Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 343 (2004) 348–358
PHYSICA A
www.elsevier.com/locate/physa

Singular hydrophobicity patterns and net charge: a mesoscopic principle for protein aggregation/folding☆

Joseph P. Zbilut[a,*], Julie C. Mitchell[b], Alessandro Giuliani[c], Alfredo Colosimo[d], Norbert Marwan[e], Charles L. Webber[f]

[a] *Department of Molecular Biophysics and Physiology, Rush University Medical Center, 1653 W. Congress, Chicago, IL 60612, USA*
[b] *Departments of Mathematics/Biochemistry, University of Wisconsin-Madison, 480 Lincoln Drive, Madison, WI 53706-1388, USA*
[c] *Health and Environment Department, Istituto Superiore di Sanitá, V.le Regina Elena 299, Rome, Italy*
[d] *Department of Human Physiology and Pharmacology, University of Rome "La Sapienza", P.le A. Moro, 5, 00185 Rome, Italy*
[e] *Nonlinear Dynamics Group, Institute of Physics, University of Potsdam, Potsdam, Germany*
[f] *Department of Physiology, Loyola University Medical Center, 2160 S. First Avenue, Maywood, IL 60153, USA*

Received in revised form 16 April 2004

Abstract

A statistical model describing the propensity for protein aggregation is presented. Only amino-acid hydrophobicity values and calculated net charge are used for the model. The combined effects of hydrophobic patterns as computed by the signal analysis technique, recurrence quantification, plus calculated net charge were included in a function emphasizing the effect of singular hydrophobic patches which were found to be statistically significant for predicting aggregation propensity as quantified by fluorescence studies obtained from the literature. These results suggest preliminary evidence for a mesoscopic principle for protein folding/aggregation.

PACS: 87.17.−d; 87.14.Ee; 87.15.He; 87.15.Cc; 87.10.+e

Keywords: Singularities; Protein; Aggregation; Folding; Recurrence analysis; Mesoscopic

☆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.

*Corresponding author. Tel.: +1-312-942-6008; fax: +1-312-942-8711.
E-mail address: joseph_P_Zbilut@rush.edu (J.P. Zbilut).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2004.05.081

1. Introduction

It has been noted 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 [1,2]. 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 [3]. It is interesting to note that such observations are gradually providing justification for a synthetic view of protein dynamics [4], not typically appreciated by more reductionistic approaches. At the basis of these synthetic views is the emphasis that biopolymer dynamics follow broad established physical principles. Yet, what factors specifically promote the formation of aggregates as opposed to native folds under relatively normal conditions remain unclear.

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 [5]. A significant finding included a correspondence between short deterministic patches of hydrophobicity distribution along the amino-acid sequence, what we term laminarity, [LAM, (L)], with 3D "unstructured" portions of acylphosphatase (AcP) [6]. 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 [7]. Beyond this, a counterpoint was defined as the degree of laminarity. Specifically, in an analysis of protein engineering experiments [8] it was shown that aggregation sensitive zones vs. folding sensitive zones were distinguished by the two complementary concepts of trend/laminarity. The implication is that these areas may be inherently unstable, and somehow involved in the promotion of (at least) partial unfolding and aggregation. The degree to which these conditions exist probabilistically determines the propensity for aggregation. What was not determined is the effect of total charge on the probabilities.

2. Recurrence hydrophobicity signal analysis

2.1. Recurrences

Recurrences are not new. Poincaré is perhaps the most famous for describing them in the context of dynamical systems as points which visit a small region of phase space. Also, the statistical literature points out that recurrences are the most basic of relations. In this respect, it is important to reiterate the fact that calculation of recurrence, unlike other methods such as Fourier, Wigner–Ville or wavelets, requires no transformation of the data, and can be used for both linear and nonlinear systems. Because recurrences are simply tallies, they make no mathematical assumptions. Given a reference point, $\mathbf{X}_0$, and a ball of radius r, a point is said to recur if

$$B_r(\mathbf{X}_0) = \{\mathbf{X} : \|\mathbf{X} - \mathbf{X}_0\| \leqslant r\} \,. \tag{1}$$

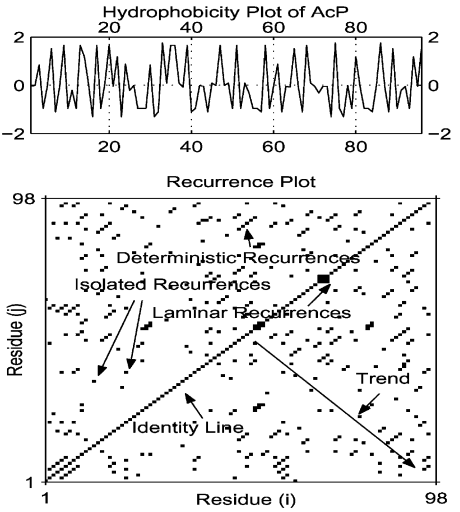

Fig. 1. Recurrence plot for AcP.

A trajectory of size N falling within $B_r(\mathbf{X}_0)$ is denoted as

$$S_1 = \{\mathbf{X}_{t_1}, \mathbf{X}_{t_2}, \ldots, \mathbf{X}_{t_i} \ldots\} , \tag{2}$$

where S is some signal, and t is some time point, and the recurrence defined as

$$T_1(i) = t_{i+1} - t_i, \quad i = 1, 2, \ldots, N , \tag{3}$$

where T is some recurrence time (not to be confused with T indicating TREND below). We note that although recurrences are usually calculated for temporal series, it is also possible to use any ordered series, as is the case here for protein amino-acid sequences.

2.2. Recurrence plots

Given a scalar series $\{x(i) = 1, 2, 3, \ldots\}$ an embedding procedure will form a vector, $\mathbf{X}_i = (x(i), x(i + del), \ldots, x(i + (m - 1)del))$ with m the embedding dimension and del the delay (not to be confused with a differential operator). $\{\mathbf{X}_i = 1, 2, 3, \ldots, N\}$ then represents the multi dimensional process of the series as a trajectory in m-dimensional space. Recurrence plots (RP) are symmetrical $N \times N$ arrays in which a point is placed at (i, j) whenever a point $\mathbf{X}_i$ on the trajectory is close to another point $\mathbf{X}_j$. The closeness between $\mathbf{X}_i$ and $\mathbf{X}_j$ is expressed by calculating the Euclidian distance between these two normed vectors, i.e., by subtracting one from the other: $\|\mathbf{X}_i - \mathbf{X}_j\| \leqslant r$ where r is a fixed radius. If the distance falls within this radius, the two vectors are considered to be recurrent, and graphically this can be indicated by a dot (Fig. 1).

An important feature of such matrixes is the existence of short line segments parallel to the main diagonal, which correspond to sequences (i, j), $(i+1, j+1), \ldots, (i+k, j+k)$ such that the piece of $\mathbf{X}(j), \mathbf{X}(j+1), \ldots, \mathbf{X}(j+k)$, is close to $\mathbf{X}(i), \mathbf{X}(i+1), \ldots, \mathbf{X}(i+k)$ in series which are deterministic. The absence of such patterns suggest randomness [9].

2.3. Recurrence quantification

Because graphical representation may be difficult to evaluate, RQA was developed to provide quantification of important aspects revealed by the plot. Recurrent points which form diagonal line segments are considered to be deterministic (as distinguished from random points which form no patterns). Unfortunately, beyond general impressions of drift and determinism, the plots of themselves provide no quantification. As a result, an algorithm was developed using several strategies to quantify features of such plots [10]. Hence, the quantification of recurrences leads to the generation of seven variables including: REC (percent of plot filled with recurrent points); DET (percent of recurrent points forming diagonal lines, with a minimum of two adjacent points); ENT (Shannon information entropy of the line length distribution); MAXLINE, length of longest line segment (the reciprocal of which is an approximation of the largest positive Liapunov exponent and is a measure of system divergence); and TREND (measure of the paling of recurrent points away from the central diagonal); LAM (percent of points forming vertical line structures; and trapping time (TT), the average length of the vertical line segments. These seven recurrence variables quantify the deterministic structure and complexity of the plot.

Specifically, for the variables used here

$$REC = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j}\,, \tag{4}$$

where R is a recurrent point

$$DET = \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\sum_{i,j}^{N} R_{i,j}}\,, \tag{5}$$

where $P(l)$ is the histogram of lengths l of diagonal line segments

$$LAM = \frac{\sum_{\upsilon=\upsilon_{\min}}^{N} \upsilon P(\upsilon)}{\sum_{\upsilon=1}^{N} \upsilon P(\upsilon)}\,, \tag{6}$$

where $P(\upsilon)$ is the histogram of lengths υ of vertical line segments

$$TT = \frac{\sum_{\upsilon=\upsilon_{\min}}^{N} \upsilon P(\upsilon)}{\sum_{\upsilon=\upsilon_{\min}}^{N} P(\upsilon)} \tag{7}$$

and

$$TREND = \frac{\sum_{i=1}^{N-2} [i-(N-2)](REC_i - \langle REC_i\rangle)}{\sum_{i=1}^{N-2} [i-(N-2)/2]^2}\,. \tag{8}$$

In order to follow changes of these variables in sequence, a "windowed" version of RQA can be performed (Fig. 2), such that for a series $(s_1, s_2, \ldots, s_n)$, where $(s_j = j\tau_s)$ and $\tau_s =$ sampling interval. For an N point long series

$$E_1 = (s_1, s_2, \ldots, s_N)\,,$$

$$E_2 = (s_{1+w}, s_{2+w}, \ldots, s_{N+w})\,,$$

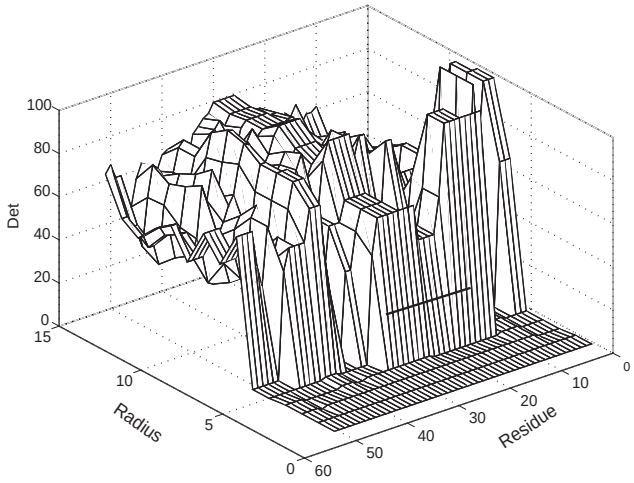

Fig. 2. Windowed RP for AcP showing singularities.

$$E_3 = (s_{1+2w}, s_{2+2w}, \ldots, s_{N+2w}) ,$$

$$\vdots$$

$$E_p = (s_{1+(p-1)w}, s_{2+(p-1)w}, \ldots, s_{N+(p-1)w}) \tag{9}$$

with $w=$ the offset, and E_p the number of epochs (windows), E_p, satisfies the relation, $N+(p-1)w \leqslant n$. Thus the algorithm is, in principle, similar to the Fourier transforms devoid of its drawbacks.

Analogous to cross power spectral analysis, cross recurrence analysis is also possible. For the series $\mathbf{X}_i=(x(i), x(i+del), \ldots, x(i+(m-1)del))$, another series, $\mathbf{Y}_i=(y(i), y(i+del), \ldots, y(i+(m-1)del))$ can be compared for recurrences by the relation $\|\mathbf{X}_i - \mathbf{Y}_j\|$ for a given r. Windowed versions are similarly possible.

The data obtained can also be used to obtain estimations of local Liapunov exponents, information entropy, or simply plotted as $N_{recurrences}$ vs. period; i.e., a histogram of recurrence times. In the case of histograms, strictly periodic points demonstrate instrumentally sharp peaks; whereas chaotic or nonlinear systems reveal more or less wider peaks depending upon the radius chosen and noise effects. RQA can also be combined with other statistical techniques to gain more information. Use of RQA for protein analysis has been extensive [18,19].

3. Methods

3.1. Data set

The data which inspired our model are in a seminal paper by Chiti et al. [8], and concern the effect of different mutations on acylphosphatase (AcP) aggregation

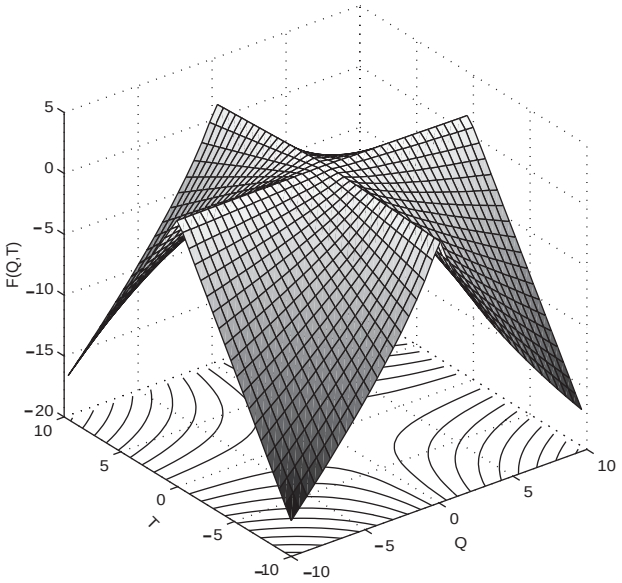

Fig. 3. Surface plot of function.

propensity. A second Chiti et al. [11] data set dealt with the effects of charge; wherein it was suggested that the net charge was an important consideration. Specifically, the authors suggested that preservation of charge prevented molecular attraction and vice versa. We used the data of this paper to evaluate its conclusions vis-a-vis our previous findings.

3.2. Model

The recurrence variables were evaluated statistically and found to fit a nonlinear function such that

$$Agg(rate) = Const. + a(|T|^{L^{TT}}|Q|)\,, \tag{10}$$

where *Const.* is a constant, a is an adjustable parameter, T is the TREND, L is the LAMINARITY, TT the TRAPPING TIME, and Q the net molecular charge. The equation should 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 formula conveys the idea of statistical "singularity" and is also mathematically singular, insofar as it does not admit continuous derivatives near $T = 0$ (Fig. 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. The charge effect is very similar in that zero charge is inadmissable.

The simple function $F(Q,T) = |T|^{L}|Q|$ 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 $Q>0$, $A'(Q)=-1$ 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 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

$$\begin{aligned} \mathrm{d}F/\mathrm{d}Q &= |T|^L A'(Q) \\ &= |T|^L \frac{|Q|}{Q} \\ &= \left(\frac{1}{Q}\right)(|T|^L|Q|)\,. \end{aligned} \tag{11}$$

Then $\mathrm{d}F/\mathrm{d}Q$ is not Lipschitz 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 is unbounded for $Q=0$, $T\neq 0$.

Looking at the partial derivative with respect to T, we see that $\mathrm{d}F/\mathrm{d}T$ is continuous as a function of Q but unbounded as a function of T near $T=0$ (for $0<L<1$)

$$\begin{aligned} \mathrm{d}F/\mathrm{d}T &= (L|T|^{L-1}A'(T))|Q| \\ &= \left(L|T|^{L-1}\frac{|T|}{T}\right)|Q| \\ &= \left(\frac{L}{T}\right)(|T|^L|Q|)\,. \end{aligned} \tag{12}$$

Near $T=0$, $\mathrm{d}F/\mathrm{d}T$ is not Lipschitz (for any value of Q), and initial value problems for the equation

$$\nabla F = \left(\frac{\mathrm{d}F}{\mathrm{d}Q}, \frac{\mathrm{d}F}{\mathrm{d}T}\right) = \left(\frac{1}{Q}, \frac{s}{T}\right) F \tag{13}$$

no longer have unique solutions. This is specifically the case for any trajectory emanating from $T=0$.

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 is further amplified by charge, Q, as it approaches zero.

4. Results and discussion

The data sets of acylphosphatase with their aggregation rates were compared to the formula, and were found to be significantly related to the function ($r=0.79$; $p=0.0002$; Fig. 4).

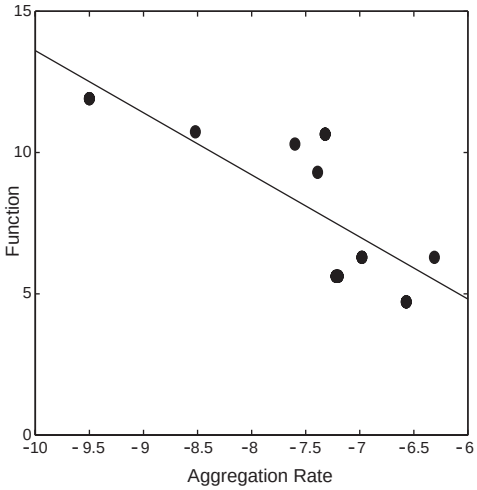

Fig. 4. Regression results.

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 indirectly quantifying the "unstructured" element of protein sequence.

In this perspective, Eq. (10) is unique in that it shows a singularity at its zero point (Fig. 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. [11] 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 environmental conditions (pH, temperature, type of ionic solvent, etc.). This view is in line with the one adopted by Dobson [3] pointing to the stochastic character of aggregation process. Indeed, this is the implication of phase diagrams exploring protein aggregation [12]. Independently, Munishkina et al. [13] have presented additional evidence for this hypothesis with work on α-synuclein.

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 [5]. An explanation for this increased flexibility relates to the fact that $|T|^{L^{TT}}$, quantifies the differential density of patches. This hypothesis is further supported by our earlier work in the analysis of rubredoxins [14]. 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 (Fig. 5). Presumably, this

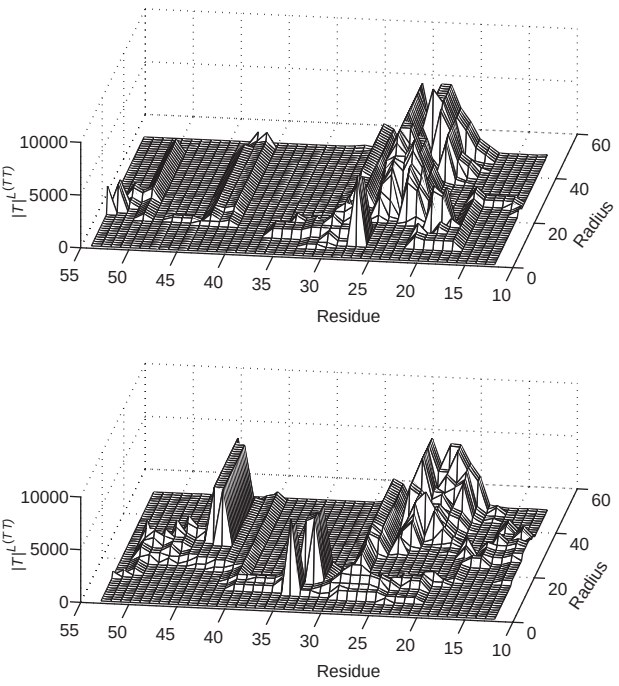

Fig. 5. Comparison of Rubr Clopa (top) with Rubr Pyrfu (bottom). Note the uneven distribution of the function for Rubr Clopa; whereas Robr Pyrfu demonstrates a more "even" distribution.

is at least one cause for the increased flexibility of the thermophilic rubredoxin over the mesophilic [15].

The charge effect in this context 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. with huntingtin-exon 1 [16]. 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.

A paper by Hans Frauenfelder and Peter Wolynes [4] highlighted the peculiarity of the sequence/structure relation: the need to have microscopic physics principles, of "simple" systems like atoms, cooperatively interacting to produce macroscopic principles which describe qualitatively the complex systems of protein architecture. While we do have an accurate knowledge of potentials (hydrophobic interactions, hydrogen bonding, size constraints, etc.) acting at microscopic levels, the "mesoscopic" principles needed to understand protein folding/aggregation remain essentially unknown [17]. The present study suggests that such principles are not that remote from our understanding.

References

[1] J.M. Zimmerman, N. Eliezer, R. Simha, The characterization of amino acid sequences in proteins by statistical methods, J. Theor. Biol. 21 (1968) 170–201.
[2] J. Kyte, R.F. Doolitle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[3] C.M. Dobson, Protein folding and disease: a view from the first Horizon Symposium, Nat. Rev. Drug Discov. 2 (2003) 154–160.
[4] H. Frauenfelder, P. Wolynes, Proteins: where physics of simplicity and complexity meet, Phys. Today 47 (1994) 58–61.
[5] J.P. Zbilut, A. Colosimo, F. Conti, M. Colafranceschi, C. Manetti, M.C. Valerio, C.L. Webber Jr., A. Giuliani, 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. 85 (2003) 3544–3557.
[6] V.N. Uversky, Natively unfolded proteins: a point where biology waits for physics, Protein Sci. 11 (2002) 739–756.
[7] K. Dunker, C.J. Brown, D. Lawson, L.M. Iakoucheva, Z. Obradovic, Intrinsic disorder and protein function, Biochemistry 41 (2002) 6573–6582.
[8] F. Chiti, N. Taddei, F. Baroni, C. Capanni, M. Stefani, G. Ramponi, C.M. Dobson, Kinetic partitioning of protein folding and aggregation, Nat. Struct. Biol. 9 (2002) 137–143.
[9] J.P. Eckmann, S.O. Kamporst, D. Ruelle, Recurrence plots of dynamical systems, Europhys. Lett. 4 (1987) 973–977.
[10] C.L. Webber, J.P. Zbilut, Dynamical assessment of physiological systems and states using recurrence plot strategies, J. Appl. Physiol. 76 (1994) 965–973.
[11] F. Chiti, M. Calamai, N. Taddei, M. Stefani, G. Ramponi, C.M. Dobson, Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases, Proc. Natl. Acad. Sci. USA 99 (2002) 16419–16426.
[12] R.I. Dima, D. Thirumalai, Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics, Protein Sci. 11 (2002) 1036–1049.
[13] L.A. Munishkina, J. Henriques, V.N. Uversky, A.L. Fink, Role of protein–water interaction and electrostatics in α-synuclein fibril formation, Biochemistry 43 (2004) 3289–3300.
[14] A. Giuliani, R. Benigni, P. Sirabella, J.P. Zbilut, A. Colosimo, Nonlinear methods in the analysis of protein sequences: a case study in rubredoxins, Biophys. J. 78 (2000) 136–149.
[15] A. Grottessi, M.-A. Ceruso, A. Colosimo, A. Di Nola, Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin, Proteins: Struct. Funct. Genet. 486 (2002) 287–294.
[16] M.G. Burke, R. Woscholski, S.N. Yaliraki, Differential hydrophobicity drives self-assembly in Huntington's disease, Proc. Natl. Acad. Sci. USA 100 (2003) 13928–13933.
[17] R.B. Laughlin, D. Pines, G. Schmalian, P. Wolynes, The middle way, Proc. Natl. Acad. Sci. USA 97 (2002) 32–37.

[18] A. Giuliani, R. Benigni, J. Zbilut, C.L. Webber, P. Sirabella, A. Colosimo, Nonlinear signal analysis methods in the elucidation of protein sequence/structure relationships, Chem. Rev. 102 (2002) 1471–1492.
[19] J.P. Zbilut, P. Sirabella, A. Giuliani, C. Manetti, A. Colosimo, C.L. Webber, Review of nonlinear analysis of proteins through recurrence quantification, Cell Biochem. Biophys. 36 (2002) 67–87.