Lecture

AGGREGATION PROPENSITY OF PROTEINS QUANTIFIED BY HYDROPHOBICITY PATTERNS AND NET CHARGE

Joseph P. Zbilut (a), Julie C. Mitchell (b), Alessandro Giuliani (c), Alfredo Colosimo (d), Norbert Marwan (e), Mauro Colafranceschi (c, d), Charles L. Webber, Jr. (f)
(a) Department of Molecular Biophysics and Physiology, Rush University Medical Center, Chicago, IL USA
(b) Departments of Mathematics and Biochemistry, University of Wisconsin-Madison, Madison, USA
(c) Dipartimento di Ambiente e Connessa Prevenzione Primaria, Istituto Superiore di Sanitá, Rome, Italy
(d) Department of Human Physiology and Pharmacology, University of Rome "La Sapienza", Rome, Italy
(e) Nonlinear Dynamics Group, Institute of Physics, University of Potsdam, Potsdam, Germany
(f) Department of Physiology, Loyola University Medical Center, Maywood, IL USA

Introduction

It has been well appreciated that the native state fold of proteins is in some way dependent upon the physico-chemical properties of their amino acid sequence, most notably, hydrophobicity (1-3). More recently it has been recognized that the actual folding process is of a stochastic nature, and also includes the possibility of forming aggregates that ultimately can be physiologically harmful. A growing body of evidence suggests that this involves partially or completely unfolded proteins (4). Yet, what factors specifically promote the formation of aggregates as opposed to native folds under relatively normal conditions remain undecided.

Recently, we have proposed that some key features of protein hydrophobicity patterns analyzed by a nonlinear signal processing technique, recurrence quantification analysis (RQA), provide some necessary conditions for aggregation. A significant finding included a correspondence between short deterministic patches of hydrophobicity distribution along the sequence, what we term *laminarity,* [LAM, (*L*)], 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" index of proteins (5). Beyond this, a counterpoint was defined as the degree of laminarity. Specifically, in an analysis of the protein engineering experiments of Chiti *et al*. (6) it was shown that aggregation sensitive zones vs. folding sensitive zones were distinguished by the two complementary concepts of trend/laminarity (7). The implication is that these areas may be inherently unstable, and somehow involved in the promotion of (at least) partial unfolding and aggregation. Indeed, it has been shown that regions tend to be involved in binding/folding events (8). The degree at which these conditions exist probabilistically determines the propensity for aggregation. What was not determined is the effect of total charge on the probabilities. Additionally, we suggested that these features may be related mathematically to provide some correlation with aggregation behavior.

In a follow-up study, Chiti, *et al*. (9) addressed the question of charge *a propos* of aggregation again using AcP mutants which minimally affected hydrophobicity, α-helical and β-sheet propensities. Thus, using selected mutations and the results of their previous work, these authors came to the conclusions that increase of net charge (not solely local) aids in avoiding aggregation. In our analysis, we confirm, in part, their observations and make further distinctions on the basis

of: 1) RQA variables and 2) a quantitative model of aggregation propensity, taking into account hydrophobicity patterns and charge. Surprisingly, the model, even if inspired by a specific problem (effects of mutations on aggregation propensity of AcP) is correlated significantly with aggregation propensities in a diverse set of proteins, suggesting a phenomenological index. Moreover, the formula predicting the effect of mutations on aggregation propensity, is also able to locate, at the extreme of a statistical distribution, all the proteins giving rise to highly structured DNA-protein assemblies (histones) and RNA binding proteins present in a large set of 1141 proteins randomly selected from the SWISS-PROT data base.

Materials and methods

Data sets

The data which inspired our model are in a seminal paper by Chiti *et al*. (6) (hereafter referred to as C1), and concern the effect of different mutations on acylphosphatase (AcP) aggregation propensity. A second Chiti *et al.* (9) data set (C2) dealt with the effects of charge. The model was also applied to a set of diverse peptides and proteins whose aggregation propensity was known by Chiti *et al*. (10) (C3). This set was composed of data from the literature specifying aggregation rates by different techniques, and were normalized by wild type aggregation values to allow for comparative analyses.

The protein population used as a test for the aggregation formula included 1141 proteins randomly chosen from the SWISS-PROT repository in order to avoid any selection bias (11) and constituting a representative sample of all known eukaryotic sequences. These proteins are available at ftp://ftp.ebi.ac.uk/pub/contrib/swissprot/testsets/signal. We utilized a subset made of eukaryotic proteins that are not secreted and thus do not have the bias of an N-terminal signal peptide, potentially imposing an externally driven correlation of protein sequences.

Recurrence plots

Eckmann *et al*. (12) introduced a tool which can visualize the recurrence of states $\mathbf{x}_i$ in a phase space. Usually, a phase space has a higher dimension than can be readily visualized. 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 sequence 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 squared matrix with black and white dots, where black dots mark a recurrence, and both axes are the ordered sequences. This representation is called recurrence plot (RP). Such an RP can be mathematically expressed as:

$$\mathbf{R}_{i,j} = \Theta\,(\varepsilon - \|\,\mathbf{x}_i - \mathbf{x}_j\|),\ \mathbf{x}_i \in \Re^m,\ i,j=1,\ldots,N \qquad [1]$$

where $\mathbf{R}_{i,j}$ is the recurrence, N is the number of considered states $\mathbf{x}_i$, ε is a threshold distance, m the embedding dimension, $\|\cdot\|$ a norm and $\Theta(\cdot)$ the Heaviside function. (The threshold distance, ε, determines if a given point is considered recurrent).

The initial purpose of RPs was the visual inspection of higher dimensional correlations. The advantage of RPs is that they can also be applied to rather short, nonlinear, and even nonstationary data. The RPs exhibit characteristic large scale and small scale patterns. The former patterns were denoted by Eckmann *et al*. (12) as *typology* and the latter as *texture*. The typology offers a global impression which can be characterized as *homogeneous*, *periodic*, *drift* and *disrupted*.

Recurrence quantifiers in recurrence plots

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 combinations of vertical and horizontal lines form rectangular clusters of recurrence points). 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.
- 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. The direction of these diagonal structures is parallel to the Line Of Identity (LOI), represented by the main diagonal in RPs, indicating the parallel running of sequences for the same evolution.
- A *vertical (horizontal) line* $\mathbf{R}_{i,j+k} = 1$ (for $k = 1, \ldots, v$, 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.

Because visual inspection is difficult, and dependent upon the resolution of the output device (monitor/printer), a quantification of recurrence plots was developed by Zbilut and Webber (13, 14) and extended with new measures of complexity by Marwan *et al*. (15). In the original definition of the RPs, the neighborhood is a ball (i.e. L_2-norm is used) and its radius is chosen in such a way that it contains a fixed amount of states $\mathbf{x}_j$ (12). (The original Eckmann *et al*. algorithm used a nearest neighbor method to choose the recurrences and the resultant RP was not symmetrical.) The most commonly used neighborhood is that with a fixed radius ε. For RPs this neighborhood was first used by Zbilut and Webber (14). A fixed radius means that $\mathbf{R}_{i,j} = \mathbf{R}_{j,i}$, resulting in a symmetric RP (see Table 1 for a formulation of the variables).

These measures can be computed in windows along the main diagonal, which allows for a study of their spatial dependence, and can be used to detect state transitions. Another possibility is to define these measures for each diagonal parallel to the main diagonal separately (16). Windowed versions are also available.

Recurrence in protein sequences

The numerical series studied in this work are protein sequences coded by the hydrophobicity of the constituent residues. Discrete time and spatial series (like non branching polymers) are completely congruent mathematical objects, given they are both linear arrangements of discrete subsequent elements with a fixed and well defined ordering. Switching from time series to protein primary structures, the dynamical concept of "state" corresponds to a patch of consecutive monomeric units of length equal to the embedding dimension.

Each protein sequence was coded by means of the Miyazawa-Jernigan hydrophobicity scale (MJ) of aminoacid residues (17). This scale corresponds to the first eigenvalue of the contact energy matrix as reported at: http://us.expasy.org/tools/pscale/Hphob.Miyazawa.html, a choice dictated by our previous analysis of a 1141 random sample of protein sequences from the SWISS-PROT Database (see above). 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.

Table 1. RQA Measures

Measure	Definition
Recurrence, *REC*	Percentage of recurrence points in an RP: $REC = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}$
Determinism, *DET*	Percentage of recurrence points which form diagonal lines: $DET = \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\sum_{i,j}^{N} R_{i,j}}$ P(l) is the histogram of the lengths l of the diagonal lines.
Laminarity, *LAM*	Percentage of recurrence points which form vertical lines: $LAM = \frac{\sum_{\upsilon=\upsilon_{\min}}^{N} \upsilon P(\upsilon)}{\sum_{\upsilon=1}^{N} \upsilon P(\upsilon)}$ P(v) is the histogram of the lengths v of the vertical lines.
Trapping time, *TT*	Average length of vertical lines: $TT = \frac{\sum_{\upsilon=\upsilon_{\min}}^{N} \upsilon P(\upsilon)}{\sum_{\upsilon=\upsilon_{\min}}^{N} P(\upsilon)}$
Longest diagonal line, L_{max}	Length of the longest diagonal line: $L_{\max} = \max\left(\{l_i; i = 1,\ldots,N_l\}\right)$
Longest vertical line, V_{max}	Length of longest vertical line: $V_{\max} = \max\left(\{\upsilon_l; l = 1,\ldots,L\}\right)$
Divergence, *DIV*	Inverse of L_{max} : $DIV = \frac{1}{L_{\max}}$ Related to the largest positive Lyapunov exponent, but does not correspond to it.
Entropy, *ENT*	Shannon entropy of the distribution of the diagonal line lengths p(l): $ENT = -\sum_{l=l_{\min}}^{N} p(l) \ln p(l)$
Trend, *TREND*	Paling of the RP towards its edges: $TREND = \frac{\sum_{i=1}^{N-2} [i - (N-2)](REC_i - \langle REC_i \rangle)}{\sum_{i=1}^{N-2} [i = (N-2)/2]^2}$

The application of RQA implies an *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 (at embedding 3) as well as our previous analyses, the above parameters were set to: embedding dimension = 3; radius = 6**,** and line = 2. The radius was determined by finding the shelf singularity of hydrophobicity as defined by the RQA variable DET (17-21). Because of the size of the SWISS-PROT Database, the radius was set to obtain approximately 1% REC values, which, in practice, ensures achievement of the singularity.

Results

Empirical refinements and model validation

In order to develop a general empirical model of aggregation we first focused attention on the data of C2 which was used to evaluate the effect of net charge on AcP folding. Our aim was to determine a simple functional relationship between RQA variables, charge (*Q*), and aggregation indexes.

As indicated by Chiti *et al*. (9), an inverse relationship emerged between aggregation rate and net charge. To determine whether such a relationship was also exhibited by any recurrence variable, six RQA variables (REC, DET, ENT, MAXL, TREND and LAM) 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 the most significant and, moreover, TREND explained the majority of variance ($p = 0.000001$). As a result, TREND was chosen to explore the charge/aggregation dependency.

When a general linear model, based only upon charge and TREND, was applied to aggregation data, a statistically significant interaction term between TREND and *Q* was found ($r = 0.802$, $p = 0.0002$) suggesting that a straightforward linear model was inappropriate (16). Careful consideration of the relationship between TREND and LAM point to the fact that LAM is a modifier of TREND, namely that repetitive deterministic patches affect the overall TREND calculation. This view is also supported by the finding that short deterministic patches, termed "singular," are an important factor in protein folding. Thus, one possible formulation of these considerations could be:

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

LAM, in turn, can be further specified by the Trapping Time (TT), i.e. the average length of the laminar segments as a "weighting" term. Thus, the relationship among RQA variables was formulated as a tower function, *|T|^(L^TT)*, with LAM being expressed as decimal fraction. This function, which conveys the idea of statistical "singularity" and is also mathematically singular, insofar as it does not admit continuous derivatives near T = 0, was included in the following empirical formula:

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

The singularity near $T = 0$ is particularly bad, as $0 < L^{TT} < 1$ implies that the derivative blows up near this point, rather than simply being undefined. Notice that in expressions [2] and [3] *a* is an adjustable parameter and *T* the absolute value is taken, without loss of generality. *Q* is the value of the net charge associated with the protein sequence and was calculated from the total number of positively and negatively charged residues at neutral pH. This was done in order to avoid presumptions of e.g., lower pH as might be obtained in specific experimental conditions, and to standardize the calculation. In this respect, it should be noted that the data sets of C1 and C2 were obtained with a pH of 5.5 involving primarily a single positive change relative to the histidine residue.

Again, *Q* was found to be significantly related to the *|T|^(L^TT)* function via significant interaction ($r = 0.79$; $p = .0002$; Figure 1). Note that the calculated function is nonlinear, and is plotted as a linear graph for convenience. To further validate this function, the original AcP data set (C1) was added to the regression evaluation for the entire set of tested mutations in addition to mutations known to be significant for aggregation. With this addendum, the results were an *r* of 0.798, and $p = 0.001$ (Figure 2).

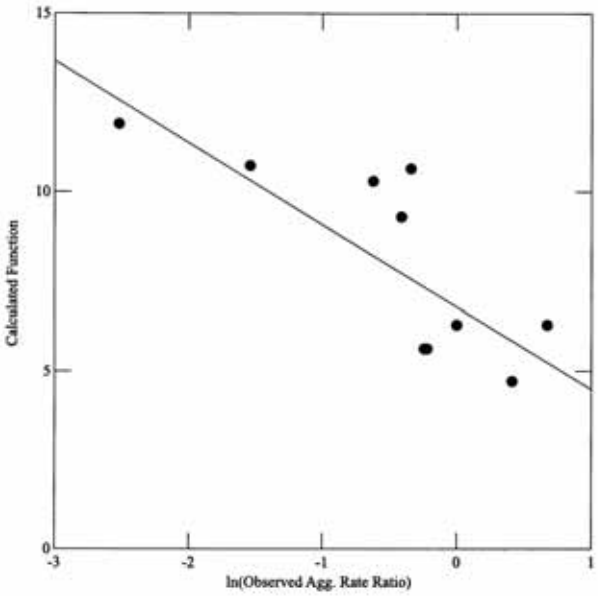

Figure 1. Calculated *vs* observed aggregation rates of AcP variants (data set C2)

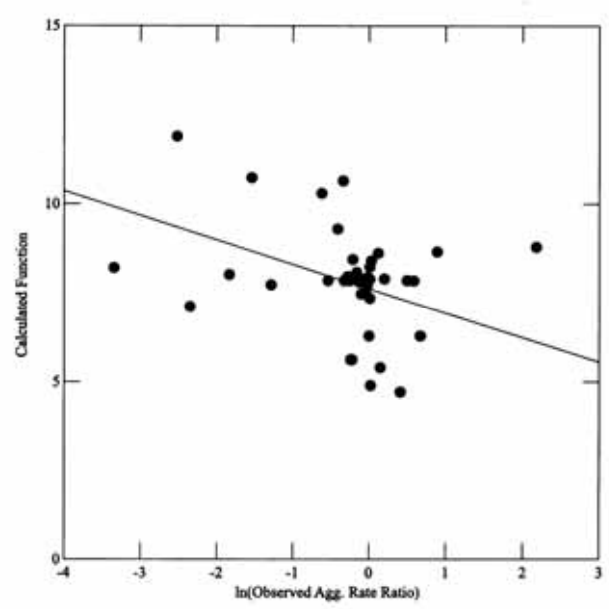

Figure 2. Calculated *vs* observed aggregation rates of AcP variants (data set C2) and mutants (data set C1) combined

Subsequently, the findings were extended to a new set of peptides whose aggregation rates are available from the literature (C3). By fitting a model using change scores of hydrophobicity, α coil and β sheet free energy, as well as charge to the prediction of aggregation rate they obtained an $r = 0.860$, $p < 0.0001$. Using Eq. 3, we obtained $r = 0.642$, $p = 0.0003$ (Figure 3).

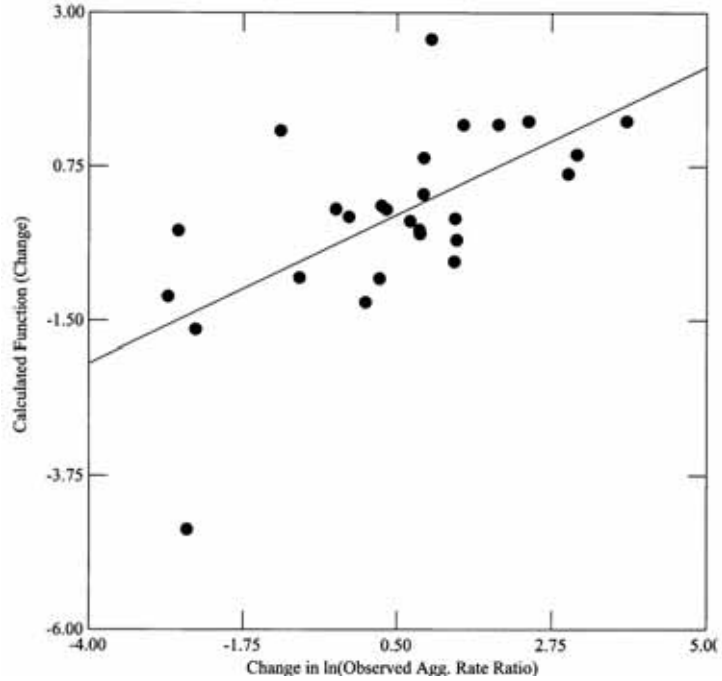

Figure 3. Data from C3 data set (see details in the text)

It is important to note that although our model the *r* value is less than that obtained by the authors for C3; it is calculated solely on the basis of hydrophobicty patterns and calculated net charge at neutral pH and, perhaps more importantly, is an *ab initio* model. This is to emphasize that the C3 data used three factors, whereas our model is based on two. In fact it was based on an independent data set (the AcP system) totally unrelated to C3 data set made of a diverse set of peptides from the literature.

Mathematical considerations

The following section outlines mathematical details of the aggregation model. Although not essential to understanding the remainder of the paper, these details will be of interest to many readers. The crucial point to be made in analyzing the partial derivatives of the recurrence model is that the aggregation propensity, *F*, is non-Lipschitz as a function of the TREND variable, *T*. The significance of this is that the rate of change in aggregation is unbounded as a function of *T* near *T = 0*. Thus, a small perturbation of *T* can result in radically different aggregation behavior, and inherent randomness in the biological system can cause instability and unpredictability. This supports the essential stochastic nature of the process by the inherent instability of TREND as suggested by our previous paper.

The simple function (ignoring the weighting factor), $F(Q,T) = |T|^L \cdot |Q|$ illustrates the differentiability of the aggregation model. First, consider the derivative of $A(Q) = |Q|$. The function $|Q|/Q$ is a good way to represent $A(Q)$, because $A'(Q) = 1$ for $Q > 0$, and $A'(Q) = -1$ for $Q < 0$, and the value is undefined at $Q = 0$.

The function $sign(A)$ is identical to *A'*, *on the domain of A'*, however $sign(A)$ is defined and equal to zero at $Q = 0$. This is not true of the derivative of *A'*. The domain of definition is relevant near the axes $Q = 0$ and $T = 0$, and we will see it is a somewhat subtle issue.

Looking at the symbolic derivatives, we have

$$\begin{aligned}\frac{dF}{dQ} &= |T|^L \cdot A'(Q) \\ &= |T|^L \cdot \frac{|Q|}{Q} \\ &= \left(\frac{1}{Q}\right)\left(|T|^L \cdot |Q|\right)\end{aligned} \quad [4]$$

Then $\frac{dF}{dQ}$ is non Lipschitz (no continuous derivatives) as a function of *T*, and it is discontinuous as a function of *Q*. Because $|Q|/Q$ remains bounded near $Q = 0$, the formula extends continuously at (0,0), but it is undefined for $Q = 0$, $T \neq 0$. Looking at the partial derivative with respect to *T*, we see that $\frac{dF}{dT}$ is continuous as a function of *Q* but unbounded as a function of *T* near $T = 0$ (for $0<L<1$):

$$\begin{aligned}\frac{dF}{dT} &= (L \cdot |T|^{L-1} \cdot A'(T)) \cdot |Q| \\ &= (L \cdot |T|^{L-1} \cdot \frac{|T|}{T}) \cdot |Q| \\ &= \left(\frac{L}{T}\right)\left(|T|^L \cdot |Q|\right)\end{aligned} \quad [5]$$

Near $T = 0$, $\frac{dF}{dT}$ is non Lipschitz (for any value of *Q*), and initial value problems for the equation:

$$\nabla F = \left(\frac{dF}{dQ}, \frac{dF}{dT}\right) = \left(\frac{1}{Q}, \frac{L}{T}\right) \cdot F \quad [6]$$

no longer have unique solutions. This is specifically the case for any trajectory emanating from $T = 0$ (Figure 4). The practical implication is that $F = 0$ is an improbable state for globular proteins.

General applications of the model

Although these calculations may be significant, actual understanding of the involved mechanisms may be deceiving since they reflect "change scores" (22). This is to say that certain putatively important contextual variables are ignored in favor of examining the variables of interest. In other words, only the change of the mutant amino acid relatively to the wild type as characterized by the formula elements is considered. Thus, while the variables of interest may

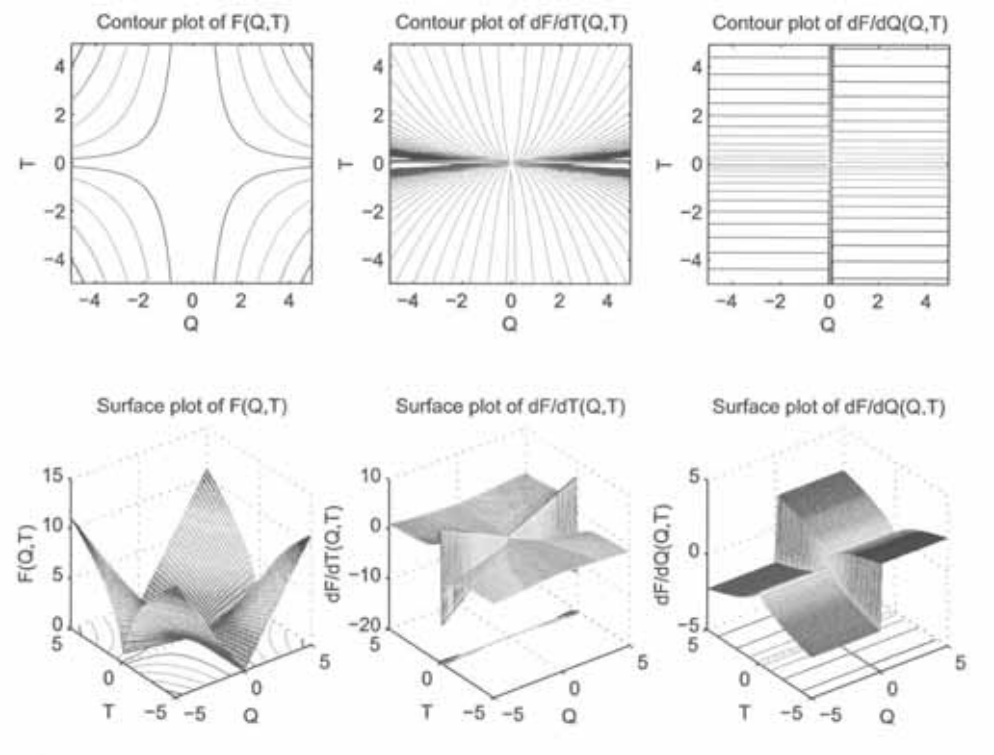

Figure 4. Graphs of the nonlinear function describing the singularities at (0,0), and related derivatives, for *L* = 0.5

be found to have noteworthy effects, ignored variables are not examined, although "controlled," given the fact that they have not changed. Another common application of change scores is the basis for "repeated measures" analysis of variance. Clearly, the most obvious ignored variable in the present case is peptide length. Consequently, a canonical analysis for the observed fluorescence by length and charge, partialing out the effects of Eq. 2 was performed and demonstrated a significant effect for peptide length ($p = 0.04$). These results, however, should be taken with caution since there is a restriction in range with 16 of the 27 mutations being equal to or less than 42 amino acids in length.

Thus a naive inclusion of protein length in a function may be premature. However, based on 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; or that there are specific point charge effects. However, in this respect, we were guided by the experience of Plaxco *et al*. (23) and Ivankov *et al*. (24) 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. To this end, Eq. 3 was normalized for length and recurrences by the relation:

$$Agg = \frac{|T|^{(L^{TT})}}{REC} * \frac{|Q|}{Length^2} \qquad [7]$$

As a first test of this normalization procedure, the C3 data were recalculated according to Eq. 7. Interestingly, the *r* dropped to 0.136 (p = NS). As was previously noted, however, 16 of the 27 mutants are equal to or less than 42 residues in length (n = 16, mean = 34; SD = 10.48). Upon restriction of the analysis to these short proteins the *r* for the original formula remained approximately the same (r = 0.63, p = 0.01); however Eq. 7 demonstrated an *r* at 0.17 (p = NS). Upon choosing the remaining proteins (n = 11, mean = 250.18, SD 151.29), the situation was reversed: the original formula dropped the *r* to 0.44 (p = NS); whereas Eq. 7 demonstrated an *r* of 0.65 (p = 0.03). Again, because of the restriction in range, these results should be viewed cautiously, but they do suggest that there is a slightly different effect for charge of very short proteins. Indeed, it may be that net charge is attenuated at very short scales by the stronger effect of local charge.

To get a better sense of the performance of Eq. 7, it was applied to a data set from the SWISS-PROT repository (mean length = 347.61, SD = 303.04; none < 50) (see above) in order to check its ability to single out peculiar functional classes allowing us to obtain better insight into the mathematical modelling of aggregation process. Functionally, the proteins were classified as reported in Table 2 (Menne *et al*. database) (11).

Table 2. Functional classification of proteins used to test Eq. 7

Protein class (subclass)	Number in class	Number in subclass
Nuclear proteins (N)	184	
Histones (h)		55
Regulative (r)		114
Other nuclear proteins (a)		15
Enzymes (E)	296	
Monomeric (m)		144
Oligomeric (o) (<4 subunits)		105
Polymeric (p) (		46
Other enzymes (a)		1
Ca++/metal – binding (C)	57	
RNA binding (RNA)	68	
Cell-Cycle proteins (CC)	42	
Membrane proteins (M)	153	
Structural proteins (S)	74	
Neurotransmitters/Transport (NT)	5	
Cornifins (K)	5	
Other proteins (A)	257	
Total	1141	

The analysis of the 1141 proteins shows that most values of Eq. 7 are near or at zero; however, a considerable number also appear to be widely dispersed. It is difficult to evaluate the cases since the taxonomy is based on varying, and/or very general criteria. Thus, many proteins could putatively be assigned to several categories (see Table 2). As a consequence, the strategy was to single out proteins and classifications which appeared to be unambiguous, ignoring details which might be considered unfairly biasing in a post-hoc analysis.

An immediate result of the computation was the finding that 54 cases exhibited values of exactly zero: a highly unlikely result for globular proteins given the nature of the model. Careful examination of these cases indicated that all of them were membrane related peptides which interact to generate complex structures (transient or stable), many of whom are immobilized in membranes (Table 3). It can be argued that having a value of zero is totally consistent (and beneficial) given their functions, and their need to have large, multi-domain structures not repelling each other.

Table 3. Proteins exhibiting formula values of zero (abbreviations from Table 2)

Code	Length	Class	Subclass	Description
P52915	406	A		Proteasome subunit
Q92524	389	A		"
P54815	357	M		Integral membrane (mitochondrial)
P46467	444	A		Intracellular protein transport
Q09143	622	M		Integral membrane (cationic aminoacid transport)
P45594	148	S		Actin-binding
O88813	683	E	m	Lipid Biosynthesis
P10107	345	C		Annexin (exocytosis)
P07150	345	C		"
P04272	338	C		"
P17785	338	C		"
P16587	180	M		ADP-ribosylation factor (subunit)
P18085	179	M		"
P36403	179	M		"
Q94231	178	M		"
Q19705	200	A		"
P91924	182	N		Transcriptional repressor
P36405	182	A		ADP-ribosylation factor
P37996	182	A		"
P32121	409	C		Beta-adrenergic receptor-binding
P51164	291	M		Potassium-transport ATP-ase
Q29473	499	E	m	Cytochrome P450
P27003	96	A		Annexin-binding (tetramer)
P08206	96	A		"
P31949	105	C		S-100 protein (dimer)
P40124	474	M		CAP protein
Q03503	176	E	m	Acetyltrasferase
P32320	146	E	o	Deaminase
P41089	223	E	m	Chalcone isomerase
O45405	273	M		Hypothetical protein (membrane channel ?)
Q64448	416	M		Gap-junction protein
O35089	160	M		Embryo development, membrane (potential)
Q63532	152	K		Cornifin
Q14061	62	A		Copper chaperone
P24878	365	M		Cytochrome b (subunit)
P22781	480	E		Decarboxylase
P38866	373	A		Hypothetical protein
P14942	222	E		Dimer
P24472	222	E		"
P46433	210	E		"
P53795	130	A		Hypothetical protein
Q00288	457	N	r	Cellular differentiation regulation
P46871	742	S		Hypothetical protein
Q05315	141	E	m	Carbohydrate-binding
P25791	158	N	a	Basic protein-binding
P80367	65	M		Metallothionein
P04734	64	C		"
Q05935	728	N	r	Transcriptional activator
Q07016	163	N	r	Transcription regulation (dimer)
P08235	984	M		Ribosomal protein
P11658	292	E	p	Subunit
P30044	161	A		Peroxisomal antioxidant enzyme
P49197	88	R		Ribosomal protein
P26490	475	E	p	Subunit
P51668	147	A		Ubiquitin-conjugating enzyme
P29595	81	A		Ubiquitin-like (subunit)

In order to check the consistency of our definition of the equation as "aggregation propensity" quantitation, we selected the 22 proteins reported in Table 4 (11), constituting the highest 2% of values for Eq. 7, and hence the least "aggregating prone." Surprisingly, they were all made of histones and histone-like proteins, such as RNA-binding proteins. It is worth noting that at the extreme of the ranking, namely P42129 and P42132 are two sperm protamins whose role is to pack chromatin, i.e., forming particularly dense aggregates.

Table 4. Proteins exhibiting the highest (2%) values for Eq. 7 (abbreviations from Table 2)

Code	Class	Subclass
P42132	N	h
P42129	N	h
P19757	N	h
P13275	N	h
P40631	N	h
P17502	N	h
Q05831	N	h
P07978	N	h
P06144	N	h
P10922	N	h
P02254	N	h
P02259	N	h
P07305	N	h
P17268	N	h
P40262	N	h
P43278	N	h
Q09821	N	H
P15870	N	H
P14798	R	
P06894	N	H
P15796	N	H
P11020	N	H

Histones and RNA-packing proteins are typically involved in the construction of highly specific DNA/RNA protein aggregates. Such supramolecular structures must respond to finely tuned signals (e.g. acetylation) inducing a reversible aggregation/disaggregation of DNA essential for gene expression regulation. Moreover, histones are probably one of the most conserved protein families: there are only very small differences between the same histone proteins across different species. This implies these are "almost deterministic" supramolecular machines exhibiting different states. Given their important work, protection from unwanted aggregation may be crucial. The fact that they are identified as a "minimum" of aggregation propensity suggests they are shielded from the possibility of aggregation, and is thus a strong proof of the biological plausibility of Eq. 7.

As a final check to evaluate the significance of Eq. 7 relative to covariation with length, charge, TREND, LAM, and TT, an analysis of variance was performed for the identified groups. The equation was in all cases significant, as were the covariates (p<.001), except for TT (p = NS).

Discussion and conclusions

These results support our previous finding that TREND and LAM are important determinants of aggregation in conjunction with net molecular charge. What is more surprising is that these variables are solely based on the hydrophobicity patterns of protein sequences. Notwithstanding this, for some time hydrophobicity has been identified as a major determinant of protein dynamics (e.g. 25, 26), it has been difficult, however, to quantitatively describe hydrophobicity patterns able to evoke basic principles. The present findings demonstrate the utility of RQA in this effort, as well as the importance of a correct formulation of TREND, LAM and charge interplay.

In the above perspective, Eq. 3 (the basic functional) is unique in that it shows a singularity at its zero point (Figure 4). At $Q = 0$ the derivatives are not continuous for T. In practice, this means that zero charge is disallowed, and supports the conjecture of Chiti *et al*. (9) that charge is an important factor to maintain intramolecular repulsive forces, thus avoiding aggregation. In the long run, whether a given protein will go to its native fold or an aggregation, may depend upon its characterization by TREND and/or LAM. The probabilities themselves are governed by the boundary conditions (pH, temperature, etc.). This view is in line with the one adopted by Dobson (4) pointing to the stochastic character of aggregation process. Indeed, this is the implication of phase diagrams exploring protein aggregation (27).

We have previously suggested that hydrophobicity segments broken by laminar patches may tend to be disordered, and exhibit more conformational variability (flexibility), thus tending to avoid aggregation. An explanation for this increased flexibility relates to the fact that the *|T|^(L^TT)* quantifies the differential density of patches. This hypothesis is further supported by our earlier work in the analysis of rubredoxins (28). In this study we used RQA to determine features 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 is distributed over the whole backbone. A graph of this finding using a windowed version of RQA demonstrates this more strikingly (Figure 5). Presumably, this is at least one cause for the increased flexibility of the thermophilic rubredoxin over the mesophilic (29).

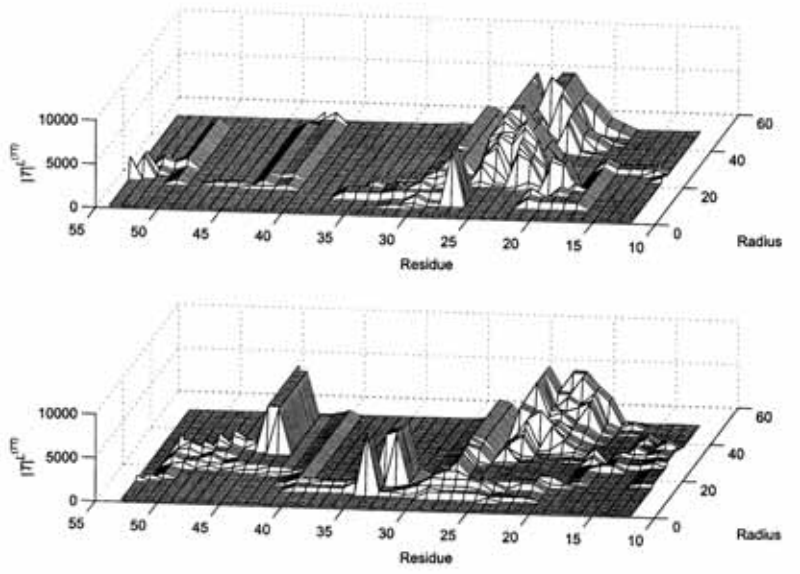

Figure 5. Comparison of Rubr Clopa (top) with Rubr Pyrfu (bottom)

Other putative causes involve the observation that amyloidogenic propensity is associated with a defect in hydrogen bonding exposed to water, making them "sticky" (30-32). 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 (33). Previously (34), we have suggested that such instabilities may occur in the form of molecular motions not associated with traditional modes analysis.

Finally, we note the obvious difference between the C3 formulation with ours is their inclusion of the free energy changes based on beta sheet and coil propensities. It may be that our quantification of patches of laminarity may be characterizing a similar phenomenon. Beta sheets and coils in some sense typify types of "patch." In our studies, we have noted a correlation; however, this 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 in line with recent results obtained by Burke *et al*. (35) with huntingtin-exon 1. The final observation is 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 (boundary conditions) such as pH. Thus, clearly, hydrophobic patterning is a necessary condition for understanding aggregation propensity, but it is insufficient without consideration of charge. 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.

Irrespective of the cause, the present results suggest that to understand the aggregation probability for a given protein sequence, unique hydrophobicity patterns need to be considered. This probability may be linked to fixed discrete patches in conjunction with net dynamic electrostatic effects.

Acknowledgements

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. JPZ thanks Prof. F. Chiti for kindly providing details of experiments and calculations.

References

1. Kyte J, Doolitle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982; 157:105-32.
2. Mandell AJ, Selz KA, Shlesinger MF. Protein binding predictions from amino acid primary sequence hydrophobicity. J Mol Liquids 2000;86:163-71.
3. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21:170-201.

4. Dobson CM. Protein folding and disease: a view from the first Horizon Symposium. Nat Rev Drug Discov 2003;2:154-60.
5. Dunker K, Brown CJ, Lawson D, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573-82.
6. Chiti F, Taddei N, Baroni F, Capanni C, Stefani M, Ramponi G, Dobson CM. Kinetic partitioning of protein folding and aggregation. Nat Struc Biol 2002;9:137-43.
7. Zbilut JP, Colosimo A, Conti F, Colafranceschi M, Manetti C, Valerio MC, Webber CL Jr., Giuliani A. Protein aggregation/folding: the role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and Aβ(1-40). Biophys J 2003;85:3544-57.
8. Ortiz AR, Skolnick J. Sequence evolution and the mechanism of protein folding. Biophys. J 2000;79:1787-99.
9. Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G, Dobson CM. Studies of the aggregation of mutant proteins *in vitro* provide insights into the genetics of amyloid diseases. Proc Natl Acad Sci USA 2002;99:16419-26.
10. Chiti F, Stefani M, Taddei N, Ramponi FG, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 2003;424:805-8.
11. Menne KML, Hermjakob H, Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 2000;16:741-2.
12. Eckmann JP, Kamporst SO, Ruelle D. Recurrence plots of dynamical systems. Europhys Letts 1987;4:973-7.
13. Webber CL Jr., Zbilut JP. Dynamical assessment of physiological systems and states using recurrence plot strategies. J Appl Physiol 1994;76:965-73.
14. Zbilut JP, Webber CL Jr. Embeddings and delays as derived from quantification of recurrence plots. Phys Lett A 1992;171:199-203.
15. Marwan N, Wessel N, Meyerfeldt U, Schirdewan A, Kurths J. Recurrence plot based measures of complexity and its application to heart rate variability data. Phys Rev E 2002;66:026702-1-026702-7.
16. Trulla LL, Giuliani A, Zbilut JP, Webber CL Jr. Recurrence quantification analysis of the logistic equation with transients. Phys Lett A 1996;223:225-60.
17. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structure: quasi-chemical approximation. Macromolecules 1985;18:534-52.
18. Giuliani A, Benigni R, Zbilut JP, Webber CL Jr, Sirabella P, Colosimo A. Nonlinear signal analysis methods in the elucidation of protein sequence structure relationships. Chem Rev 2002;102:1471-91.
19. Strait BJ, Dewey TG. The Shannon information entropy of protein sequences. Biophys J 1996;71:148-55.
20. Weiss O, Jimenez-Montano MA, Herzel H. Information content of protein sequences. J Theor Biol 2000;206:379-86.
21. Zbilut JP, Sirabella P, Giuliani A, Manetti C, Colosimo A, Webber CL Jr. Review of nonlinear analysis of proteins through recurrence quantification. Cell Biochemistry and Biophysics 2002;36:67-87.
22. Vickers J, Altman DG. Analyzing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123-4.
23. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985-94.
24. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. Protein Sci 2003;12:2057-62.

25. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 1980;136:225-70.
26. Rose GD. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 1978;272;586-90.
27. Dima RI, Thirumalai D. Exploring protein aggregation and self-propagation using lattice models: Phase diagram and kinetics. Protein Science 2002;11:1036-49.
28. Giuliani A, Benigni R, Sirabella P, Zbilut JP, Colosimo A. Nonlinear methods in the analysis of protein sequences: A case study in rubredoxins. Biophys J 2000;78:136-49.
29. Grottesi A, Ceruso M-A, Colosimo A, Di Nola A. Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. Proteins: Structure, Function and Genetics 2002;486:287-94.
30. Fernández A, Berry RS. Proteins with H-bond packing defects are highly interactive with lipid bilayers: implications for amyloidogenesis. Proc Natl Acad Sci USA 2003;100:2391-6.
31. Fernández A, Scheraga HA. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc Natl Acad Sci USA 2003;100:113-8.
32. Fernández A, Scott R. Dehydron: a structurally encoded signal for protein interaction. Biophys J 2003;85:1914-28.
33. Mingaleev SF, Gaididei YB, Christiansen PL, Kivshar YS. Nonlinearity-induced conformational instability and dynamics of biopolymers. Europhys. Lett 2002;59:403-9.
34. Manetti C, Giuliani A, Ceruso MA, Webber CL Jr., Zbilut JP. Recurrence analysis of hydration effects on nonlinear protein dynamics: multiplicative scaling and additive processes. Phys Lett A 2001;281:317-23.
35. Burke MG, Woscholski R, Yaliraki SN. Differential hydrophobicity drives self-assembly in Huntington's disease. Proc Natl Acad Sci USA 2003;100:13928-33.