767
TRAPT	0.031590
charge	**−0.504134**
formula3	**−0.589689**

[a] The table reports the canonical coefficients of the different variables. The canonical coefficients are proportional to the relative contribution of the single variables to the canonical variate and thus can be considered as a ranking of the contribution of each descriptor to the discrimination. The significant loadings are in bold.

Table 4. Prediction of Folding Mechanisms Based on Aggregation Algorithm

	predicted class	
observed class	A	B
A	94	10
B	7	32
statistics: $\chi^2 = 71.74$; $p < 0.0001$		

Our analysis revealed that Class A has relatively high values of CAN1 (0.566), whereas Class B corresponds to low values of CAN1 (−1.585). Thus, simply putting class A = CAN1-higher-than-general-mean and class B = CAN1-lower-than-general-mean we can check how many correct predictions on the analyzed data set our algorithm can attain. Overall, this model catches 94 correct predictions out of 104 (90.4%) for the group A and 32 correct predictions out of 39 (82.1%) for the group B (see Table 4 and Figure 6). It is worth noting that our model,[44] as well as the Chiti group analyses,[46−48] predicts zero net charge as a singularity of the aggregation formulas characterized by a very strong propensity to aggregate. Thus, we were forced to

not include these zero-charged proteins to the analysis for the need to avoid singularities. However, if we simply consider, according to the model, these zero-net-charge proteins as nontwo step folders, we increase the accuracy of our model to predict the folding behavior of the proposed proteins to 105 correct predictions out of 115, corresponding to 91.3% for the set A. On the other hand, it would be misleading to cite 91.3% as an overall accuracy for the prediction, since this value is based on the true positives alone on one set only. The analysis of data presented in Table 4 shows that our model gives 8.7% false positive predictions for the proteins whose unfolding is accompanied by the accumulation of equilibrium intermediate states (subset A) (10 of 115 proteins in the subset A are predicted to fold without a partially folded intermediate when they are nontwo state folders); 17.9% false negative predictions for the subset B (7 of 39 proteins in set B are predicted to fold with an intermediate, whereas they are two state folders); and 11.0% for the entire set (overall, 17 of 154 proteins in the combined A+B set are not properly predicted). A better performance of our model on set A in comparison with set B could be explained by some uncertainties associated with the creation of set B. In fact, we have used solely the criteria of published unfolding mechanism to put a given protein into a given set. Obviously, set A is more certain, as to be added to this set, a protein should only be able to form an equilibrium partially folded intermediate at *any* experimental conditions studied. The request for being classified as a member of set B is much more restricted, proteins should not form a partially folded intermediate at *any* experimental conditions. The conclusion on whether a given protein has or has not an equilibrium intermediate is based on the analysis of multiple unfolding data from several biophysical techniques, with the noncoincidence of unfolding curves detected by different methods being considered as an indication of an intermediate accumulation.[72] However, in some cases, the unfolding curves are considerably overlapped, making the assignment of proteins to one or another set questionable. The other source of uncertainties for the set B creation might be related to the incomplete knowledge of a conformational space available for some proteins. In fact, if a protein was shown to unfold according to the two-state mechanism at some conditions (i.e., pH or urea), it does not necessarily mean that it will be unable to form an intermediate at other conditions. All this shows that the further work is absolutely necessary in the direction of further improvement of the training sets.

Importantly, Figure 6 clearly shows that two-state folders can be discriminated from proteins whose folding is accompanied by the accumulation of partially folded intermediates based solely on their differences in predisposition to aggregate, which is encoded in charge/hydrophobicity distribution patterning along their sequences. This justifies the link between aggregation propensity (note, FORMULA1, FORMULA2, and FORMULA3 came from the modeling of aggregation of AcP and they were inserted by the statistical program into the canonical variate) and the presence of partially folded intermediates. On a general ground, it is necessary to emphasize that the charge descriptor enters the canonical variate structure with a negative coefficient. This means that a high value of canonical variate (which correspond to pertaining to the A group of proteins forming partially folded intermediates) is favored by a decrease in the net charge. This is consistent with the effect of charge on aggregation, pointing to the overlapping between these two phenomena.

At the next step, we have submitted the data set to a partial correlation analysis in order to check the single contributes of length, mean hydrophobicity, charge and formula to canonical function. To this end, the correlation coefficient of each single significant descriptor with the first canonical variate (representing the best least-squares collective model of folding behavior) was partialled out of the influence of the others and checked for the presence of a specific contribute, not dependent with the linear interaction with another descriptor. We have established that LENGTH and CHARGE were the only two descriptors for which a relevant cross-talking was detected: length contribution (as expressed by correlation coefficient with CAN1) when partialled out of charge effect dropped from around 0.50 to 0.30 and the same was observed for the reversed analysis (correlation coefficient of charge partialled out of length contribution). On the contrary, both FORMULA3 and MEAN (hydrophobicity) were demonstrated to exert a peculiar, genuine effect on discrimination, not mediated by confounding with other elements of the discrimination function. This is particularly interesting for FORMULA3 that incorporates both hydrophobicity and charge into its formulation, but these two descriptors are incorporated in a nonlinear manner that prevents any linear confounding with these two descriptors. In other words, it looks like that the parameters studied are intertwined (given their nonlinear nature) in such a way that a simple, straightforward explanation of their relationship is not easily understood at this time. However, overall results indicate a specific role for our formula in folding behavior discrimination independent of protein length/size considerations.

We have previously suggested that hydrophobicity segments broken by laminar patches may tend to be disordered and exhibit more conformational variability (flexibility), thus having tendency to avoid aggregation. An explanation for this increased flexibility relates to the fact that $|T|^{(L^{TT})}$ quantifies the differential density of patches.[73] The motivation for this hypothesis was suggested by our earlier work in the analysis of rubredoxins.[74] In this study we used RQA to help understand factors differentiating the function of thermophilic vs mesophilic forms. An important finding was that in the *Rubr. clopa* (mesophilic) case, the concentration of deterministic patches occurred in unequally distributed areas, whereas in the *Rubr. pyrfu* case (thermophilic), there is no preferentially populated area and the concentration of deterministic patches is distributed over the whole backbone. It was hypothesized that this is at least one cause for the increased flexibility of the thermophilic rubredoxin over the mesophilic counterpart.[75]

Other putative causes involve the observation that amyloidogenic propensity is associated with a defect in hydrogen bonding exposed to water, making them "sticky".[76−78] In fact, it has been recently emphasized that a high density of backbone hydrogen bonds exposed to water attack in monomeric structure (i.e., increased amount of defects of hydrogen bond "wrapping") represents a structural characteristic indicating amyloidogenic propensity of a protein retaining some of its native structure.[84] On the basis of these observations, a diagnostic tool based on the identification of hydrogen bonds with a paucity of intramolecular dehydration or "wrapping" has been proposed.[84] Clearly, if a folding intermediate is enriched in such "wrapping" defects, then it will be more prone to aggregate than a soluble protein with stable structure. Thus, it may be that the singular functions address the amount of molecular "patchiness" which may be an inverse indicator of hydrophobic cores. Another view suggests that biopolymers

may develop instability and collapse due to soliton-like nonlinear excitations at bends, or patches.[79] Previously,[80] we have suggested that such instabilities may occur in the form of molecular motions not associated with analysis of traditional modes.

Another possible explanation is the difference between the Chiti et al. model[46−48] with ours is their inclusion of the free energy changes based on β-sheet and coil propensities. It is possible that our quantification of patches of laminarity may characterize a similar phenomenon. β-Sheets and coils in some sense typify a kind of "patch." In our studies, we have noted a correlation, which is, however, is not a perfect one. We are currently pursuing additional investigation into this area.

The charge effect in such an explanation takes on a more complex role than that of an indicator of general repulsion between molecules. This is to say that if the patches are sequestered unequally along the series, the inequality may set up a "screening" effect for net charge: the non patchy areas may be related to "blocks" with contrasting solubilities which can, depending upon their size, modify the net charge effect. Given a change in pH, which alters a charge, a protein's probability to aggregate may become enhanced. This is consonant with recent results obtained by Burke et al. with huntingtin-exon 1.[81] Thus, one may speculate that the deterministic patches constitute a static factor involved in folding; whereas the net charge effect is a "dynamic" component often modulated by circumstantial factors such as pH. It might be of significance to understand the customary milieus of proteins: environments which expose proteins to different pHs may carry a greater likelihood of aggregation as opposed to those which perform their work in relatively circumscribed settings (e.g., proteins which are predominantly found in the nucleus).

Interestingly, Papoian et al. have recently re-emphasized the importance of water mediated long-range interactions between hydrophilic surfaces for protein folding and binding: long-range water-mediated pairing of hydrophilic groups, being an integral part of protein architecture and representing a universal feature of biomolecular recognition landscapes in both folding and binding, might guide protein folding, smooth the underlying folding funnels and facilitate nativelike packing of supersecondary structural elements.[82] This means that in addition to the hydrophobic interactions, known to drive the folding of proteins into compact globular structures, long-range hydrophilic interactions also have to be taken into account since their role in the guiding the processes of protein folding and binding. Thus, the aqueous environment has a more active role in protein dynamics and stability than what it was traditionally imagined and may have many applications.[83] It is necessary to emphasize that these observations are in a good agreement with our findings. In fact, we have established that charge of a polypeptide chain represents one of the strongest discriminators between two sets analyzed, two-state folders and proteins that unfold via the partially folded intermediate, with two-state folders being enriched in charged amino acid residues. On the other hand, Papoian et al. clearly showed that the long-range water-mediated hydrophilic interactions are able to smooth the folding funnels and significantly stabilize the nativelike structure, as the enthalpy gain from water-mediated contacts is greater than the entropic cost that must be paid for immobilizing interfacial water.[82] In other words, such long-range water-mediated hydrophilic interactions might help polypeptide chain to eliminate some local traps, partially folded intermediates. One can assume that the larger number of polar groups (i.e., the larger number of potential long-range water-mediated hydrophilic contacts), the lesser chance to form partially folded intermediates would be, which was in fact observed for the proteins from group B.

Conclusion

The 'partially folded aggregates' and "interaction between hydrophobicity patterning and charge" theories of protein aggregation were demonstrated to be the two faces of the same coin. This unification was carried out in a purely data-driven manner by demonstrating the statistical significance in the discrimination between 'two-steps folders' and 'continuous folders' of the same set of sequence descriptors of hydrophobicity and charge modification that were demonstrated useful in the prediction of the aggregation behavior of a set of AcP mutants. If a model built for quantifying the mutual interaction between hydrophobicity and charge in the definition of the aggregation propensity of a specific protein system works in the discrimination between different folding mechanisms, then we can safely state both the relevance of different folding mechanisms in the aggregation propensity and the importance of a fine-tuning between hydrophobicity patterning and charge in the protein folding.

Acknowledgment. This work was supported by a joint DMS/NIGMS initiative to support mathematical biology, from the National Science Foundation and National Institutes of Health, (NSF DMS #0240230); J. P. Zbilut, Principal Investigator

Supporting Information Available: A table contains the full list of proteins used in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kelly, J. W. The alternative conformations of amyloidogenic proteins and their multistep assembly pathways, *Curr. Opin. Struct. Biol.* **1998**, *8*, 101−106.
(2) Bellotti, V.; Mangione, P.; Stoppini, M. Biological activity and pathological implications of misfolded proteins, *Cell. Mol. Life Sci.* **1999**, *55*, 977−991.
(3) Dobson, C. M. Protein misfolding, evolution and disease, *Trends Biochem. Sci.* **1999**, *24*, 329−332.
(4) Uversky, V. N.; Talapatra, A.; Gillespie, J. R.; Fink, A. L. Protein deposits as the molecular basis of amyloidosis. I. Systemic amyloidoses, *Med. Sci. Monitor* **1999**, *5*, 1001−1012.
(5) Uversky, V. N.; Talapatra, A.; Gillespie, J. R.; Fink, A. L. Protein deposits as the molecular basis of amyloidosis. II. Localized amyloidosis and neurodegenerative disordres, *Med. Sci. Monitor* **1999**, *5*, 1238−1254.
(6) Rochet, J. C.; Lansbury, P. T., Jr. Amyloid fibrillogenesis: themes and variations, *Curr. Opin. Struct. Biol.* **2000**, *10*, 60−68.
(7) Uversky, V. N.; Fink, A. L. Conformational constraints for amyloid fibrillation: The importance of being unfolded. *Biochim. Biophys. Acta* **2004**, *1698*, 131−153.
(8) Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: reassessing the protein structure−function paradigm, *J. Mol. Biol.* **1999**, *293*, 321−331.
(9) Uversky, V. N.; Gillespie, J. R.; Fink, A. L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415−427.
(10) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z. Intrinsically disordered protein, *J. Mol. Graph. Model.* **2001**, *19*, 26−59.
(11) Uversky, V. N. What does it mean to be natively unfolded? *Eur. J. Biochem.* **2002**, *269*, 2−12.
(12) Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **2002**, *11*, 739−756.

(13) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry* **2002**, *41*, 6573−6582.
(14) Dunker, A. K.; Brown, C. J.; Obradovic, Z. Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* **2002**, *62*, 25−49.
(15) Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527−533.
(16) Dyson, H. J.; Wright, P. E. Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol.* **2002**, *12*, 54−60.
(17) Uversky, V. N. Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell Mol. Life Sci.* **2003**, *60*, 1852−1871.
(18) Sunde, M.; Serpell, L. C.; Bartlam, M.; Fraser, P. E.; Pepys, M. B.; Blake, C. C. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* **1997**, *273*, 729−739.
(19) Dobson, C. M. The structural basis of protein folding and its links with human disease. *Philos. Trans. R. Soc. Lond B Biol. Sci.* **2001**, *356*, 133−145.
(20) Fandrich, M.; Fletcher, M. A.; Dobson, C. M. Amyloid fibrils from muscle myoglobin. *Nature* **2001**, *410*, 165−166.
(21) Pertinhez, T. A.; Bouchard, M.; Tomlinson, E. J.; Wain, R.; Ferguson, S. J.; Dobson, C. M.; Smith, L. J. Amyloid fibril formation by a helical cytochrome. *FEBS Lett.* **2001**, *495*, 184−186.
(22) Goers, J.; Permyakov, S. E.; Permyakov, E. A.; Uversky, V. N.; Fink, A. L. Conformational prerequisites for alpha-lactalbumin fibrillation. *Biochemistry* **2002**, *41*, 12546−12551.
(23) Munishkina, L. A.; Fink, A. L.; Uversky, V. N. Formation of amyloid fibrils from the core histones in vitro. *J. Mol. Biol.* **2004**, in press.
(24) Marston, F. A. O. The purification of eukaryotic polypeptides synthesized in *Escherichia coli*. *Biochem. J.* **1986**, *240*, 1−12.
(25) Schein, C. H. Solubility as a function of protein structure and solvent components. *Biotechnology* **1990**, *8*, 308−315.
(26) Fink, A. L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Folding & Design* **1998**, *3*, 9−15.
(27) Wetzel, R.; Chrunyk, B. A. Inclusion body formation by interleukin-1 beta depends on the thermal sensitivity of a folding intermediate. *FEBS Lett.* **1994**, *350*, 245−248.
(28) Wetzel, R. In. *Stability of Protein Pharmaceuticals, Part B.; In Vivo Pathways of Degradation and Strategies for Protein Stabilization*; Ahern, T. J., Manning, M. C., Eds.; Plenum Press: New York, 1992; pp 43−88.
(29) Mitraki, A.; King, J. Protein folding intermediates and inclusion body formation, *Biotechnology* **1989**, *7*, 690−697.
(30) Lougheed, W. D.; Woulfe-Flanagan, H.; Clement, J. R.; Albisser, A. M. Insulin aggregation in artificial delivery systems. *Diabetologia* **1980**, *19*, 1−9.
(31) Tanford, C. Protein denaturation, *Adv. Protein Chem.* **1968**, *23*, 121−282.
(32) London, J.; Skrzynia, C.; Goldberg, M. E. Renaturation of *Escherichia coli* tryptophanase after exposure to 8 M urea. Evidence for the existence of nucleation centers. *Eur. J. Biochem.* **1974**, *47*, 409−15.
(33) Clark, A. H.; Saunderson, D. H.; Suggett, A. Infrared and laser-Raman spectroscopic studies of thermally induced globular protein gels, *Int. J. Pept. Protein Res.* **1981**, *17*, 353−364.
(34) DeYoung, L. R.; Fink, A. L.; Dill, K. A. Aggregation of globular proteins, *Acc. Chem. Res.* **1993**, *26*, 614−620.
(35) DeYoung, L. R.; Dill, K. A.; Fink, A. L. Aggregation and denaturation of apomyoglobin in aqueous urea solutions. *Biochemistry* **1993**, *32*, 3877−3886.
(36) Eliezer, D.; Chiba, K.; Tsuruta, H.; Doniach, S.; Hodgson, K. O.; Kihara, H. Evidence of an associative intermediate on the myoglobin refolding pathway. *Biophys. J.* **1993**, *65*, 912−7.
(37) Fink, A. L. Molten globules, *Methods Mol. Biol.* **1995**, *40*, 343−60.
(38) Ptitsyn, O. B. Molten globule and protein folding. *Adv. Protein Chem.* **1995**, *47*, 83−229.
(39) Uversky, V. N.; Segel, D. J.; Doniach, S.; Fink, A. L. Association-induced folding of globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5480−5483.
(40) Kuznetsova, I. M.; Biktashev, A. G.; Khaitlina, S. Y.; Vassilenko, K. S.; Turoverov, K. K.; Uversky, V. N. Effect of self-association on the structural organization of partially folded proteins: inactivated actin. *Biophys. J.* **1999**, *77*, 2788−2800.
(41) Uversky, V. N.; Karnoup, A. S.; Khurana, R.; Segel, D. J.; Doniach, S.; Fink, A. L. Association of partially folded intermediates of staphylococcal nuclease induces structure and stability. *Protein Sci.* **1999**, *8*, 161−173.
(42) Bushmarina, N. A.; Kuznetsova, I. M.; Biktashev, A. G.; Turoverov, K. K.; Uversky, V. N. Partially folded conformations in the folding pathway of bovine carbonic anhydrase II: A fluorescence spectroscopic analysis, *Chembiochem*. **2001**, *2*, 813−821.
(43) Kuznetsova, I. M.; Stepanenko, O. V.; Turoverov, K. K.; Zhu, L.; Zhou, J. M.; Fink, A. L.; Uversky, V. N. Unraveling multistate unfolding of rabbit muscle creatine kinase, *Biochim. Biophys. Acta* **2002**, *1596*, 138−155.
(44) Zbilut, J. P.; Colosimo, A.; Conti, F.; Colafranceschi, M.; Manetti, C.; Valerio, M.; Webber, C. L., Jr.; Giuliani, A. Protein aggregation/folding: the role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and Abeta(1−40), *Biophys J.* **2003**, *85*, 3544−3557.
(45) Chiti, F.; Taddei, N.; White, P. M.; Bucciantini, M.; Magherini, F.; Stefani, M.; Dobson, C. M. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding, *Nat. Struct. Biol.* **1999**, *6*, 1005−1009.
(46) Calloni, G.; Taddei, N.; Plaxco, K. W.; Ramponi, G.; Stefani, M.; Chiti, F. Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding, *J Mol. Biol.* **2003**, *330*, 577−591.
(47) Chiti, F.; Stefani, M.; Taddei, N.; Ramponi, G.; Dobson, C. M. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* **2003**, *424*, 805−808.
(48) Calamai, M.; Taddei, N.; Stefani, M.; Ramponi, G.; Chiti, F. Relative influence of hydrophobicity and net charge in the aggregation of two homologous proteins. *Biochemistry* **2003**, *42*, 15078−15083.
(49) Lansbury, P. T., Jr. Evolution of amyloid: what normal protein folding may tell us about fibrillogenesis and disease. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 3342−3344.
(50) Zerovnik, E. Amyloid-fibril formation. Proposed mechanisms and relevance to conformational disease. *Eur. J. Biochem.* **2002**, *269*, 3362−3371.
(51) Chiti, F.; Calamai, M.; Taddei, N.; Stefani, M.; Ramponi, G.; Dobson, C. M. Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases, *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16419−16426.
(52) Eckmann, J. P.; Kamphorst, S. O.; Ruelle, D. Recurrence plots of dynamic-systems. *Europhys. Lett.* **1987**, *4*, 973−977.
(53) Zbilut, J. P.; Webber, C. L. Embeddings and delays as derived from quantification of recurrence plots. *Phys. Lett. A* **1992**, *171*, 199−203.
(54) Webber, C. L.; Zbilut, J. P. Dynamical assessment of physiological systems and states using recurrence plot strategies. *J. Appl. Physiol.* **1994**, *76*, 965−973.
(55) Marwan, N.; Thiel, M.; Nowaczyk, N. R. Cross recurrence plot based synchronization of time series. *Nonlin. Proc. Geophys.* **2002**, *9*, 325−331.
(56) Marwan, N.; Wessel, N.; Meyerfeldt, U.; Schirdewan, A.; Kurths, J. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data, *Phys. Rev. E* **2002**, *66*, 026702−1−026702−8.
(57) Miyazawa, S.; Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal-structures−Quasi-chemical approximation. *Macromolecules* **1985**, *18*, 534−552.
(58) Strait, B. J.; Dewey, T. G. The Shannon information entropy of protein sequences. *Biophys. J.* **1996**, *71*, 148−155.
(59) Weiss, O.; Jimenez-Montano, M. A.; Herzel, H. Information content of protein sequences. *J. Theor. Biol.* **2000**, *206*, 379−386.
(60) Giuliani, A.; Benigni, R.; Zbilut, J. P.; Webber, C. L.; Sirabella, P.; Colosimo, A. Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. *Chem. Rev.* **2002**, *102*, 1471−1491.
(61) Zbilut, J. P.; Sirabella, P.; Giuliani, A.; Manetti, C.; Colosimo, A.; Webber, C. L. Review of nonlinear analysis of proteins through recurrence quantification, *Cell Biochem. Biophys.* **2002**, *36*, 67−87.
(62) Uversky, V. N. Cracking the folding code. Why do some proteins adopt partially folded conformations, whereas other do not? *FEBS Lett.* **2002**, *514*, 181−183.
(63) Kleinbaum, D. G.; Kupper, L. L. *Applied regression analysis and other multivariable methods*; Duxbury Press: Boston, 1978.
(64) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277*, 985−994.
(65) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* **2003**, *12*, 2057−2062.

(66) Appel, R. D.; Bairoch, A.; Hochstrasser, D. F. A new-generation of information-retrieval tools for biologists—the example of the Expasy WWW server. *Trends Biochem. Sci.* **1994**, *19*, 258−260.
(67) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **1994**, *27*, 49−54.
(68) Jackson, S. E. How do small single-domain proteins fold?, *Folding & Design* **1998**, *3*, R81−R91.
(69) Fersht, A. R. Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism, *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1525−1529.
(70) Eaton, W. A.; Munoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys Biomol. Struct.* **2000**, *29*, 327−359.
(71) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* **2003**, *51*, 162−166.
(72) Uversky, V. N. A multiparametric approach to studies of self-organization of globular proteins. *Biochemistry (Mosc.)* **1999**, *64*, 250−266.
(73) Zbilut, J. P.; Mitchell, J. C.; Giuliani, A.; Colosimo, A.; Webber, C. L. Singular hydrophobicty patterns and net charge: a mesoscopic principle for protein aggregation/folding. *Physica A* **2004**, in press.
(74) Giuliani, A.; Benigni, R.; Sirabella, P.; Zbilut, J. P.; Colosimo, A. Nonlinear methods in the analysis of protein sequences: a case study in rubredoxins. *Biophys J.* **2000**, *78*, 136−149.
(75) Grottesi, A.; Ceruso, M. A.; Colosimo, A.; Di Nola, A. Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin, *Proteins* **2002**, *46*, 287−294.
(76) Fernandez, A.; Scheraga, H. A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions, *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 113−118.
(77) Fernandez, A.; Berry, R. S. Proteins with H-bond packing defects are highly interactive with lipid bilayers: Implications for amyloidogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2391−2396.
(78) Fernandez, A.; Scott, R. Dehydron: A structurally encoded signal for protein interaction. *Biophys. J.* **2003**, *85*, 1914−1928.
(79) Mingaleev, S. F.; Gaididei, Y. B.; Christiansen, P. L.; Kivshar, Y. S. Nonlinearity-induced conformational instability and dynamics of biopolymers. *Europhys. Lett.* **2002**, *59*, 403−409.
(80) Manetti, C.; Giuliani, A.; Ceruso, M. A.; Webber, C. L.; Zbilut, J. P. Recurrence analysis of hydration effects on nonlinear protein dynamics: multiplicative scaling and additive processes. *Phys. Lett. A* **2001**, *281*, 317−323.
(81) Burke, M. G.; Woscholski, R.; Yaliraki, S. N. Differential hydrophobicity drives self-assembly in Huntington's disease. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13928−13933.
(82) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. Water in protein structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3352−3357.
(83) Levy, Y.; Onuchic, J. N. Water and proteins: A love-hate relationship. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3325−3326.
(84) Fernandez, A.; Kardos, J.; Scott, L. R.; Goto, Y.; Berry, R. S. Structural defects and the diagnosis of amyloidogenic propensity. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6446−6451.
(85) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

PR049883+