Protein Aggregation/Folding: The Role of Deterministic Singularities of Sequence Hydrophobicity as Determined by Nonlinear Signal Analysis of Acylphosphatase and Aβ(1–40)

Joseph P. Zbilut,* Alfredo Colosimo,† Filippo Conti,‡ Mauro Colafranceschi,†¶ Cesare Manetti,‡ MariaCristina Valerio,‡ Charles L. Webber, Jr.,§ and Alessandro Giuliani¶

*Department of Molecular Biophysics and Physiology, Rush Medical College, Chicago, Illinois; †Department of Human Physiology and Pharmacology, University of Rome "La Sapienza," Rome, Italy; ‡Department of Chemistry, University of Rome "La Sapienza," Rome, Italy; §Department of Physiology, Loyola University Medical Center, Maywood, Illinois; and ¶Health and Environment Department, Istituto Superiore di Sanitá, Rome, Italy

ABSTRACT The problem of protein folding vs. aggregation was investigated in acylphosphatase and the amyloid protein Aβ(1–40) by means of nonlinear signal analysis of their chain hydrophobicity. Numerical descriptors of recurrence patterns provided the basis for statistical evaluation of folding/aggregation distinctive features. Static and dynamic approaches were used to elucidate conditions coincident with folding vs. aggregation using comparisons with known protein secondary structure classifications, site-directed mutagenesis studies of acylphosphatase, and molecular dynamics simulations of amyloid protein, Aβ(1–40). The results suggest that a feature derived from principal component space characterized by the smoothness of singular, deterministic hydrophobicity patches plays a significant role in the conditions governing protein aggregation.

INTRODUCTION

Protein aggregation has attracted much attention, driven by the discovery of the involvement of protein misfolding (with consequent formation of polymer insoluble aggregates) in degenerative pathologies such as Alzheimer's, Huntington's, and prion diseases (Dobson, 2003). Moreover, the discovery of the possibility of forming potentially harmful aggregates after posttranslational modifications of proteins has motivated scientists to consider proteins as a direct ''toxicological target'' without the intermediate agency of DNA modifications, and to reconsider general ideas regarding their mechanisms of pathogenesis.

As clearly stated by Chiti et al. (2002), the possibility of forming aggregates is intrinsic to any protein. Consequently, it may be difficult to delineate a clearcut border separating aggregating and nonaggregating proteins. Any modification of environmental conditions (pH, temperature, ionic strength, etc.) could in principle drive any protein structure to shift from an isolated globular existence in solution to the formation of multimeric aggregates and the eventual precipitation exiting the solvent-solute equilibrium. This possibility is implicit in the character of the hydrophobic interaction. The main driving force shaping protein tertiary structures is the need to be soluble in water. For this task to be accomplished, the protein must fold in such a way as to hide hydrophobic residues, while exposing polar residues (Bryngelson et al., 1995).

On the same basis, the protein-protein interaction can be considered as another aspect of the same phenomenon. In general, the search for aggregation cores is not basically different from the search for folding cores, and aggregation can be simply considered an *alternative folding*. The choice between *correct* (autonomous) and *incorrect* (multimeric) folding is a matter of relative preponderance (in energetic terms for given boundary conditions) of the two possible ways. Thus, the choice between alternative foldings is a stochastic matter and the boundary conditions can dramatically alter the relative probabilities of predominance. This is to say that an understanding of this process lies in a statistical, i.e., probabilistic characterization. In the case of protein folding, the relative probability of the two is driven by the balance of hydrophobic charge and steric effects of sequence/environment interaction (Dobson, 2003). This implies the possibility of recognizing the relative propensity for aggregation by means of an efficient chemicophysical representation of proteins.

The mainly hydrophobic character of folding processes motivated us to look at the hydrophobicity coding of protein sequences as the first step for such a study. The coding of single monomers by means of chemicophysical properties, while allowing for a mechanistic interpretation of the observed results, turns protein sequence investigation into a classical numerical signal analysis problem.

The consideration of proteins as numerical time series has a long history, dating back to the pioneering work by Zimmerman et al. (1968) and Kyte and Doolitle (1982). These initial studies, while providing useful insights, were limited by the use of signal analysis methods not completely suitable for protein sequences. In fact Fourier analysis and linear autocorrelation functions (the basic methods used) have strong limitations for protein sequence studies, since they assume sequence stationarity and signals with a length much higher than an average protein. Thus, although strictly periodic features can be identified, complicated, less obvious features are easily missed.

Submitted July 14, 2003, and accepted for publication August 7, 2003.

Address reprint requests to Joseph P. Zbilut. Tel.: 312-942-6008; Fax: 312-942-8711; E-mail: joseph_p_zbilut@rush.edu.

0006-3495/03/12/3544/14 $2.00

The interest in nonlinear systems in the eighties allowed for a reinitiation of time-series-style analysis of protein sequences with new mathematical methods independent of data length and stationarity. This resurgence of interest was marked by numerous successes: the demonstration of a correlation between hydrophobicity patterning of peptides and their relative receptors by Mandell et al. (2000); hydrophobicity energy patterns (Selz et al., 1998); the demonstration of a ''signature'' in terms of hydrophobicity patterning of different classical three-dimensional motifs (Murray et al., 2002); and demonstration of the predictability of protein stability and protein-protein interaction patterns by our group (Zbilut et al., 2000; Giuliani and Tomasi, 2002). For reviews of the second-wave of time series analysis of protein sequences, see Giuliani et al. (2002), and Zbilut et al. (2002).

In the present article, we report the results from a nonlinear signal analysis approach to hydrophobicity patterning, using both a *static* and a *dynamic* approach. The static approach was based both on the search for singularities of the distribution of hydrophobicity along amino acid sequences of aggregating protein systems, and on the classification of different folding behaviors relative to their hydrophobicity patterns. (The term *singularity* has several definitions depending upon a discipline's perspective. Here, we use the term in a general, nonformal sense of a uniquely occurring pattern. Different patterns can emerge depending upon the analytic tool used. In the present case recurrence quantification was employed; see below.) The protein groups were chosen on the basis of available experimental evidence describing their folding tendencies. The dynamic approach was based on the molecular dynamics simulation of amyloid β-peptide Aβ(1–40), at different pH values known to have a different permissivity in terms of fibril formation.

Both approaches give a general picture of aggregation mechanisms in terms of the relative propensity for undergoing structural order-disorder transitions of the intervening structures acting at hydrophobicity-singular points along the amino acid chain. Thus, more than being linked to a particular structural feature, the ability to form intermolecular aggregates appears to be correlated with conformational flexibility. Moreover, the mechanism governing the formation of protein polymers like collagen or silk is shown to be completely different from the one governing the formation of aggregates typical of misfolding diseases, which has been shown to be more similar to the one governing the formation of multimeric enzyme complexes, thus stressing the crucial role played by small aggregates in misfolding diseases as suggested by Dobson (2003).

MATERIALS AND METHODS

Static approach

Strategy of analysis

Each studied protein sequence was coded by means of Miyazawa-Jernigan hydrophobicity (MJ) of the constituent residues (Miyazawa and Jernigan, 1985). This scale corresponds to the first eigenvalue of the contact energy matrix as reported at the URL, http://us.expasy.org/tools/pscale/Hphob.Miyazawa.html. The choice of MJ as a hydrophobicity score was dictated by our analysis for a 1141 random sample of protein sequences from the Swiss-Prot Database for which we demonstrated that the MJ was the code that exhibited the largest separation in distance space for obtained patterns, as compared to a random assortment of amino acids (in preparation). (The Proteome Analysis database, http://www.ebi.ac.uk/proteome/, lists 460 residues as the average *Homo sapiens* protein length, with a lower range value of 3.) This finding suggests some potentially important ''syntactic rule,'' shaping the amino acid distribution along the chain, perhaps a *folding rule*. The MJ-coded sequences were submitted to recurrence quantification analysis (RQA; see below) to obtain sensible quantitative indices of the degree (and quality) of hydrophobicity autocorrelation along the sequence.

We applied this kind of approach to two different situations: 1) the discrimination of different folding behavior of different proteins, and 2) the identification of crucial ''hotspots'' for aggregation behavior along the acylphosphatase (AcP) protein sequence.

In the first situation, 90 protein sequences of specific structural and functional classes were used for analysis: *A*, mainly α-helical structures; *B*, mainly β-sheet structures; *C*, proteins giving rise to long extracellular polymeric structures; *D*, self-aggregating systems (amyloid, serpins); *E*, natively unfolded proteins; *F*, proteins involved in DNA processing through supramolecular structures; *G*, artificial α-helices with very regular patterning of amino acids (Kamtekar et al., 1993); *H*, artificial β-sheets with very regular patterning of amino acids (West et al., 1999); and *I*, proteins known to posses π-helix structures (Fodje and Al-Karadaghi, 2002). (A note regarding the inclusion of π-helices: although the first eight groups are derived from a logical classification of protein groups, group *I* was included to address the concern that π-helices are an underreported structure of some structural as well as functional significance. It was also noted in the MD simulation of APP Aβ(1–40) with aggregation-prone conditions; see Table 1).

In the second situation, we carefully investigate a single system (AcP) to overcome the simple statistical correlation between folding behavior and recurrence spectrum and identify a particular ''aggregation signature'' in terms of hydrophobicity patterning along the given sequence. Use was made of the experimental data by Chiti et al. (2002) who identified, by means of site-directed mutagenesis experiments, the zones which significantly modify aggregation behavior.

Recurrence quantification analysis

RQA is a nonlinear time-series analysis method (Webber and Zbilut, 1994) which, in addition to the application of protein sequence analysis, has been adopted with success in a number of other fields ranging from physiology, to theoretical physics, to the analysis of reaction mechanisms. The method has been documented extensively, but briefly:

The basis of the method is the projection of the original monodimensional series into a multidimensional space constituted by subsequently lagged copies of the original sequence. This corresponds to the generation of the so-called embedding matrix (EM). The EM columns are, in order: *a*), the original series; *b*), the series shifted by one amino acid; *c*), the series shifted by two amino acids; *d*), etc. . .until a dimension variable from three to eight consecutive shifts is reached. Thus, the EM is a multivariate matrix whose rows (statistical units) are subsequent patches (or sliding windows) of amino acids with length equal to the embedding dimension, and whose columns (statistical variables) are the whole sequence lagged by subsequent delays. EM is an $M \times N$ matrix, with M being the number of amino acids minus the embedding dimension (the last amino acids are eliminated by the shifting of the series due to the embedding procedure), and N the embedding dimension.

The notion of recurrence, at the basis of this technique, is well established (Kac, 1959). For any ordered series (temporal or spatial), a recurrence is defined as a point which repeats itself. Because recurrences are simply

TABLE 1 Groups used for analysis

Group *A* (almost pure α-helix)	
PDB Code	
1A4F	Bar-headed goose hemoglobin
1CLL	Calmodulin homo sapiens
1AIN	Annexin 1 (Human)
2ABK	Endonuclease III
1GH0	Phycocyanin α (Spirulina)
101M	Sperm Whale myoglobin
1A2F	Cytochrome c peroxidase
3ATJ	Horseradish peroxidase
1FXKA	Prefoldin chain A (methanobacter termoautrophicum)
3PGHA	Cyclooxigenase-2- (*mus musculus*)
Group *B* (almost pure β-sheet)	
PDB Code	
1A4A	Azurin
1QR4	Tenascin
1HOE	α-amylase inhibitor
1PLC	Plastocyanin
1K5J	Nucleoplasmine
1G13A	Ligand binding protein
1K42	Xylanase
1l5C	Neurophysin1
1F53	Toxin yeast
1G6E	Antifungal
Group *C* (mainly α-helix polymerizing; not necessarily fibrillating)	
Swiss-Prot Code	
P02453	Collagen α 1(I) chain
P04258	Collagen α 1(III) chain
P12107	Collagen α 1(XI) chain
P14106	Complement C1q subcomponent, B chain
P17657	Cuticle collagen dpy-13
P34804	Cuticle collagen 40
Q16987	Fibroin-3
Q16988	Fibroin-4
Q26427	Fibroin light chain
Q99050	Fibroin heavy chain
Group *D* (aggregating system; i.e., amyloids, serpins)	
Swiss-Prot Code	
O35430	Amyloid β A4 precursor protein-binding family A member 1
P30740	Leukocyte elastase inhibitor
P01012	Ovalbumin
P05120	Plasminogen activator inhibitor-2
P50453	Cytoplasmic antiproteinase 3
P29508	Squamous cell carcinoma antigen 1
P48595	Bomapin
Q95241	Amyloid β A4 protein
O00213	Amyloid β A4 precursor protein-binding family B member 1
O96018	Amyloid β A4 precursor protein-binding family A member 3
Group *E* (natively unfolded proteins)	
Swiss-Prot Code	
P10622	Acidic ribosomal protein P1 (133069)
1G5MA	bcl2 antiapoptotic protein (1378693)
KAB0SB	α-5 casein (107620)
P38963	Cyclin-dependent kinase inhibitor p21 (729143)
P17639	Embryonic abundant protein from carrot (119316)
P06302	Prothymosin α (135836)
P25912	Max protein (126776)
P04637	P53 (129361)
P27088	DNA repair protein XP-A cells (139817)
P15340	Protamine (123705)
Group *F* (protein undergoing a lot of interactions with other proteins; i.e., DNA repair systems)	
Swiss-Prot Code	
P46100	Transcriptional regulator ATRX
Q9UUA2	DNA repair and recombination protein pif1
P07271	DNA repair and recombination protein PIF1 (yeast)
P12689	DNA repair protein REV1
P06839	DNA repair helicase RAD3
Q00578	DNA repair helicase RAD25
P19447	TFIIH basal transcription factor complex helicase XPB subunit
O08811	TFIIH basal transcription factor complex helicase subunit
P13010	ATP-dependent DNA helicase II, 80 kDa subunit (human)
P27641	ATP-dependent DNA helicase II, 80 kDa subunit (mouse)
Group *G* (α-helices, synthetic)	
Group *H* (β-sheets, synthetic)	
Group *I* (Greek π-helix)	
PDB Code	
1HRK:A	Human ferrochelatase, chain a
1HRK:B	Human ferrochelatase, chain b
1QGO	Anaerobic cobalt chelatase
1QH8:A	Nitrogenase Mo-Fe protein, chain a
1QH8:B	Nitrogenase Mo-Fe protein, chain b
1QH8:C	Nitrogenase Mo-Fe protein, chain c
1QH8:D	Nitrogenase Mo-Fe protein, chain d
1YGE	Lipoxygenase-1
1A8E	Human serum transferrin
1DOZ	Ferrochelatase

tallies, they make no mathematical assumptions. Given a reference point, $\mathbf{X}_0$, and a ball of radius r, in an N-dimensional space, a point is said to recur if

$$\mathbf{B}_r(\mathbf{X}_0) = \{\mathbf{X} : \|\mathbf{X} - \mathbf{X}_0\| < r\}. \quad (1)$$

The pairwise distances in hydrophobicity space between all the N-amino-acids-long subsequent windows (the M rows of EM) are computed, and all the distances smaller than r are scored as recurrent. The radius, r, thus seeks to group approximately similar values. The procedure can be considered as a search for similar patches in terms of hydrophobicity along the chain as measured by calculating the Euclidean distance for each pairwise comparison. The application of this computation produces a recurrence plot (RP), i.e., a symmetrical $M \times M$ array in which a point is placed at (i, j) whenever a point $\mathbf{X}_i$ is close to another point $\mathbf{X}_j$. Graphically this can be indicated by a dot. Thus RPs simply correspond to the distance matrix between the different epochs (rows of EM) filtered by the action of the radius to a binary 0/1 matrix, with 1 (dot) for distances falling below the radius and 0 for distances greater than the radius. Fig. 1 reports the RP for AcP, with the identification of typical features (see below).

Because graphical representations may be difficult to evaluate, Webber and Zbilut (1994) developed several strategies to quantify features of such plots originally pointed out by Eckmann et al. (1987). The quantification of recurrences consists in the generation of six variables:

1. REC: percent of plot filled with recurrent points.
2. DET: percent of recurrent points forming diagonal lines, with a minimum of adjacent points equal to the predefined parameter line.
3. ENT: Shannon information entropy of the line length distribution.
4. MAXL: length of longest deterministic segment.
5. TREND: measure of the rate of recurrent points away from the central diagonal expressed as the slope of the linear function linking identity in time and number of recurrences.
6. LAM: percent of recurrent points forming vertical lines (Marwan et al., 2002), which are exactly repeating points, indicating laminar transition points.

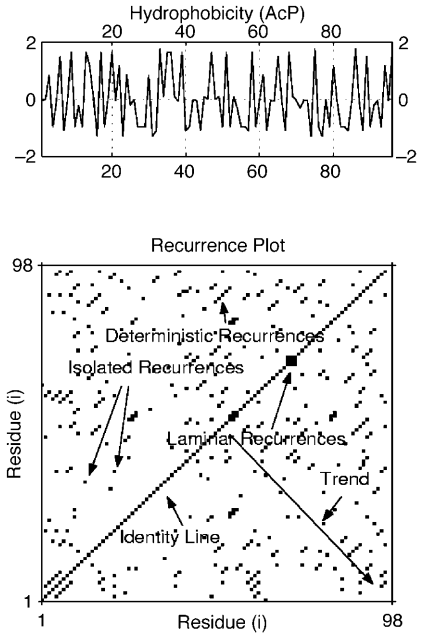

FIGURE 1 Recurrence plot of AcP (*bottom*) with typically quantified features, compared to a regular plot of hydrophobicity (*top*).

These six indexes give a summary of the autocorrelation structure of the series.

The application of RQA implies the a priori setting of the measurement 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: *embedding dimension*, 3; *radius*, 6 (first minimum of DET as determined by a plot of the radius from 0 to 100; see Fig. 8, below), and *line*, 2 (Strait and Dewey, 1996; Giuliani et al., 2002; Zbilut et al., 2002; Weiss et al., 2000).

RQA can be applied either in a global mode (as in Fig. 1), taking into consideration all the recurrences of a given sequence at once; or in a sliding window mode, by which local changes in RQA values can be measured. Another possibility, exploited in our analysis of different protein classes, is the computation of the recurrence spectrum, i.e., the scoring of the number of recurrences found between windows centered at amino acids separated by different lags along the chain. In this case, analogous to a Fourier spectrum, the data of interest are the profiles of recurrences recruited at different distances that correspond to the periodicities (waves of correlation) of hydrophobicity distribution along the chain. This mode is called recurrence quantification intervals (RQI). The RQA software can be freely downloaded from http://homepages.luc.edu/~cwebber/.

The dynamic approach

Although the statistical information from the static analysis can be suggestive, an attempt to compare the information with a clearly aggregating protein was sought via molecular dynamics simulations (MDS). Aβ(1–40) is such a protein, and, moreover, its relatively short length permits relatively facile MD manipulation and analysis.

MD simulations

The amyloid β-peptide, Aβ(1–40) (Serpell, 2000), was investigated in aqueous solutions at low, medium (pH range 2–4 and 5–6, respectively), and neutral pH by three molecular dynamics simulations in an N (number of particles), V (volume), T (temperature) ensemble at normal conditions. The starting configuration was taken from the eighth nuclear magnetic resonance (NMR) model, obtained from PDB, entry code 1BA4, which is the closest to the average NMR structure (Coles et al., 1998).

The different pH environments were created by changing the protonation state of the ionizable residues according to their pK_a. Thus, Glu and Asp residues were negatively charged at medium and neutral pH, and His residues were positively charged at low and medium pH. Moreover, Lys and Arg residues were positively charged under different pH conditions. The site of protonation of all His residues was based on an analysis performed by the program WHATCHECK (Hooft et al., 1996) To select the protonated nitrogen, the structures were checked for the presence of possible hydrogen bonds by looking at the closest hydrogen bond receptor. Histidines 6 and 13 were protonated at the N_ε position and residue 14 at the N_δ position.

Each peptide was immersed in a rectangular box of pre-equilibrate SPC water molecules (Berendsen et al., 1981). Periodic boundary conditions were adopted and the simulations have been performed at constant temperature using the Berendsen thermal coupling (Berendsen et al., 1984). To achieve charge neutrality of systems, counterions, Na and Cl, were added by replacing water molecules at the most negative and positive, respectively, electrical potential. In Table 2 the composition of the simulated systems is reported.

The SHAKE algorithm was used to constrain bond lengths (Ryckaert et al., 1977) The long-range nonbonding interactions were treated with the particle-mesh Ewald method (Darden et al., 1993). First, the solvent was subject to 1000 cycles of minimization using steepest descent method, followed by 2000 steps of MD with harmonic positions restraints applied to all heavy atoms. Next, the position restraints were removed and each of the systems were minimized and then gradually heated from 50 to 300 K in a stepwise manner for 50 ps. After the preparatory steps described above, each of the systems were simulated for 10 ns using a 2-fs integration timestep. All MD trajectories were generated by using the GROMACS software packages and the GROMOS96 force field (van der Spoel et al., 1995).

Individual trajectories are defined as follows: Aβ(1–40) simulations at low, medium, and neutral pH are referred to as AB40L, AB40M, and AB40N, respectively.

Strategy of analysis

To compare the different simulations, the structures were classified according to Jarvis-Patrick method by projections into root-mean-square deviation (RMSD) space (Jarvis and Patrick, 1973). The obtained structures were clustered by means of the Jarvis-Patrick algorithm as applied to their RMSD values. The method allocates two structures into the same cluster if they are reciprocal *first neighbors* and share at least three *common neighbors*. The criterion for two structures to be considered as first-neighbors is simply being among the first 10 structures with the most similar RMSD. Each simulation is sampled every 30 ps and lasts 10,000 ps, and the results are expressed in terms of the subsequent visits of the trajectory to the clusters. This allows for an immediate appreciation of the relative configurational stability of the studied trajectory. An MD simulation remaining for all the simulation period in the same (or few) cluster points to a very stable situation and consequently to a very low number of configurational transitions. On the contrary, an MD simulation characterized by an elevated number of clusters and a rich dynamics between different

TABLE 2 Composition and definitions of the systems

Peptide	pH	Ionization state of residues	Water molecules	Counterions	Definitions
Aβ(1–40)	2–4	Asp^0, Glu^0, His^+	2744	−6	AB40L
Aβ(1–40)	5–6	Asp^-, Glu^-, His^+	2752	–	AB40M
Aβ(1–40)	7	Asp^-, Glu^-, His^0	2750	+3	AB40N

configurations (clusters) during the simulation period points to a very flexible system.

To characterize the structural differences among different trajectories, the RMSD per residue with respect to the NMR structure and secondary structure content were calculated within each cluster. The analysis of secondary structure was done with the DSSP program (Kabsch and Sanders, 1983).

RESULTS AND DISCUSSION

The static approach

Discrimination of different folding behavior on the basis of recurrence quantification intervals spectrum

As has been indicated, evidence that hydrophobicity distribution along a protein sequence, and thus the possibility of inferring both structural and functional features of the proteins by considering the sequence as a time series of some carefully selected chemicophysical property of the constituent amino acids, has had a long history. The great majority of these studies exploited single homogenous series of proteins, trying to predict, by the analysis of hydrophobicity patterning along the sequence, some differential property of the series members (e.g., thermal stability, enzymatic efficiency). In the present case we decided to test the potential of the method as ''structural/functional class identifier'' in an heterogeneous set of proteins. The recognition of the ability of the method to discriminate between different protein classes is not only the demonstration of the practical utility of the method to assign a putative function (or structure) to an unknown sequence but, more importantly, a clue into the possible mechanism of protein folding and aggregation. For these reasons we tested our method on the classes listed in Table 1. The classes were chosen with the goal of comparing, in terms of hydrophobicity distribution, the self-aggregating systems (group *D*) with other general models of protein-protein aggregation as well as with specific secondary structure motifs that were described as typical of aggregating systems.

Both artificial α and β structures (groups *G* and *H*) were selected as *synthetic* (and thus extremely clean in terms of hydrophobicity distribution) examples of the two main secondary structure motifs. The relative similarity of the amyloid system profile to one of these two poles may have the meaning that one of these two ''ideal'' motifs should be more fibrillating-prone than the other. The same reasoning is at the basis of the choice of groups *A* and *B* but with more of an accent on real, natural, α and β structures that are much more noisy (in terms of hydrophobicity distribution) than the artificial polypeptides. Group *C* is made of proteins giving rise to large supramolecular structures such as natural silk or collagen and represent a clear example of self-aggregating systems. At odds with amyloidlike structures, these structures are mainly extracellular and are made of an extremely high number of monomers. Polymerizing systems are made of very repetitive patterns of amino acids whereas the majority of natural proteins have quasirandom sequences.

Group *E* proteins come from the Dunker list of natively unfolded systems (Dunker et al., 2002). They are proteins completely unfolded in solution and perform their physiological roles through order-disorder transitions. The Dunker group has noted that the great majority of these unfolded systems is involved in many protein-protein or protein-DNA interactions. If amyloid systems should have a recurrence spectrum analogous to these systems we could hypothesize a link between molecular flexibility and the propensity to form intermolecular links. The same reasoning holds true for group *F* proteins that, at odds with other groups, is not a structural but a functional classification: proteins involved in DNA repair and, in general with DNA transcription regulation, are known to work through the formation of aggregates of different protein species. The similarity of group *D* amyloid-forming species with proteins of this group could be another indication of a common ''signature'' of aggregation.

When submitted to RQI, these proteins gave the results indicated in Fig. 2. From the figure it is evident how self-aggregating systems (group *D*) are very similar to DNA processing proteins (*F*), and (*I*) π-helix proteins characterized by a scale-free flat spectrum devoid of major peaks. The apparent connection between amyloid and DNA-related enzymes may be linked to the capability of forming multimeric aggregates of globular proteins. The extracellular polymerizing systems (*C*) point to a completely different pattern of hydrophobicity distribution. The two synthetic groups (*G* and *H*) show very clearcut peaks. This is consistent with the aim of the formation of these artificial

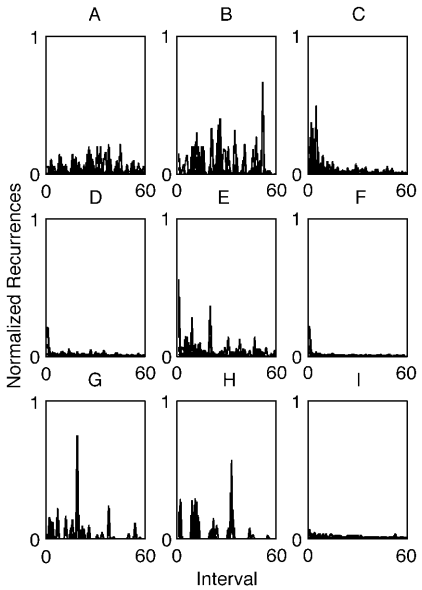

FIGURE 2 Results (histograms) for interval analysis of groups. The *y*-axis represents (unit normalized) tallies. The *x*-axis indicates interval length. *A*, Natural α-helix; *B*, natural β-sheet; *C*, polymerizing; *D*, amyloid; *E*, natively unfolded; *F*, DNA repair proteins; *G*, synthetic α-helix; *H*, synthetic β-sheet; and *I*, π-helix.

proteins, based on exactly repetitive motifs forming α-helices and β-sheets, respectively. To substantiate these qualitative observations it is important to demonstrate that the recurrence spectrum differences are sufficient to achieve a good discrimination between classes of proteins by means of a quantitative procedure. This was actually the case, when submitted to a canonical discriminant analysis (CDA).

CDA is a variant of a general linear model which has been used extensively in the natural sciences (Ortiz and Skolnick, 2000), such that $\mathbf{Y} = \mathbf{XB} + \mathbf{e}$, where $\mathbf{Y}$ is a matrix of dependent variables, $\mathbf{X}$ is a matrix of independent variables, $\mathbf{B}$ is the matrix of regression coefficients, and $\mathbf{e}$ is a matrix of random errors. This model is used to calculate the variables which can best separate cases into predefined groups. These variables are called canonical variates and are new synthetic variables, orthogonal to each other, that maximize the between groups distance.

The first 60 recurrence intervals (this being the shortest molecule), with tallies being normalized as a percent of total recurrences, were submitted to CDA. The result was an almost perfect discrimination of the 90 proteins as depicted in Table 3.

The canonical variables are extracted in order of discrimination power, thus, concentrating on the first four canonical variates we can have an idea of the mutual relationships of the groups in the recurrence spectrum space. Fig. 3 reports the two most important axes for protein discrimination: basically score 1 is a measure of shape of the recurrence periodicity. *G* and *B* classes, respectively corresponding to synthetic α-helices and natural β-sheets, are situated at the extremes of the axis: both situations correspond to very regular arrangement of hydrophobic/hydrophilic residues. Natural β-sheets, with longer periodicities than αs, however, are broken up in their natural setting. The *H* group (artificial βs) is clearly unique in that it is the most periodic, and clearly defines score 2 as spanning a regularity index. This opposition is of scarce interest for natural amyloid proteins that are posited in the center of the axis (*D*). This first analysis confirms the hypothesis that fibril-forming systems are not particularly specialized (like synthetic β-sheets or collagen-like systems)

TABLE 3 Classification matrix (cases in row categories classified into columns)

	A	B	C	D	E	F	G	H	I	% Correct
A	10	0	0	0	0	0	0	0	0	100
B	0	10	0	0	0	0	0	0	0	100
C	0	0	10	0	0	0	0	0	0	100
D	0	0	0	8	0	2	0	0	0	80
E	0	0	0	0	10	0	0	0	0	100
F	0	0	0	2	0	8	0	0	0	80
G	0	0	0	0	0	0	10	0	0	100
H	0	0	0	0	0	0	0	10	0	100
I	0	0	0	0	0	0	0	0	10	100
Total	10	10	10	10	10	10	10	10	10	96

Wilk's Lambda, $p < 0.00001$.

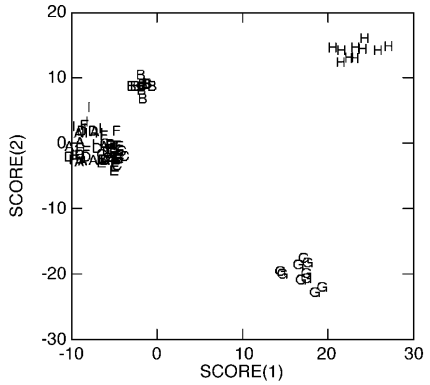

FIGURE 3 Results of group classification based on first two CDA scores.

but represent relatively normal proteins (Dobson, 2003; Chiti et al., 2002).

Score 3 (Fig. 4) models the opposition between collagen-like and artificial α systems at one hand and all the other proteins at the other hand. Again there is an opposition between ''extremely regular'' and ''irregular'' distributions, under the point of view of α-helix signature of hydrophobicity distribution.

Score 4 represents a sort of fine-tuning of different spectra: amyloid systems are particularly near to both DNA repair systems and synthetic β-sheet peptides that in fact have the possibility to fibrillate (West et al., 1999). Interestingly enough, natively unfolded systems (*E*) are at the opposite extreme of this axis with respect to natural α-helices (*A*). This suggests that the recipe for building a natural amyloid system should mix some features of both α-helices and natively unfolded systems. Basically this recipe is not different from the recipe of proteins that for their normal behavior generate supramolecular complexes like DNA processing systems (*F*).

To synthesize all these observations, we computed a *k*-means cluster analysis on the data set constituted by the 90 proteins as statistical units and the first four canonical variates as variables. *K*-means clustering splits a set of objects into a selected number of groups by maximizing

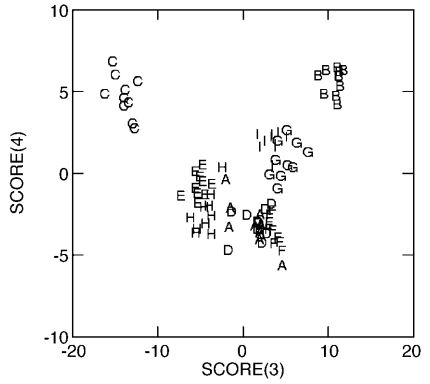

FIGURE 4 Classification results based on third and fourth CDA scores.

between-cluster variation relative to within-cluster variation. It is similar to doing a one-way analysis of variance where the groups are unknown and the largest F value is sought by reassigning members to each group. K-means starts with one cluster and splits it into two clusters by picking the case farthest from the center as a seed for a second cluster and assigning each case to the nearest center. It continues splitting one of the clusters into two (and reassigning cases) until a specified number of clusters are formed. K-means reassigns cases until the within-groups sum of squares can no longer be reduced. The Euclidean distance was used to determine distance variation. The results are presented in Table 4. The natural clusters in the four dimensional space share a strict relation with the a priori folding groups. What is interesting, is the fact that the D group of amyloidlike proteins goes together with natural α-helices (A), DNA repair enzymes (F), and π-helix proteins (I).

The four-canonical-variates solution is not able to fully discriminate the protein space, the full discrimination being attainable with seven canonical variates; but, given the hierarchical character of canonical analysis (with the subsequent canonical variates explaining progressively finer details of the entire picture), it allows us to get a picture of the relative similarities between classes. The fact that the analysis fuses together the amyloid-forming systems with pure α-helices (A), and DNA processing enzymes (F) gives us a clear message on the nature of protein-protein aggregation process taking place in the amyloidlike systems:

1. The β-sheet pattern is not a prerequisite for amyloid-forming systems but is probably a consequence of the aggregation process.
2. Oligomerization has a pivotal role in amyloid-forming system and has a specific signature common with systems undergoing the formation of oligomers for their physiological activity like DNA repair enzymes.
3. The formation of high polymers of the collagen type, although having an obvious resemblance with protein-protein aggregation in amyloidosis, probably follow different mechanistic pathways.

These evidences come from a statistical analysis of protein ensembles, and as suggestive as they are, they are not completely clear. The clustering of *ADFI* suggests some commonalities, but not beyond obvious initial classification. To understand these general conclusions, we now shift to a more local view.

TABLE 4 Composition of clusters

Cluster	Group
One	A,D,F,I
Two	G
Three	H
Four	B
Five	C
Six	E

Identification of aggregation hotspots

In a previous article we obtained preliminary evidence for the possibility of equating singularities in determinism along the sequence to *aggregation hotspots*. This link between aggregation hotspots and deterministic singularities came from the analysis of two prion-like 36-mers where it was demonstrated that the scaling of determinism with radius had a very clear peak at very low radius corresponding to the presence of a very high interaction probability confined to a very specific portion of the sequence (Zbilut et al., 2000). This peak was present in the case of aggregation-prone peptides and suddenly disappeared when sequence was randomly shuffled (Fig. 5). The presence of such singularities in determinism scaling was demonstrated for Syrian hamster PrP protein as well.

The implications would seem straightforward: contrary to the impressions of hydrophobicity plots, which suggest no remarkable features, RQA demonstrates that there is a definite, pronounced structure in terms of hydrophobicity patterning (high values of DET). This structuring should be understood in terms of a repetitive hydrophobic/hydrophilic pattern, and not simply as a region of uniform hydrophobicity values. What is more striking is the narrowness of the shelf and concomitant dropoff. This would imply that local contacts predominate.

An almost ideal model system to confirm these findings is represented by human AcP, whose aggregation propensity was carefully analyzed by Chiti et al. (2002). These authors demonstrated the presence of mutational *aggregation zones* along AcP corresponding to the sequence in the 16–31 and 87–98 residues range: only mutations intervening in these portions of the sequence are capable of significantly influencing the aggregation behavior of the protein. When looking at the hydropathy profile of AcP, no unique feature of the curve characterizes this site. On the contrary, when submitted to the windowed version of RQA with *delay* as 1, *emb* as 3, *epoch* (window) as 28, *overlap* as 27, *shift* = 1, *scaling*

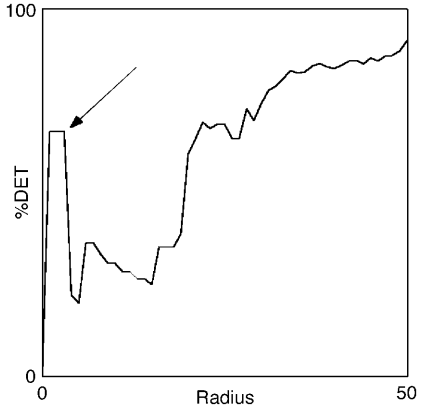

FIGURE 5 Singularity for PrP. The singularity is defined as a narrow, relatively high value for DET in the low radius region, followed by a steep dropoff for subsequent DET values as the radius increases.

as unit normalization, and *radius* as 30, a unique determinism peak is evident at one of the aggregation-sensitive portions of the sequence (Fig. 6).

This result was augmented by the analysis of the mutations enumerated by Chiti et al. (2002): using a strategy we have previously employed (Zbilut et al., 1998), the change in recurrence values for TREND, ENT, MAX, and LAM were computed for all given mutations, and found to be significantly different between the aggregation zones and folding zones ($p < 0.05$; *t*-test). When submitted to CDA, the results were also significant ($p = 0.0006$; Wilk's Lambda) for the variables DET, ENT, TREND, and LAM. These results were tested with a jackknifed procedure to confirm robustness of the procedure (Table 5). The plot (Fig. 7) of these factors makes clear that TREND and LAM are approximately opposed to each other at the first factor axis, with fine-tuning occurring in the second factor with DET and ENT. Thus an interpretation of TREND/LAM opposition suggests a *smoothness* dimension. The *smoothness* here refers to nearly identically repeated patches.

These results point to a clear role of hydrophobicity structuring of sequence as a major determinant of aggregation behavior of molecules. This structuring intervenes both in terms of repetitions of complex patterns of hydrophobicities in different portions of the sequence (DET, ENT) and in terms of repetitions of particular motifs (LAM) and presence of singularities.

Chiti et al. (2002) were able to finely tune, through mutational analysis, the zones of the molecule relevant for folding and the zones relevant for aggregation behavior. In an attempt to interpret their data in light of the relative order/disorder status of the different zones along the sequence by making use of the Dunker et al. (2002) PONDR computation for the determination of disordered areas of AcP (Romero et al., 2001). The results obtained with the PONDR unfolding predictor algorithm were compared with the AcP RP in Fig. 8. What becomes immediately apparent is that the ''disordered'' zone approximately encompasses residues 30–75, which excludes the aggregationally important zones, but approximates the zone relevant for folding. Moreover, the RP highlights these same areas by darkened patches of laminarity.

TABLE 5 Classification matrix (cases in row categories classified into columns)

			% Correct
	Classification matrix		
1	19	3	86
2	0	12	100
Total	19	15	91
	Jackknifed classification matrix		
1	19	3	86
2	0	12	100
Total	19	15	91

Additionally, RQI was performed on both groups of folding- and aggregation-affecting mutations to determine if there were any preferred interval lengths that affect aggregation behavior. This result is immediately related to the statistical comparisons described in the first section, allowing us to go from a general statistical perspective to a particular mechanistic one.

Fig. 9 reports the recurrence spectra relative to folding and aggregation mutants of AcP. As can be seen from the figure, the two behaviors correspond to different characteristic recurrence intervals. CDA analysis as performed for the previous nine groups was repeated for AcP. The procedure was able to distinguish the folding/aggregation zones again to a significant level ($p = 0.03$) (Table 6).

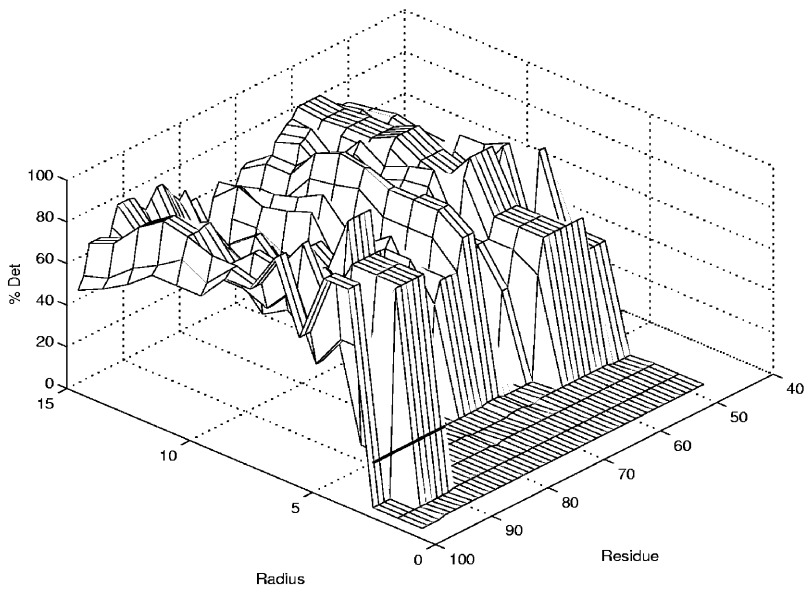

FIGURE 6 To evaluate the local change of determinism along a protein sequence, a form of RQA can be performed similar to the windowing procedure common in spectral analysis. Windows of variable size are stepped through the sequence, overlapping one residue at a time. This results in a three-dimensional graph of *radius vs. DET vs. residue number*. Shown is such a plot of DET for AcP. The ''aggregation zones'' as identified by Chiti et al. (2002) are identified by the solid bar. Note that they are characterized by singularities. Here, residues 87–98 are highlighted.

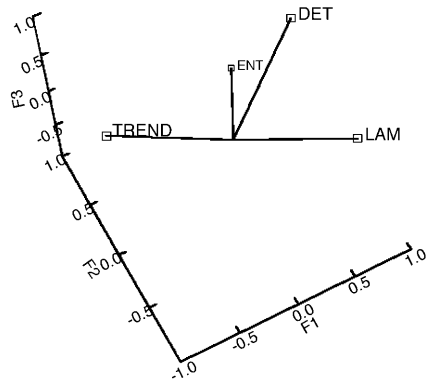

FIGURE 7 Factor loadings plot. Note near opposition between TREND and LAM.

Dynamic approach

We analyzed the Aβ(1–40) peptide in three different pH situations known to have different propensities for the aggregation process (Kirkitadze et al., 2001). To confirm the basic points raised by the sequence analysis, we should demonstrate that the *aggregation-favoring* condition results in a richer conformational flexibility with respect to the other conditions. The measure of conformational flexibility is the number of conformational clusters typical of each trajectory and the number of flipping between different clusters (see Material and Methods section for the computation of conformational clusters along the trajectories). Tables 7–9 summarize the results of the different simulations, whereas Fig. 10 expresses the information as a time series.

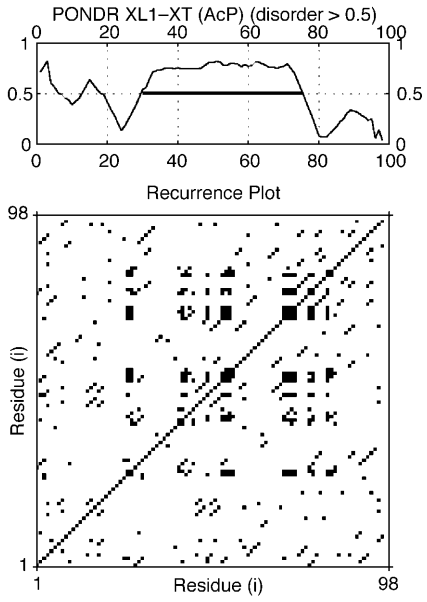

FIGURE 8 RP of AcP (*bottom*) compared with PONDR results (*top*). Note the approximate concordance between the patchy areas of the RP and the disordered area of the PONDR plot.

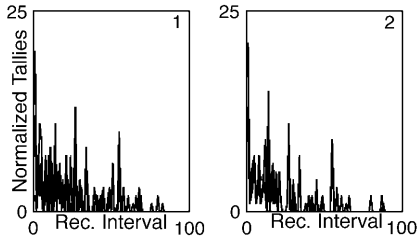

FIGURE 9 Comparison of RQI interval histograms for folding (1) vs. aggregating areas (2) of AcP according to Chiti et al. (2002). The *y*-axis reports normalized tallies.

As evident from the tables, the situation with by far the highest propensity for fibrillation (AB40M) presents the richest conformational dynamics: the state space of the trajectory can be subdivided into 10 clusters among which the system oscillates during the 10,000 ps of the simulation. On the contrary, the most prohibitive environment for fibrillation (AB40N) has only one cluster and the AB40L situation is characterized by five clusters which the trajectory progressively visits without any oscillation.

It is worth noting that the AB40M trajectory, at odds with the other two simulations, presents as a dominant structure (at approximately the same frequency as the α-helix), the structure called π-helix. This is considered a meta-stable helix which was already noted in the Aβ(1–40) peptide and marks systems with an high propensity for conformational transitions (Fodje and Al-Karadaghi, 2002). Thus the dynamical approach is globally in line with the results of the static analysis relative to the nexus between aggregation/folding and conformational flexibility. The conformational flexibility adds another degree of stochastic variability.

The other point to emphasize in the case of Aβ(1–40) is the previously identified correspondence between disordered areas with determinism punctuated by short laminar areas. As can be seen from Fig. 11 and the record of the molecular dynamics simulation, laminar areas of the RP (roughly 5–15; 20–30) exhibit the largest numbers of conformational changes. This is not a contradiction to the above statement regarding AB40M as the richest source of conformational dynamics. This is to say that irrespective of the overall cluster dynamics, the patchy areas are significant sources for

TABLE 6 Classification matrix (cases in row categories classified into columns)

			% Correct
	Classification matrix		
1	20	2	91
2	0	12	100
Total	20	14	94
	Jackknifed classification matrix		
1	20	2	91
2	3	9	71
Total	23	11	85

TABLE 7 Average number of residues in a given secondary structure of Aβ(1–40) calculated within each cluster of AB40L trajectories, according to DSSP software

		Secondary structure					
Clusters	Time period	Coil	Bend	Turn	α-helix	π-helix	3_{10}-helix
	Starting structure	9	5	4	22		
I	0	12.4 (1.7)	8.8 (2.5)	4.9 (2.7)	12.8 (3.9)	0.6 (1.6)	
II	6090	11.6 (1.7)	10.1 (1.7)	5.7 (2.6)	11.0 (0.6)		1.1 (1.4)
III	8670	11.5 (1.7)	9.5 (1.7)	5.0 (1.0)	11.0 (0.0)		3.0 (0.0)
IV	8730	10.7 (0.9)	10.3 (1.7)	4.6 (2.2)	11.0 (0.5)	0.7 (1.7)	2.7 (0.9)

Standard deviations are given in parentheses. The corresponding reference values in the NMR structure (referred to as starting structure) are also reported.

the motions. The conjecture was confirmed by a Pearson correlation analysis of the three MD simulations with the laminar patches (AB40L, $r = 0.735$, $p = 0.001$; AB40M, $r = 0.555$, $p = 0.007$; AB40N, $r = 0.569$, and $p = 0.005$; Bonferroni adjusted probabilities, see also Fig. 12). This tends to substantiate the recent observation of Satheeshkumar and Jayakumar (2003), that the prion protein (113–127) exhibits polymorphic behavior, especially at acidic pH. In this respect, the correlation between deterministically laminar regions with MDS histograms is highest for the acidic environment. Thus segmental motion does not necessarily translate into structural homogeneity. Of note is the appreciation of a π-helix in the 27–35 region which is emphasized by a plot of laminarity (Fig. 13). This strong possibility is supported by research suggesting kinetic intermediate helical structures (Gzit, 2002).

Additionally, the salt bridge suggested in the middle of the peptide (24–27) which binds both sheets, acting possibly as a hinge mechanism, is also easily seen in terms of a laminar singularity of the hydrophobicity distribution in Fig. 14 (Petkova et al., 2002; Thompson, 2003; Ma and Nussinov, 2002). These results constitute a further link between punctuated laminarity and disordered regions.

CONCLUSIONS

For some time it has been recognized that hydrophobicity remains an important physicochemical variable in the protein folding problem. A major difficulty has been the appropriate characterization of this variable. Hydrophobicity plots, although suggestive and useful for strictly periodic structures, have been less informative for intermittent patches. The difficulty has been compounded by lack of systematic analysis of specific proteins for their folding/aggregating properties derived from mutational analysis.

In the present study, both these limitations have been obviated. Specifically, use was made of a signal analysis technique which overcomes the limitations of traditional techniques dependent upon requirements of stationarity and relative periodicity. Secondly, the elegant work of the Oxford/Florence groups on AcP by means of site-directed mutagenesis, has presented specific evidence for the existence of preferential areas important for folding vs. aggregation. We used this information to systematically analyze hydrophobicity patterns via a ''static'' approach, and carrying over these observations to another molecule important in aggregation and fibril formation via a ''dynamic'' approach.

TABLE 8 Average number of residues in a given secondary structure of Aβ(1–40) calculated within each cluster of AB40M trajectories, according to DSSP software

		Secondary structure					
Clusters	Time period	Coil	Bend	Turn	α-helix	π-helix	3_{10}-helix
	Starting structure	9	5	4	22		
I	0	7.5 (2.1)	7.4 (2.2)	3.0 (2.6)	19.7 (2.7)	2.3 (2.9)	
II	2790	8.2 (1.0)		6.7 (1.4)	2.2 (1.8)	14.2 (1.8)	8.5 (2.6)
III	3990	7 (0.0)	8 (0.0)	2 (0.0)	14 (0.0)	9 (0.0)	
IV	4020	7.3 (1.0)	7.5 (1.2)	3.0 (1.7)	16.2 (2.1)	5.9 (2.5)	
V	5250	8.4 (1.0)	6.0 (1.5)	3.7 (2.5)	12.3 (3.8)	9.5 (4.9)	
VI	7500	7 (0.0)	7 (0.0)	4 (0.0)	12 (0.0)	10 (0.0)	
VII	7530	8.5 (0.5)	6.0 (1.4)	3.5 (2.2)	10.0 (2.5)	12.0 (2.7)	
V	7590	8.4 (1.0)	6.0 (1.5)	3.7 (2.5)	12.3 (3.8)	9.5 (4.9)	
VIII	8490	8 (0.0)	8 (0.0)	1 (0.0)	11 (0.0)	12 (0.0)	
VII	8520	8.5 (0.5)	6.0 (1.4)	3.5 (2.2)	10.0 (2.5)	12.0 (2.7)	
V	8580	8.4 (1.0)	6.0 (1.5)	3.7 (2.5)	12.3 (3.8)	9.5 (4.9)	
IX	9150	8 (0.0)	6 (0.0)	2 (0.0)	7 (0.0)	17 (0.0)	
V	9180	8.4 (1.0)	6.0 (1.5)	3.7 (2.5)	12.3 (3.8)	9.5 (4.9)	
X	9990	7 (0.0)	9 (0.0)	1 (0.0)	9 (0.0)	14 (0.0)	

Standard deviations are given in parentheses. The corresponding reference values in the NMR structure (referred to as starting structure) are also reported.

TABLE 9 Average number of residues in a given secondary structure of Aβ(1–40) calculated within each cluster of AB40N trajectories, according to DSSP software

		Secondary structure						
Clusters	Time period	Coil	β-Sheet	β-Bridge	Bend	Turn	α-helix	π-helix
	Starting structure	9			5	4	22	
I	0	11.3 (1.3)	4.2 (2.8)	0.7 (0.8)	3.8 (1.6)	6.5 (3.1)	11.6 (5.1)	1.9 (3.5)

Standard deviations are given in parentheses. The corresponding reference values in the NMR structure (referred to as starting structure) are also reported.

The results have confirmed the importance of hydrophobicity relative to its patterning along a protein chain. Specifically, an important criterion is what we have labeled as its *smoothness* along the TREND/LAM dimension (Table 10). We derived this observation first upon a comparative analysis of broadly classified protein groups, which was confirmed by further analysis based on the Chiti et al. (2002) results. It becomes apparent that deterministic singularities can become a nucleation center dependent upon how laminar or smooth the singularities are. If they are relatively smooth with respect to deterministic patches, they tend to become important for the possibility of aggregation. This may be due to the fact that residues in such patches maintain a hydrophobic profile which is relatively stable favoring nearby contacts (perhaps based on hydrophobic cores), breaking only at the termini of the patches. Folding zones, on the other hand, are characterized by broken patches of laminarity (made of the repetition of internally very diverse patches in terms of hydrophobicity). These areas are more likely to form connections beyond their immediate (short) patches. Interestingly enough, this profile corresponds roughly to the identification of disordered zones as identified by the work of Dunker et al. (2002) and their PONDR index. The more *patchy* zones, however, tend to be more liable to change (i.e., from disorder to order) because a mutation has a larger probability to stabilize two short deterministic patches.

What are the consequences of the presence of short, singular patches in the hydrophobicity pattern of consecutive residues for the dynamics of the protein structure from a theoretical point of view? We have argued that there exist many physical and biological motions which may be better modeled by non-Lipschitz (nonsmooth) differential equations whereby there is no single solution to the equation. In plain terms, a nonsmooth (and thus non-Lipschitz) approach implies the existence of discrete and identifiable turning-points in which the observed phenomenon instead of approaching in an infinite time to a limit (as in classical differential equations), on the contrary, is abruptly stopped. Examples of these kinds of dynamics are the ''rip'' of a flag or the discharge of a seismic wave. In all these systems we can identify singular points in which the dynamics abruptly ''forget the past'' and re-start with a completely stochastic choice of direction. In this situation the relative probability of moving one way or another (in terms of folding/aggregation)

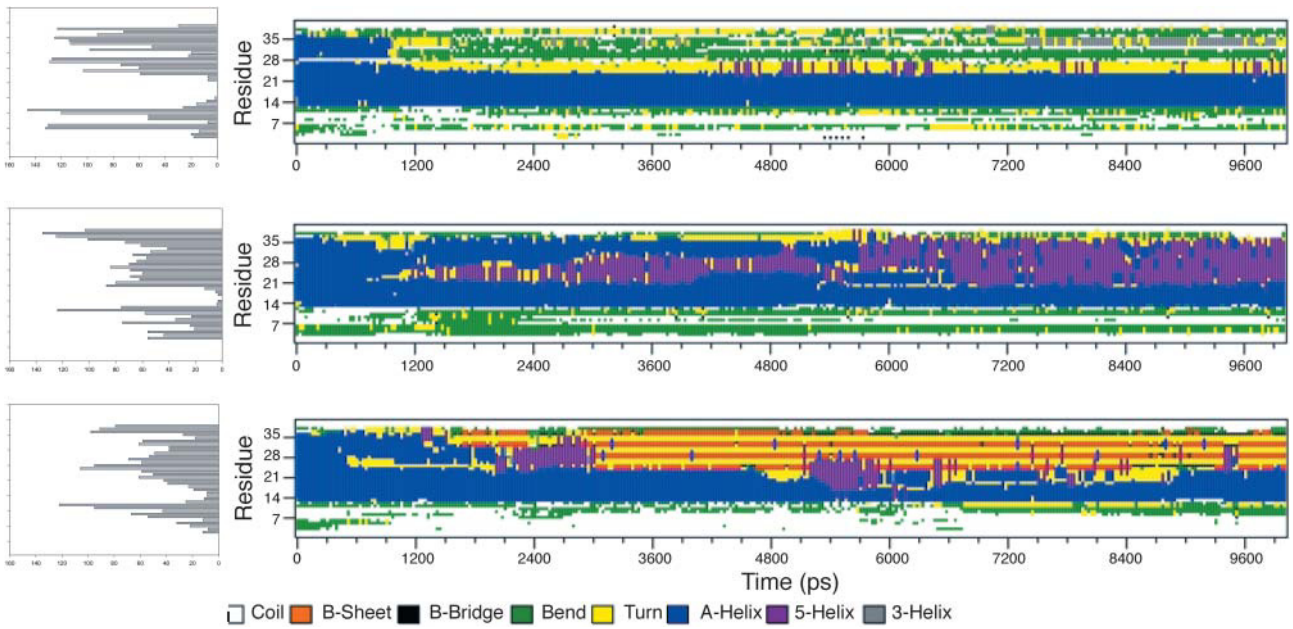

FIGURE 10 Results from molecular dynamics simulations of Aβ(1–40) at low (*top*), medium (*middle*), and normal (*bottom*) pH.

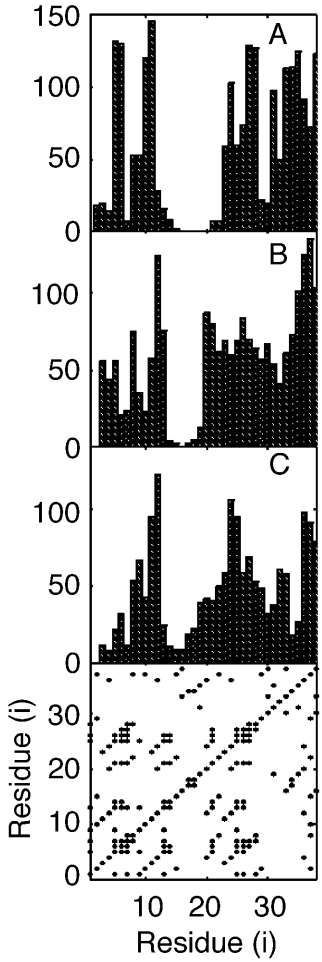

FIGURE 11 RP of Aβ(1–40) (*bottom*), and the three histograms from Fig. 10 above (low pH, *A*; medium pH, *B*; and neutral pH, *C*). Note the approximate correspondence between the most deterministic laminar areas (residues 5–30, with intervening empty patch at ~18-25).

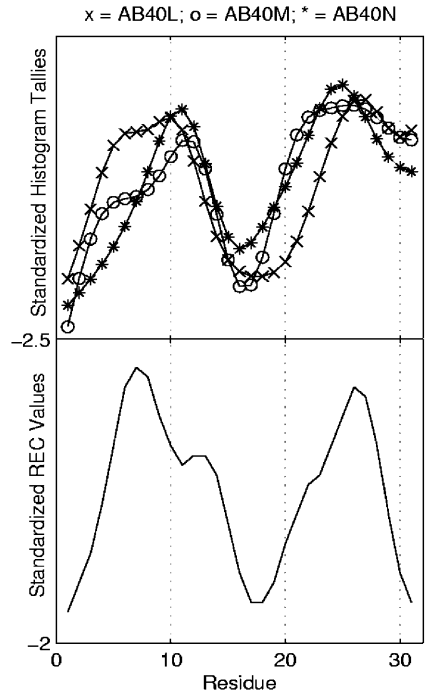

FIGURE 12 Comparison of histograms with deterministic areas. Because the histograms include inherent noisy fluctuation, they were smoothed with a three-point rectangular moving average window, passed 4×. A numerical slice from the RP incorporating their deterministic points was obtained, being coded 1/0 for each residue position depending on the existence of a recurrence. Because this also incorporates a certain degree of fluctuation realized by the radius function for recurrence calculation, these values were similarly passed through the moving average filter. The resulting values were used for the correlation calculations (see text).

becomes a combinatorial one with a resultant *nondeterministic* dynamics, and a *stochastic attractor*. This is to say that there exist unstable singularities with an associated probability distribution regulated by factors unique to the given system. In the protein case, narrow patches of high determinism and/or laminarity at low radius are the basis of this singularity. Although in the present case the discussion centers around a topological feature, without loss of generalization, it can be argued that the energy functions of molecular dynamics behave in a similar way. As a matter of fact hydrophobicity is a partition function governing the relative propensity of being solvated and thus the hydrophobicity profile can be considered as the static recipe giving rise to dynamical behavior of the protein in solution. This is to say that if the process is described as a minimization of the energy function through potential minima, the boundaries can be described as unstable singularities as well (Zak et al., 1997; Chikishev et al., 1998). In the long run, the associated probabilities that may govern whether a given protein will go to its native fold or an aggregation, may depend upon its TREND/LAM characterization. The probabilities themselves are governed by the boundary conditions; i.e., pH, temperature, etc. This view is in line with the one adopted by

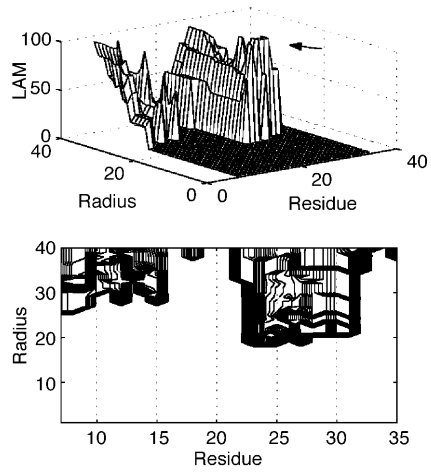

FIGURE 13 RQE plot (as in Fig. 6) (*top*, with its contour version *below*) with increasing radius for LAM. Note singularity (*arrow*) coinciding with appearance of Greek π-helix in MDS.

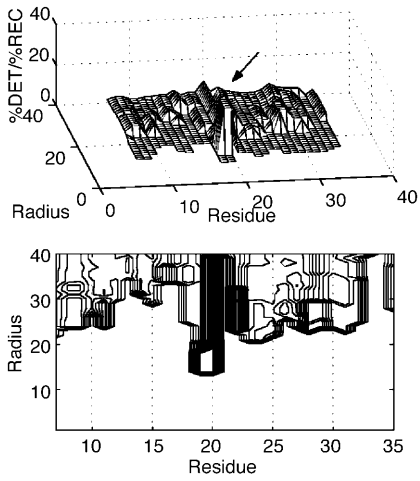

FIGURE 14 RQE plot (as in Fig. 6) (*top*, with its contour version *below*) with increasing radius for DET. Note singularity appearing approximately at area of salt bridge (*arrow*).

Dobson (2003), pointing to the stochastic character of the aggregation process. Indeed, this is the implication of phase diagrams exploring protein aggregation (Dima and Thirumalai, 2002). The suggestion is that segments broken by laminar patches may tend to be disordered, and exhibit more conformational variability, given proper circumstances.

To prove these theoretical statements in the actual case of proteins we need to shift from the sequence analysis perspective to the study of the actual behavior of aggregating systems in solution in terms of both molecular dynamics simulations (MDS) and experimental spectroscopic data. Thus we must shift from what we have called the ''static'' to the ''dynamic'' approach.

These observations are further confirmed by application to the Aβ(1–40) peptide. Specifically, the patchy laminar (folding), relatively smooth deterministic (aggregating) observation continues. Two aggregating areas are identified, with a central disordered area containing a unique deterministic singularity. Additionally, the unequal hydrophobic cores forming fibrils, are further distinguished by the possibility that the longer patch is responsible for an intermediate π-helix as evidenced by MDS.

Thus the picture emerges that the laminarity of protein deterministic patches are a key in determining folding tendencies. This is supported by our previous work implying that mutations have variable effects depending upon the net effect upon the patch (i.e., maintaining the laminarity versus breaking it; Zbilut et al., 1998). Additionally, our work with MDS and the tendencies to form *laminar blocks* is also suggestive (Manetti et al., 2001). Clearly, however, the present results, although highly evocative, are not complete. The results are specific to the studied proteins, and generalization cannot be immediately assumed. Also, additional research in both AcP and Aβ(1–40) have questioned the role of charge or other electrostatic measures (Tycko, 2003; Massi et al., 2002). Further investigation into these areas using descriptive differential equations are currently being evaluated. This investigation could give important clues not only for the prediction, starting from the sequence, of the aggregation propensity of a given system and thus suggesting possible targets for drugs but, on a more speculative but nevertheless very important dimension, for understanding the dynamics of the so-called misfolding diseases. Along this path, Kellershohn and Laurent (2001) clearly demonstrated the non-Lipschitz character of the prion infection and the consequent structural transition, albeit qualitatively. Similarly, Harrison et al. (2001) have proposed a model of ''glassy'' behavior of alternative states for folding behavior. These data are in agreement with our models for AcP and Aβ(1–40), allowing us to speculate the existence of a common mechanism underlying a number of apparently disperse and heterogenous folding/aggregation phenomena.

TABLE 10 Distinctive features of folding vs. aggregation relative to RQA variables

RQA variable	Folding	Aggregation
Laminarity	Patchy	Little or none
Trend	Rugged	Relatively smooth
Determinism	High	Varies

J.P.Z. thanks Prof. C.M. Dobson and Prof. F. Chiti for kindly providing details of experiments, and Dr. Julie C. Mitchell, University of California at San Diego, for useful discussions.

This work was supported by a joint Division of Mathematical Sciences/National Institute of General Medicine Sciences initiative to support mathematical biology, from the National Science Foundation and National Institutes of Health (NSF DMS 0240230); J. P. Zbilut, Principal Investigator.

REFERENCES

Berendsen, H. J. J., J. P. M. Postma, W. F. van Gusteren, A. Di Nola, and J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

Berendsen, H. J. J., J. P. M. Postma, W. F. van Gusteren, and J. Hermans. 1981. Interaction models for water in relation to protein hydration. *In* Intermolecular Forces. B. Pullman, editor. D. Reidel Publishing Company, Dordrecht, The Netherlands. 331–342.

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins.* 21:167–195.

Chikishev, A. Y., N. V. Netrebko, Y. M. Romanovsky, W. Ebeling, L. Schimansky-Geier, and A. V. Netrebko. 1998. Stochastic cluster dynamics of macromolecules. *Int. J. Bifurc. Chaos.* 8:921–926.

Chiti, F., N. Taddei, F. Baroni, C. Capanni, M. Stefani, G. Ramponi, and C. M. Dobson. 2002. Kinetic partitioning of protein folding and aggregation. *Nat. Struct. Biol.* 9:137–143.

Coles, M., W. Bicknell, A. A. Watson, D. P. Fairlie, and D. J. Craik. 1998. Solution structure of amyloid β-peptide(1–40) in a water-micelle environment. Is the membrane-spanning domain where we think it is? *Biochemistry.* 37:11064–11077.

Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an N*log(N) method for computing Ewald sums. *J. Chem. Phys.* 98:10089–10092.

Dima, R. I., and D. Thirumalai. 2002. Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci.* 11:1036–1049.

Dobson, C. M. 2003. Protein folding and disease: a view from the first Horizon Symposium. *Nat. Rev. Drug Discov.* 2:154–160.

Dunker, K., C. J. Brown, D. Lawson, L. M. Iakoucheva, and Z. Obradovic. 2002. Intrinsic disorder and protein function. *Biochemistry.* 41:6573–6582.

Eckmann, J. P., S. O. Kamporst, and D. Ruelle. 1987. Recurrence plots of dynamical systems. *Eur. Phys. Lett.* 4:973–977.

Fodje, M. N., and S. Al-Karadaghi. 2002. Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Eng.* 15:353–358.

Giuliani, A., R. Benigni, J. P. Zbilut, C. L. Webber, Jr., P. Sirabella, and A. Colosimo. 2002. Nonlinear signal analysis methods in the elucidation of protein sequence structure relationships. *Chem. Rev.* 102:1471–1491.

Giuliani, A., and M. Tomasi. 2002. Recurrence quantification analysis reveals interaction patterns in paramyxoviridae envelope glycoproteins. *Proteins.* 46:171–176.

Gzit, E. 2002. A possible role for π-stacking in the self-assembly of amyloid fibrils. *FASEB J.* 16:77–83.

Harrison, P. M., H. S. Chan, S. B. Prusiner, and F. E. Cohen. 2001. Conformational propagation with prion-like characteristics in a simple model of protein folding. *Protein Sci.* 10:819–835.

Hooft, R. W. W., G. Vriend, C. Sander, and E. E. Abola. 1996. Errors in protein structures. *Nature.* 381:272.

Jarvis, R. A., and E. A. Patrick. 1973. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Computers.* C22:1025–1034.

Kabsch, W., and C. Sanders. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.

Kac, M. 1959. Probability and Related Topics in Physical Sciences. Wiley Intersciences, New York.

Kamtekar, S., J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht. 1993. Protein design by binary patterning of polar and non-polar amino acids. *Science.* 262:1680–1685.

Kellershohn, N., and M. Laurent. 2001. Prion diseases: dynamics of the infection and properties of the bistable transition. *Biophys. J.* 81:2517–2529.

Kirkitadze, M. D., M. M. Condron, and D. B. Teplow. 2001. Identification and characterization of key kinetic intermediates in amyloid β-protein fibrillogenesis. *J. Mol. Biol.* 312:1103–1119.

Kyte, J., and R. F. Doolitle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.

Ma, B., and R. Nussinov. 2002. Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers ($A\beta_{16-22}$, $A\beta_{16-35}$, and $A\beta_{10-35}$): sequence effects. *Proc. Natl. Acad. Sci. USA.* 99:14126–14131.

Mandell, A. J., K. A. Selz, and M. F. Shlesinger. 2000. Protein binding predictions from amino acid primary sequence hydrophobicity. *J. Mol. Liquids.* 86:163–171.

Manetti, C., A. Giuliani, M.-A. Ceruso, C. L. Webber, Jr., and J. P. Zbilut. 2001. Recurrence analysis of hydration effects on nonlinear protein dynamics: multiplicative scaling and additive processes. *Phys. Lett. A.* 281:317–323.

Marwan, N., N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths. 2002. Recurrence plot based measures of complexity and its application to heart rate variability data. *Phys. Rev. E.* 66:026702-1–026702-7.

Massi, F., D. Klimov, D. Thirumalai, and J. E. Straub. 2002. Charge states rather than propensity for β-amyloid peptide compared to E22Q Dutch mutant. *Protein Sci.* 11:1639–1647.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structure: quasi-chemical approximation. *Macromolecules.* 18:534–552.

Murray, K. B., D. Gorse, and J. M. Thornton. 2002. Wavelet transforms for the characterization and detection of repeating motifs. *J. Mol. Biol.* 316:341–363.

Ortiz, A. R., and J. Skolnick. 2000. Sequence evolution and the mechanism of protein folding. *Biophys. J.* 79:1787–1799.

Petkova, A. T., Y. Ishii, J. J. Balbach, O. N. Antzutkin, R. D. Leapman, F. Delaglio, and R. Tycko. 2002. A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl. Acad. Sci. USA.* 99:16742–16747.

Romero, P., Z. Obradovic, X. Li, E. Garner, C. Brown, and A. K. Dunker. 2001. Sequence complexity of disordered proteins. *Proteins.* 42:38–48.

Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* 23:327–341.

Satheeshkumar, K. S., and R. Jayakumar. 2003. Conformational polymorphism of the amyloidogenic peptide homologous to residues 113–127 of the prion protein. *Biophys. J.* 85:473–483.

Selz, K. A., A. J. Mandell, and M. F. Shlesinger. 1998. Hydrophobic free energy eigenfunctions of pore, channel, and transporter proteins contain beta-burst patterns. *Biophys. J.* 75:2332–2342.

Serpell, L. C. 2000. Alzheimer's amyloid fibrils: structure and assembly. *Biochim. Biophys. Acta.* 1502:16–20.

Strait, B. J., and T. G. Dewey. 1996. The Shannon information entropy of protein sequences. *Biophys. J.* 71:148–155.

Thompson, L. K. 2003. Unraveling the secrets of Alzheimer's β-amyloid fibrils. *Proc. Natl. Acad. Sci. USA.* 100:383–385.

Tycko, R. 2003. Insights into the amyloid folding problem from solid-state NMR. *Biochemistry.* 42:3151–3159.

van der Spoel, D., H. J. C. Berendesen, A. R. van Buuren, M. E. F. Apol, P. J. Meulenhoff, and A. L. T. M. Sijbers. 1995. GROMACS User Manual. AG Groningen, Nijenborgh, The Netherlands.

Webber, C. L., Jr., and J. P. Zbilut. 1994. Dynamical assessment of physiological systems and states using recurrence plot strategies. *J. Appl. Physiol.* 76:965–973.

Weiss, O., M. A. Jimenez-Montano, and H. Herzel. 2000. Information content of protein sequences. *J. Theor. Biol.* 206:379–386.

West, M. W., W. Wang, J. Patterson, J. D. Mancias, J. R. Beasley, and M. H. Hecht. 1999. De novo amyloid proteins from designed combinatorial libraries. *Proc. Natl. Acad. Sci. USA.* 96:11211–11216.

Zak, M., J. P. Zbilut, and R. E. Meyers. 1997. From instability to intelligence: complexity and predictability in nonlinear dynamics. *In* Lecture Notes in Physics: New Series m49. Springer-Verlag, Berlin.

Zbilut, J. P., A. Giuliani, C. L. Webber, Jr., and A. Colosimo. 1998. Recurrence quantification analysis in structure function relationships of proteins: an overview of a general methodology applied to the case of TEM-1 β-lactamase. *Protein Eng.* 11:87–93.

Zbilut, J. P., P. Sirabella, A. Giuliani, C. Manetti, A. Colosimo, and C. L. Webber, Jr. 2002. Review of nonlinear analysis of proteins through recurrence quantification. *Cell Biochem. Biophys.* 36:67–87.

Zbilut, J. P., C. L. Webber, Jr., A. Colosimo, and A. Giuliani. 2000. The role of hydrophobicity patterns in prion folding as revealed by recurrence quantification analysis of primary structures. *Protein Eng.* 13:99–104.

Zimmerman, J. M., N. Eliezer, and R. Simha. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* 21:170–201.