Nonlinear Signal Analysis Methods in the Elucidation of Protein Sequence–Structure Relationships

Alessandro Giuliani,*[,†,] Romualdo Benigni,[†] Joseph P. Zbilut,[‡] Charles L. Webber, Jr.,[§] Paolo Sirabella,[||] and Alfredo Colosimo[||]

Istituto Superiore di Sanità, TCE Laboratory, V.le R. Elena 299, 00161 Roma, Italy, Department of Molecular Biophysics and Physiology, Rush University, 1653 West Congress Parkway, Chicago, Illinois 60612, Department of Physiology, Loyola University Medical Center, 2160 South First Avenue, Maywood, Illinois 60153, and Department of Biochemical Sciences, University of Rome, "La Sapienza", P.le A. Moro, 5-00185 Roma, Italy

Received July 30, 2001

Contents

I. Introduction

a. The Basic Problem

The relationship between sequence-embedded information and folding behavior of proteins is currently a dominant concern of both theoretical and applied biochemical research. More specifically, this concern redounds to areas such as (a) sequence-based functional predictions, (b) 3D structure-based functional predictions, and (c) folding mechanism elucidation.

The above issues are regarded as key points in basic research, and all of them have immediate applicable spin-offs in biotechnology, where the elucidation of new protein structures is of main interest for areas ranging from pharmaceutical industry to electronics.[1] Moreover, the completion of the sequencing phase of the human genome project shifted the attention of the scientific community to so-called "structural genomics" where the structural consequences of genome data in terms of protein structure and activity are exploited.[2,3] This new phase, called "post-genomic", in contrast with the previous one is of direct interest to chemists.

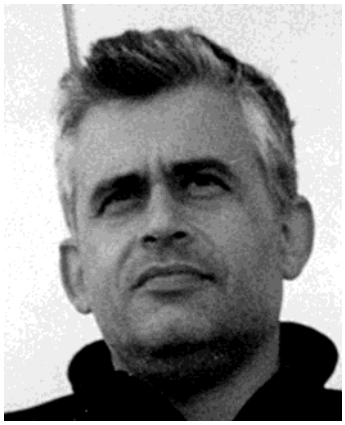
In a fundamental paper published in 1994 entitled "Proteins: where physics of simplicity and complexity meet" by Hans Frauenfelder and Peter Wolynes[4] focused on the peculiarity of the sequence−structure relation and on the need to have microscopic (and in principle very accurate) physics principles of "simple" systems (like atoms) cooperatively interacting to produce macroscopic principles qualitatively describing 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,[5] the "mesoscopic" principles needed to predict the 3D structure of proteins[6] remain essentially unknown. This blend of microscopic principles and macroscopic consequences has been a typical feature of chemical sciences in the last 150 years as well as a leit-motif in the present review.

Proteins occupy a unique position in the hierarchy of natural systems, since they lie in a gray region between chemistry and biology.[7] Proteins are large, complicated molecules that any polymer chemist would have difficulty in modeling. From the biological side, although any single protein would not be considered as alive, it does not take many of them (plus a bit of nucleic acid) before life-like behavior begins to emerge. For example, some of the smallest viruses, such as HIV, which might be considered on the borderline of life, are endowed with only 10 different types of proteins.[7]

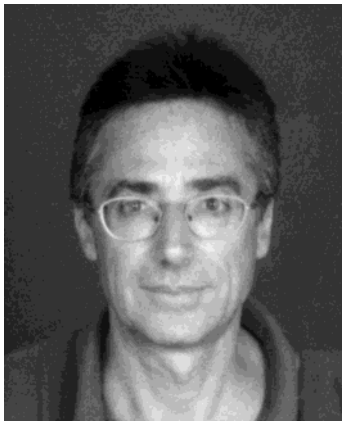
From a chemical viewpoint, proteins are linear heteropolymers that, unlike most synthetic polymers, are formed of basically nonperiodic sequences of 20 different monomers. While artificial polymers are generally very large extended molecules forming a matrix, the majority of proteins fold as self-contained structures determined by the sequence of monomers.

* To whom correspondence should be addressed. E-mail: alessandro.giuliani@iss.it.
† Istituto Superiore di Sanità.
‡ Rush University.
§ Loyola University Medical Center.
|| University of Rome.

10.1021/cr0101499 CCC: $39.75
Published on Web 03/23/2002

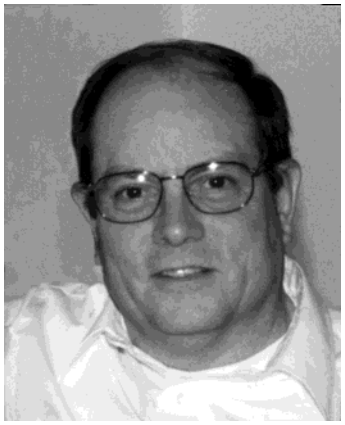


Alessandro Giuliani is Senior Scientist at the Istituto Superiore di Sanità (Italian National Institute of Health) in Roma. He received his Ph.D. degree in Biological Sciences at the University of Rome, "La Sapienza", specializing in Biostatistics. His research is devoted to the exploration of multidimensional statistical methods in the study of complex systems like elucidation of structure–activity relations, nonlinear dynamics, physiology, and ecology. Recently he studied the application of RQA in the study of sequence–structure relationships of proteins.



Romualdo Benigni received his education in Chemistry at the University of Rome "La Sapienza". He is now the director of the Structure–Activity Relationships unit at the Istituto Superiore di Sanità (Italian National Institute of Health) in Roma. He worked experimentally in the field of molecular biology and environmental chemical mutagenesis. In the 1980s he turned his attention to the statistical modeling of toxicological data and to the study of the relationships between the structure of organic compounds and their toxicological properties (mainly mutagenesis and carcinogenesis).

Joseph P. Zbilut received his doctorate degree in 1973 from Northwestern University and one from Rush University in 1987 specializing in time series analysis of physiological phenomena. Currently he is a Professor in the Department of Molecular Biophysics and Physiology at Rush Medical College. In the last two decades his interest has centered at methods which deal with nonlinear, nonstationary data and culminated in the development of recurrence analysis in collaboration with Charles L. Webber, Jr. Another component resulted in a new formulation of dynamics which admits stochastically via singularities. In the last few years, he has been engaged in adapting recurrence quantification analysis to the spatial and dynamical requirements of proteins.



Charles L. Webber, Jr. is Professor of Physiology at Loyola University Chicago, where he received his Ph.D. degree in Systems Physiology. His research interests include the multidimensional analysis of biological rhythms, proteomics, and genomics. He and Joseph Zbilut are the authors of Recurrence Quantification Analysis (RQA), which has demonstrated utility in multiple fields of study including physiology, physics, and mathematics. He has received the Teacher-of-the-Year Award five times over a quarter-century of medical teaching.

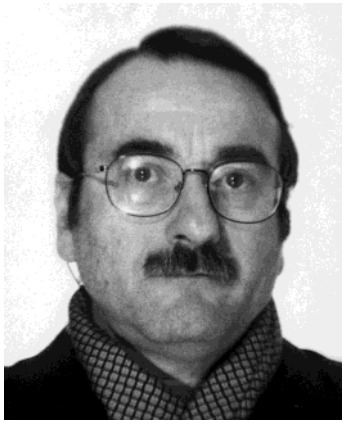
Thus, we can consider the particular linear arrangement of amino acids as a sort of "recipe" for making a water-soluble polymer with a well-defined three-dimensional architecture.[7−9]

It is important to stress this dynamical perspective. "Well defined three-dimensional structure" should not be intended as "fixed architecture": many proteins appear as partially or even totally disordered when analyzed with spectroscopic methods;[10] however, this apparent disorder corresponds to an efficient organization as for protein physiological function.

The task of being water soluble while maintaining the structural specificity necessary for a physiologically motivated activity is not easy, and only a relative minority of linear amino acid arrangements can actually accomplish this. Thus, the most basic problem in the sequence−structure puzzle is "What particular linear arrangement of amino acids makes a real protein?" This can be rephrased as "Is it possible to discriminate between amino acid sequences that in water, acquire a well-defined three-dimensional structure, and sequences that never will?"

While only specific sequences are able to generate functional proteins, there are only weak departures of real protein sequences from random strings.[8,11] A "random string" should be intended in the information theoretic sense, as a series whose autocorrelation structure remains substantially invariant after random shuffling of the positions of its constituent elements.[12] This obviously has nothing to do with the fact that particular sequence motifs carry a very peculiar "meaning" in both structural and physiological terms. As a matter of fact, even written texts that have obvious meaningful motifs corresponding to words appear as random strings of letters. The

Paolo Sirabella received his doctorate degree in Physics and another doctorate degree in Biochemistry at the University of Roma "La Sapienza". He is now the recipient of a postdoctoral contract in the Department of Biochemical Sciences at the same University. The two main research interests of P. Sirabella are (i) the development of neural networks and in general parallel computation devices for solving complex biological problems and (ii) the study of protein sequence–structure relationships.



Alfredo Colosimo graduated in Pure Chemistry at the University of Roma "La Sapienza", where he has served as Full Professor of Biophysics in the Medical Faculty since 1989. The most recent scientific interests of A. Colosimo focus on (i) set-up and use of innovative hardware and software tools in handling some computationally heavy problems related to proteins and nucleic acids and (ii) studies on the dynamics of complex phenomena of biomedical interests by statistical techniques.

linkage between words and meaning in human languages is purely "external" and arises from the historical evolution of languages: the English word "dog" has the same (null) relation with the actual animal as the Italian word "cane" (which in turn, incidentally, has a completely different meaning in English). In the case of protein sequences, there is no "external reader" attaching a 3D translation to monodimensional sequences of amino acids; thus, the sequences must "embed" their own code. The case of chaperonins (proteins helping other proteins to fold inside cells) is not general and only shifts the problem (who folds the folders?). The possibility of sequentially denaturing/refolding a given protein clearly indicates that the three-dimensional structure is in some way encoded in its amino acid order.[9,13]

Given these premises, the fact that protein sequences are only slightly different from random strings corresponds to the notion that the "code" linking a sequence to a particular structure is not emerging from simple periodicities in the amino acids' occurrence.[14]

Additionally, we know that the basic biological roles of proteins are mediated by their three-dimensional structure, thus giving the sequence−structure puzzle the character of a crucial knot.[15]

Whereas at a first sight there appear to be no peculiar periodicities in amino acid ordering that discriminate real protein sequences from random strings,[16] if one considers huge collections of protein primary structures, hydrophobicity profiles of sequences have been shown to significantly correlate with 3D structural properties and have a weak (but still statistically significant) departure from randomness.[12,17] The peculiar character of hydrophobicity is probably linked to its prominent role in stabilizing protein structures and points to the existence of specific constraints in the arrangement of hydrophobic/hydrophilic patterns along chains, leading to foldable structure. Such constraints, however, are not unique and do not allow for an efficient discrimination between foldable and nonfoldable sequences in a majority of very different sequences. In a sense, following the metaphor of the "code", we can equate these statistical regularities to the empirical laws linguists discovered in human languages such as the so-called Zipf's law[18] linking by an inverse logarithmic function the number of occurrences and the words length. The presence of such correlation laws characterize a given text as written in a human language without revealing its meaning. The presence of such regularities in protein's hydrophobicity profiles, taken together with the chemico-physical principles, points to hydrophobicity as the chemico-physical feature of choice for the disambiguation of the sequence−structure code.[11,19]

In analogy with linguistics, the rationale of this choice is the consideration that statistical correlations emerging from an otherwise quasi-random baseline point to important constraints of the studied code.

b. Possible Approaches

There is voluminous literature dealing with theoretical models following ab initio approaches to the sequence−structure puzzle.[15,20] The great majority of the theoretical models adopt a statistical physics perspective based on proteins considered as lattices, i.e., squared grids in which each residue is considered as interacting with the same number of neighbors. These interactions are modeled by carefully chosen potential functions, generally based on hydrophobicity or related properties,[21,22] associating an energetic score to each amino acid pair interaction. The wide popularity of these simplified models stems from the failure of more complex (and in some sense realistic) ones, like those incorporating motion equations.[23] Such complex models, in fact, most often failed in capturing the salient behavior of heteropolymers. For example, a so-called "glassy" behavior has been observed which typically does not occur in proteins at temperatures of interest. As a result, minimal models have been devised, and lattice models are chief among these. Lattice models had already been used for a long time in polymer physics[8] and, in the context of proteins, were first introduced by Go and co-workers.[24] The most popular Go model considers

only interactions between residues occurring in the native state; the authors, however, did not use it to generate general principles of protein folding. To be able to distill general principles of protein folding, Dill and co-workers[25] suggested a simplified lattice model and used it to obtain minimal modeling of polypeptide chains.[26] To account for the major interactions in proteins, the latter authors argued that the naturally occurring amino acids can be divided into hydrophobic and polar categories. Chan and Dill emphasize that their model can capture many important features of proteins, including cooperativity, folding kinetics, and structural properties.[26]

Other groups (actually the vast majority) adopt an alternative view of the sequence−structure puzzle: instead of looking for "universals", i.e., general laws linking sequence and structure across all protein families, they apply a purely local statistical approach.[27,28] A protein whose primary structure only is known is compared, in terms of relative sequence alignment, with a large number of proteins whose three-dimensional structure is solved. Scoring a significant (>30%) superposition between the query sequence and those of already solved proteins allows for structural and consequently functional (this point requires further consideration, due to the nonperfect one-to-one mapping between structure and function) inferences.[29,30] This line of research is followed by many scientists with the impetus of postgenomic projects, whose basic aim is obtaining from genetic information structural models of the encoded proteins. A large number of research groups are presently investigating new algorithms to maximize the ability in discovering even relatively "remote" homologues of a given leader sequence. The pursuit of this goal has fueled the development of new sequence alignment techniques and represents one of the basic pillars of the new "bioinformatics" science.[20,30−33]

The major advances in sequence alignment techniques of the past decade in terms of both refinement of information technologies and biological consequences are surely one of the most important topics in contemporary science.[33]

The inspection of larger and larger databases of protein sequences with more and more sophisticated bioinformatics methods is the most crowded avenue of research aimed at finding statistical regularities in the sequence databases for solving the sequence−structure puzzle.

The above field is represented by significant literature. The present review instead is devoted to a scarcely populated but potentially extremely interesting field of computational biochemistry: the use of signal analysis methods typical of engineering and physics to describe protein sequences as monodimensional series. The protein sequences are described by means of a vector of numerical invariants that summarize the autocorrelation structures of the analyzed series. In this way, the "atomic level" of protein sequences description shifts from the pairwise alignment of structures to a self-consistent numerical description of the SINGLE sequence.

The basic signature of the methods described here is the production of self-consistent numerical indexes that parametrize the protein sequences as a whole in terms of amount and profile of periodicities in the hydrophobicity distribution along the chain. This is similar to the quantitative structure activity relationships (QSAR) analyses widely used in medicinal chemistry.[34,35] By analogy with QSAR, the investigated molecules (proteins in this case, organic compounds in the case of QSAR) are described by means of an array of numerical features parametrizing various chemico-physical properties of the molecules. These properties act as regressors (independent variables) for modeling a given biological activity, which in turn acts as a dependent variable. The biological activities most often modeled by QSAR are pharmacological or toxicological potencies, while the properties modeled so far for proteins are protein/peptide interactions, folding behavior, and thermal stability.

This "QSAR-like" signal analysis approach includes elements coming from both the "statistical-mechanics" and "bioinformatics" points of view. From the theoretical side comes the consideration of protein sequence as a unitary system embedded into a global force field based on hydrophobicity[36] (the signal analysis step ends with one number deriving from a computation extended over the whole sequence). From the statistical side comes the local approach and the use of soft data analysis methods with no peculiar distributional constraints.[37]

The main steps of the signal analysis approach can be summarized as follows: (a) use of hydrophobic code for primary structures;[38,39] (b) treatment of the hydrophobicity distribution along the sequence like a time series, with the corresponding use of nonlinear signal analysis techniques to underpin fine position-dependent properties of the hydrophobicity profiles;[40−42] these position-dependent properties are summarized by means of self-consistent numerical descriptors at the level of single sequences; (c) adoption of a local approach for both intersequence (within homologous series of proteins) comparisons and intrasequence (among short patches along the same sequence) analyses as a starting point for periodicity detection.[42] A short description of the rationale of the QSAR analyses will help clarify these points.

c. The QSAR Strategy

The search for the relationship between the structure of chemicals and their biological effects has been continual during the entire development of organic chemistry. For a long period of time these attempts produced essentially qualitative results until the 1960s. The foundation of modern quantitative structure−activity relationships (QSAR) science came about when Corwin Hansch[35,43] found the way to bring together two areas of science which had seemed far apart for many years: physical organic chemistry and the study of chemicals−life interaction.

A cornerstone of physical organic chemistry is the Hammett equation[35,44]

$$\log k = \rho\sigma + \text{constant}$$

which models the reaction mechanisms of organic

chemicals: k is a rate or equilibrium constant, ρ is a measure of the sensitivity of the reaction to substituents changes, and σ is a parameter characteristic of each chemical. The Hammett equation was extended by Taft, who considered also steric factors.[45] Subsequently, Hansch[43] showed that this type of model can also be used for biochemical reactions by introducing a hydrophobic parameter

$$\log k = f(\text{electronic, steric, hydrophobic})$$

It should be underlined that this type of approach was developed from and applied to sets of congeneric chemicals, i.e., chemicals structurally similar and acting by the same mechanism of action (better than the same rate-limiting step). Moreover, the model considers only variations of activity (i.e., the potency of active compounds). In other words, there must be a very clear definition of the applicability domain of the model (class of chemicals to which it applies). Finally, the model is derived from the statistical analysis of a (training) set of chemicals.[46] This approach has worked for an enormous number of biological problems,[47,48] and the success obtained is indicated by the fact that QSAR is routinely used in the industrial production of new chemicals.

At present, QSAR is one of the basic tools of modern drug and pesticide design and has an increasing role in environmental sciences.[49] Also, from a theoretical point of view, it has a great interest because it is one of the few areas of biology where the passage from qualitative to quantitative has been fully accomplished.

A basic ingredient we import from QSAR success is the clear definition of the physico-chemical (hydrophobic, electronic, steric) forces involved in the modeled process. The comparative analysis of thousands of QSARs,[35,50] even assigning a role to other forces in particular cases, evidenced the predominant role of hydrophobicity in determining the biochemical processes. However, the most important lesson derived from QSAR is, in our opinion, of methodological and not chemico-physical nature. In QSAR analyses a set of molecules is placed in a metric space, where the different chemicals are the statistical objects and the chemico-physical parameters are the axes (or variables). In such a space, one can use a large mathematical and statistical toolset and the comparison among different chemicals is put on a firm and rigorous basis. The consideration of protein sequences as time series allows this kind of approach to be applied to proteins that can be unequivocally described by a set of descriptors parametrizing the autocorrelation structure of the hydrophobicity distribution along the chain.

Last but not least, the practice of restricting each QSAR analysis to a set of congeneric chemicals was crucial for the success of QSAR. A congeneric set of chemicals is made of chemicals with the same basic structure, which provoke the same biological effect and act through the same mechanism of action (possibly having the same rate-limiting step). This very clear definition of the applicability domain of the model allows the investigator to focus the analysis on the structural differences responsible for the difference in biological activity. This particular point has been studied in depth, and it has been shown that all QSAR models have an applicability domain centered around the range of parameter values used; beyond this range, any relationship will fade in an unpredictable way.[51] Illustrative of the crucial importance of the correct selection of the chemical set is, for example, the striking difference between the good fitting of QSARs for individual classes of chemical carcinogens and the much lower fitting of models based on large databases of chemicals that provoke the same biological effect (carcinogenesis) through different mechanisms.[49]

When we shift from the analysis of an organic series to the analysis of a set of proteins, all the above considerations (local character of the extracted models, extensive use of multivariate analysis, study of a series of homogeneous chemical entities) remain valid. The passage from classical QSAR to protein-QSAR is characterized by (a) the substitution of the molecular descriptors with self-consistent indexes derived from time-series analysis for parametrizing hydrophobicity distribution and (b) the substitution of biological properties such K_i or IC_{50} values with protein structural and/or physiological properties such as thermal stability, protein/peptide interaction, folding behavior, or any other "global property" that can be measured for proteins as a whole.

II. Mathematical Methods

a. How Proteins Appear from a Signal Analysis Perspective

When coded as monodimensional arrays of hydrophobicity values corresponding to the amino acid sequence,[52] the primary structure of a protein can be considered as a numerical discrete series equivalent to a time series with the amino acid order playing the role of subsequent time intervals (Figure 1).

Thus, on a purely formal point of view, any one of the myriads of techniques routinely used for signal analysis in electronics as well as in physiology or meteorology[53] could be profitably applied to protein hydrophobicity sequences. From a practical viewpoint, the fact that protein sequences are short with respect to the signals analyzed in other fields (even if there is no reasonable physical limit to the length of a polypeptide chain, the greater part of naturally occurring proteins have less than 1000 residues[7]) and in some cases extremely short (e.g., rubredoxins, a class of proteins we will discuss below are made up of around 50 amino acids[54]) drastically limits the range of signal analysis techniques usable in this context. Moreover, protein hydrophobicity profiles are basically nonstationary signals displaying different statistical and correlation features along the chain, which does not favor classical techniques such as Fourier analysis.

The ideal method for approaching signal analysis of protein sequences should be nonlinear, independent of any stationary assumptions, and able to deal with very short series.[53] Methods satisfying these constraints are those approaching the analyzed series from a purely correlative point of view, with no a

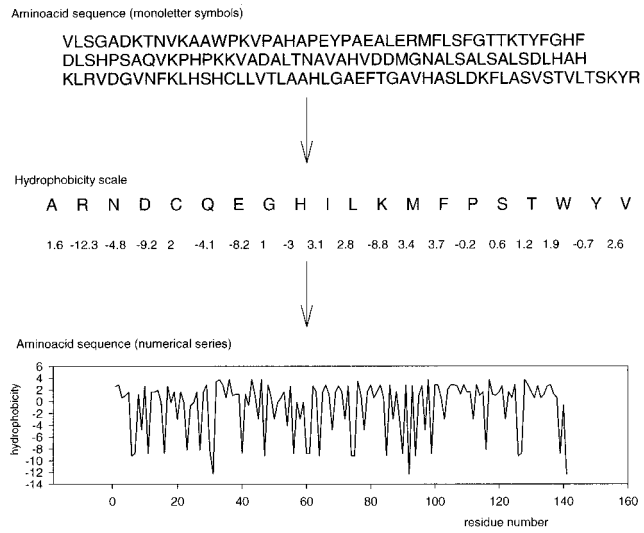


Figure 1. Hydrophobicity profile of a protein sequence. The procedure transforming a linear symbolic sequences of amino acids (here described by the single-letter code) to a numerical ordered series of hydrophobicity values is depicted in the figure. The hydrophobicity scale adopted here corresponds to the Shneider and Wrede scale,[52] and the amino acid/hydrophobicity translation scheme is reported in the center of the figure. High values correspond to hydrophobic amino acids, while negative values point to hydrophilic amino acids. The described sequence is that of the human haemoglobin α subunit.

priori distributional and/or physical assumption. The only aim of these methods is to look for autocorrelation patterns along the series, i.e., for the recurrences of particular short motifs along the chain (like in recurrence quantification analysis, RQA) or for periodicities of no predefined functional form spanning all the studied sequences (like in singular value decomposition, SVD).[53−56] At the basis of all these methods is the transformation of the original series into its "embedding matrix" with the method of delays.[53,54,56]

The embedding procedure consists of building an n-column matrix (in the example below $n = 4$) out of the original linear array by shifting the series by a fixed lag. For example,[54] given the series 10, 11, 21, 32, 41, 35, 40, 19..., the corresponding 4-dimensional embedding space at lag = 1 (the discrete character of amino acid sequences dictates this choice) is

10	11	21	32
11	21	32	41
21	32	41	35
32	41	35	40
41	35	40	19
35	40	19	
40	19		
19			

The rows of the embedding matrix (EM) correspond to subsequent windows of length 4 (embedding dimension) along the sequence. Notice that the last n values are eliminated from the analysis as an obvious consequence of shifting the series for the embedding. RQA is based on the computation of the Euclidean distance matrix (DM) between the rows (epochs) of the EM,[54] looking for epochs close to each other (recurrences). SVD is based on the correlation matrix among EM columns, whereas wavelets are based on the correlation of the rows of the EM with user-defined kernels.[53,55,56]

The choice of the embedding dimension corresponds to the choice of the scale at which the autocorrelation structure of hydrophobicity pattern is estimated, which varies across different techniques and problems. All the signal analysis techniques used in research on proteins give a global picture of the series in terms of degree of complexity (relative order/disorder of hydrophobicity distribution along the series), presence of singularities (regions within the sequence strongly different in terms of hydrophobicity pattern), and specific periodicities. In general, they quantitatively describe the shape of the hydrophobicity profiles[54,55,57] by appropriate numerical indicators.

b. Algorithms Used for the Analysis of Hydrophobicity Sequences

1. Recurrence Quantification Analysis (RQA)

Recurrence quantification analysis is a relatively new nonlinear technique, originally developed by Eckmann et al.[58] as a purely graphical method and then made quantitative by Webber and Zbilut.[59] It was successfully applied to different fields ranging from physiology[59,60] to molecular dynamics[61] and the study of chemical reactions.[62] Only in relatively recent times RQA was investigated by our group for its ability to deal with protein sequences.[54,63−66]

The concept of recurrence is straightforward: for any ordered series (time or spatial), a recurrence is simply a point which repeats itself. In this respect, the statistical literature points out that recurrences are the most basic of relations[67] shaping a given system, since they are strictly local and independent of any mathematical assumption regarding the system itself. Furthermore, it is worth stressing that calculation of recurrences, 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.[59,60]

The concept of a recurrence can be expressed as follows: given a reference point, X_0, and a ball of radius r, a point X is said to recur (with reference to X_0) if

$$B_r(X_0) = \{X:\ ||X - X_0|| \leq r\}$$

In the case of a time series, i.e., of a system occupying in different times different positions along a trajectory in a suitable state space, the recurrences correspond to the time points where the system passes nearby to already visited states. In the case of protein sequences, time corresponds to the amino acid order and the recurrences are patches, with a length equal to the embedding dimension, sharing their hydrophobicity profile with other patches along the chain. The number and relative positions of recurrences are expressed by recurrence plots (RP) that are symmetrical $N \times N$ arrays in which a point is placed at (i, j) whenever a point X_i on the trajectory

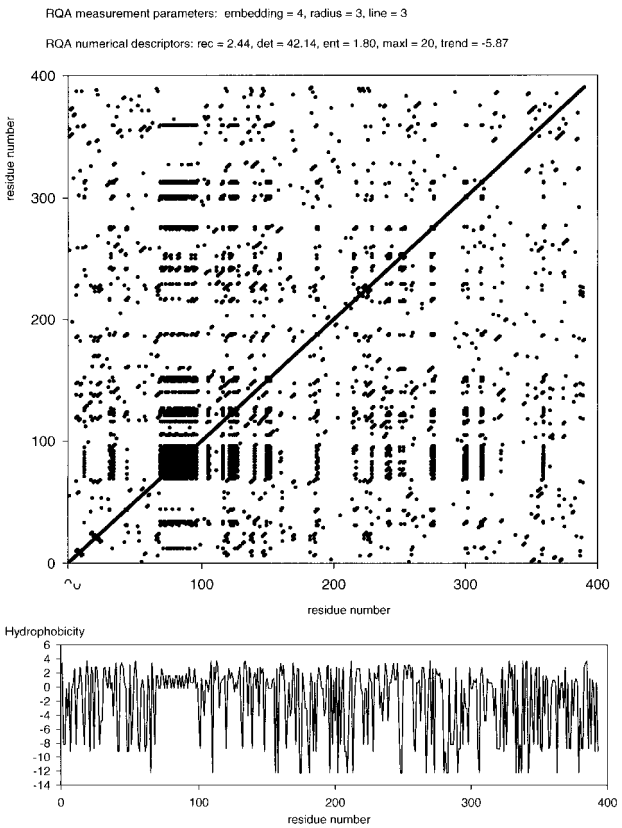


Figure 2. Recurrence plot of human P53 protein. The recurrence plot of human P53 protein is reported together with the corresponding hydrophobicity plot (bottom). The presence of an extremely deterministic ordering of amino acids between residues 61 and 98 is clearly evident in the figure in terms of its consequences on the recurrence plot. This highly deterministic portion is "resembled" by other segments along the sequence. This observation is not clear by the simple inspection of the hydrophobicity plot but is made evident by the recurrence plot: the "resemblances" correspond to linear (or alternatively horizontal given the symmetrical character of recurrence plot) banding of the plot. The RQA numerical descriptors corresponding to the plot have been reported together with the chosen measurement settings (see text for further details).

is close to another point X_j. The closeness between X_i and X_j is expressed by calculating the Euclidian distance between these two normed vectors, i.e., by subtracting one from the other obtaining the expression $||X_i - X_j|| \leq r$ 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 (Figure 2).

Thus, recurrence plots simply correspond to the distance matrix between the different epochs (rows of the embedding matrix) filtered, by the action of the radius, to a binary 0/1 matrix having a 1 (dot) for distances falling below the radius and a 0 for distances greater than radius. Distance matrixes are demonstrated[68] to convey all the relevant information for the global reconstruction of a given system. An important feature of such matrixes is the existence of short line segments parallel to the main diagonal, which correspond to sequences (i, j), $(i + 1, j + 1)$, ..., $(i + k, j + k)$ such that the fragment $X(j)$, $X(j + 1)$, $X(j + k)$ is close to $X(i)$, $X(i + 1)$, ..., $X(i + k)$. The absence of such patterns suggests randomness.[58] For protein sequences these deterministic lines correspond to contiguous patches of similar hydrophobic/hydrophilic patterns.

Because graphical representations may be difficult to evaluate, Zbilut and Webber[59] developed several strategies to quantify features of such plots originally pointed out by Eckmann et al.[58] Hence, the quantification of recurrences leads to the generation of five 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 Lyapunov exponent and is a measure of system divergence[69]), and TREND (measure of the paling of recurrent points away from the central diagonal). These five recurrence variables quantify the deterministic structure and complexity of the plot. The application of these simple statistical indexes to the recurrence plots gives rise to a five-dimensional representation of the studied series. This five-dimensional representation gives a summary of the autocorrelation structure of the series and has been demonstrated, by means of a psychometric approach,[63] to correlate with the visual impression a set of unbiased observers derive from the inspection of an ensemble of recurrence plots.

When one needs to appreciate possible changes in the autocorrelation structure at the level of single elements of the series, it is not possible to rely solely on the "holistic" summaries given by the direct application of RQA to the global sequence, and it is preferable to get a local measure of the degree of order of hydrophobicity distribution at the level of single zones along the chain. In these cases a "windowed" version of RQA can be performed, such that a time series $(y_1, y_2, ..., y_N)$ is fragmented into p subsequent epochs. Each epoch is treated by the algorithm as a complete series and the relative RQA descriptors computed in the usual way. Thus, the "windowed" version of RQA[59,60] produces monodimensional series corresponding to the distribution of RQA parameters along the sequence (Figure 3).

2. Singular Value Decomposition (SVD)

In contrast to RQA, singular value decomposition (SVD) is a well-established method frequently used in physical as well as in social and biological sciences.[37] SVD roughly corresponds to PCA (principal component analysis), which is perhaps the most widely used method in chemometrics.[70] The term SVD is preferred to the term PCA in physical applications and, in general, when dealing with dynamical phenomena. As in PCA, the aim of SVD is to project an originally multidimensional phenomenon onto a reduced set of new orthogonal axes, representing the basic modes explaining the analyzed data set.[37,56] When applied to a time (or spatial) series that is originally monodimensional, SVD necessitates that the original series is represented on a multidimensional space by the agency of the embedding procedure. This "expansion" of the original mono-

dimensional series on a multidimensional support made by the time-lagged copies of the original series allows for the autocorrelation structures of the series to be appreciated.[56]

The EM can be thought of as a multivariate matrix having subsequent patches of amino acids of length equal to the embedding dimension as statistical units (rows) and the whole sequence lagged by subsequent delays as variables (columns). Thus, the EM can be considered as an $M \times N$ matrix, with M being the chain length minus the embedding dimension (the last amino acids are eliminated by the shifting due to the embedding procedure) and N the embedding dimension.

A basic theorem in linear algebra states that each $M \times N$ matrix, **X**, can be expressed as

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{1}$$

where the matrixes **U** and **V** are of dimensions $M \times K$ and $N \times K$, respectively, and fulfill the relations $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = 1$. The $K \times K$ matrix **S** (typically the covariance matrix) is diagonal and has its diagonal elements (singular values) arranged in descending order $s_1 > s_2 > s_3 ... > s_k > 0$.

In intuitive terms this means that the original data can be projected into a new set of coordinates US (principal component scores or eigenfunctions) such that no original information is lost, given that each element of **X** is immediately reconstructible by the equation

$$X_{ij} = \sum U_{ik} S_k V_{jk}$$
$$k = 1 \text{ to } N \tag{2}$$

The new coordinates are orthogonal by construction (i.e., statistically independent), each representing an independent aspect of the data set.

PCA (and equivalently SVD) has an optimal property which has made this method one of the most widespread modeling techniques in diverse science fields: with the expansion truncated to A terms (with $A < N$), one obtains the summation

$$\mathrm{X}_{ij} = \sum U_{ik} S_k V_{jk} + E_{ij}$$
$$k = 1 \text{ to } A \tag{2a}$$

where the squared error term $\sum E_{ij}^2$ is a minimum. What differentiates eq 2 from eq 2a is the presence of the error term E_{ij} and the summation limited to a lower number of coordinates with respect to the original data set. The fact that the error term is a minimum means that the projection of the original data on the new component space spanned by a smaller number of dimensions ($A < N$) is optimal in a least-squares sense. This implies that we can save the meaningful (signal-like) part of the information retained by the first principal components and discard the noise in the error term.[56] In other words, the most correlated (in terms of coordinated variation of hydrophobicity along the chain) portion of information is retained by the first components, while all the singularities are discarded in the minor components.

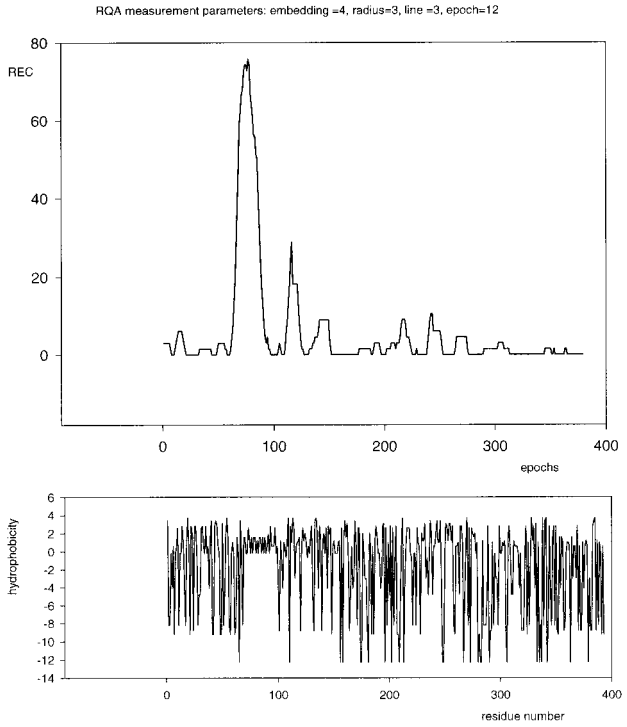


Figure 3. Recurrence analysis by sliding windows. The windowed version of RQA as applied to human P53 sequence is reported. The singular deterministic ordering of the 61−98 portion is here evident as a major peak in the recurrence distribution, while the "bands" of Figure 2 here appear as minor peaks.

Typically, in protein studies SVD is computed with an embedding between 8 and 10 and the first three eigenfunctions are selected.[55,57]

The reconstruction of the original hydrophobicity plot by means of the first components thus corresponds to smoothing of the original series to eliminate high-frequency noise due to the insertion of "spurious amino acids" while keeping alive slower rhythms reminiscent of protein structural secondary and supersecondary structures.[55,57,71] The SVD-smoothed series is then analyzed for the presence of periodicities and general patterns not evident in the original series. These periodic structures are put in evidence by computation of all-poles, maximum-entropy power spectra[57] on the eigenfunctions convolution or by the application of wavelet analysis.[72]

Figure 4 represents the effect of SVD on an hydrophobicity plot: SVD acts like a filter for correlated information and permits periodicities in the amino acid along the chain to come to light by eliminating the "disturbance" caused by singularities (e.g., a hydrophilic amino acid embedded in an hydrophobic fragment). Thus, SVD provides a "global" view of the hydrophobicity distribution, while RQA, especially in its windowed version, presents a "local" view on the same pattern. This implies a complementary character of the two techniques that were in fact used in combination in one of the applications described in this review.[54]

3. Wavelet Analysis

Wavelets are related to Fourier methods and have found popularity among a variety of scientists. The

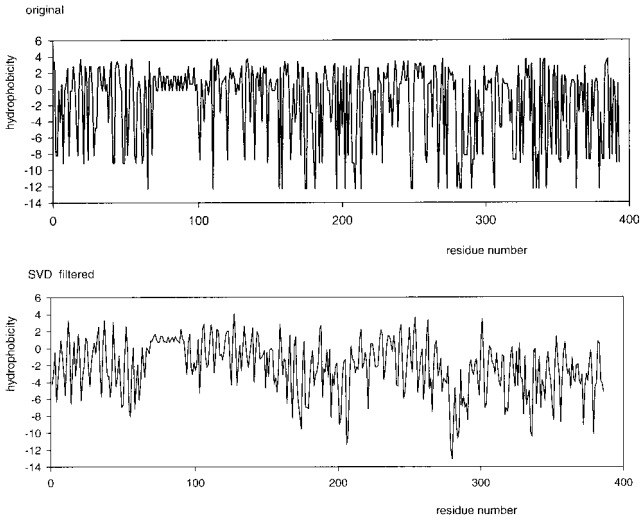


Figure 4. SVD analysis of a protein sequence. The hydrophobicity plot relative to human P53 protein (top panel) is compared with its SVD filtered version (bottom panel). It is evident from the figure that SVD demonstrates long-range periodicities completely hidden in the raw hydropathy plot. These periodicities were shown to be correlated with secondary and supersecondary structures.[55,57] In the case of P53, the quasi-harmonic wave approximately going from residue 100 to 280 corresponds to a β-sheet structuring of the protein.

wavelet approach is essentially an adjustable window Fourier spectral analysis with the following general definition

$$W(a,b; X,\psi) = \sqrt{|a|} \int X(t)\psi^*((t - b)/a)\, dt$$

in which ψ^* is the basic wavelet function where a is the dilation factor and b is the translation of the origin. Although time and frequency do not appear explicitly in the transformed result, the variable $1/a$ gives the frequency scale and b the temporal location of the event. By subsequently applying the above formula on the studied signal, we obtain the value of the "superposition" of the wavelet with the actual signal. An intuitive physical explanation of the above formula is that W is the "energy" of X at scale a at $t = b$.

Unlike Fourier transforms which are able only to decompose a signal into its constituent frequencies, wavelets also have the ability to provide scale analysis by continuously modifying the window length at which the analysis is performed.

Fourier demonstrated that any 2π-periodic function $f(x)$ is the sum of a series of sine and cosine functions. This theorem implicitly considers the analyzed series as stationary and completely ignores the actual location (on the time axis) of the extracted periodicities: the signal is expressed in a frequency space, globally averaging the hidden time information. Wavelets, on the other hand, emphasize scale and are thus able to localize a feature in the time domain.[72] The procedure is to adopt a wavelet prototype function, called an *analyzing wavelet*, corresponding to a particular pattern and is continuously shifted on the analyzed signal in order to look for the correlation between the signal and the wavelet. Analysis of the ordered time series is performed with a contracted, high-frequency version of the prototype wavelet, while frequency analysis is performed with a dilated, low-frequency version of the same wavelet. The analyzing wavelet ($\Phi^{(x)}$) is defined as

$$2^{-s/2}\Phi(2^{-s}x - l)$$

where s and l are integers that scale and dilate the analyzing wavelet. A source of difficulty is choosing the appropriate analyzing wavelet, depending upon the given task.[73] Also, care must be taken regarding other important choices such as orthogonality and coefficients.

The output of a wavelet analysis is a representation of the original series in terms of a linear combination of forming functions (wavelets): this representation, analogously to that offered by SVD, corresponds to a noise-filtered version of the original series potentially able to reveal otherwise hidden periodicities.

In addition, the local character of wavelets allows for the identification of possible singularities (change-points) along the series.

III. Applications

a. Transmembrane Helix Locations

While a large portion (10−35%) of proteins in a genome encodes membrane proteins, elucidating the structure of transmembrane (TM) proteins is a very difficult task for both nuclear magnetic resonance spectroscopy and X-ray spectroscopy.[9] On the other hand, TM proteins have a fundamental importance in practical applications because a large fraction of these proteins play key functional roles as drug receptors, immunological recognition targets, etc.[9] A number of algorithms, designed to identify putative TM helices, i.e., the portions of primary structures embedded into the lipid membrane phase, have been developed.[74] This problem was historically the first to be approached from a signal analysis perspective[75,76] since it allows for an immediate link between signal analysis and chemico-physical principles. The most important driving force of protein structure is the hydrophobic potential, i.e., the tendency of exposing toward the water environment the hydrophilic residues while hiding in the interior the hydrophobic ones.[39] Whereas the "inside" phase for globular protein is the so-called hydrophobic core,[9] for TM proteins the out-of-water phase is represented by the lipid membrane.

This behavior is translated into a signal analysis perspective by means of cluster analysis:[75] the identification of relatively hydrophilic/hydrophobic continuous clusters along the chain is made to correspond to the identification of exposed and HTM (buried in the lipid membrane) residues. Figure 5 reports the hydropathy plot of a TM protein with the water-exposed and TM portions assigned by a cluster analysis approach. As evident from the figure, the continuity of the selected clusters is interrupted by the insertion of hydrophilic residues inside mainly hydrophobic patches or vice versa. The first solution to this problem made use of moving averages[77,78] in

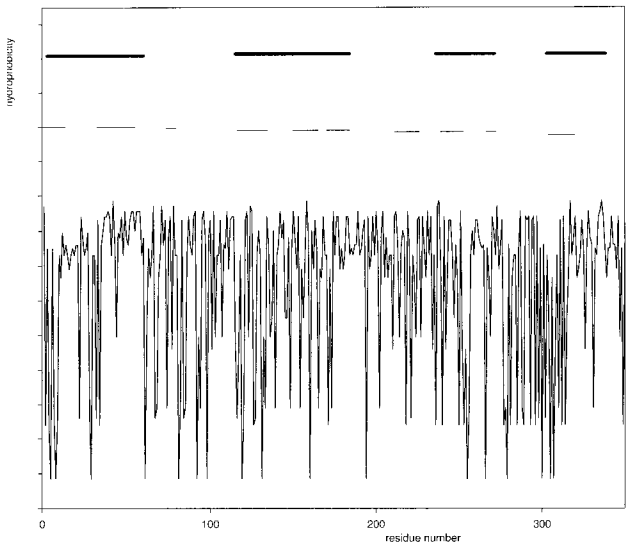


Figure 5. Transmembrane regions in HN protein from Sendai virus. The hydropathy plot of a transmembrane protein (Haemoagglutinin-Neuraminidase (HN) protein from Sendai virus) is reported together with the indication of its effective (thick bars) and presumptive (short thin bars) transmembrane portions. The presumptive transmembrane portions were assigned by a cluster analysis method.

order to eliminate the sharp peaks corresponding to the insertion of "outliers" into homogeneous regions, the hydrophobicity value for each residue inside the window was summed, and the average value was given for a residue in the middle of the window.[75] While, with coarse-graining, this method attained statistically significant results in the identification of HTMs, the fact that the value calculated for each amino acid residue is greatly affected by neighboring residues in a way strictly dependent upon the chosen averaging window size makes it difficult to assign to each single residue a reliable interior/exterior state. To overcome this problem, two main directions were followed that greatly enhanced the precision of the buried/exposed boarders detection: the first one is based upon enriching the pure hydropathy information by other information relevant for the assignment of the interior/exterior state.[79] Examples of this complementary information are the assignment of a "helicity propensity scale" (based on the observation that TM portions are in nearly 90% of cases arranged as alpha-helices) or the derivation of hydrophobicity scales by means of high-pressure liquid chromatography for short peptides mimicking the TM patches.[79] Other authors[57,74] followed the alternative direction of exploring sophisticated signal analysis techniques in their ability to identify interior/exterior change points on the sole basis of single-residue hydrophobicity information. The application of wavelet analysis to an hydropathy scale coding of TM proteins was demonstrated to be particularly efficient in the detection of change points.[74] As explained in the methods section, at the basis of wavelet decomposition of a given signal is the convolution of the original series with a predefined pattern (the wavelet) at different scales. This operation decomposes the signal into its wavelet coefficients at the various scales. When wavelet analysis is used to filter out noise from the series, the wavelet coefficients are separated into large (signal-like) series that are retained and small (noise) coefficients that are discarded. The scale attaining the higher values of the signal coefficients with respect to noise is then used to reconstruct a relatively noise-free signal. This technique, known as wavelet shrinkage,[80] was adapted to change-point detection; after the above-mentioned shrinkage, the resulting filtered representation of the original signal is used to identify change points with a simple threshold approach: once the denoised profile is centered, all regions with value above zero can be interpreted as HTMs.[74] The application of wavelet technique to the prediction of HTM segments resulted in a very good 98.3% accuracy per segment in a test set of 83 proteins.

The wavelet approach was applied to the similar problem of detecting the hydrophobic core of globular proteins.[81] In this case, while still giving statistically significant results (around 70% accuracy at cross-validation tests), the procedure was by far less efficient than in the case of TM prediction. In the authors' opinion, this was linked to the need for sketching different models for different protein families, in analogy to the dependence of congeneric series in classical QSAR, and to the related difficulty of defining the concept of "homologous series" of proteins. In any case, when applied to noncongeneric sets, the wavelet technique has a performance comparable to that of the classical sequence alignment approach[30] and the advantage of not requiring known structural homologues of the studied protein.

Other scientists approached the problem of detecting HTMs through the use of SVD followed by all-poles, maximum-entropy spectral analysis. In this case, they also were able to identify the location and extension of TM patches even if with a slightly lower accuracy than wavelets. Furthermore, they were able to detect the signature of secondary and super-secondary structures (like the so-called β-bursts peculiar arrangements of subsequent β-sheets or α-helices) as well identified peaks in the power spectrum of the SVD-filtered representation of hydrophobicity plots.[57]

As a matter of fact, both wavelets and SVD approaches are based on the same basic assumption: denoising the original hydrophobicity series to allow the "real" hydrophobic/hydrophilic patterning of the original sequence (corresponding to the lipid/water exposition of the protein) to come to light.

b. Protein/Peptide Interactions

The importance of the sequential arrangements of amino acid side chain hydrophobicities in the determination of peptide and protein secondary structures is well established.[80] Significant roles are played by two kinds of hydrogen bond energies. One involves a restricted range of local, side chain independent, sterically allowed, main chain peptide bond rotations represented in Ramachandran plots.[71] The other, more prominent in aqueous environments, also regulates secondary structural turn formation but is dominated by in line, surface minimizing attraction

Table 1. Dominant Frequencies in Terms of Maximum Entropy All-Poles Spectrum of SVD Filtering of Relative Hydrophobicity Plots Are Reported for Six Ligand/Receptor Couples Pointing to the Strict Resemblance between the Hydrophobicity Periodicities of the Ligand and the Corresponding Receptor

system	receptor-dominant frequency	ligand-dominant frequency
Kappa Opioid	3.31	3.46
CRF-1	2.22	2.18
Orphan Opioid	3.09	2.99
Somatostatin-5	2.90	2.83
Neuropeptide Y	3.77	3.63
Bombesin-3	5.71	5.60

between hydrophobic phase coherent patches of amino acid side chains.[71] These hydrophobic effects emerge from nonlocal cooperative interactions of hundreds to thousands of hydrogen bonds of the surrounding water solvent.[71,82]

Peptide−receptor interactions are of the same nature as the solvent−protein and protein−protein interactions; it was demonstrated that the substitution of hydrophobically equivalent amino acids in peptide ligands maintains the potency of their cell membrane-mediated actions.[83] Other scientists demonstrated that the binding of bovine growth hormone to the extracellular domains of its receptor was more related to common helical, loop, and/or disordered secondary structure than to specific amino acid sequences or the local geometry of tertiary structures.[84] In the same fashion, the ability of the prion protein, as well as of other protein species, to aggregate many other protein molecules and eventually precipitate (the basis of the so called prion-like behavior) was recently demonstrated to rely upon a peculiar mainly β-sheet organization of secondary structures.[85]

Thus, a picture seems to emerge of protein/protein (and analogously protein/peptide) interactions driven by the reciprocal similarity of their secondary and supersecondary structure organization that in turn can be (at least presumptively) identified in terms of similarity in their hydrophobic profile sequential arrangement. An application of SVD followed by an all-poles, maximum-entropy spectral analysis and a wavelet analysis was able to confirm this hypothesis in the case of four receptor−ligand pairs: neurotensin (NT), cholecystokinin (CK), dopamine D2 receptor ((DA)D2), and dopamine transporter (DT).[71] In all these cases the hydrophobicity periodicity along the chain as highlighted by SVD was able to unequivocally match the correct receptor/ligand pair.[71] Analogous results were obtained for other receptor/ligand (peptide/protein) pairs by the same authors[55] with the same data analysis approach. They were able to generate a relevant match between ligand/receptor pairs for six receptor systems on the simple basis of the dominant frequency of the respective eigenfunctions representations (Table 1). More importantly, a mutant of the CRF-1 receptor unable to bind efficiently to CRF-1 was shown to have lost the peculiar "ligand" frequency of its hydrophobicity plot.[55]

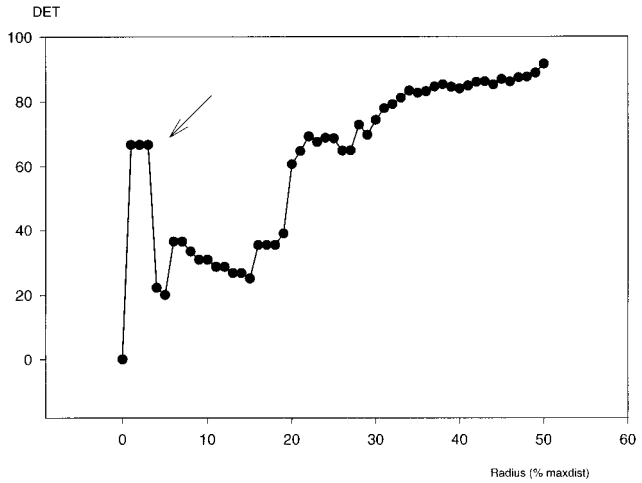
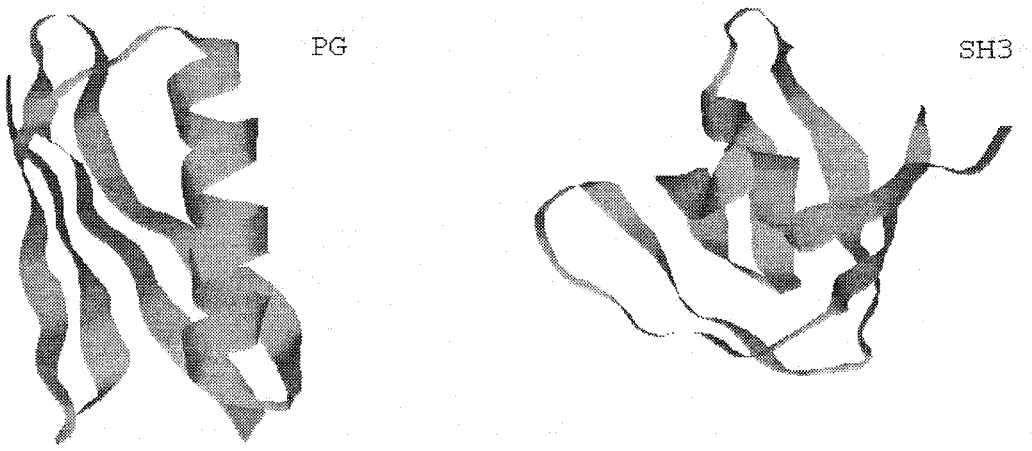


Figure 6. Determinism structure of Syrian Hamster prion protein. The determinism scaling of Syrian hamster PrP protein is reported. The "shelf-like" singularity, corresponding to the nucleation center, is indicated by an arrow. The scoring of a singularity like the one shown points to the presence of an extremely deterministic hydrophobicity structuring localized in a well-defined portion of the sequence. Beyond the singularity, the determinism drops off as new recurrences are recruited by increasing the radius, since these recurrences, being spread all along the entire sequence, are not contiguous (it is worth recalling that determinism corresponds to the proportion of recurrent points out of total recurrences that are contiguous, see Methods). At increasing values of radius, progressively all the portions of the protein become recurrent and the determinism saturates (if all the points become recurrent there is no room for isolated points) with a curve resembling the "average" characteristics of the protein hydrophobicity distribution.

Obviously these evidences are still episodic. A key point in the success of the above procedure relies on the fact that the leading mode of the protein structure was the one responsible for binding its peptide effector. This by no means suggests a generalization: the proteins are similar to microscopic "machines", as they must concomitantly perform many tasks such as remaining soluble in water, being attached to a given cellular scaffold, binding in different and specific portions of the structure diverse peptides and/or small molecules. The entire protein structure is fitted to perform the entire spectrum of tasks, but we cannot anticipate if only a particular portion of the sequence is involved in (and thus optimized for) a particular task. This implies that we have no guarantee the activity we are investigating is the principal "shaping agent" of the studied system, and thus, we cannot be sure that a measure computed over the entire sequence like SVD spectrum retains the important information for the particular activity to be modeled. Nevertheless, these analyses demonstrate the possibility of using a statistically motivated "drug design" procedure for modeling protein/peptide interactions.

c. Protein Folding

Protein folding dynamics is one of the most fascinating natural phenomena, at the crossroads between different fields such as statistical physics, chemistry, biology, and theoretical complex systems



1pgb - 56 AA
IMMUNOGLOBULIN BINDING PROTEIN
PROTEIN G (B1 IGG-BINDING DOMAIN)
(STREPTOCOCCUS, LANCEFIELD GROUP G) STRAIN GX7809

MET THR TYR LYS LEU ILE LEU ASN GLY LYS THR LEU LYS
GLY GLU THR THR THR GLU ALA VAL ASP ALA ALA THR ALA
GLU LYS VAL PHE LYS GLN TYR ALA ASN ASP ASN GLY VAL
ASP GLY GLU TRP THR TYR ASP ASP ALA THR LYS THR PHE
THR VAL THR GLU

1shg - 62 AA
CYTOSKELETON
ALPHA SPECTRIN (SH3 DOMAIN)
CHICKEN (GALLUS GALLUS) BRAIN

MET ASP GLU THR GLY LYS GLU LEU VAL LEU ALA LEU TYR
ASP TYR GLN GLU LYS SER PRO ARG GLU VAL THR MET LYS
LYS GLY ASP ILE LEU THR LEU LEU ASN SER THR ASN LYS
ASP TRP TRP LYS VAL GLU VAL ASN ASP ARG GLN GLY PHE
VAL PRO ALA ALA TYR VAL LYS LYS LEU ASP

Figure 7. Primary and tertiary structures of PG and SH3 proteins. The ribbon diagrams summarizing the three-dimensional structures together with the amino acid sequences of PG and SH3 proteins have been reported.

science.[86] Here we will describe two approaches linked to this problem at two different variability scales: the scale of the single protein[65] and the scale of an ensemble of protein sequences generated by the fusion of patches coming from two largely different parent protein structures.[63]

It has been already mentioned that simple models of protein folding on a lattice proved useful in understanding the basic principles of protein folding dynamics.[87] One important idea gleaned from these studies is the conjecture posed by one group that polymers may have multiple ground states and thus may fold into different structures[88] and then extended this idea to explain the conformational flips of prions. They designed sequences, based on lattice models, which exhibited two different conformations of equal energy corresponding to a global energy minimum.[88] Folding simulations demonstrated that one of these ground states was much more accessible than the other. A critical factor in determining the accessibility was the number and strength of local contacts in the ground state conformation. Although it is recognized that this may not be the only factor involved in such a phenomenon, it does provide some basic understanding of the process. To explore this possibility as well as the feasibility of deriving an empirical, hydrophobicity based phenomenology, RQA of hydrophobicity values was applied[65] along the sequence of the two given model 36-mers described in ref 88. The results were compared to the recombinant prion protein (PrP) of the Syrian hamster, *sh*PrP (PDB ID code 1B10), which corresponds to the infectious fragment of the scrapie isoform.[89] The rationale was that the transition state corresponding to the most kinetically accessible state should be characterized by an extremely high number of local contacts (between nearby residues) originating from a nucleation seed which would drive the subsequent folding of the entire structure. This in view of the fact that the contact order (number of contacts in the native structure between nonadjacent residues weighted for their relative closeness) was unequivocally singled out as the main determinant of the folding rate.[86] The general relationship holding between similar hydrophobicity profiles of different portions of the sequence and their contact probability (see the above cases of hydrophobic core prediction and peptide/protein interactions) should provide a signature of this nucleation zone in terms of hydrophobicity patches characterized by a highly deterministic structure.

This could allow for a sketch of folding dynamics prediction on the basis of pure sequence information.

To check this hypothesis, we computed the amount of determinism at various choices of radius parameter. This procedure called "determinism scaling" allowed for the detection of the presence of "nucleation zones" in terms of a determinism peak at very low radius. This peak corresponds to the presence, in a well-defined segment of the chain, of a very strong periodic structuring of the hydrophobicity distribution.

In the present analysis,[65] the determinism (%DET, percentage of recurrent points forming line segments in recurrence plots) was calculated for a radius from 1% to 100% (the maximum; distances being rescaled on the unit interval) with an embedding of 3 to simulate a chemical environment in which each residue "views" adjacent residues in simulated three dimensions. It should be emphasized that this dimensional perspective is but a result of the mathematical "embedding" procedure and should not be

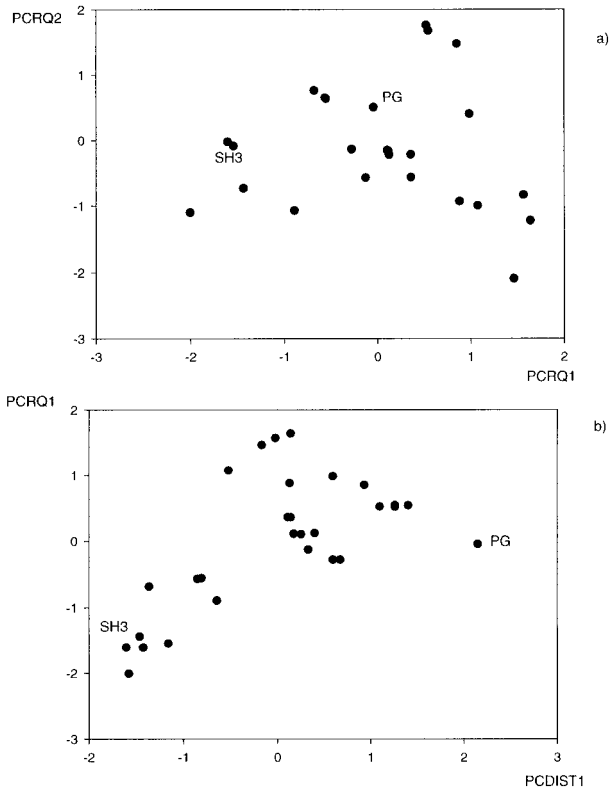


Figure 8. Distribution of chimeric sequences from PG and SH3 proteins in a principal component space. (a) The distribution of chimeric as well parent (indicated by the code) sequences in the plane spanned by the two first principal components of the RQA representation of the sequences. It is evident how the two parent structures, while being at the extremes of the distribution in terms of amino acid order and composition (all the chimeric proteins are made by mixing in different proportion segments coming from the two parent structures), are by no means at the extremes of the RQA space, pointing to the nonlinear character of the procedure. This nonlinear character is evident from the inspection of panel b, where the first component of sequence alignment space (PCDIST1) is contrasted with the first component of the RQA space (PCRQ1). The two variables, while showing a certain degree of correlation (after all, RQA is based on the amino acid sequence), clearly indicate a marked departure from linearity. (Reprinted with permission from ref 63. Copyright 2000 Oxford University Press.)

confused with real coordinates. As a a matter of fact, the two 36-mers designed to show a "fast folding" behavior were demonstrated to have a marked singularity in their determinism scaling.[65] As a control, the sequences were randomized 25 times, with a resultant loss of the above-mentioned singularity.

A similar analysis was also performed for the Syrian hamster PrP sequence, obtaining equivalent results: the determinism scaling of PrP is reported in Figure 6, where a shelf-like change in relatively linear constant %DET values (low radius region), which quickly drop off to become exponentially increasing, is evident. Exploiting the same information present in Figure 6 by a windowed RQA along the sequence indicated the highly deterministic segment responsible of the "shelf" in the scaling between residues 127 and 149.[65] This structuring should be understood not simply as a region of uniformly high hydrophobicity values but as a region of highly "concentrated" determinism, large enough (with respect to the whole protein) to influence the global scaling of the system. What is more striking is the narrowness of the shelf and of the subsequent drop. This would imply that local contacts predominate in agreement with Abkevich and colleagues.[88]

In this respect it is important to emphasize that the local character of the contacts means that residues close in space, in the usual 3D Euclidean space, constitute a nucleation center driving the subsequent folding of the entire protein. Here we add another dimension to the closeness in the Euclidean geometrical space: the closeness in the hydrophobic distribution space which is investigated by RQA. The nucleation center is identified by RQA as a localized (Euclidean space) singularity in the hydrophobic ordering of residues (chemico-physical space), thus allowing for a mechanistic interpretation of the observed folding behavior. This interpretation stems directly from the character of the %DET descriptor: this index measures the presence of "contiguous" (along the sequence) repetitive patterns of recurrences in the hydrophobicity space that correspond to a deterministic hydrophobic structuring of nearby amino acids in the usual Euclidian space, since the embedding dimension of 3 does not permit pulling away in 3D space of amino acids pertaining to the same row of the EM. Thus, a segment with an unusual deterministic profile in a windowed RQA (Figure 3) could play an important role in the folding process as nucleation center.

In the case of PrP, the individuated area may be termed singular, insofar as after the drop the %DET values increase slowly with no unique profile. In one sense this singularity is unstable. In the presence of destabilizing perturbations (e.g., ΔpH, Δtemperature, or mutations), the observed ordered hydrophobicity can be easily destroyed. A different folding could then develop with "access" to %DET patterns beyond the shelf. Presumably, this would increase the time to reach such a different state.[88]

Having discussed the hints the signal analysis perspective can give to the theoretical protein folding problem, let us shift to the more practical perspective of examining the foldability of sequences. The solution of this problem, beside the theoretical appeal, has significant practical applications in biotechnology: a current trend in biotechnological industries is generating new potentially useful proteins by the so-called "heterologous recombination"[1] consisting of the production of chimeric sequences by the fusion of DNA regions coding for two different parent enzymes, both having interesting properties from an industrial point of view. The possibility to, at least statistically, predict which of the large number of possible fusion products will effectively fold in a protein-like shape could be crucial in reducing the burden of a random search. The similarity of this problem with the class of problems encountered in combinatorial chemistry is evident.

The basic material of the present application[63] included 27 chimeric structures made up by different patches coming from two parent proteins: the α-spectrin SH3 domain (PDB code = 1shg) and the protein

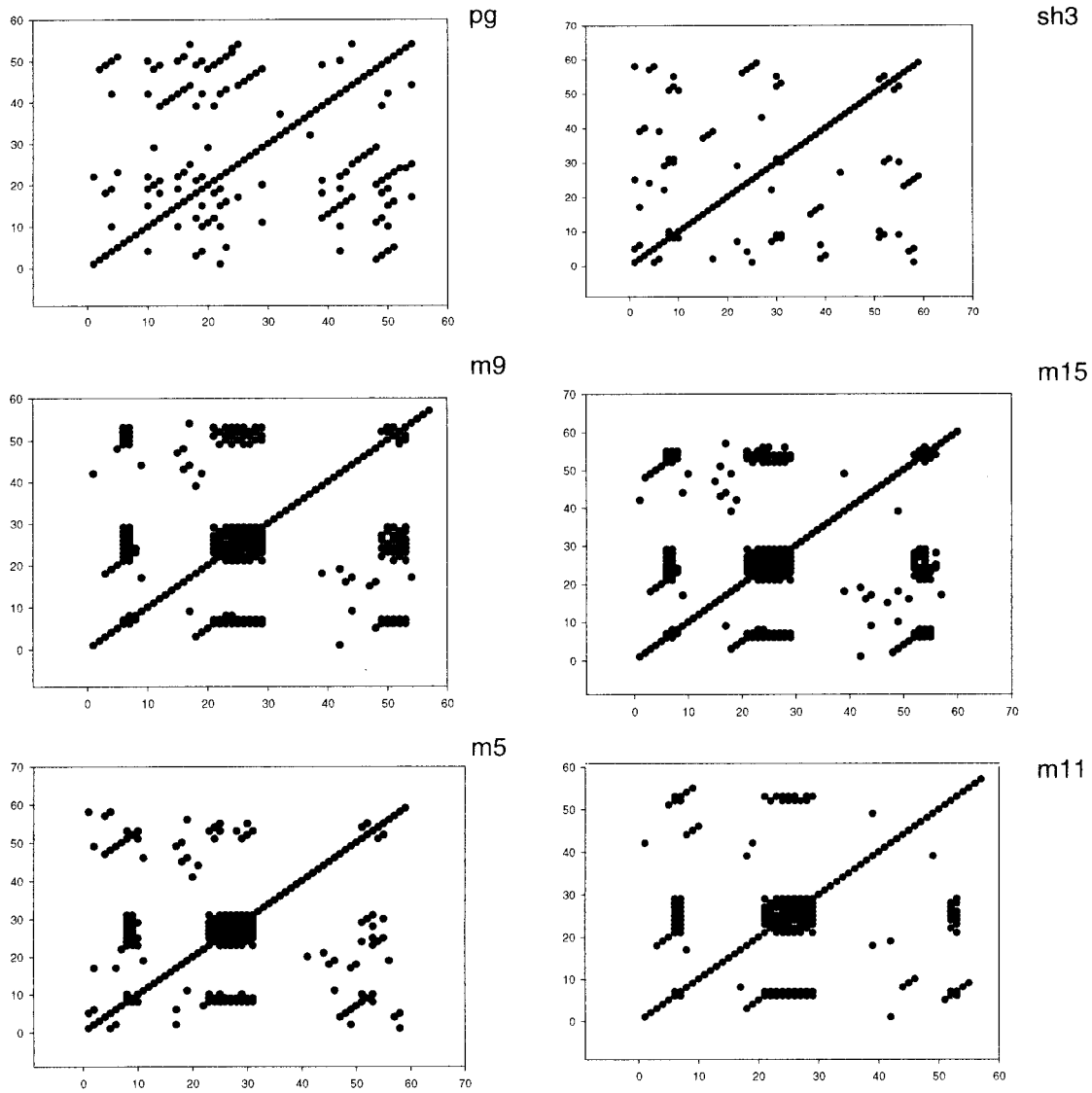


Figure 9. Recurrence plots of chimeric sequences from PG and SH3 proteins. Recurrence plots of parent (first row) and chimeric structures. It is evident how the selected chimeric sequences display a hydrophobic patterning completely different from both the parent structures. (Reprinted with permission from ref 63. Copyright 2000 Oxford University Press.)

G B1 domain (PDB code = 1pgb), having very different fold and primary structures (Figure 7).

The 27 chimeric polypeptides were generated[90] using different fractions of the two parent sequences. This implies a displacement of the chimeras along a linear axis going from 1shg to 1pgb corresponding to the relative degree of superposition of the chimeras with the parent structures. The chimeric structures were checked for their ability to fold to protein-like ordered structures. The results showed a globally unpredictable and nonlinear folding behavior along the sequence alignment axis. We checked the ability of RQA to rationalize this behavior, separating out hydrophobicity patterns (at least statistically) correlated with the folding behavior of sequences.

The hydrophobic profiles of the 27 chimeric sequences were submitted to an RQA procedure,[63] and the descriptors computed for each sequence were %REC, %DET, ENT, MAXL, TREND. Such descriptors were associated with mean and standard deviation hydrophobicity values in a 7-dimensional space which, filtered by principal component analysis, generated the first and second principal components (PCRQ1, PCRQ2), explaining about 89% of the total variability.

The distribution of the chimeras in the space spanned by PCRQ1 and PCRQ2 (Figure 8) demonstrates the type of information emerging from RQA: while 1shg and 1pgb are obviously the extremes of the sequence alignment space, they are by no means the extremes of the RQ space. This implies that by mixing patches from two parent hydrophobicity profiles, it is possible to observe a much greater diversity than that observed in the sequence space. This diversity is registered by RQA description thanks to the nonlinear character of the technique.

On a biological perspective, the generation of brand new hydrophobicity arrangements from the linear recombination of preexisting sequences could be at the basis of evolutionary changes not explainable by means of point mutation accumulation.

Figure 9 makes evident the presence of hydrophobicity patterns completely unforeseeable by the 1shg and 1pgb original patterns.

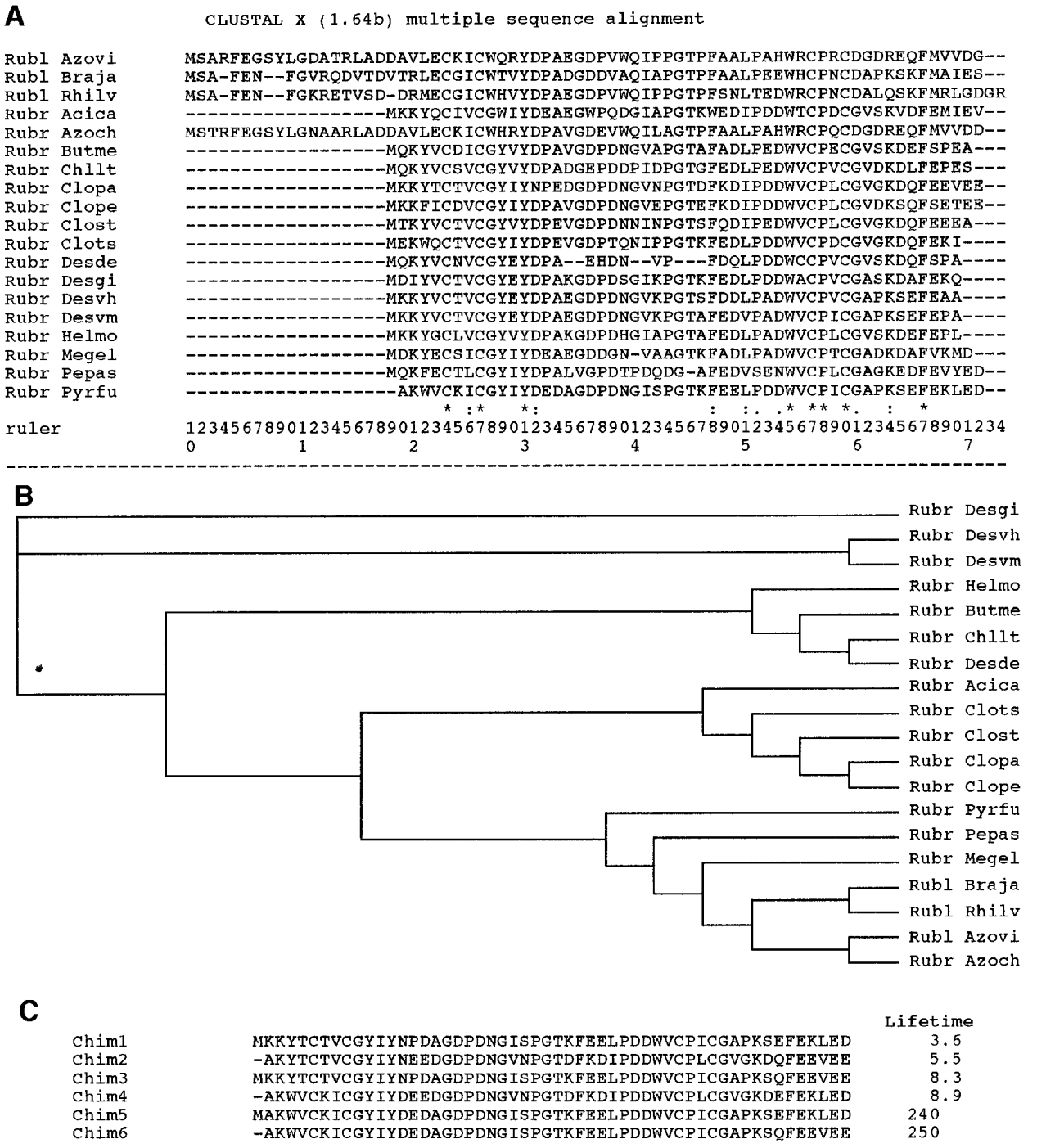


Figure 10. Sequence and thermal stability of native and chimeric rubredoxins. (a) The analyzed rubredoxins mutually aligned in order to maximize the sequence similarity according to the CLUSTAL procedure.[30] (b) The classification tree corresponding to the exploiting of sequence alignment similarities between rubredoxins. (c) The six analyzed chimeras together with their half-life (hours) at 92 C°. (Reprinted with permission from ref 54. Copyright 2000 Biophysical Society.)

From a sequence−structure prediction perspective it was important to check whether the nonlinearity with respect to the sequence alignment space introduced by RQA parametrization correlates with the folding behavior of the chimeras as revealed by the NMR spectroscopy.[90] This was demonstrated to be the case, at least on statistical grounds, given that the two major principal components of the RQ space were able to operate a statistically significant discrimination between *folders* and *nonfolders*.[63] On methodological grounds this result reinforces the idea that RQA adds new dimensions to classical sequence alignment strategies by allowing the identification of similarities and differences not apparent from sequence alignment methods (departures from the linear 1shg−1pgb axis). The demonstration that these new dimensions are linked to 3D structural properties allows one to consider the possible use of RQA in driving the synthesis of artificial enzymes by heterologous recombination of parent, natural structures.

d. Thermal Stability

This case study[59] has been directly inspired by the classical QSAR approach. It concerns a set of natural congeners (rubredoxins, a small family of bacterial

Table 2. Multivariate Matrix Relative to the 19 Rubredoxins Analyzed by Means of the Nonlinear Signal Analysis Methodologies Described in the Review[a]

name	MDIST	REC	DET	ENT	MAXL	LZ	R	SD	FD	SP1	SP2	SP3	SP4
Azovi	14.11	1.54	8.33	0	3	1.28	0.14	5.27	0.41	−0.41	−0.27	0.88	0.10
Braja	12.89	2.75	30.51	0	3	1.33	0.12	4.82	0.40	-0.28	−0.25	0.91	−0.21
Rhilv	13.79	1.27	21.43	0	3	1.23	0.01	5.11	0.39	−0.21	−0.20	0.79	−0.24
Acica	13.06	4.23	37.04	0.92	4	1.17	0.08	4.90	0.42	0.20	0.20	−0.15	−0.15
Azoch	13.70	2.00	19.15	0	3	1.37	0.04	5.10	0.13	−0.34	−0.11	0.13	−0.09
Butme	12.22	5.31	46.15	1.41	5	1.19	0.08	4.56	0.17	0.17	0.43	−0.14	−0.16
Chllt	13.07	3.92	37.50	1.37	5	1.19	0.07	4.83	0.07	0.92	0.17	−0.25	0.21
Clopa	12.90	3.37	53.49	0.86	4	1.17	0.10	4.86	0.29	0.27	0.89	−0.29	−0.21
Clope	13.24	2.51	12.50	0	4	1.28	0.01	4.96	0.30	0.04	0.87	−0.17	−0.28
Clost	12.22	4.33	45.28	1.15	5	1.30	0.15	4.58	0.06	0.84	0.35	−0.32	0.15
Clots	12.77	2.38	10.71	0	3	1.42	0.04	4.79	0.20	0.08	0.18	0.15	0.03
Desde	12.07	2.67	26.09	0	3	1.22	0.21	4.43	0.27	0.79	0.27	−0.23	0.09
Desgi	13.02	5.19	40.98	1.38	5	1.31	0.05	4.86	0.28	0.16	0.59	−0.29	0.14
Desvm	12.33	5.02	44.07	1.15	6	1.31	0.08	4.60	0.44	0.27	−0.18	−0.18	0.98
Desvh	12.37	5.44	46.88	1.66	7	1.31	0.10	4.63	0.44	0.16	−0.18	−0.15	0.99
Helmo	12.09	6.04	56.34	1.04	5	1.31	0.10	4.64	0.07	0.86	0.03	−0.21	0.19
Megel	13.10	5.02	27.12	0.72	4	1.31	0.02	4.97	0.07	0.67	−0.22	−0.21	0.13
Pepas	12.70	4.82	33.90	0.92	4	1.30	0.08	4.78	0.08	0.83	0.36	−0.49	0.13
Pyrfu	13.33	3.84	51.06	0.99	4	1.19	0.01	4.97	0.30	0.18	0.89	−0.24	−0.17

[a] Name = organism codes; MDIST= mean distance between rows of the EM; REC = percent recurrence; DET = percent determinism; ENT = entropy of the determinism lines; MAXL = length of the longest determinism line; LZ = Lempel−Ziv complexity; R = correlation coefficient between adjacent residues (absolute value); SD = standard deviation of hydrophobicity values; FD = dominant frequency (MEM spectrum) of the SVD filtered signal; SP1−SP4 = correlation coefficient between the MEM spectrum of the protein and the MEM spectrum of the modal class (from 1 to 4).

enzymes), the members of which (analogously to medicinal chemistry congeneric series) show the same property (thermal stability) at different degrees and can be easily described by a set of chemico-physical variables, namely, the various numerical descriptors of the hydrophobicity profiles generated by the different methods described in this review.

From a chemico-physical point of view, as pointed out by Plaxco et al.,[86] the problems of protein folding and stability are very closely related, being in some sense the kinetic and thermodynamic side of the same coin. Thus, deriving an efficient model to predict protein thermal stability from pure sequence information is another facet of the general sequence−structure puzzle.

Figure 10 summarizes the data set used in our study. The sequences of 19 bacterial rubredoxins (all those available at the time of the study) were retrieved, 18 pertaining to mesophilic organisms and 1 (Rubr-Pyrfu) relative to a thermophilic species (Pyrococcus Furiosus) living in thermal springs at a temperature near to 100 °C. The thermophilic sequence (panel B) was not emerging, by usual structure alignment methods, as an outlier with respect to the others. In addition to the natural rubredoxins, six chimeric sequences (panel C) were generated by using segments of two natural sequences having a practically identical three-dimensional structure but different thermal stability (Rubr-Pyrfu, Rubr-Clopa). The last column of panel C reports the half-life (expressed in hours) of each chimeric structure at 92 °C estimated from the 490 nm absorbance.[91] This point deserves further discussion: at odds with the previously discussed chimeric study, all the artificial molecules took on a well-defined 3D structure in solution. This stemmed from the fact that the two parent structures give rise to the same 3D fold, and thus, both Rubr-Pyrfu and Rubr-Clopa sequences encrypt the same solution to the sequence−structure puzzle. This makes the two solutions interchangeable and prone to combinatorial scrambling of residues in corresponding positions, maintaining virtually unaltered the 3D general solution. On the contrary, when combining structurally unrelated sequences, like SH3 and PG folds in the previous section, the probability to give rise to a realistic protein-like structure drastically lowers and brand new folds are expected by mixing two vastly different protein sequences.

Notwithstanding the close 3D structural resemblance among chimeras and between parent structures, the resulting thermal stability of the fusion products were drastically different and divided into two sharp classes of thermophilic (Rubr-Pyrfu, Chim5, Chim6) and mesophilic (Rubr-Clopa, Chim1, Chim2, Chim3, Chim4) structures (Figure 10). This implies that the serial relation from sequence to structure to properties is not generally linear in proteins and that "short cuts" directly linking sequence to properties without necessarily passing through the 3D structure are possible. This is exactly the case where practically identical 3D structures, reached by means of different amino acid linear ordering, display very different properties.

The aim of the present study was 2-fold: (a) to control whether the signal analysis methods, at odds with sequence alignment strategies, were able to recognize the peculiar character of the Rubr-Pyrfu sequence with respect to the mesophilic natural rubredoxins and (b) to predict the thermophilic/mesophilic character of the chimeric sequences and their parent structures.

Task a was approached by the simultaneous application of all the signal analysis methods described in this review to the set of 19 rubredoxins sequences coded for the hydrophobicity values of their residues. This step gave rise to the matrix reported in Table 2, i.e., the usual unit/variable matrix used in QSAR to collect the chemico-physical variables describing the studied series.

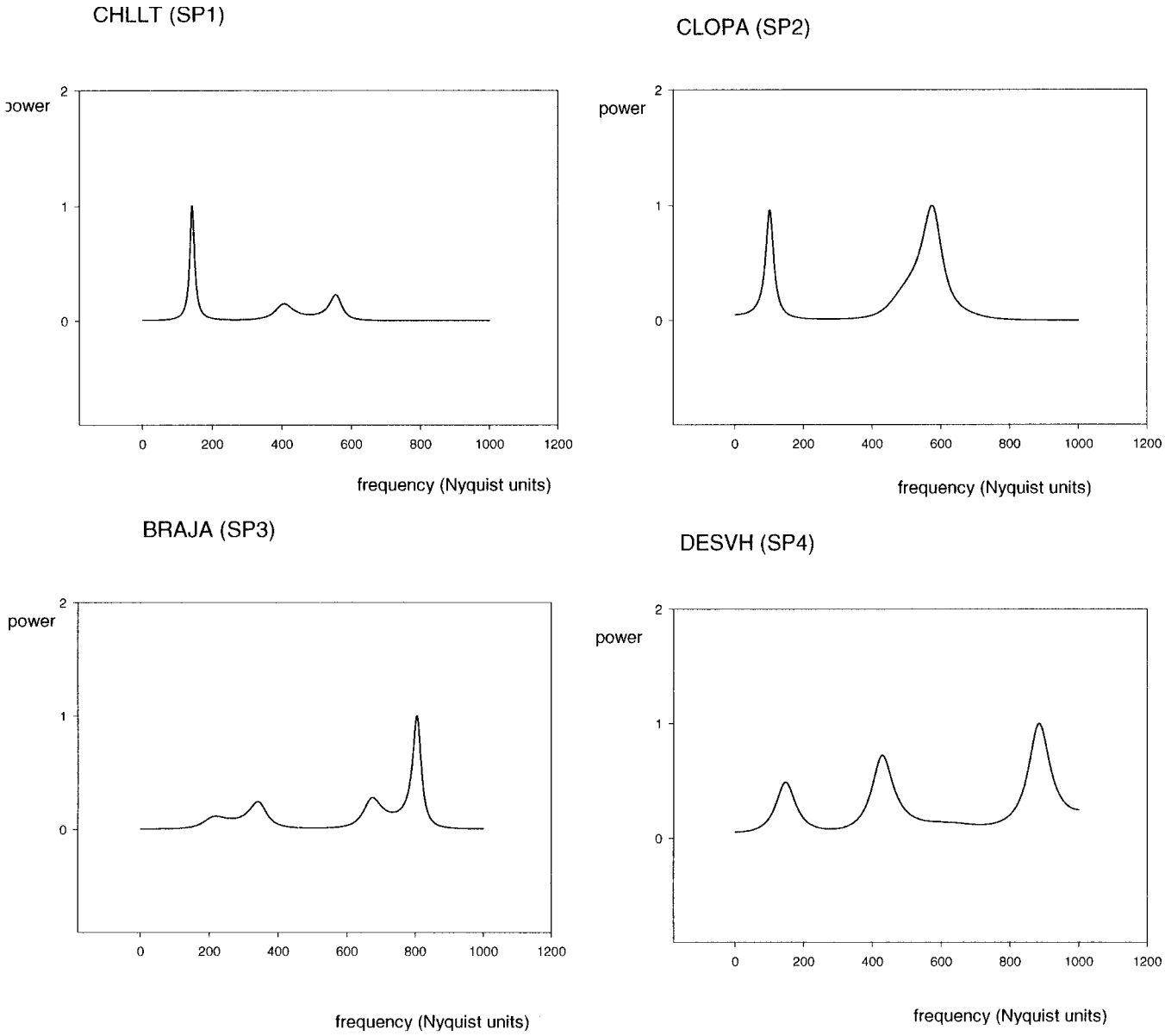


Figure 11. Maximum entropy spectra from SVD analysis of rubredoxins. The four maximum entropy all-poles spectra constituting the modal classes of the SVD spectra of the analyzed rubredoxins. The power spectra have in the abscissa the Nyquist frequency (×1000). The Nyquist critical frequency is the maximum frequency that can be resolved by spectral methods and corresponds to the reciprocal of twice the interval between data points (here set equal to 1000, corresponding to two residues). The slowest frequency corresponds to one-half the length of the entire signal (here set to 0 corresponding to more or less 25 residues). The power relative to the different frequencies is expressed in arbitrary units. It is worth noting how SVD was able to turn a noisy series (giving rise to a uniform distribution of peaks) into a smooth series with a clearly distinguishable spectrum. (Reprinted with permission from ref 54. Copyright 2000 Biophysical Society.)

The first five columns of Table 2 come from RQA and were all discussed in the methods section with the only exclusion being that of MDIST, i.e., simply the mean Euclidian distance between the rows of the EM corresponding to the protein sequence. The other variables include the LZ index (a general sequence complexity index),[92] *R* (the absolute value of the Pearson correlation coefficient between the hydrophobicities of subsequent residues), SD (the standard deviation of the hydrophobicity series), and FD (the dominant frequency of the all-poles maximum-entropy spectrum of the SVD results).

SP1−SP4 variables are the correlation coefficients between the SVD-based spectra and four prototypical classes of spectra (respectively, SP1−SP4) automatically singled out by the application of a pattern recognition method (oblique principal component analysis[93]) to the digitized 1000 points representation of the 19 rubredoxins spectra. Generally speaking, these variables measure the relative resemblance of each particular SVD spectrum with the four "modal" spectra reported in Figure 11.

To check the general consistency of the different views offered by the various methods to the description of hydrophobicity profiles, Table 2 was analyzed by means of a PCA on both the entire set and on variously reduced subsets of variables. This produced very reliable and congruent descriptions relative to the coarse graining of the amount of deterministic structuring of protein sequences pointing to a general

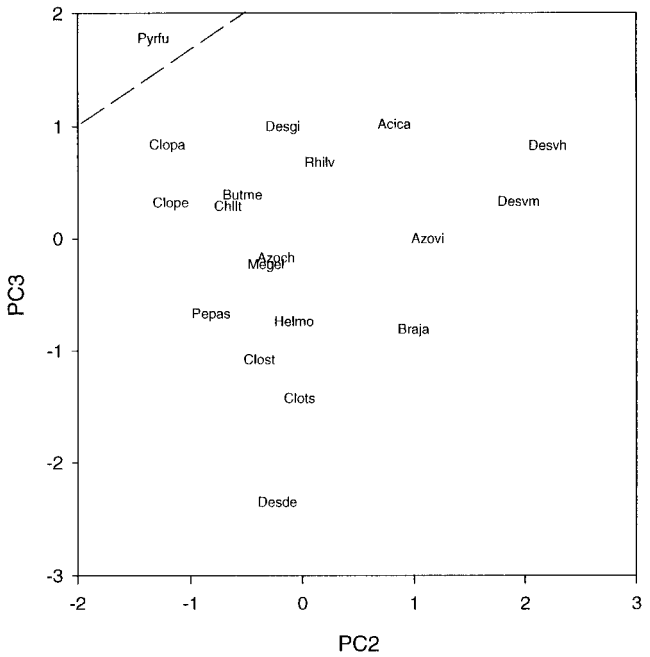


Figure 12. Distribution of rubredoxins in principle component space. The space spanned by the second and third components of the multivariate matrix having as statistical units the rubredoxins and as variables the nonlinear descriptors of their hydrophobicity plots is reported. The peculiar position of the thermophilic protein (Pyrfu) is now evident, at odds with the classical sequence alignment approach. (Reprinted with permission from ref 54. Copyright 2000 Biophysical Society.)

commonality of the different signal analysis methods. PCA of the whole data set highlighted four principal components (PC1−PC4) as necessary to attain a sufficiently rich picture of the set. In particular, the second and third components (PC2, PC3) were able to catch the peculiar position of the thermophilic sequence located at an extreme of the component space (Figure 12).

Moreover, the 3D structural resemblance between Rubr-Pyrfu and Rubr-Clopa is mirrored in the component plane by the relative closeness of the two proteins (Figure 12).

In the case of the local analysis of the chimeric space, since the two parent structures Rubr-Pyrfu and Rubr-Clopa were located in the same portion of the component space, to exploit the chimeric space between the two a much finer approach was needed than the one described above.

We tackled the problem directly using the information embedded in recurrence plots, which can be considered as a sort of lattice model in the hydrophobicity space without the filter of RQA indexes.

By this method, fine differences in hydrophobicity patterning related to thermal stability could come to light: in Figure 13 a sharp difference between thermophilic and mesophilic structure is apparent. While thermophilic structures have a broad distribution of deterministic lines, mesophilic structures display a marked concentration of deterministic lines mutually linking the two patches from residue 5 to 10 and from 37 to 42, respectively. This different patterning of deterministic lines distribution allowed for a clear-cut resolution of thermophilic and mesophilic structures.[54] It is worth noting that this discrimination is feasible only when directly examining the recurrence plots, given that Rubr-Pyrfu and Rubr-Clopa have very similar values for general RQA indexes (Table 2).

Thus, the pattern of hydrophobicity distribution rules the thermal stability character of rubredoxins in a partially independent way from general 3D structure. This result allows direct appreciation of the inherent complexity of proteins where the same linear array (amino acid sequence) simultaneously encodes for many chemico-physical properties not necessarily reflected by three-dimensional structure.

Recently Romero et al.[94] explicitly addressed the problem of the relationship between the amount of complexity of protein sequences and the degree of three-dimensional structuring. It is well-known that a number of proteins remain as flexible ensembles under physiological conditions and yet exhibit function when assayed. Such proteins have been called “natively denatured” or “natively unfolded”. Many other proteins are not intrinsically disordered throughout but rather have functionally significant regions of disorder. The application of measures of amino acid sequence complexity to the problem of detecting such “natively unfolded” proteins, while effectively demonstrating a statistically significant decrease in the average sequence complexity going from more complex to more periodic three-dimensional structures along the line, globular protein > coiled coil > collagen > silk, did not show any marked change in complexity which could be used as an identifier of disordered proteins. On the contrary, while a lower bound seems to exist for the sequence complexity of an ordered structure,[94] “intrinsically disordered” structures are found at both the lower and higher ends of the sequence complexity spectrum. These results force us to enlarge the usual concept of three-dimensional structure as an essentially fixed ensemble of coordinates and mutual distances between atomic centers, including that of a dynamical system engaging a continuous exchange with the solvent environment. In this respect, the difference between structured and unstructured proteins can be simply interpreted in terms of the amount of motions the two proteins undergo while coping with their environment. Thus, the sequence−structure puzzle cannot be simply interpreted in terms of going from amino acid sequence to a set of three-dimensional coordinates but in terms of a statistical ensemble of relations between atoms in solution. In other words, the sequence−structure puzzle has molecular dynamics as a basic ingredient. In the case of thermal stability, molecular motions are expected to play a major “structural” role, since the thermophilic character entails the generation of a dynamical ensemble able to cope with wilder motions than those experienced under mesophile conditions.

IV. Conclusions

We tried to present the general meaning and scope of the application of signal analysis methods to protein sequence−structure relationships. In our opinion, the dominant character of this nascent field

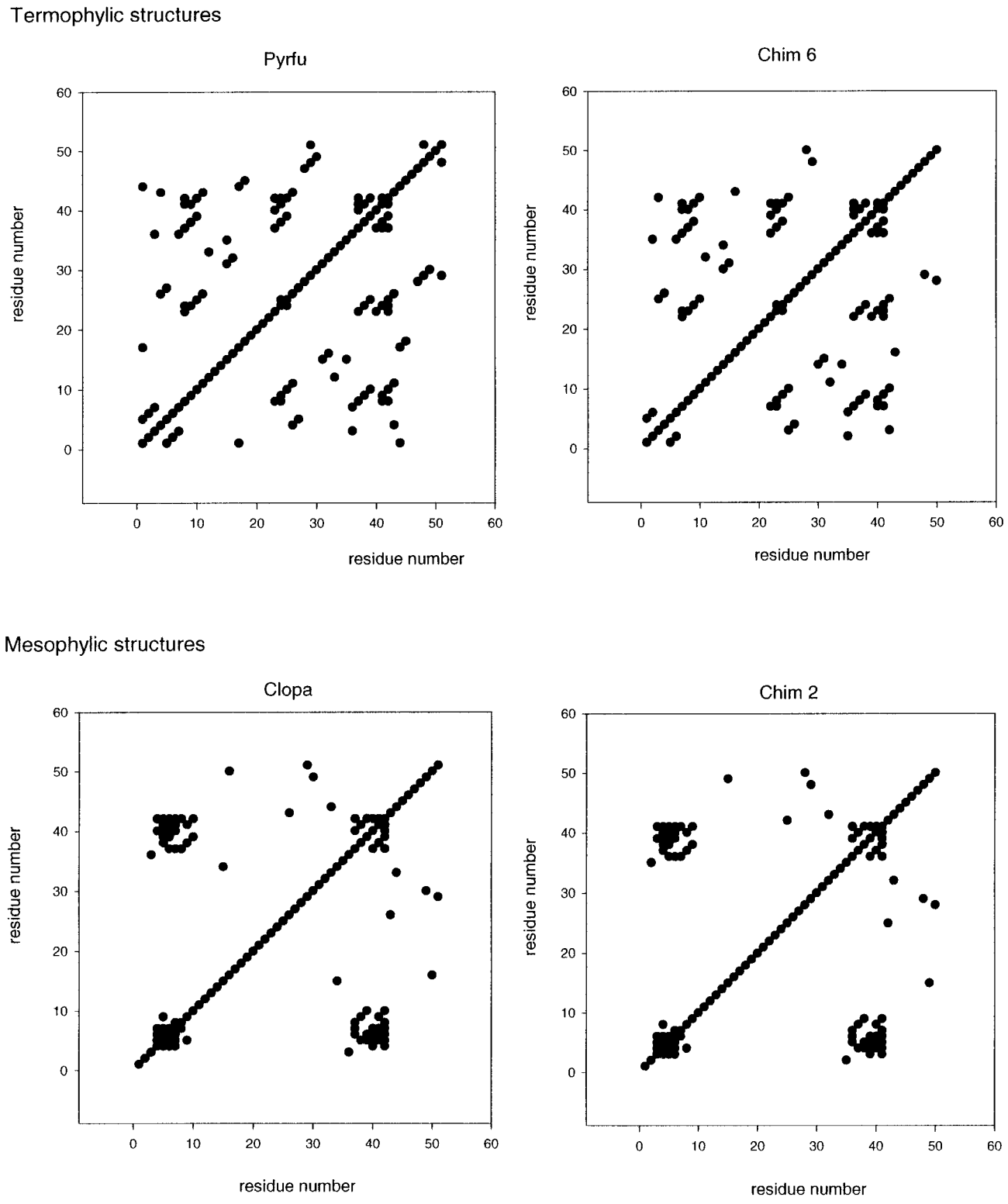


Figure 13. Recurrence plots of thermophilic and mesophilic rubredoxins. The first row reports the recurrence plots of thermophilic proteins, while the second row report the corresponding plots of mesophilic structures. It is worth recalling that all the reported proteins have practically equivalent three-dimensional structures; thus, the different patterns of recurrences is immediately linked to thermal stability without the mediating agency of 3D structure. (Reprinted with permission from ref 54. Copyright 2000 Biophysical Society.)

concerns the number of the contributing disciplines, both in terms of methodological (the data analysis techniques come directly from engineering, applied mathematics, computational physics) and theoretical influence (the basic assumptions informing the various approaches coming from molecular biology, evolutionary genetics, physical chemistry, biochemistry). As eloquently discussed in the Laughlin et al. paper entitled "The Middle Way",[6] the frontiers of science are rapidly shifting from the investigation of the basic bricks of matter to the elucidation of mesoscopic principles of organization. This represents a dramatic

change with respect to the usual way we look at the mutual relations between science.[6] It is becoming more and more evident that the most dramatic advancements in our knowledge of nature will come from those fields at the interface between different disciplines and, more important, strictly linked with practical applications.

The study of proteins is perhaps the most typical "mesoscopic" investigation field with its mixing of basic physical laws, empirical results, and qualitative descriptions.[4] In this review we described a particular approach that adopts an empirical style typical of medicinal chemistry, letting general principles emerge from the practical solution of local cases. The convergence between the study of sequence−structure relationships of proteins and QSAR is mainly methodological in nature, being based on the common use of self-consistent numerical properties of the investigated objects (organic molecules in the case of QSAR, protein sequences in the proposed approach) as starting material for prediction models.

This aspect clearly differentiates the approach from classical sequence comparison strategies: the derivation of self-consistent numerical indexes, nonlinearly related to the autocorrelation structure of hydrophobicity distribution along the sequence, allows for the direct comparison of nonhomologue proteins and for the extraction of order-dependent analogues of the amino acid chemico-physical properties (which are per se scalar).

Signal analysis techniques offer a global and "coarse-grained" view of primary structures as opposed to the local and detailed view given by sequence alignment methods. While detailed sequence alignment strategies are expected to outperform coarse-grained approaches in cases such as homology-based threading of particular structural motifs or construction of phylogenetic trees, the proposed global view could be very important in investigating nonlocal effects of single mutations on protein structures (for example, see ref 40) and, in general, in all those cases in which general properties of the protein, not confined to particular motifs and/or substructures, are expected to play a dominant role like the above-described case of rubredoxins thermal stability.

The application of a signal analysis perspective to the study of protein sequence−structure−properties relations is still in its infancy, and albeit encouraging, the thus far obtained results are still preliminary. Besides the utility of the method in solving specific, well-defined problems, what is still uncertain is the relevance of the general knowledge we can derive about protein structure and behavior from the application of time-series analysis methods. It should be stressed that we cannot think of proteins as "fully optimized" systems from which it is possible to derive the "inescapable" rules of sequence−structure relations. Simply, these general rules do not exist. This can be appreciated by a simple information theoretic reasoning: all the existent natural proteins are efficient (but surely not absolutely optimal) solutions to different but somewhat related optimization problems. These solutions have arisen from billions of years long history of natural selection; nevertheless, the number of possible proteins far exceeds the possible trials nature (intended as an organic chemist) has executed. As a matter of fact, if we think that proteins are typically about 50−500 monomers long and 20 types of monomers are used to build proteins, there are 20^n possible sequences of proteins of length n. Neither the material on the earth nor the time since the earth or the universe was formed is sufficient to exhaustively try all.[7,8] Therefore "good folders" could not have been selected by an unbiased search through the space of sequences. This is not surprising as protein sequences are the result of evolution, and thus, by definition they have a memory of the previous "synthetic steps".[8] Thus, we can imagine that the possible "folds" explored by the actual proteins are a relative minority of all the possible ones, and there is the possibility of singling out a relatively limited number of "protein structural classes" out of the myriads of actual protein structures.[7,86,95] At the same time, the basic environmental constraints shaping the "optimality landscape" of proteins are expected to be quite similar for different proteins deriving from basic chemico-physical principles (partition effects, hydrogen bonding).[96] This means that there is the possibility of finding principles at the "coarse-grain" level but not inescapable rules out of the solution of different local problems of the kind described above. These "coarse-grained" statistical regularities are expected to be approachable by means of the signal analysis perspective that operates at the same level of the principles that are expected to govern protein behavior. This kind of perspective is analogous to the one adopted by other biological fields (e.g., the study of heartbeat or electroencephalography) in which the application of nonlinear signal analysis techniques to physiological time series have provided very important insights into system functioning.

V. Abbreviations

EM	embedding matrix
RQA	recurrence quantification analysis
RP	recurrence plot
PCA	principal component analysis
QSAR	quantitative structure−activity relationships
SVD	singular value decomposition
HTM	transmembrane helix
PDB	Protein Data Bank

VI. References

(1) Lutz, S.; Benkovic, S. J. *Curr. Opin. Biotechnol.* **2000**, *11*, 319.
(2) Skolnick, J.; Fetrow, J. S.; Kolinski, A. *Nat. Biotechnol.* **2000**, *18*, 283.
(3) Teichmann, S. A.; Murzin, A. G.; Chothia, G. *Curr. Opin. Struct. Biol.* **2001**, *11*, 354.
(4) Frauenfelder, H.; Wolynes, P. *Phys. Today* **1994**, *47*, 58.
(5) Li, H., Tang, C.; Wingreen, N. S. *Phys. Rev. Lett.* **1997**, *79*, 765.
(6) Laughlin, R. B.; Pines, D.; Schmalian, G.; Wolynes, P. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 32.
(7) Taylor, W. R.; May, A. C.; Brown, N. P.; Aszodi, A. *Rep. Prog. Phys.* **2001**, *64*, 517.
(8) Pande, V. S.; Grosberg, A. Y.; Tanaka, T. *Rev. Mod. Phys.* **2000**, *72*, 259.
(9) Branden, C. I.; Tooze, J. *Introduction to Protein Structure*; Garland: New York, 1991.
(10) Adkins, J. N.; Lumb, K. J. *Proteins: Struct. Funct. Genet.* **2002**, *1*, 1.
(11) Senno, C. F.; Micheletti, A.; Maritan, A.; Banavar, J. R. *Phys. Rev. Lett.* **1998**, *80*, 2237.

(12) Weiss, O.; Herzel, H. *J. Theor. Biol.* **1998**, *190*, 341.
(13) Anfinsen, C. B. *Science* **1973**, *181*, 223.
(14) Guharay, S.; Hunt, B. R.; Yorke, J. A.; White, O. R. *Physica D* **2000**, *146*, 388.
(15) Dobson, C. M.; Karplus, M. *Curr. Opin. Struct. Biol.* **1999**, *9*, 92.
(16) Chakraborty, A. K. *Phys. Rep.* **2001**, *342*, 1.
(17) Grigoriev, I. V.; Kim, S. H. *Proteins: Struct. Funct. Genet.* **1999**, *96*, 14318.
(18) Nowak, M. A. *J. Theor. Biol.* **2000**, *204*, 179.
(19) Irback, A.; Sandelin, E. *Biophys. J.* **2000**, *79*, 2252.
(20) Simons, K. T.; Strauss, C.; Baker, D. *J. Mol. Biol.* **2001**, *306*, 1191.
(21) *Old and New views of protein folding*; Kuwajima, K., Arai, M., Eds.; Elsevier: Amsterdam, 1999.
(22) Dill, K. A. *Biochemistry* **1990**, *29*, 7133.
(23) *Monte Carlo approach to biopolymers and protein folding*; Grassberger, P., Barkema, G. T., Nadler, W., Eds.; World Scientific: Singapore, 1997.
(24) Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445.
(25) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *56*, 561.
(26) Chan, H. S.; Dill, K. A. *J. Chem. Phys.* **1989**, *90*, 493.
(27) Chasman, D.; Adams, R. M. *J. Mol. Biol.* **2001**, *307*, 683.
(28) Sander, C.; Schneider, R. *Proteins: Struct. Funct. Genet.* **1991**, *9*, 56.
(29) Guex, N.; Diemand, A.; Peitsch, M. C. *Trends Biochem. Sci.* **1999**, *24*, 364.
(30) *Computational Methods in Molecular Biology*; Salzberg, S. L., Searls, D. B., Kasif, S., Eds.; Elsevier: Amsterdam, 1998.
(31) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291.
(32) Ortiz, A. R.; Kolinski, A.; Rotkiewicz, P.; Ilkowski, B.; Skolnick, J. *Proteins: Struct. Funct. Genet.* **1999**, *37*, 177−185.
(33) Baker, D.; Sali, A. *Science* **2001**, *294*, 93.
(34) Hansch, C. *Acc. Chem. Res.* **1993**, *26*, 147.
(35) Hansch, C.; Hoekman, D.; Gao, H. *Chem. Rev.* **1996**, *96*, 1045.
(36) Banavar, J. R.; Maritan, A. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 433.
(37) Benigni, R.; Giuliani, A. *Am. J. Physiol.* **1994**, *266*, R1697.
(38) Sweet, R. M.; Eisenberg, D *J. Mol. Biol.* **1983**, *171*, 479.
(39) Makhadatze, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307.
(40) Sinha, N.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 3139.
(41) von Heijne, G. *J. Mol. Biol.* **1982**, *159*, 537.
(42) Pearson, W. R.; Lipman, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444.
(43) Fujita, T.; Iwasa, J.; Hansch, C. *J. Am. Chem Soc.* **1964**, *86*, 5175.
(44) Hammett, L. P. *Chem. Rev.* **1935**, *17*, 125.
(45) Taft, R.; Newman, M. S.; Verhoek, F. H. *J. Am. Chem. Soc.* **1950**, *72*, 4511.
(46) Martin, Y. *J. Med. Chem.* **1981**, *24*, 229.
(47) Hansch, C.; Leo, A. *Exploring QSAR 1. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, D.C., 1995.
(48) Fujita, T. *Quant. Struct.-Act. Relat.* **1997**, *16*, 107.
(49) Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. *Chem. Rev.* **2000**, *100*, 3697.
(50) Lipophilicity in Drug Action and Toxicology. In *Methods and Principles in Medicinal Chemistry*; Pliska, W., Testa, B., van de Waterbeeemd, H., Eds.; WCH: Weinheim, 1996; Vol. 4.
(51) Hansch, C. In *Biological Activity and Chemical Structure*; Keverling Huismann, G. A., Ed.; Elsevier: Amsterdam, 1977; pp 47−61.
(52) Palliser, C.; Parry, A. D. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 243.
(53) Schreiber, T. *Phys. Rep.* **1999**, *308*, 1.
(54) Giuliani, A.; Benigni, R.; Sirabella, P.; Zbilut, J. P.; Colosimo, A. *Biophys. J.* **2000**, *78*, 136.
(55) Mandell, A. J.; Selz, K.; Shlesinger, M. F. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 13576.
(56) Broomhead, D. S.; King, G. P. *Physica D* **1986**, *20*, 217.
(57) Selz, K. A.; Mandell, A. J.; Shlesinger, M. F. *Biophys. J.* **1998**, *75*, 2332.
(58) Eckmann, J. P.; Kamphorst, S. O.; Ruelle, D. *Europhys. Lett.* **1987**, *4*, 324.
(59) Webber, C. L.; Zbilut, J. P. *J. Appl. Physiol.* **1994**, *76*, 965.
(60) Giuliani, A.; Piccirillo, G.; Marigliano, V.; Colosimo, A. *Am. J. Physiol.* **1998**, *275*, H1455.
(61) Manetti, C.; Ceruso, M. A.; Giuliani, A.; Webber, C. L.; Zbilut, J. P. *Phys. Rev. E* **1999**, *59*, 992.
(62) Rustici, M.; Caravati, C.; Patretto, E.; Branca, M.; Marchettini, N. *J. Phys. Chem. A* **1999**, *103*, 6564.
(63) Giuliani, A.; Sirabella, P.; Benigni, R.; Colosimo, A. *Protein Eng.* **2000**, *13*, 671.
(64) Zbilut, J. P.; Giuliani, A.; Webber, C. L.; Colosimo, A. *Protein Eng.* **1998**, *11*, 87.
(65) Zbilut, J. P.; Webber, C. L.; Colosimo, A.; Giuliani, A. *Protein Eng.* **2000**, *13*, 99.
(66) Webber, C. L.; Giuliani, A.; Zbilut, J. P.; Colosimo, A. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 292.
(67) Feller, W. *An Introduction to Probability Theory and Its Applications*; Wiley: New York, 1968; Vol. 1.
(68) Rao, C. R.; Suryawanshi, S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 12132.
(69) Trulla, L. L.; Giuliani, A.; Zbilut, J. P.; Webber, C. L. *Phys. Lett. A* **1996**, *223*, 225.
(70) Oprea, T. I.; Gottfries, J. *J. Comb. Chem.* **2001**, *3*, 157.
(71) Mandell, A. J.; Owens, M. J.; Selz, K. A.; Morgan, W. N.; Shlesinger, M. F.; Nemeroff, C. B. *Biopolymers* **1998**, *46*, 89.
(72) Kaiser, G. *A Friendly Guide to Wavelets*; Birkhauser: Boston, 1994.
(73) Meyer, Y. *Wavelets: Algorithms and Applications*; Society for Industrial and Applied Mathematics: Philadelphia, 1993.
(74) Lio, P.; Vannucci, M. *Bioinformatics* **2000**, *16*, 376.
(75) Kite, J.; Doolitle, R. F. *J. Mol. Biol.* **1982**, *157*, 105.
(76) Fasman, G. D; Gilbert, W. A. *Trends Biochem. Sci.* **1990**, *15*, 89.
(77) Von Heijne, G. *J. Mol. Biol.* **1992**, *225*, 487.
(78) Rose, G. D. *Nature* **1978**, *272*, 586.
(79) Deber, C. M.; Wang, C.; Liu, L. P.; Prior, A. S.; Agrawal, S.; Muskat, B.; Cuticchia, A. J. *Protein Sci.* **2001**, *10*, 212.
(80) Donoho, D.; Johnstone, I. *Ann. Stat.* **1998**, *26*, 879.
(81) Hirakawa, H.; Muta, S.; Kuhara, S. *Bioinformatics* **1999**, *15*, 141.
(82) Pratt, L. R.; Chandler, D. *J. Chem. Phys.* **1977**, *67*, 3683.
(83) Kaiser, E. T.; Kezdy, F. S. *Science* **1984**, *223*, 249.
(84) Beattie, J.; Shand, J.; Flint, D. J. *Eur. J. Biochem.* **1996**, *239*, 479.
(85) Harrison, P. M.; Chan, H. S.; Prusiner, S. B.; Cohen, F. E. *Protein Sci.* **2001**, *10*, 819.
(86) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. *Biochemistry* **2000**, *39*, 1177.
(87) Klimov, D. K.; Thirumalai, D. *Fold. Des.* **1998**, *3*, 127.
(88) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Proteins: Struct. Funct. Genet.* **1998**, *31*, 335.
(89) James, T. L.; Lin, H.; Ulyanov, N. B.; Farr-James, S.; Zhang, H.; Daure, D. G.; Kaneko, K.; Groth, D.; Mehlon, I.; Prusiner, S. E.; Cohen, F. E. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10086.
(90) Blanco, F. J.; Angrand, I.; Serrano, L. *J. Mol. Biol.* **1999**, *285*, 741.
(91) Eidsness, M.; Richie, K. A.; Burden, A. E.; Kurtz, D. M.; Scott, R. A. *Biochemistry* **1997**, *366*, 10406.
(92) Kaspar, F.; Shuster, K. G. *Phys. Rev. A* **1987**, *36*, 842.
(93) Harman, H. H. *Modern Factor Analysis*, 3rd ed.; Chicago University Press: Chicago, 1976.
(94) Romero, P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, A. K. *Proteins: Struct. Funct. Genet.* **2001**, *42*, 38.
(95) Tiana, G.; Broglia, R. A.; Shakhnovich, E. *Proteins: Struct. Funct. Genet.* **2000**, *39*, 244.
(96) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.

CR0101499