Nonlinear Methods in the Analysis of Protein Sequences: A Case Study in Rubredoxins

Alessandro Giuliani,* Romualdo Benigni,* Paolo Sirabella,† Joseph P. Zbilut,‡ and Alfredo Colosimo†
*TCE Laboratory, Istituto Superiore di Sanitá, 00161 Roma, Italy; †Department of Biochemical Sciences, University of Roma "La Sapienza," 00185 Roma, Italy; and ‡Department of Molecular Biophysics and Physiology, Rush University, Chicago, Illinois 60612 USA

ABSTRACT Two computational methods widely used in time series analysis were applied to protein sequences, and their ability to derive structural information not directly accessible through classical sequence comparisons methods was assessed. The primary structures of 19 rubredoxins of both mesophilic and thermophilic bacteria, coded with hydrophobicity values of amino acid residues, were considered as time series and were analyzed by 1) recurrence quantification analysis and 2) spectral analysis of the sequence major eigenfunctions. The results of the two methods agreed to a large extent and generated a classification consistent with known 3D structural characteristics of the studied proteins. This classification separated in a clearcut manner a thermophilic protein from mesophilic proteins. The classification of primary structures given by the two dynamical methods was demonstrated to be basically different from classification stemming from classical sequence homology metrics. Moreover, on a more detailed scale, the method was able to discriminate between thermophilic and mesophilic proteins from a set of chimeric sequences generated from the mixing of a mesophilic (*Rubr Clopa*) and a thermophilic (*Rubr Pyrfu*) protein. Overall, our results point to a new way of looking at protein sequence comparisons.

INTRODUCTION

The relationship of the physiological role of proteins with their primary structure is a crucial issue in molecular biology. It is well known that both the 3D structure and the physiological role of proteins is heavily dependent upon the particular linear arrangements of amino acid residues along polypeptide chains, or primary structures (Anfinsen, 1973; Dayhoff et al., 1978; Sweet and Eisenberg, 1983). Despite the recent progress in the field, fully documented in the series of Critical Assessment of Techniques of Proteins Structure Prediction (CASP) meetings (Murzin and Patthy, 1999), attempts to derive general rules in predicting 3D structure and physiological features based on protein sequences cannot be considered fully satisfactory as yet (Micheletti et al., 1998). Thus, we decided to tackle this problem with a local approach: instead of looking for general rules, we tried to develop a methodology able to derive local structure/sequence relationship models, with the ultimate goal of predicting physiological properties by means of sequence information within properly selected sets of proteins. This approach is similar to one used in medicinal chemistry (Hansch, 1993; Martin, 1981), where quantitative models for predicting pharmacological properties of organic molecules from their physicochemical and structural properties have been routinely used for the past three decades (Hansch et al., 1962; Hansch and Leo, 1995). The success of quantitative structure activity relationship (QSAR) models has been demonstrated (Martin, 1981; Hansch and Leo, 1995; Hansch et al., 1996) to be closely linked to the use of strictly congeneric molecules, i.e., of the same class.

From an operational point of view, a QSAR procedure is based on the correct selection of two components in a prediction model: 1) a meaningful set of "x" variables (regressors of the model, descriptors spanning the physicochemical space in which the molecules are located); and 2) an adequate statistical technique to quantitatively tackle the problem of both physicochemical space description and biological activity prediction (Martin, 1981; Franke, 1984). In the present case, the above elements are 1) hydrophobicity, as the physicochemical descriptor of amino acid residues, and 2) time series analysis as a general strategy to describe the proteins' primary structures. [We note that from a practical standpoint, spatially ordered series are equivalent to time-ordered series for the purpose of analysis; see (Zbilut et al., 1998a)].

The relevance of hydrophobicity for protein folding is well known (Anfinsen, 1973; Li et al., 1997; Micheletti et al., 1998; Sweet and Eisenberg, 1983). Additionally, hydrophobicity is the only physicochemical feature that displays a nonrandom ordering along protein sequences (Weiss and Herzel, 1998; von Heijne, 1982), and may thus be considered the best candidate for application of time series analysis methods. The amino acid residues along protein chains were coded in terms of their hydrophobicity, expressed as log P, P being the partition coefficient between octanol and water (Franke, 1984). Thus, our analysis focused on the hydrophobicity series with the idea of using hydrophobicity as the "semantics" attached to the residues' ordering along a chain. [The semantic character of a physicochemical property should reflect both energetic interchange with elements present in the environment (e.g., water molecules) and entropic rearrangements induced in them.] An additional goal

Received for publication 31 December 1998 and in final form 20 October 1999.

Address reprint requests to Dr. Alfredo Colosimo, Department of Biochemical Sciences, University of Roma "La Sapienza," 00185 Roma, Italy. Tel.: 39-06-499-10957; Fax: 39-06-444-0062; E-mail: a.colosimo@caspur.it.

0006-3495/00/01/136/13 $2.00

was to check whether our approach was able to generate different, complementary information for protein classification, with respect to standard sequence comparison methods (Pearson and Lipman, 1988; Wilbur and Lipman, 1983; Thompson et al., 1997).

With respect to computational aspects, the ability of recurrence quantification analysis (RQA) to deal with a sequence/function relationship problem has been recently demonstrated (Zbilut et al., 1998b), while almost at the same time, Mandell and his co-workers (1997; 1998; Selz et al., 1998) reported that singular value decomposition (SVD) in concert with spectral analysis might be able to provide useful structural 3D information from sequence data. Thus, we combined the two approaches to obtain a global representation of hydrophobicity patterns along a sequence. These two main methods were supplemented with three other relatively simple descriptors, namely 1) standard deviation (SD), 2) absolute value of the Pearson correlation coefficient between adjacent residues (R), and 3) algorithmic complexity of the series (Kaspar and Schuster, 1987) as estimated by the Lempel-Ziv algorithm (LZ).

The entire set of descriptors was filtered by principal component analysis (PCA) (Harman, 1976) in order to obtain a set of orthogonal axes on which to project the studied sequences (dynamical space). [We recognize the essential mathematical equivalence between SVD and PCA. In the present context, we use the term PCA to distinguish this step in the algorithm from the SVD plus spectral analysis step.] The chosen "congeneric series" of proteins consists of 19 rubredoxins (Sieker et al., 1994), and corresponds to all the bacterial rubredoxins whose primary structures were known at the time of our analysis. The biological feature to be predicted is the exceptional stability of the rubredoxin from *Pyrococcus furiosus*, a thermophile bacterium living at very high temperatures.

Finally, the thermostability of rubredoxins was also investigated at a much finer level of detail by using our approach to evaluate the results by Eidsness et al. (1997), who synthesized six chimeric proteins originating from a thermophilic sequence and a mesophilic one. These authors observed the splitting of the six chimeric proteins into a thermophilic subset and a mesophilic subset, although a specific mechanistic explanation for this behavior could not be found (Eidsness et al., 1997). Even under such stringent sensitivity requirements, the predictive abilities of our approach were successful.

MATERIALS AND METHODS

The biochemical problem

Rubredoxins are probably the simplest members of the ubiquitous and huge family of redox metalloenzymes (Sieker et al., 1994) and consist of a relatively short polypeptide chain (~53 AA) endowed with a prosthetic group in the form of a ferrous/ferric ion tetrahedrically bound to four cysteine (Cys) residues. Even though their exact metabolic role in anaerobic cells has not yet been fully clarified, their structural features are fairly well known, and in Fig. 1 *A* the primary structures of 19 bacterial rubredoxins are reported: they are quite similar, and >20% of the residues are strictly conserved, among which are the four Cys residues of the active site and the five aromatic residues that constitute the hydrophobic core of the proteins. Such a high level of homology, however, does not find a match in the huge spectrum of thermal stability of the bacterial strains from which they are extracted. In particular, it has been impossible up to now to find a rationale, on the basis of the primary as well as of the tertiary structures, for the fact that the rubredoxin from *Pyrococcus furiosus* (*Rubr Pyrfu*), which lives normally at 90°C, has a half-time of thermal denaturation of 400 h at 92°C, as compared to 6 h of *Clostridium pasteurianum* (*Rubr Clopa*) rubredoxin, which is the most similar to it in terms of 3D structure. In the same figure the phylogenetic tree of rubredoxins corresponding to their best alignment and the structure of the chimeric rubredoxins obtained by Eidsness et al. (1997) starting from *Rubr Clopa* and *Rubr Pyrfu* are also reported.

Data analysis

Dynamical methods

Each sequence was coded in terms of the amino acid hydrophobicities (Franke, 1984). The numerical discrete series corresponding to the protein sequence was then submitted to a 4D embedding by the method of delays (Broomhead and King, 1986). The embedding procedure consists in building an n-columns matrix (in our case $n = 4$) out of the original linear array, by shifting the series by a fixed lag. For example, given the series 10, 11, 21, 32, 41, 35, 40, 19. . . the corresponding 4D embedding space at lag of 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. The choice of the embedding dimension was dictated by a balance between the need for having a window large enough to keep track of between-residues interactions and on relying—at the same time—on a sufficiently long series (Broomhead and King, 1986; Webber and Zbilut, 1994). Notice that the last n values are eliminated from the analysis as an obvious consequence of shifting the series for the embedding. Moreover, the four-residues window was demonstrated (Strait and Dewey, 1996) through the application of a formalism derived from information theory, to constitute an upper limit for the information content of protein sequences. [We note that Kumosinski and Liebman (1994) have previously explored the use of a somewhat related methodology.] After the common step of building an EM, RQA and SVD diverge. While RQA, being based on the EM rows, assumes a local viewpoint over the time series (Giuliani et al., 1998; Zbilut et al., 1998a) SVD, being based on EM columns and dealing with average regularities, provides a global view of the series (Broomhead and King, 1986; Mandell et al., 1997). The former and the latter method are then particularly sensitive (Trulla et al., 1996), and relatively insensitive, respectively, to small perturbations and provide a complementary view of the autocorrelation structure of the time series (Zbilut et al., 1998a).

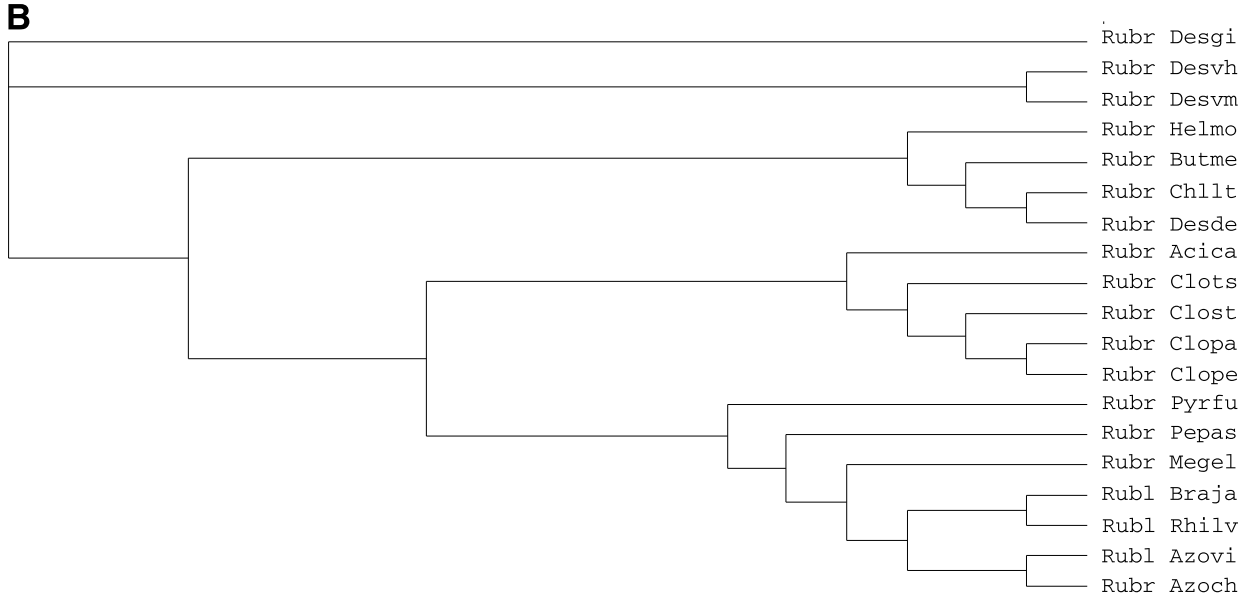
RQA-based descriptors. RQA was first introduced in physics by Eckmann et al. (1987) as a purely graphical technique (see the Appendix).

A

```
                CLUSTAL X (1.64b) multiple sequence alignment

Rubl Azovi      MSARFEGSYLGDATRLADDAVLECKICWQRYDPAEGDPVWQIPPGTPFAALPAHWRCPRCDGDREQFMVVDG--
Rubl Braja      MSA-FEN--FGVRQDVTDVTRLECGICWTVYDPADGDDVAQIAPGTPFAALPEEWHCPNCDAPKSKFMAIES--
Rubl Rhilv      MSA-FEN--FGKRETVSD-DRMECGICWHVYDPAEGDPVWQIPPGTPFSNLTEDWRCPNCDALQSKFMRLGDGR
Rubr Acica      ------------------MKKYQCIVCGWIYDEAEGWPQDGIAPGTKWEDIPDDWTCPDCGVSKVDFEMIEV--
Rubr Azoch      MSTRFEGSYLGNAARLADDAVLECKICWHRYDPAVGDEVWQILAGTPFAALPAHWRCPQCDGDREQFMVVDD--
Rubr Butme      ------------------MQKYVCDICGYVYDPAVGDPDNGVAPGTAFADLPEDWVCPECGVSKDEFSPEA---
Rubr Chllt      ------------------MQKYVCSVCGYVYDPADGEPDDPIDPGTGFEDLPEDWVCPVCGVDKDLFEPES---
Rubr Clopa      ------------------MKKYTCTVCGYIYNPEDGDPDNGVNPGTDFKDIPDDWVCPLCGVGKDQFEEVEE--
Rubr Clope      ------------------MKKFICDVCGYIYDPAVGDPDNGVEPGTEFKDIPDDWVCPLCGVDKSQFSETEE--
Rubr Clost      ------------------MTKYVCTVCGYVYDPEVGDPDNNINPGTSFQDIPEDWVCPLCGVGKDQFEEEA---
Rubr Clots      ------------------MEKWQCTVCGYIYDPEVGDPTQNIPPGTKFEDLPDDWVCPDCGVGKDQFEKI----
Rubr Desde      ------------------MQKYVCNVCGYEYDPA--EHDN--VP---FDQLPDDWCCPVCGVSKDQFSPA----
Rubr Desgi      ------------------MDIYVCTVCGYEYDPAKGDPDSGIKPGTKFEDLPDDWACPVCGASKDAFEKQ----
Rubr Desvh      ------------------MKKYVCTVCGYEYDPAEGDPDNGVKPGTSFDDLPADWVCPVCGAPKSEFEAA----
Rubr Desvm      ------------------MKKYVCTVCGYEYDPAEGDPDNGVKPGTAFEDVPADWVCPICGAPKSEFEPA----
Rubr Helmo      ------------------MKKYGCLVCGYVYDPAKGDPDHGIAPGTAFEDLPADWVCPLCGVSKDEFEPL----
Rubr Megel      ------------------MDKYECSICGYIYDEAEGDDGN-VAAGTKFADLPADWVCPTCGADKDAFVKMD---
Rubr Pepas      ------------------MQKFECTLCGYIYDPALVGPDTPDQDG-AFEDVSENWVCPLCGAGKEDFEVYED--
Rubr Pyrfu      ------------------AKWVCKICGYIYDEDAGDPDNGISPGTKFEELPDDWVCPICGAPKSEFEKLED--
                                        * :*   *:               :  :. .* ** *.  :  *
ruler           12345678901234567890123456789012345678901234567890123456789012345678901234
                0         1         2         3         4         5         6         7
```

B



C

		Lifetime
Chim1	MKKYTCTVCGYIYNPDAGDPDNGISPGTKFEELPDDWVCPICGAPKSEFEKLED	3.6
Chim2	-AKYTCTVCGYIYNEEDGDPDNGVNPGTDFKDIPDDWVCPLCGVGKDQFEEVEE	5.5
Chim3	MKKYTCTVCGYIYNPDAGDPDNGISPGTKFEELPDDWVCPICGAPKSQFEEVEE	8.3
Chim4	-AKWVCKICGYIYDEEDGDPDNGVNPGTDFKDIPDDWVCPLCGVGKDEFEKLED	8.9
Chim5	MAKWVCKICGYIYDEDAGDPDNGISPGTKFEELPDDWVCPICGAPKSEFEKLED	240
Chim6	-AKWVCKICGYIYDEDAGDPDNGISPGTKFEELPDDWVCPICGAPKSQFEEVEE	250

FIGURE 1 Primary structures of rubredoxins used in this work. (*A*) The best alignment of the primary structures of 19 rubredoxins as provided by the *Clustal X* program (see text). The reported sequences are all the known bacterial rubredoxins at the time of the completion of the present work (Sept. 1998) and derive from the Swiss Prot data base (see also Table 2). (*B*) The phylogenetic tree corresponding to the alignment in (*A*) as given by the neighbor joining method of Saitou and Nei (1987) and (Page, 1996). (*C*) The sequence of the six chimeric rubredoxins obtained by Eidsness et al. (1997) starting from *Rubr Pyrfu* and *Rubr Clopa*. The correspondence between the nomenclature in the original paper and in this one (from *chim1* to *chim6*) is the following: Cp15|Pf, [M1 ,K2A, P15E]Cp, Cp15|Pf47|Cp, Pf15|Cp47|Pf, [1M]Pf, Pf47|Cp. The last column in (*C*) reports the spectroscopic lifetime of each chimeric structure at 92°C estimated from 490 nm absorbance changes by Eidsness et al. (1997).

Subsequently, Zbilut and Webber (1992) enhanced the technique by defining five nonlinear quantitative descriptors of the recurrence plot that were found to be diagnostically useful in the quantitative assessment of time series structure in fields ranging from molecular dynamics to physiology (Giuliani and Manetti, 1996; Webber and Zbilut, 1994; Faure and Korn, 1997).

The RQA descriptors used in this work are:

1. MDIST. Average Euclidean distance between the rows of EM;
2. REC (percent recurrence). This measure quantifies the fraction of the plot filled by recurrent points. It corresponds to the fraction of recurrent pairs over all the possible pairs of epochs or, equivalently, to the fraction of pairwise distances below the chosen radius among all the computed distances;
3. DET (percent determinism). This is the percentage of sequential recurrent points that form diagonal line structures in the distance matrix. DET corresponds to the amount of patches of similar hydrophobic/hydrophilic characteristics along the sequence;
4. ENT (entropy).The entropy is defined here in terms of the Shannon-Weaver formula for information entropy computed over the distribution of length of the lines of recurrent points and measures the richness of deterministic structures of the series;
5. MAXLINE (maximal line).This index is simply the length (in terms of consecutive points) of the longest recurrent line in the plot, and is inversely related to the largest positive Ljapunov exponent.

In Fig. 2 the recurrence plots of four rubredoxin sequences are shown. In addition to the above-mentioned descriptors, the distribution of recurrent points in the plot was quantified by recurrence displacement histograms (see Fig. 3), which report the average recurrence of the plot as a function

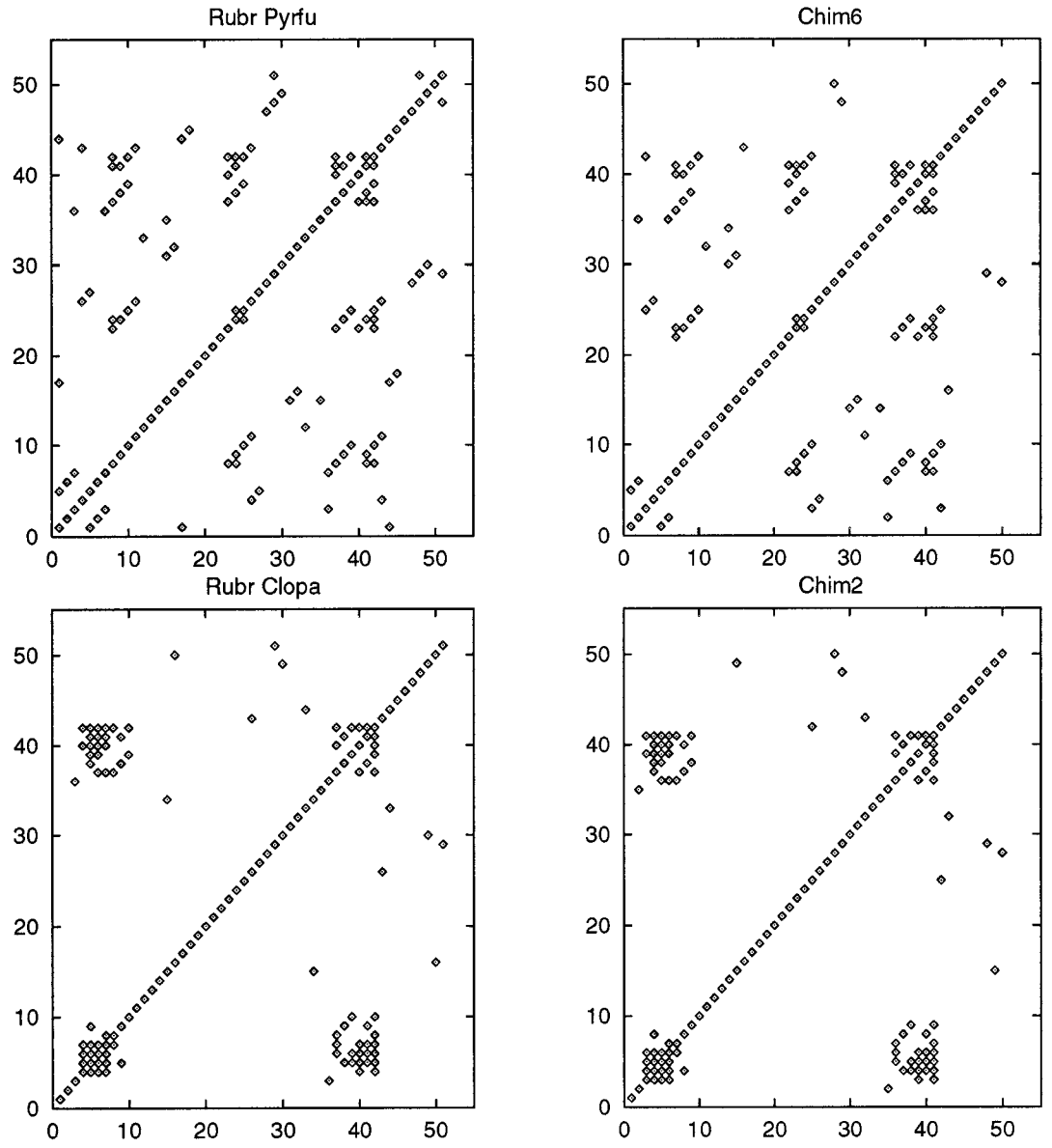


FIGURE 2 Recurrence plots of four rubredoxin sequences. The top line reports the recurrence plots of two thermophilic proteins (*Rubr Pyrfu* and *chim6*, see text for details) while in the bottom line the plots of two mesophilic proteins are shown (*Rubr Clopa* and *chim2*). The axes of the plots correspond to the residues' numbering along the polypeptide chain. Each time a recurrence is scored between a residue pair, the corresponding location is darkened. The darkened main diagonal corresponds to the trivial recurrence between each residue with itself. The plot derives its symmetric character from the symmetry of the distance operator. The different recurrence patterns between thermophilic and mesophilic sequences is common to all the analyzed structures (see text for details).

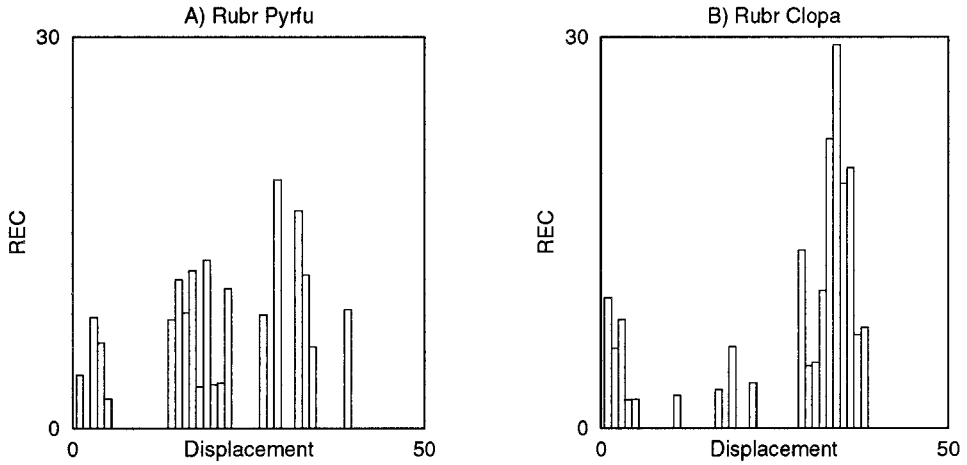


FIGURE 3 Recurrence displacement histograms of *Rubr Pyrfu* and *Rubr Clopa*. The average recurrence of the sequence is displayed as a function of the displacement from the main diagonal of the recurrence plot. Given that the main diagonal of the recurrence plot corresponds to the identity in time (i.e., sequence position), the displacement histograms can be equated to a sort of local autocorrelation integral function. The kurtosis of recurrence displacement was used to quantitatively discriminate mesophilic and thermophilic structures.

of the displacement from the identity line (main diagonal). These histograms highlight the presence of regularities (quasi-periodic structures) in the distribution of recurrences along the series.

SVD-based sequence descriptors. SVD is a well known technique of time series analysis (Broomhead and King, 1986) and we refer to other papers for its implementation in the case of hydrophobicity plots (Mandell et al., 1997, 1998; Selz et al., 1998). Here it is sufficient to say that the technique involves the computation of the first three eigenfunctions (principal components) of the EM and the subsequent projection of the original series on the component space. The projection is then submitted to a complex pole maximum entropy spectral analysis.

In order to derive quantitative descriptors of these spectra, oblique principal components analysis (OPC, VARCLUS procedure in SAS) (Anderberg, 1973; Harman, 1976) was applied to the digitized SVD spectra of the 19 sequences under study (see the Appendix). OPC subdivided the 19 spectra into four clusters. Consequently, for each sequence, four coefficients (SP1-SP4) were computed, which correspond to the Pearson correlation coefficients of the respective spectrum with the four clusters that provide a global description of spectral shape. SP1-SP4 were used to parameterize the spectral information.

Other sequence descriptors. The SD of the residue hydrophobicities, the absolute value of the Pearson correlation coefficient (R) of the hydrophobicity of adjacent residues, and the LZ parameter (Kaspar and Schuster, 1987) of the sequences constituted the last three descriptors of the sequence space (Table 1). SD measures the relative heterogeneity of the amino acid hydrophobicity, and is a purely statistical (shuffling-resistant, order independent) index; in contrast, R is order-dependent and estimates the short-range regularities in the sequence. The LZ index is an easily calculable estimate of the algorithmic complexity of a symbolic sequence (Kaspar and Schuster, 1987), and is strictly dependent upon the ordering of amino acids in the primary structure (see the Appendix).

TABLE 1 Dynamical variables used in this work

Name	Description
MDIST	Average distance between epoch P pairs
REC	Percentage of recurrence
DET	Percentage of determinism
ENT	Shannon entropy of deterministic line distribution
MAXL	Maximal length of deterministic lines
LZ	Lempel-Ziv complexity
R	Pearson correlation coefficient between adjacent residues pairs (abs. val.)
SD	Standard deviation
FD	Dominant frequency of SVD spectrum
SP1	Correlation coefficient between digitized spectrum and cluster 1
SP2	Correlation coefficient between digitized spectrum and cluster 2
SP3	Correlation coefficient between digitized spectrum and cluster 3
SP4	Correlation coefficient between digitized spectrum and cluster 4

Sequence homology methods

The final output of any quantitative sequence comparison method is a distance matrix whose elements contain the estimated divergence between the sequences in the corresponding row and column (Pearson and Lipman, 1988). A popular graphical representation of such matrices, especially useful to visualize evolutionary relationships, is in the form of dendrograms (Sneath and Sokal, 1973). The distance matrices relative to symbolically coded primary structures used in this work have been generated through the *Clustal X* (Thompson et al., 1997) program. The standard alignment algorithm in *Clustal X* can be enriched by two optional refinements: the NoGap (NG) and the multiple substitution (MS) options. The former does not allow any gap in the global alignment: this guarantees that "like" is always compared to "like," although it may exclude much of the information in the global alignment if many gaps are present. The latter (MS) is particularly useful for largely divergent sequences. In such a case, in fact, it becomes increasingly likely (as a result of equally time-spaced random events) that more than one substitution will have happened at the same site (Saitou and Nei, 1987; Higgins and Sharp, 1989). In any case, the strong correlation we observed among all the three sequence homology metrics in our data set (HOMOL1-HOMOL3 in Table 6), points to a substantial stability and algorithm-independence of the sequence alignments for rubredoxins.

RESULTS

The RQA and SVD methods generated 13 variables (Table 1) describing the protein hydrophobicity sequences from a dynamical perspective. In order to reduce the number of variables to a more manageable size, and to identify the effective (orthogonal) axes spanning the space under study, the data set was analyzed by PCA. In order to check for consistency and robustness of the dynamical descriptors, two separate PCAs were performed using 1) all 13 variables and 2) a "reduced" set lacking the SP1-SP4 variables.

TABLE 2 Dynamical descriptors of rubredoxins used in this work

Name	MDIST	REC	DET	ENT	MAXL	LZ	R	SD	FD	SP1	SP2	SP3	SP4
Rubl Azovi	14.11	1.54	8.33	0	3	1.285	0.137	5.27	0.41	−0.414	−0.272	0.877	0.103
Rubl Braja	12.89	2.75	30.51	0	3	1.328	0.119	4.82	0.40	−0.277	−0.254	0.908	−0.209
Rubl Rhilv	13.79	1.27	21.43	0	3	1.226	0.009	5.11	0.39	−0.207	−0.200	0.794	−0.245
Rubr Acica	13.06	4.23	37.04	0.92	4	1.172	0.085	4.90	0.42	−0.148	−0.146	0.197	0.291
Rubr Azoch	13.70	2.00	19.15	0	3	1.371	0.004	5.10	0.13	−0.340	−0.109	0.135	−0.090
Rubr Butme	12.22	5.31	46.15	1.41	5	1.189	0.083	4.56	0.17	0.172	0.428	−0.138	−0.158
Rubr Chllt	13.07	3.92	37.50	1.37	5	1.189	0.073	4.83	0.07	0.920	0.170	−0.249	0.210
Rubr Clopa	12.90	3.37	53.49	0.86	4	1.172	0.100	4.86	0.29	0.269	0.887	−0.291	−0.213
Rubr Clope	13.24	2.51	12.50	0	4	1.278	0.010	4.96	0.30	0.040	0.871	−0.173	−0.283
Rubr Clost	12.22	4.33	45.28	1.15	5	1.297	0.146	4.58	0.06	0.837	0.354	−0.322	0.155
Rubr Clots	12.77	2.38	10.71	0	3	1.425	0.038	4.79	0.20	0.080	0.185	0.154	0.03
Rubr Desde	12.07	2.67	26.09	0	3	1.220	0.206	4.43	0.27	0.789	0.275	−0.234	0.087
Rubr Desgi	13.02	5.19	40.98	1.38	5	1.315	0.050	4.86	0.28	0.157	0.586	−0.289	0.145
Rubr Desvm	12.33	5.02	44.07	1.15	6	1.315	0.085	4.60	0.44	0.269	−0.181	−0.183	0.985
Rubr Desvh	12.37	5.44	46.88	1.66	7	1.315	0.101	4.63	0.44	0.162	−0.178	−0.150	0.989
Rubr Helmo	12.09	6.04	56.34	1.04	5	1.315	0.100	4.64	0.07	0.856	0.031	−0.212	0.189
Rubr Megel	13.10	5.02	27.12	0.72	4	1.315	0.017	4.97	0.07	0.672	−0.223	−0.209	0.129
Rubr Pepas	12.70	4.82	33.90	0.92	4	1.297	0.085	4.78	0.08	0.833	0.356	−0.487	0.130
Rubr Pyrfu	13.33	3.84	51.06	0.99	4	1.188	0.006	4.97	0.30	0.178	0.890	−0.236	−0.173

The first column contains the Swiss Prot code of the rubredoxins whose dynamical descriptors are in columns 2–14. The full names of the descriptors and their meaning are provided in Table 1 and in the text, respectively. The sources' full names are reported in the same order: Azotobacter vinelandii, Bradyrhizobium Japonicum, Rhizobium Leguminosarum, Acinetobacter calcoaceticus, Azotobacter chroococcum Mcd 1, Butyribacterium methylotrophicum, Chlorobium limicola F.Sp. thiosulfatophilum, Clostridium pasteurianum, Clostridium perfringens, Clostridium sticklandii, Clostridium thermosaccharolyticum, Desulfovibrio desulfuricans, Desulfovibrio gigas, Desulfovibrio vulgaris, Desulfovibrio vulgaris (strain Miyazaki), Heliobacillus mobilis, Megasphaera elsdenii, Peptostreptococcus asaccharolyticus, Pyrococcus furiosus.

The first four principal components extracted from the complete set of variables were used to define a suitable space in which the structure-function relationships of interest could be identified. The correlation between classical sequence comparison and the dynamical methods was performed by a simple Pearson coefficient between the distance matrices corresponding to 1) each of the three investigated sequence homology metrics, and 2) the 4D space of the major principal components extracted from the dynamical descriptors. The data matrix for the global dynamical characterization of the 19 sequences is reported in Table 2.

The set of variables shown in Table 2 without the SP1-SP4 variables (thus mainly based on RQA), and submitted to PCA, produced a four-component solution explaining 92% of the total variability: such a compression is due to the high correlation among descriptors. The factor loading matrix, containing the correlation coefficients between original variables and the four components, is reported in Table 3. Sketching an interpretation of the components is made possible by considering which original variables obtain the larger loadings on each component (see the legend to the table).

The SP1-SP4 variables, excluded from the above analysis, deal with the classification of the sequences into four well separated families of spectra (Fig. 4) and, from a computational point of view, are based on a completely different approach (see the Appendix). Moreover, since FD provides a poor representation of the spectral shapes, including SP1-SP4 in the PCA analysis implies the use of brand-new information. This is true even from a purely statistical point of view, given that SP1-SP4 are "almost" mutually orthogonal (see Methods). The "spectral shape" information has approximately the same dimensionality (four) as the PC1r, . . . , PC4r space. Thus its addition to the reduced set of variables could, in principle, strongly perturb its PCA solution. If, however, the PCA solution of the complete set of variables remains substantially the same,

TABLE 3 Factor loadings of the reduced set of dynamical variables

	PC1r	PC2r	PC3r	PC4r
MDIST	−0.841	0.479	0.105	−0.008
REC	0.901	0.224	−0.214	0.051
DET	0.852	0.264	0.184	−0.137
ENT	0.871	0.428	0.018	0.022
MAXL	0.832	0.298	−0.027	0.364
LZ	−0.192	−0.278	−0.752	0.529
R	0.368	−0.727	0.409	−0.045
SD	−0.805	0.522	0.019	−0.006
FD	−0.215	0.041	0.690	0.671

PC1r, which explains 51% of the total variability, is linked to the amount of recurrence and determinism, and might be considered a general quantifier of the rule-obeying character of the hydrophobicity profiles. PC2r explains 17% of the total variability and is linked to the local heterogeneity (negative loading with R) of the hydrophobicity patterns. PC3r and PC4r explain 15% and 10%, respectively, of the total variability and, since they include the LZ- and FD-linked information, can be defined as "shape" parameters describing the kind of periodicity displayed by the sequences.

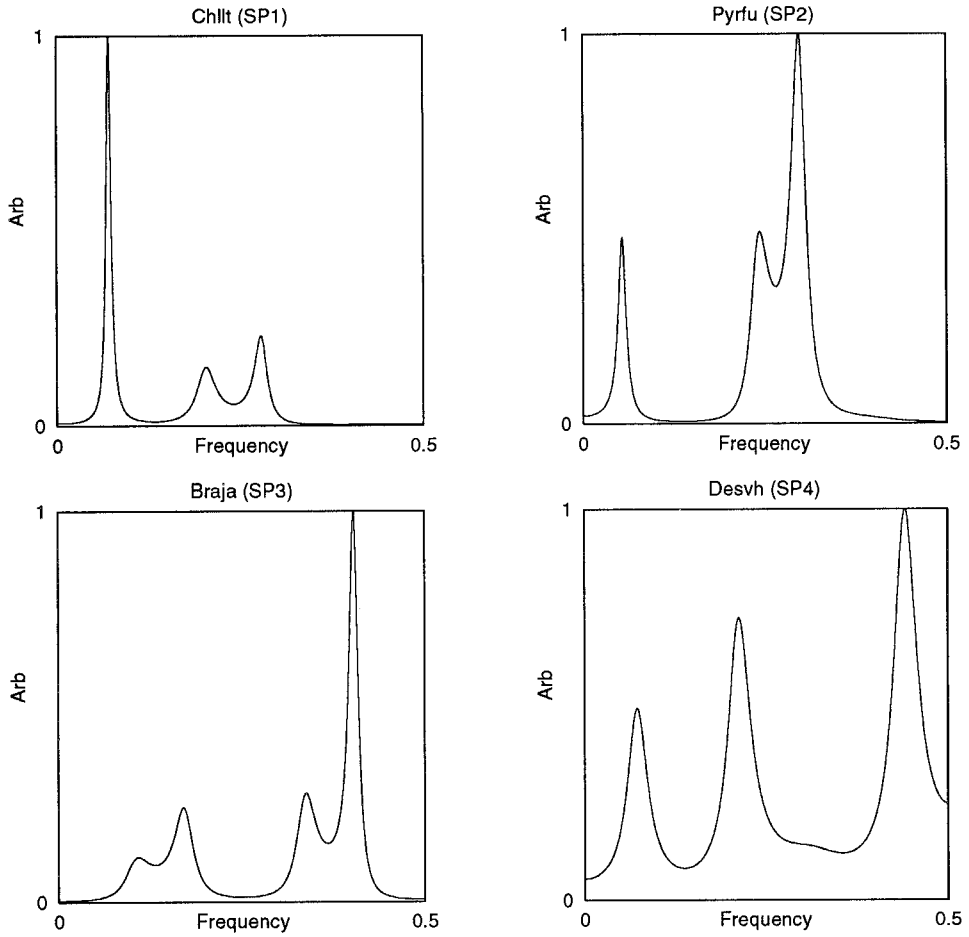


FIGURE 4 Representative elements of rubredoxins' spectral classes generated by OPC analysis. The figure reports in arbitrary units (Arb) the SVD spectra of the most representative (highest correlation coefficient) elements of the spectral clusters generated by OPC analysis over rubredoxins' primary structures. The correlation coefficients (SP1-SP4) between each spectrum and the four clusters were used as synthetic quantitative descriptors of SVD analysis of sequences together with the dominant frequency (FD).

this suggests an internal consistency of the dynamical descriptors.

Table 4 shows that PCA of the reduced set of variables produces a practically equivalent solution as compared to the complete set. The latter solution explains 86% of total variability and shows a one-to-one correspondence (Pearson r) with the components of the reduced set (Table 5). Hence, the new set of components (PC#f) can be faithfully used as dynamical descriptors of the protein data set (for the meaning of the components see the legend to Table 4).

TABLE 4 Factor loadings of the full set of dynamical variables

	PC1f	PC2f	PC3f	PC4f
MDIST	−0.836	−0.093	0.160	0.410
REC	0.893	0.091	0.172	0.214
DET	0.804	−0.046	0.344	−0.204
ENT	0.852	0.080	0.437	0.081
MAXL	0.811	0.338	0.348	0.159
LZ	−0.162	0.320	−0.446	0.697
R	0.329	0.301	−0.529	−0.618
SD	−0.800	−0.105	0.420	0.245
FD	−0.325	0.626	0.386	−0.453
SP1	0.748	−0.347	−0.371	0.089
SP2	0.240	−0.766	0.225	−0.217
SP3	−0.795	0.432	−0.018	−0.194
SP4	0.538	0.737	0.056	0.198

Based on the factor loading value, the meaning of the components can be expressed as follows: PC1f = deterministic structuring of hydrophobicity profiles (RQA parameters); PC2f = shape component, i.e., relative SP2/SP4 character of the hydrophobicity spectrum; PC3f = local heterogeneity (R); and PC4f = complexity of the sequence (LZ).

TABLE 5 Correlation matrix between the reduced (PC#r) and the full set (PC#f) of dynamical components

	PC1r	PC2r	PC3r	PC4r
PC1f	0.981 (*)	−0.005	−0.092	0.066
	(0.0001)	(0.9839)	(0.7066)	(0.7898)
PC2f	0.103	−0.158	0.205	0.803 (*)
	(0.6746)	(0.5175)	(0.3998)	(0.0001)
PC3f	0.053	0.906 (*)	0.354	0.157
	(0.8286)	(0.0001)	(0.1370)	(0.5213)
PC4f	−0.082	0.345	−0.887 (*)	0.213
	(0.7381)	(0.1482)	(0.0001)	(0.3810)

The two dynamical spaces are strongly intercorrelated with each PC#r having its homologous PC#f marked by an asterisk. The change in order of components corresponds to the different percentage of variability explained by the same underlying variables (components) in the two spaces. The table reports the Pearson r together with the corresponding p value under the null hypothesis $r = 0$ (t-test).

The relationship between the dynamical description of proteins and their sequence homology description has been checked by correlating two sets of "between-sequences" distance matrices. One set includes three distance matrices (HOMOL1, HOMOL2, HOMOL3) generated by three different sequence homology metrics; the other set includes just one member, i.e., the Euclidean distance matrix between the same 19 sequences in the 4D PC1f/PC4f space (DYNAM). The four distance matrices are directly comparable by means of a Pearson coefficient (Table 6). The three sequence-matching algorithms are completely equivalent (Pearson $r \approx 1$), while the dynamical description is proven to be practically independent from them (Pearson $r \approx 0.22$).

The internal consistency of the dynamical description of the hydrophobicity profiles provides a solid basis for the attempt to recognize known structural and/or functional features inside the "dynamical" space. The most relevant feature to look for is, in the present case, the high thermal stability of *Rubr Pyrfu* rubredoxin. Looking at the rubredoxins' location in the PC2f-PC3f plane, shown in Fig. 5, *Rubr Pyrfu* rubredoxin is located at an extreme of both the second (PC2f) and the third (PC3f) axes of the dynamical space, whereas such a peripheral location does not find a counterpart in the sequence alignment metrics (Fig. 1 *B*) where *Rubr Pyrfu* rubredoxin is located well inside the mesophilic sequence space.

Since *Rubr Clopa* rubredoxin is the nearest neighbor of *Rubr Pyrfu* rubredoxin in the PC2f-PC3f plane, the 3D structural similarities between *Rubr Pyrfu* and *Rubr Clopa* rubredoxins is recognized by the dynamical description (Fig. 5), whereas the sequence alignment metrics fail in this respect. The modes of the hydrophobicity spectrum were interpreted (Selz et al., 1998) as secondary and supersecondary structural features of the protein molecule; from a purely computational point of view, the modes correspond to peaks in the autocorrelation structure of the series. In the RQA perspective, the presence of peaks of autocorrelation at well-defined scales can be appreciated by plotting the amount of recurrence (i.e., the number of recurrent pairs) on the relative displacement along the chain of all the residue pairs. The recurrences' displacement histograms of *Rubr Pyrfu* and *Rubr Clopa* reported in Fig. 3 provide a much closer look at the recurrence structure of the two sequences as compared to REC or DET. We relied on these histograms to tackle the "high resolution" part of our analysis; i.e., the discrimination between mesophilic and thermophilic structures in the space spanned by *Rubr Clopa*, *Rubr Pyrfu* and the six chimeras (Eidsness et al., 1997). This problem is on a much more detailed scale than the previous one, since all the sequences are located in a small fraction of the component space (Fig. 6). Thus we needed a finer look at the recurrence plots in order to solve it.

A simple inspection of Fig. 2 reveals a macroscopic difference between "thermophilic" and "mesophilic" recurrence plots: while thermophilic hydrophobicity series display a regular distribution of recurrences along lines parallel to the main diagonal, reminiscent of a quasi-periodic distribution of similar hydrophobicity patterns; mesophilic series display a dense clustered distribution of recurrences. Such a difference is present in all the examined structures and, in order to quantify it, we computed the kurtosis of the recurrence displacement distribution for all the sequences. The results, reported in Fig. 7, show that the thermophilic proteins have a low kurtosis corresponding to a quasi-normal distribution of recurrences (a normal distribution has a kurtosis equal to 3) as compared to the relatively higher kurtosis of mesophilic structure corresponding to a peaked (or clustered) distribution of recurrences. Thus the kurtosis of recurrence displacement distributions is a simple index allowing for a clearcut separation of the two stability behaviors on a quantitative, completely data-driven, basis.

From a structural point of view, this means that the thermophilic sequences have a "multiple scale" correlation structure made up of both short and long range correlations as compared to the single-scale correlation (just one peak of recurrences) in mesophilic sequences. The 3D structural counterpart of such a pattern are shown in Fig. 8 and are further discussed below.

TABLE 6 Correlation between sequence homology and dynamical variables distance matrices of rubredoxins

	HOMOL1	HOMOL2	HOMOL3	DYNAM
HOMOL1	1.000	0.978	0.965	0.2206
	(0.0000)	(0.0001)	(0.0001)	(0.0038)
HOMOL2	0.978	1.000	0.988	0.232
	(0.0001)	(0.0000)	(0.0001)	(0.0023)
HOMOL3	0.966	0.988	1.000	0.215
	(0.0001)	(0.0001)	(0.0000	(0.0048)
DYNAM	0.220	0.232	0.215	1.000
	(0.0038)	(0.0023)	(0.0048)	(0.0000)

Correlation coefficients between the distance matrices corresponding to classical alignment metrics computed with three different algorithms (HOMOL#) and to the Euclidean distances in the dynamical component space (DYNAM), together with the corresponding *p*-value under the null hypothesis of $r = 0$ (*t*-test).

DISCUSSION

The main result of the present study is that consideration of the physicochemical semantics (hydrophobicity profiles) of protein primary structures, together with the computation of order-dependent dynamical descriptors, generates completely novel information with respect to classical homology analysis, and indicates new exploration pathways for studying sequence/function relationships of proteins. This has been demonstrated by showing that 1) at difference with classical best-alignment methods, a classification of 19 bacterial rubredoxin primary structures is able to unequivocally single out the only thermophilic element of the set; and that 2) thermal sensitivity can be modeled, on a finer scale of six

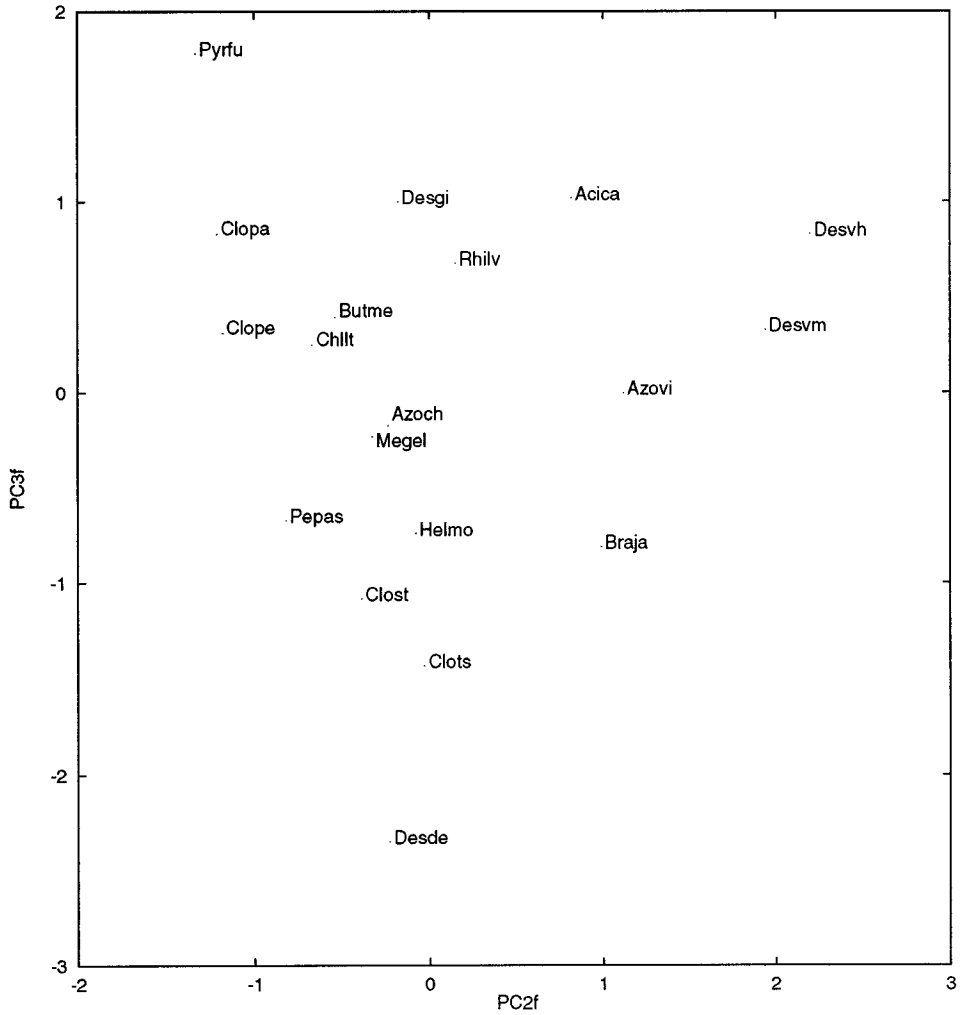


FIGURE 5 Distribution of rubredoxins' sequences in a principal components' space. Notice the "unique" character of *Rubr Pyrfu* and the structural resemblance between *Rubr Clopa* and *Rubr Pyrfu* which are relatively close in the PC2f-PC3f plane.

chimeric sequences derived from a thermophilic and a mesophilic protein, by the analysis of recurrence plots of the corresponding hydrophobicity profiles.

It is worth noting the different assumptions at the basis of sequence alignment and dynamical metrics for the classification of protein sequences. Sequence alignment metrics generate a "phylogenetic space" (Kimura, 1983; Saitou and Nei, 1987) in which proteins are compared in terms of the number and location of mutations needed to go from one structure to another and, in so doing, implicitly measure the underlying evolutionary process. On the contrary, the dynamical approach compares sequences in terms of their resemblance in the global ordering of hydrophobicity values along a chain that 1) emphasizes a physiological, synchronous view; and 2) implies that a similar kind of ordering can be achieved even by a relatively diverse use of amino acid residues in different proteins. These features may explain the statistical independence of the two metrics in the considered data set.

The premise at the basis of the dynamical approach is that the primary structure "encodes" the folding features of proteins and that the "folding code" can be reconstructed by order-dependent descriptors of polypeptide sequences (Mandell et al., 1997; Selz et al., 1998). In this respect, the dynamical approach can be considered a statistical "mean field" approximation, which, through the agency of time series analysis methods, phenomenologically captures an "equilibrium" folding state. This assumes that the individual amino acid residues are part of an interacting system with long-range correlations, which results in scaling and critical behavior. [Some evidence for such a view are demonstrated by Manetti et al. (1999).] In such a frame, hydrophobicity has been considered the variable of election for energy minimization, due to the vast amount of experimental and theoretical evidence (Weiss and Herzel, 1998; Li et al., 1997; von Heijne, 1982; Sweet and Eisenberg, 1983) pointing to it as the most crucial parameter in determining protein 3D structures.

Focusing on global properties of a sequence (dynamical approach), as opposed to local features (sequence homology), changes the point of view on protein sequence/function relationships. The holistic character of the dynamical

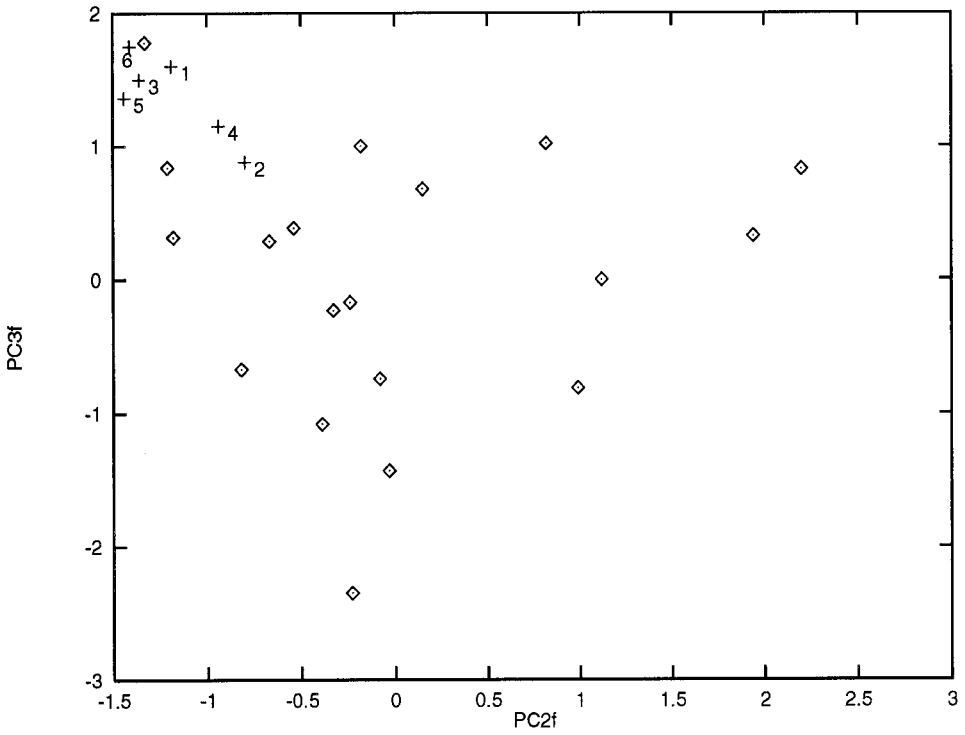


FIGURE 6 Distribution of native and chimeric rubredoxins in a principal components' space. The chimeric sequences were added as external elements (test set) to the PC2f-PC3f plane spanned by the 19 rubredoxins (training set). As expected, the chimeric sequences (+), labeled as in Fig. 1 *C*, occupy a small portion of the space corresponding to the area between *Rubr Pyrfu* and *Rubr Clopa* (their parental sequences).

approach forces the analysis to the congeneric series level; i.e., at the level of proteins having the same kind of activity and comparatively similar structures. In this common frame of sequence/activity relationships, the dynamical approach can be very useful (Zbilut et al., 1998b) to model the modulation of such relationships. The need for studying such homogeneous classes is a direct consequence of the fact that the same values of dynamical descriptors can be reached by completely different sequences pertaining to unrelated proteins, which is another important resemblance between the proposed approach and medicinal chemistry QSAR studies. Even in this field, in fact, the global values of physicochemical descriptors used to model and predict biological activity of organic molecules [e.g., octanol/water partition coefficient, highest occupied molecular orbital (HOMO), etc.] can assume the same value for very different

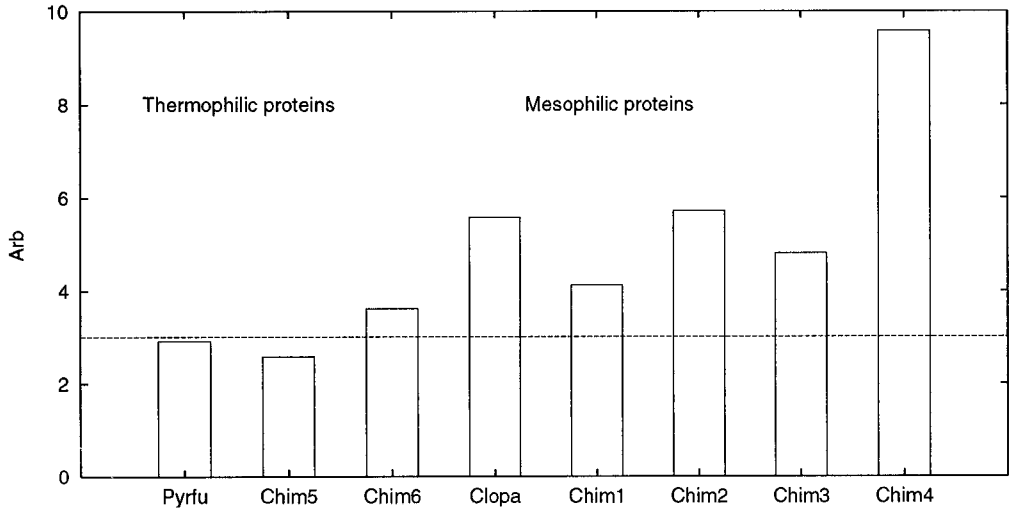


FIGURE 7 Kurtosis of the recurrences' distribution of mesophilic and thermophilic rubredoxins' recurrence plots. Thermophilic proteins display a uniformly lower kurtosis than mesophilic ones; given that kurtosis is an index of the relative "peaked" character of a distribution, this result points to a more concentrated location of recurrences in mesophilic proteins with respect to thermophilic ones. The line marks the expected kurtosis for a Gaussian distribution.

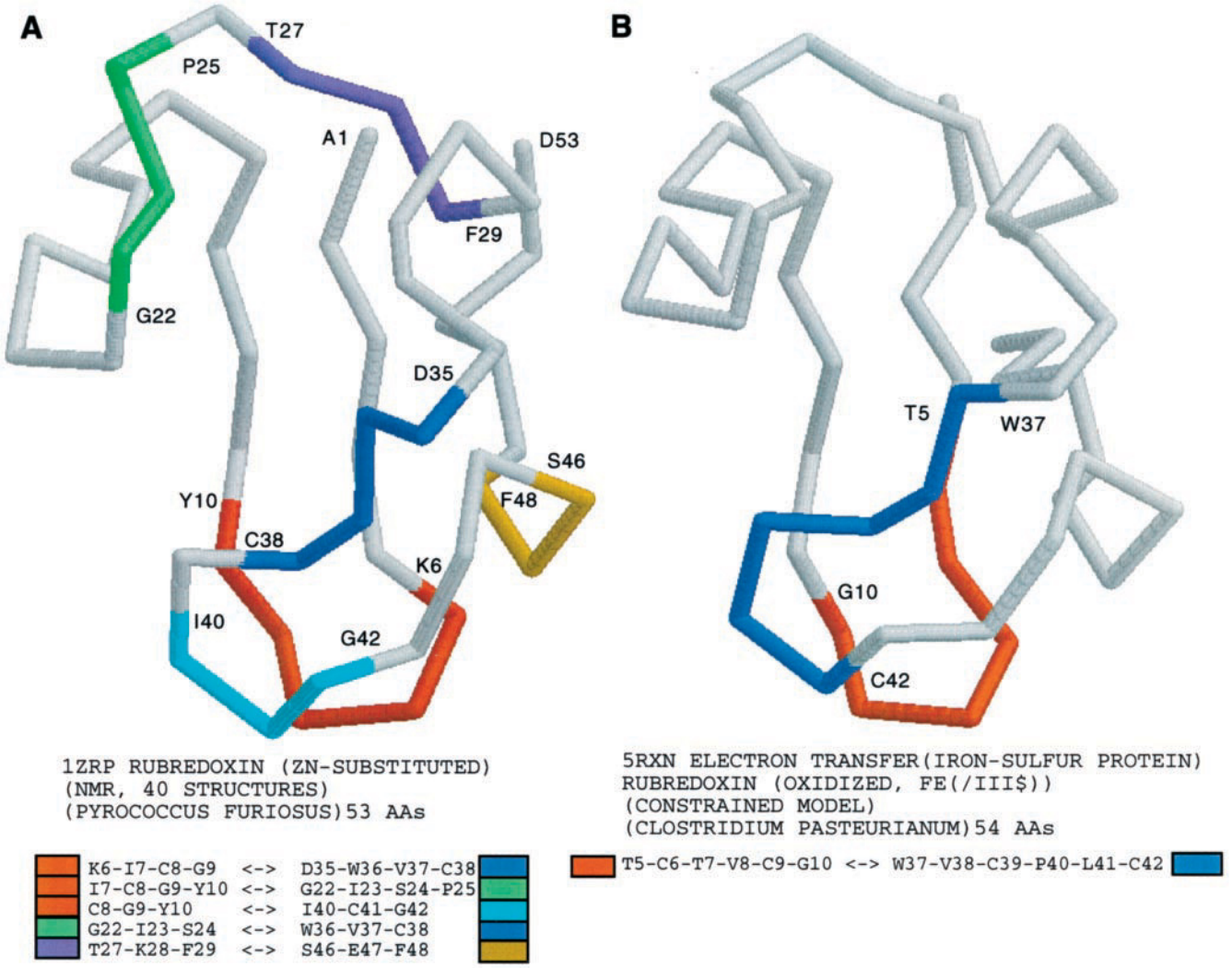


FIGURE 8 Location of recurrent fragments over tertiary structures in *Pyrfu* and *Clopa* rubredoxins. The figure can be considered the structural counterpart of Fig. 2 and reports the 3D structure of both *Rubr Pyrfu* (*A*) and *Rubr Clopa* (*B*). The colors identify in the 3D structure of both *Rubr Pyrfu* and *Rubr Clopa* the long-range recurrences forming the deterministic lines of the recurrence plots. It is apparent how very similar REC, DET values (3.37, 53.49, and 3.84; 51.06 for *Rubr Clopa* and *Rubr Pyrfu*, respectively) go hand-in-hand with a completely different distribution of recurrences in the 3D space: widespread for *Rubr Pyrfu*, concentrated for *Rubr Clopa*.

structures (Franke, 1984) and constrains the analysis within a particular congeneric series of molecules with the same basic "skeleton" and different substituents (Franke, 1984; Martin, 1981; Hansch, 1993). The congeneric series paradigm allows, however, for a nonequivocal definition of the biological activity to be modeled: all the considered organic molecules are potentially active, since all of them possess the basic determinants of that biological activity and only differ from each other in comparatively minor elements (substituents) modulating the activity. This allows for the use of observed differences in the physicochemical descriptors to model such modulation (Franke, 1984; Hansch and Leo, 1995; Martin, 1981).

In the present case we adopted the same paradigm, taking as a congeneric series the rubredoxin family, within which we modeled the thermophilic character. The problem of predicting the thermal stability of proteins could be posed in many general ways, although to our knowledge no general determinant of thermal stability, independent from a specific context, has yet been identified (Adams and Keltzin, 1996). As a matter of fact, the congeneric series paradigm showed itself to be amazingly successful even in the difficult case of the chimeric protein set, which spanned a much more limited portion of the sequence space as compared to the rubredoxin classifications (Fig. 6), and needed a much more detailed dynamical description. This was obtained by shifting from an average description of recurrence plots based upon synthetic indexes, to a more detailed description of the plots' shape (distribution of recurrences) which only required a change in the quantification of the same basic

dynamical description (recurrence plots), without shifting to a brand new parameterization. More importantly, the description maintained its "holistic" character by taking into account the entire sequence. Fig. 8 provides some help in exploiting the bulk of structural information embedded in the recurrence plots of *Rubr Clopa* and *Rubr Pyrfu* (Fig. 2), and emphasizes the different distribution of recurrent fragments, in their 3D representations, over essentially identical backbones. In the *Rubr Clopa* case, the concentration of recurrence lines in the same area of the recurrence plot (Fig. 2) is paralleled by a marked concentration of the long-range recurrences that mainly occur between two well-defined locations on the polypeptide chain (Fig. 8). In the *Rubr Pyrfu* case, there is no preferentially populated area in the recurrence plot, and Fig. 8 shows that recurrent fragments are widespread over the whole backbone. The two situations can be viewed as a different spatial distribution of the same amount of REC (3.84 and 3.37 in *Rubr Pyrfu* and *Rubr Clopa,* respectively) both in 1 and in 3D spaces. In either case, however, it is impossible to identify one (or a few) localized residues specifically responsible for the different thermal stability of the two proteins.

If it is yet difficult to infer in general terms a well-defined set of stabilizing interactions from a recurrence distribution of hydrophobicity along primary structures, it is worth noting that the information conveyed by recurrence patterns well exceeds the simple observation of repetitive chemical motifs. In the specific case of rubredoxins, in fact, the four strictly conserved short fragments centered on cysteins in positions 6, 9, 39, and 41, which bind the prosthetic groups, are all included into the deterministic fraction of recurrences both in *Rubr Clopa* and in *Rubr Pyrfu* (see Fig. 8); only in *Rubr Pyrfu*, however, similar deterministic patterns not immediately perceivable from periodic chemical identity of residues can be found in completely different regions.

In any case, our conclusions are in agreement with the work of Eidsness et al. (1997), who state:

". . . Since our results do not identify a few dominant localized interactions, we suggest that the extraordinary thermostability of *Rubr Pyrfu* may involve a precise optimal alignment of a large number of residues, whose network of interactions are very sensitive to small structural changes dictated by the context of the sequence."

This statement corresponds to the peculiar quasi-periodic pattern in the recurrence plot of *Rubr Pyrfu* (Fig. 2), which can be, in principle, destroyed by a few mutations interrupting the lines of determinism, but cannot be considered as a "local" feature (i.e., specific amino acids responsible for thermostability do not exist), since it spans the entire sequence in the form of long range correlations. On completely different, and exclusively methodological, grounds, it is worth noting that time series analysis methods, given their holistic character, pose no constraints for the relative length of the sequences to be compared. Finally, besides the possibility of deriving useful sequence/function relationships for specific situations, the independence between dynamical and sequence homology metrics opens the way to the exploration of larger data bases, looking for functional similarities among proteins by a method complementary to sequence homology analyses.

APPENDIX

Computing dynamical descriptors of hydrophobicity sequences

RQA descriptors

RQA is based on the computation of a distance matrix, DM, between any possible pair-combination of rows (epochs) of an EM. The distance matrix is then colored, darkening the pixels located at specific (*i, j*) coordinates, corresponding to distance values between *i*th and *j*th rows (epochs) lower than a predefined radius (for details see Giuliani and Manetti, 1996; Webber and Zbilut, 1994)].

The features of the distance function make the plot symmetric ($DM_{i,j} = DM_{j,i}$) and with a darkened main diagonal corresponding to the identity line ($DM_{i,j} = 0$ when $j = i$) (Fig. 2). The darkened (recurrent) points single out recurrences within the series and the plot can be considered as a global picture of the autocorrelation structure of the system (Webber and Zbilut, 1994; Giuliani et al., 1998). In other words, the recurrence plot visualizes the distance matrix between the epochs (rows) of the embedding matrix and consequently the autocorrelation present in the signal at any possible scale. Since, in fact, distances are computed for all the possible pairs of epochs, the elements close to the main diagonal of the plot correspond to short-range correlations (the diagonal marks the identity in time), while long-range correlations correspond to points distant from the main diagonal. Besides the global impression given by the graphic appearance of the plot (Fig. 2), the indexes developed by Zbilut and Webber (1992; Webber and Zbilut, 1994; Trulla et al., 1996) allow for a quantitative description of the recurrence structure of the plot (see Methods). The computation of such indexes implies the setting of a recurrence threshold (radius) that in our case was set to 3 and a line length (minimum number of subsequent recurrent points to be considered as deterministic) that in our case was set to 3.

SVD-based descriptors

SVD spectral analysis (Mandell et al., 1997; Broomhead and King, 1986) is based on the computation of a correlation matrix (CM) among the columns of an EM. The first three eigenvectors of the CM are then computed by SVD and used to estimate the original series *S* (primary structure) by a linear least-squares fit. It is worth noting that the eigenvectors are extracted in order of explained variance, so that the first vectors are more representative of the correlated, signal-like part of the information contained in *S* (Broomhead and King, 1986). The estimated series, *S′*, maintains all the general features of *S* filtered by noise, and are subsequently submitted to a maximum entropy, complex poles, power spectrum analysis (Selz et al., 1998). The result is parameterized in terms of the dominant frequency (FD) and used as a whole (with a 1000-point sampling). In order to derive quantitative synthetic parameters of the proteins' whole spectra, the 19 1000-point-long arrays corresponding to the digitized spectra were analyzed by means of oblique principal component analysis (OPC) (Anderberg, 1973; Harman, 1976). OPC analysis attempts to divide a set of variables (in our case digitized spectra) into nonoverlapping clusters such that each cluster can be considered as essentially unidimensional. The clusters are formed with the goal to include in each of them variables that are both highly intercorrelated (maximal variance explained by the cluster first component), and as much as possible independent from those included in the other clusters (minimal correlation between clusters). The procedure consists in a stepwise increase of the clusters' number until

a user-specified criterion, involving either the percentage of variation accounted for by the global solution or the between-clusters relative independence, is fulfilled. In our case OPC generated a four-cluster partition of the spectra. The correlation coefficients between the original spectra and the center of mass of the clusters (SP1-SP4) were then used to quantitatively parameterize spectral information for the subsequent analysis.

LZ parameter

LZ transforms the representation of a numerical sequence into a binary format, substituting 1 for the higher-than-median values and 0 otherwise. This binary sequence is then analyzed trying to generate any subsequent configuration of 1's and 0's from the previous one using just two operators: copy and insert acting on the initial sequence. Starting from an initial random sequence, *Sr*, the procedure progressively reconstructs any predefined series: the number of instructions (copy plus insert operations) needed to produce the series, normalized by the number of instructions needed to generate the corresponding random sequence, constitutes the LZ index (Kaspar and Schuster, 1987).

Software

RQA was performed by using the original Webber and Zbilut programs that can be freely downloaded from http://homepages.luc.edu/~cwebber.

SVD analysis was performed by CDA (Chaos Data Analyzer) software by J.C. Sprott of the University of Wisconsin and George Rowlands of the University of Warwick. The same program was used for the computation of the LZ index and of the *r* value between adjacent residues.

All statistical routines were performed by SAS (SAS Institute Inc., NY, 1990), Version 6.2 (1998) for Unix systems.

This work has been partly supported by Italian M.U.R.S.T. (40% and 60%) grants to A. Colosimo and C.N.R. (Grant CTB CNR 96.03746.CT114-INV557)

REFERENCES

Adams, M. W. W., and A. Keltzin. 1996. Oxidoreductase-Type Enzymes and Redox Proteins Involved in Fermentative Metabolisms of Hyperthermophilic Archaea. *In* Advances in Protein Chemistry, Volume 48. F. M. Richards, D. S. Eisenberg and P. S. Kim, editors. Academic Press, New York.

Anderberg, M. R. 1973. Cluster Analysis for Applications. Academic Press, New York.

Anfinsen, C. B. 1973. Principles that govern the folding of proteins. *Science.* 181:223–230.

Broomhead, D. S., and G. P. King. 1986. Extracting qualitative dynamics from experimental data. *Physica D.* 20:217–236.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. *In* Atlas of Protein Sequence and Structure, Vol. 5, supplement 3. M. O. Dayhoff, editor. NBRF, Washington. 345.

Eckmann, J. P., S. O. Kamphorst, and D. Ruelle. 1987. Recurrence plots of dynamical systems. *Europhys. Lett.* 4:324–327.

Eidsness, M., K. A. Richie, A. E. Burden, D. M. Kurtz, and R. A. Scott. 1997. Dissecting contributions to the thermostability of *Pyrococcus furiosus* rubredoxin: sheet chimeras. *Biochemistry.* 36:10406–10413.

Faure, P., and H. Korn. 1997. A nonrandom dynamic component in the synaptic noise of a central neuron. *Proc. Natl. Acad. Sci. USA.* 94: 6506–6511.

Franke, R. 1984. Theoretical Drug Design Methods. Elsevier, Amsterdam.

Giuliani, A., and C. Manetti. 1996. Hidden peculiarities in the potential energy time series of a tripeptide highlighted by a recurrence plot analysis: a molecular dynamics simulation. *Phys. Rev. E.* 53: 6336–6340.

Giuliani, A., G. Piccirillo, V. Marigliano, and A. Colosimo. 1998. A nonlinear explanation of aging-induced changes in heartbeat dynamics. *Am. J. Physiol.* 275:H1455–H1461.

Hansch, C. 1993. Quantitative structure-activity relationships and the unnamed science. *Account. Chem. Res.* 26:147–153.

Hansch, C., D. Hoekman, and H. Gao. 1996. Comparative QSAR: toward a deeper understanding of chemicobiological interactions. *Chem. Rev.* 96:1045–1075.

Hansch, C., and A. Leo. 1995. Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. American Chemical Society, Washington DC.

Hansch, C., P. P. Maloney, T. Fujita, and R. Muir. 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature.* 194:178–180.

Harman, H. 1976. Modern Factor Analysis, 3rd Ed. Chicago University Press, Chicago.

Higgins, D. G., and P. M. Sharp. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Comp. Appl. Biosci.* 5:151–153.

Kaspar, F., and K. G. Schuster. 1987. Easily calculable measure for the complexity of spatiotemporal patterns. *Phys. Rev. A.* 36:842–847.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Kumosinski, T. F., and M. N. Liebman. 1994. Molecular Modeling. American Chemical Society, Washington, DC.

Li, H., C. Tang, and N. S. Wingreen. 1997. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Lett.* 79:765–768.

Mandell, A. J., M. J. Owens, K. A. Selz, W. N. Morgan, M. F. Shlesinger, and C. B. Nemeroff. 1998. Mode matches in hydrophobic free energy eigenfuctions predict peptide-protein interactions. *Biopolymers.* 46: 89–101.

Mandell, A. J., K. Selz, and M. F. Shlesinger. 1997. Mode matches and their locations in the hydrophobic free energy sequences of peptide ligands and their receptor eigenfunctions. *Proc. Natl. Acad. Sci. USA.* 94:13576–13581.

Manetti, C., M. A. Ceruso, A. Giuliani, C. L. Webber, Jr., and J. P. Zbilut. 1999. Recurrence quantification analysis as a tool for characterization of molecular dynamics simulations. *Phys. Rev. E.* 59:992–998.

Martin, Y. 1981. A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. *J. Med. Chem.* 24: 229–237.

Micheletti, C., F. Senno, A. Maritan, and J. R. Banavar. 1998. Protein design in a lattice model of hydrophobic and polar amino acids. *Phys. Rev. Lett.* 80:2237–2240.

Murzin, A. G., and L. Patthy. 1999. Sequences and topology. From sequence to structure to function. *Curr. Opin. Struct. Biol.* 9:359–362.

Page, R. D. M. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. *Comp. Appl. Biosci.* 12:357–358.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* 85:2444–2448.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.

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

Sieker, L. C., R. E. Stenkamp, and J. LeGall. 1994. Rubredoxin in crystalline state. *Methods Enzymol.* 243:203–216.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical Taxonomy. Freeman, San Francisco.

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

Sweet, R. M., and D. Eisenberg. 1983. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* 171:479–488.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The Clustal X windows interface: flexible strategies for

multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876–4882.

Trulla, L. L., A. Giuliani, J. P. Zbilut, and C. L. Webber. 1996. Recurrence quantification analysis of the logistic equation with transients. *Phys. Lett. A.* 223:255–260.

von Heijne, G. 1982. Signal sequences are not uniformly hydrophobic. *J. Mol. Biol.* 159:537–541.

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

Weiss, O., and H. Herzel. 1998. Correlations in protein sequences and property codes *J. Theor. Biol.* 190 4:341–353.

Wilbur, W. J., and D. J. Lipman. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA.* 80:726–730.

Zbilut, J. P., A. Giuliani, and C. L. Webber. 1998a. Recurrence quantification analysis and principal components in the detection of short complex signals. *Phys. Lett. A.* 237:131–135.

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

Zbilut, J. P., and C. L. Webber. 1992. Embeddings and delays as derived from quantification of recurrence plots. *Phys. Lett. A.* 171:199–203.