Structure-Related Statistical Singularities along Protein Sequences: A Correlation Study

Mauro Colafranceschi,[†] Alfredo Colosimo,[†] Joseph P. Zbilut,[‡] Vladimir N. Uversky,[§] and Alessandro Giuliani*,[||]

Department of Human Physiology and Pharmacology - University of Rome “La Sapienza”, P.le A. Moro, 5-00185 Rome, Italy, Department of Molecular Biophysics and Physiology, Rush Medical College, Chicago, Illinois 60612, Department of Chemistry and Biochemistry, University of California, Santa Cruz, California 95064, Institute for Biological Instrumentation of the Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia, and Environment and Health Department − Istituto Superiore di Sanità, Viale Regina Elena, 299-00161 Rome, Italy

Received May 18, 2004

A data set composed of 1141 proteins representative of all eukaryotic protein sequences in the Swiss-Prot Protein Knowledge base was coded by seven physicochemical properties of amino acid residues. The resulting numerical profiles were submitted to correlation analysis after the application of a linear (simple mean) and a nonlinear (Recurrence Quantification Analysis, RQA) filter. The main RQA variables, Recurrence and Determinism, were subsequently analyzed by Principal Component Analysis. The RQA descriptors showed that (i) within protein sequences is embedded specific information neither present in the codes nor in the amino acid composition and (ii) the most sensitive code for detecting ordered recurrent (deterministic) patterns of residues in protein sequences is the Miyazawa-Jernigan hydrophobicity scale. The most deterministic proteins in terms of autocorrelation properties of primary structures were found (i) to be involved in protein−protein and protein−DNA interactions and (ii) to display a significantly higher proportion of structural disorder with respect to the average data set. A study of the scaling behavior of the average determinism with the setting parameters of RQA (embedding dimension and radius) allows for the identification of patterns of minimal length (six residues) as possible markers of zones specifically prone to inter- and intramolecular interactions.

1. INTRODUCTION

Protein sequences are, with rare exceptions (e.g. fibrous polymerizing proteins such as collagen or silk), quasi-random strings of symbols with scant evidence of order or periodicity: a reliable estimate of the entropy reduction due to the autocorrelation of residues in an average protein sequence is only about 1%.[1] Nevertheless, such quasi-random strings are the basic recipes producing refined three-dimensional structures, which sustain sophisticated dynamics along with specific physiological roles. Thus, the observed quasi-randomness may be a specious image for underlying meaning. It is interesting to note that a similar situation occurs in the case of human languages where it is almost impossible to generate meaningful texts using just periodic repetitions of symbols.[2] There is, however, a fundamental difference between linguistic rules and the rules governing sequence/structure/activity of proteins: in human languages the linkage between the strings of characters (words) and their semantic meaning is completely arbitrary and needs an external intelligent and active receiver to be decoded. Amino acid sequences, on the other hand, are translated into biologically meaningful messages in the form of proteins by the physicochemical environment (e.g., ionic strength, relative hydrophobicity, temperature, pressure).[3−5]

A focus on the numerical series of physicochemical properties of amino acid residues has provided interesting results in the study of specific protein behavior.[6,7] At the same time, the quasi-random qualification of symbolically coded protein sequences evokes the possibility of solving the sequence/structure/activity puzzle by discovering subtle, albeit crucial regularities in the juxtaposition of symbols. The importance of such regularities in decoding signals can be better appreciated if one recalls the development of cryptography during the Second World War. The decipherment of hidden information in encrypted messages was based upon the notion that any human language, despite its apparent randomness and arbitrariness, is endowed with regularities of various kinds (e.g. the relative abundance of words of given length, the juxtaposition of pairs of symbols, etc.) and that no “masking code” can obscure the code-independent features typical of the original language.[8]

In the present study we adopt the following assumption: the various physicochemical code of amino acid residues are considered as masking codes, and the distinction between “code-dependent” and “code-independent” regularities is used to highlight some statistical features of amino acid patterns. The assumption derives from the fact that a well defined set of physicochemical rules are able to unambiguously transform a given amino acid sequence into a 3D molecular structure. Hence, any physicochemical code can be consid-

* Corresponding author phone: ++39 06 49902579; fax: ++39 06 49902355; e-mail: alessandro.giuliani@iss.it.

[†] University of Rome “La Sapienza.

[‡] Rush Medical College.

[§] University of California and Institute for Biological Instrumentation of the Russian Academy of Sciences.

[||] Istituto Superiore di Sanità.

10.1021/ci049838m CCC: $30.25 © 2005 American Chemical Society
Published on Web 11/24/2004

ered as a “masking code” and the existence of code-independent syntactical features in proteins as a result of the underlying and unknown sequence-3D structure transformation rules. Finally, the most efficient masking code in detecting such features is considered to be the closest approximation to the transformation rules.

In the literature, regularities in the amino acids usage are mostly detected in the form of repeats, i.e., patches of typical length from 4 to 40 residues correlated with structural and activity features, which have been demonstrated in approximately 14% of all proteins.[9−11] In our work, detection of regularities was achieved by means of Recurrence Quantification Analysis (RQA), a nonlinear, model-independent statistical tool which has been shown to be useful for the study of protein sequences as well as in other fields.[12,13]

By applying RQA to a large number of proteins coded by seven different physicochemical properties, we located several syntactic invariants in the amino acid location along primary structures, i.e., “words” of approximately six residues on average. These syntactic rules were shown to correlate with the relatively folded/unfolded status of proteins. In such a context the Miyazawa-Jernigan hydrophobicity scale was also identified as the most useful code, in agreement with results obtained through quite different approaches.[14,15]

2. METHODS AND RESULTS

2.1. Recurrence Quantification Analysis. RQA has been extensively described in the literature.[6,12] Briefly, however, the method is based upon detecting recurrences of similar epochs along a given series, as defined by their Euclidean distances calculated in an artificially created multidimensional mathematical space. The Euclidian distances between all the possible patches of consecutive residues of predefined length are reported on a square, symmetric matrix whose rows and columns correspond to the ordering number of residues along the chain. The graphical representation of such a matrix is called a Recurrence Plot (RP), where a dot is placed if the Euclidian distance between two epochs in the space of the chosen physicochemical property fall below a predefined radius. The fraction of recurrences (dots) in the RP is called REC, and the percentage of consecutive dots forming lines of predefined length and parallel to the main diagonal is called determinism (DET). This last index was recently discussed by Marwan et al.[16] in the realm of dynamical systems theory and found to be directly linked to the presence of “rule obeying” patterns in a time series.

Both REC and DET have been used in this work. We set the length of epochs along protein sequences to 4 and the radius to 20% of the mean Euclidian distance between epochs. In the case of symbolic series, however, the length of epochs was set to 3 and the radius to zero, due to the impossibility of measuring recurrent relations other than their identity. In both numerical and symbolic series the minimum number of consecutive recurrences to be considered to be deterministic was set to 3. The setting of these parameters was driven by previous works aimed at defining the information content of the primary structure of proteins[1,17] and was further controlled by a scaling procedure reported in the text.

2.2. Data Set and Analytical Details. The analyzed “texts” were 1141 proteins randomly chosen from the Swiss-Prot repository in order to avoid any selection bias[18] and constituting a representative sample of all known eukaryotic protein sequences (ftp://ftp.ebi.ac.uk/pub/contrib/swissprot/testsets/signal).

Table 1. Pairwise Correlations between Physicochemical Codings in Amino Acids and Protein Sequences[a]

code pairs	CODES	MEAN	REC	DET
Ch/KD	0.96	0.70	0.70	0.46
Ch/MJ	0.85	0.67	0.64	0.44
Ch/mw	0.29	0.12	0.79	0.41
Ch/Po	0.79	0.63	0.52	0.50
Ch/mr	0.06	0.09	0.76	0.43
Ch/Vo	0.36	0.37	0.55	0.36
KD/MJ	0.87	0.90	0.66	0.45
KD/mw	0.27	0.23	0.63	0.41
KD/Po	0.87	0.92	0.50	0.63
KD/mr	0.07	0.04	0.71	0.37
KD/Vo	0.46	0.53	0.68	0.50
MJ/mw	0.19	0.17	0.65	0.34
MJ/Po	0.90	0.92	0.40	0.44
MJ/mr	0.48	0.40	0.58	0.33
MJ/Vo	0.62	0.69	0.53	0.34
mw/Po	0.09	0.03	0.51	0.43
mw/mr	0.84	0.82	0.86	0.43
mw/Vo	0.55	0.56	0.59	0.42
Po/mr	0.41	0.26	0.57	0.42
Po/Vo	0.61	0.60	0.62	0.51
mr/Vo	0.58	0.53	0.70	0.37

[a] The “CODES” column contains the pairwise correlations, measured in the space of the 20 natural amino acids, of the following numerical scales of physicochemical properties: Ch = Chothia hydrophobicity; KD = Kyte and Doolittle hydrophobicity; MJ = Miyazawa-Jernigan hydrophobicity; mw = molecular weight; Vo = volume; Po = polarity; mr = molar refractivity. The last three columns contain the pairwise correlations, measured in the space of the 1141 proteins in our data set (see the text and Figure 1), of the simple average values of corresponding physicochemical profiles (“MEAN” column), or of the main recurrence variables calculated over the physicochemical profiles (REC and DET columns). The Pearson correlation between the “CODES” and the “MEAN” columns is equal to 0.95; between “CODES” and “REC” is equal to −0.20; between “CODES” and “DET” is equal to 0.46.

We utilized a subset (eukaryotic negative) of eukaryotic proteins that are not secreted and hence not biased by an N-terminal signal peptide.

Each protein sequence was transformed into seven numerical profiles by means of the following physicochemical properties of amino acids: three hydrophobicity scales (Chothia,[19] Kyte and Doolittle,[20] Miyazawa-Jernigan[21]), molecular weight, volume,[22] polarity,[23] and molar refractivity.[24] As an extra coding the standard one-letter symbolic code was also used.

The analysis was based upon correlating different “translations” of the protein sequences as well as the various codings used in these translations. The message (protein sequence) in its symbolic, one-letter form goes through different coding rules producing different translations. We can define a metric to estimate the a priori similarities between different codes: since proteins are made of 20 different units (amino acids), the Pearson correlation coefficients between pairs of codes computed on the space of the 20 natural amino acids (Table 1, column 2) measure the relative similarities between codes. A Pearson coefficient close to 1 (in absolute value) implies the almost complete equivalence of the two codes in terms of the total conveyed information, irrespective of their being positively or negatively correlated.

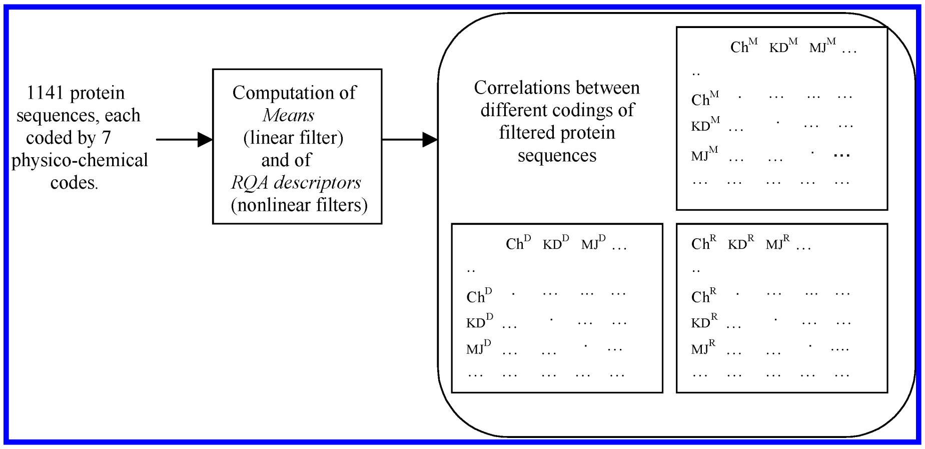
1141 protein sequences, each coded by 7 physico-chemical codes.
Computation of Means (linear filter) and of RQA descriptors (nonlinear filters)
Correlations between different codings of filtered protein sequences
Ch^M KD^M MJ^M …
Ch^D KD^D MJ^D …
Ch^R KD^R MJ^R …

Figure 1. Overview of the data analysis strategy used in this work. The profiles (numerical series) are averaged (linear filter) and submitted to Recurrence Quantification Analysis (nonlinear filter). This gives rise to three sets of vectors, each of 1141 elements, _^{M}, _^{R}, and _^{D}, for the Means and the Recurrence and Determinism (RQA descriptors), respectively. Within each set, a correlation matrix of the seven codings is computed (Table 1), and a Principal Component Analysis is carried out (Table 3).

Upon submitting the available texts (1141 protein sequences) to the seven codings we obtain seven different translations that can be compared in terms of mutual similarities (Figure 1).

More importantly, these similarities can be compared with the mutual similarities among the codes (Table 1): if the similarities between translations are correlated with the similarities between their codes, the analysis would simply return the structure of the translation rules, with no information on the possible existence of syntactic rules in the text. This is reminiscent of the situation in which a transparent object transmits, without distortion, the same light spectrum it was illuminated by. If, however, using an appropriate analytical tool, the correlations between translations are no longer related to the correlations between codes, this implies (i) that the structure of the text has modified the light spectrum and (ii) that we are looking at the "image" of some code-independent, syntactic rule present in the text.

2.3. Correlation Analysis of Codes and of Protein Sequences. Table 1 reports the between-codes correlation computed over the 20 amino acids set (CODES) and over the 1141 proteins set under the agency of both a linear (MEAN) and two nonlinear (REC and DET) operators. It is important to note that the transformation metric (CODES) simply reflects the nature of the physicochemical codes, with the hydrophobicity and polarity scales closely related to each other and independent from the other codes. These purely physicochemical relations are almost exactly maintained if one considers, for each of the 1141 proteins in our data set, the average value relative to each coding (MEAN column) but not in the case of the main RQA variables (REC and DET). In fact, the Pearson correlation between the MEAN and the CODES columns in Table 1 scores 0.95, while, in contradiction, the correlation between the DET and the CODES columns scores only 0.46, and the CODES/REC correlation drops to −0.20. Thus, through the agency of a nonlinear tool (RQA), we are looking at something which is not merely reflecting the physicochemical description of amino acids. In other words, the features of the protein systems do not merely derive from those of the physicochemical codes but may reflect some new, higher-level property. This is confirmed by the fact that very dissimilar coding schemes, like KD and mr, generate transformations not more divergent than those generated by highly correlated codes, like Ch and KD.

Table 2. Descriptive Statistics of REC and DET Variables on Protein Sequences for Different Codings[a]

code	mean	std. dev.	min.	max.
		(a)		
Ch	0.50	0.33	0.15	4.80
KD	0.77	0.51	0.25	10.24
MJ	1.78	1.05	0.54	24.01
mw	0.40	0.56	0.06	11.75
Po	0.66	0.50	0.16	8.86
mr	0.41	0.33	0.06	6.18
Vo	0.63	0.52	0.21	9.88
Sy	0.08	0.39	0	8.91
		(b)		
Ch	16.78	11.33	0	90.14
KD	20.54	10.09	0	90.14
MJ	27.46	9.49	0	84.06
mw	14.26	11.03	0	80.27
Po	19.78	11.56	0	94.89
mr	15.52	11.30	0	87.32
Vo	18.19	10.14	0	89.81
Sy	17.07	21.47	0	100.00

[a] Panels (a) and (b) contain, respectively, statistics concerning the REC and DET variables (for further explanations see the text) and calculated on the whole data set of 1141 proteins used in this work.

The above result might seem to be a consequence of the fact that while searching for recurrences in a protein, the coding is irrelevant because a given patch of residues should have the same recurrence pattern, whatever the code. This, however, is not the case. In fact, Table 2 a points to a lower value of mean recurrence for the symbolic coding (= 0.08) with respect to the numerical codings. The main question is whether protein sequences are made of repeated patches, due to evolutionary events[9] influencing structural and functional features.[25] Such repeats are not perfect since they are altered by point mutations most often introducing residues similar to the original ones. This explains why, by representing proteins through physicochemical profiles and relaxing the strict identity in favor of the weaker similarity requirement, the ability to detect recurrent "words" is enhanced.

2.4 .The Most Useful Coding and RQA Variables. Table 2 reports the average value of recurrence (REC) and determinism (DET) in our protein data set for the different codings. It is evident how the recurrence value markedly

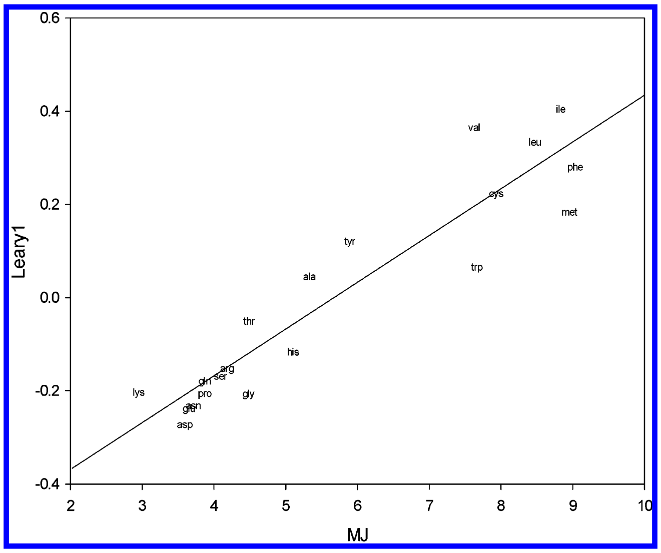

Figure 2. Codings of the natural amino acids by the MJ hydrophobicity scale and the Structural Discriminatory Index by Leary et al. (2004). The figure reports the relation between the MJ hydrophobicity scale and an index having the maximal structural discriminatory power in a set of 3876 protein sequences. This index is termed Leary1 (Leary et al. 2004) and corresponds to the linear combination of a set of 494 different indexes maximizing the discrimination of protein sequences into 174 different structural classes.

varies among codings, from 0.08 (symbolic coding) to 1.78 (MJ scale). This 20-fold difference underestimates the real difference if we consider that the symbolic coding is analyzed with a shorter epochs length as compared to other codings. MJ hydrophobicity scores have an average recurrence (Table 2a) and determinism (Table 2b) much higher than the other physicochemical scales, highlighting a peculiar position of this index in elucidating sequence/structure relations. The MJ scale derives from an investigation of the contact probability between different types of amino acid residues in a large ensemble of 3-D protein structures: it was designed as a sort of statistical potential for amino acid interactions, and only a posteriorii was it recognized as a hydrophobicity scale.[26,27] Since it has been specifically tailored to protein structures, this may explain its performance in detecting amino acid patterning along polypeptide chains. In a recent work by Leary et al.[14] the authors used a supervised learning algorithm to classify 3876 sequences from 174 structural families and defined a class of rules that assigns test sequences to structural classes based on the closest match of an amino acid index profile of a test sequence to a profile centroid for each class. A mathematical optimization procedure was then applied to determine an amino acid index of maximal structural discriminatory power. Figure 2 shows that the MJ scale is strongly correlated with the Leary et al. index ($r = 0.93$), confirming by a completely independent approach its general relevance and its peculiar ability to single out meaningful features of protein sequences.

When submitting the RQA-based representations of proteins of the various codings to a Principal Component Analysis (PCA), we should obtain, if our hypothesis of a basic code-independent structure is true, as the main mode (first Principal Component), a consensus axis collecting all the codings and representing the degree of code-independent autocorrelation structure; namely, the image of the code-independent meaning conveyed by the message. This was actually the case: all the codings were strongly loaded on the first principal component which, both for REC and DET, was the most important source of information explaining, respectively, 70% and 50% of the total variability (Table 3). It is worth noting that, the symbolic coding was highly correlated with the first component, as a further indication of the role of code-independent autocorrelation measure played by PC1.

Table 3. Loadings of the Amino Acid Codes on the First Three Principal Components from RQA Filtered Proteins[a]

(a)

code	PC1REC	PC2REC	PC3REC
Ch	0.86	−0.20	−0.07
KD	0.82	0.04	0.42
MJ	0.77	−0.29	0.37
mw	0.90	−0.23	−0.26
Po	0.69	0.60	−0.20
mr	0.91	−0.03	−0.17
Vo	0.79	0.39	0.21
Sy	0.92	−0.14	−0.25
% expl.variance	69.9	8.9	6.9

(b)

code	PC1DET	PC2DET	PC3DET
Ch	0.71	−0.04	0.38
KD	0.77	−0.30	−0.19
MJ	0.64	−0.42	0.42
mw	0.68	0.44	−0.01
Po	0.79	−0.19	−0.15
mr	0.65	0.49	0.26
Vo	0.70	0.001	−0.50
Sy	0.70	0.10	−0.12
% expl.variance	50	9.4	8.9

[a] Panels (a) and (b) refer, respectively, to REC and DET variables and contain the "loadings" (correlation values) of the original variables with the new one extracted by the PCA algorithm. The Principal Components were obtained from matrices containing as rows the 1141 proteins of our data set, and as columns the REC (panel a) and DET (panel b) descriptors of each protein calculated from the profiles in the various amino acid codings. Sy refers to the symbolic, one-letter code. The last rows in both panels contain the % of total variance explained by each component.

Relative to the aim to separate order dependent from pure compositional effects, we repeated the above analyses on the shuffled texts, i.e., looking for what remains invariant after a random scrambling of amino acid order in each protein sequence. The results showed that REC (Native) and REC (Shuffled) remain largely similar ($r = 0.76$), while in the case of DET no correlation was detected ($r = -0.14$). Analogously, the ranking of the 1141 proteins based on the first recurrence component for both shuffled and native sequences was markedly correlated [PC1REC (Native) vs PC1REC (Shuffled) ($r = 0.87$)], while the determinism rankings based on shuffled and native structures were essentially unrelated ($r = 0.2$). This result suggests that (i) REC in each protein sequence is strongly dependent on the amino acid composition and (ii) DET only depends on the order of amino acids along the chain. Since, in fact, REC is the simple count of how many times four-residue epochs are repeated (even if not perfectly) in whatsoever location along the sequence, in a quasi-random string this is expected to occur with similar frequency, by chance, both before and after scrambling. DET, on the other hand, represents the fraction of consecutive recurrent points, considering the

Table 4. Elements of the "High Determinism Tail" in the Distribution along the First Determinism Component (PC1DET) of the Protein Data Set Used in This Work

Swiss-Prot code	name	PC1DET
P35324	CORNIFIN ALPHA	8.52
Q62267	CORNIFIN B	7.79
Q63532	CORNIFIN ALPHA	6.95
Q62266	CORNIFIN A	6.19
Q07187	EM-like PROTEIN GEA1	6.05
P06144	LATE HISTON H1	5.27
P35326	SMALL PROLINE-RICH PR. 2A	4.97
P17483	HOMEOBOX PROTEIN HOX-B4	4.49
O35762	HOMEOBOX PROTEIN NKX-6.1	4.32
P37108	SIGNAL RECOGN. PART. 14 Kda	4.24
P28318	PROTEIN MRP-126	4.19
P15771	NUCLEOLIN	3.99
009116	CORNIFIN BETA	3.96
P02604	MYOSIN LIGHT CHAIN 1	3.89
P42132	SPERM PROTAMINE P1	3.79
P22793	TRICHOHYALIN	3.71
Q34522	NADH−UBIQ. OXYDORED. CHAIN 3	3.61
P17502	PROTAMINE	3.52
P42129	SPERM PROTAMINE P1	3.42
*P*22238	DESICCATION REL. PROT.	3.37
Q22053	FIBRILLARIN	3.35
P55947	COPPER-METALLOTHIONEIN	3.30
P15870	HISTONE H1-DELTA	3.21
P41139	DNA BINDING PROT. INHIB. ID-4	3.13
Q13329	COMPLEXIN 2	3.08
Q63754	BETA-SYNUCLEIN	3.07
Q01821	GUANINE NUCL. BIND.	3.07
P34618	CEC-1 PROTEIN	3.04
P06146	HISTONE H2B.2, SPERM	3.02
P09442	LATE EMBRYOG. PROT. D-11	3.01
P02292	HISTONE H2B.3, SPERM	2.99
P12950	DEHYDRIN DHN1	2.97
Q05831	SPERM-SPECIFIC PROTEIN PHI-2B	2.79
P12952	DEHYDRIN DHN2	2.77
P47928	DNA BIND. PROTEIN INHIB. ID-4	2.74
P52168	GATA-BINDING FACTOR-A	2.74
P22974	SPERM SPECIFIC PROTEIN PHI-2B	2.66
P12035	KERATIN TYPE II CYTOSKEL. 3	2.62
Q09821	SPERMATID NUCLEAR TRANS.	2.56
Q15672	TWIST RELATED PROTEIN	2.46
P90648	MYOSIN HEAVY CHAIN KINASE B	2.40
P06145	HISTONE H2B.1, SPERM	2.32
P02836	SEGMENT. POLAR. HOMEOBOX	2.31
O42105	COMPLEXIN 2	2.24
P17480	NUCLEOLAR TRANSCR. FACT. 1	2.23
P54844	TRANSCR. FACTOR MAF	2.20
P40262	HISTON H1 E	2.20
P25979	NUCLEOLAR TRANSCR. FACTOR 1	2.17
P21952	OCT. BIND. TRANSCR. FACT. 6	2.07
Q12948	FORK HEAD BOX PROTEIN C1	2.04

relative position and not the number of recurrent patches. Since any peculiar syntactic rule of amino acid patterning should be shuffling dependent, the quantification of contiguous and mutually correlated patches of hydrophobicity (DET) appears as a significant and informative descriptor of monomer distribution in protein chains.

2.5. Proteins Distribution in a Principal Component Space. To find the consequences of amino acid patterning in terms of protein structural or functional features, we inspected proteins endowed with exceedingly high values for the first determinism Principal Component (PC1DET). Protein distribution along this Component is quite asymmetric with a very small but long tail made of extremely deterministic sequences: Table 4 lists the 50 most deterministic sequences in our 1141 protein data set, having component scores greater than 2, with a maximum of 8.5. The remaining 1091 proteins are confined in the interval between −2 and +2. Keeping in mind that Principal Components are constrained by construction to have a mean equal to zero and a unitary standard deviation helps in appreciating this extremely asymmetric distribution. No enzyme or enzyme subunit is present in Table 4, with the only exception of protein Q34522 (NADH −ubiquinone oxydoreductase, chain 3). This is, however, only an apparent exception, since the Q34522 sequence is included in a much bigger functional unit working in the form of a multimeric enzyme. All the extremely deterministic proteins share the property of being involved in protein−protein or DNA−protein interactions, both for regulatory and structural purposes (e.g., histones, protamines, and trascription factors) as well as of forming polymeric assemblies (cornifin, myosin, keratin). Recently Dunker and co-workers demonstrated how the most represented class of natively unfolded structures is composed of polypeptides involved in protein−protein interactions. Moreover, the increasing evidence that low complexity sequences tend to be natively unfolded[28,29] suggested a check for the presence of an excess of natively unfolded zones in the deterministic tail of our data set. The 10 most deterministic sequences scored a percentage of estimated disorder (computed by the PONDR predictor[30]) of 66.46% against the 27.27% of the 10 proteins situated at the low determinism tail (significance of $p < 0.001$). Calculation of a foldability coefficient[28] for the highly deterministic sequences listed in Table 4 indicates that more than 75% may be classified as natively unfolded, and this figure becomes even larger if the reduced form of the −S−S− bridges present in many sequences is considered. From the above analysis the role of "deterministic spots" as crucial sites for interaction seems to gain support. Assuming that protein−protein interactions are driven by essentially the same type of forces leading to mutual recognition between different portions of the same molecule in normal folding, we can hypothesize that highly deterministic sections along the sequence mark the nucleation zones for both folding and protein−protein interactions.

Concerning the functional implications involving "quasi-repeats" of deterministic singularities along protein sequences, an intriguing conjecture may be attempted, based on the analogy between quasi-repeats and "rhymes" within otherwise prosodically unstructured texts. Such rhymes make two different texts (or two different portions of the same text) mutually recognizable and interacting. Their different and possibly function-related appearance in protein sequences is shown in Figure 3, where the recurrence plots of an enzyme molecule and of a transcription factor are reported. In the case of the enzymatic molecule the rhymes appear as faint "columns" of recurrent points along the RP (Figure 3a), while in the case of a "strongly interacting" protein (transcription factor, Figure 3b) the rhymes are clearly defined by block structures in the RPs. Conjectures of this type are perfectly amenable to be (dis)proven by analyzing a protein data set appropriately designed to include a well-balanced blend of different functional classes.

DISCUSSION

In seeking to achieve a general picture of amino acid patterning in protein sequences, (i.e., a grammar of such patterning, based upon strong order-dependent regularities),

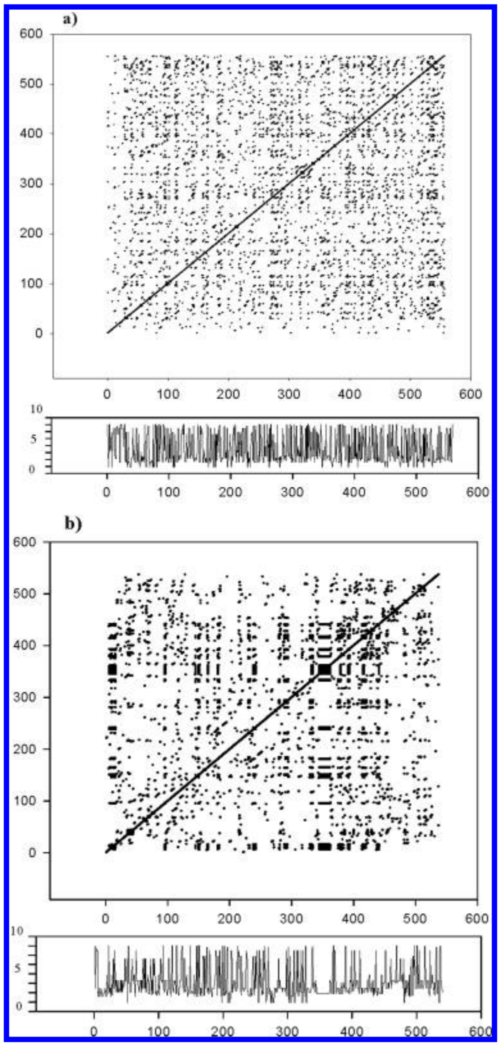

Figure 3. Recurrence plots of two representative proteins. The figure reports in panel (a) the recurrence plot and hydrophobicity series (MJ) of an enzymatic protein (polypeptide *N*-acetylgalactosaminyl transferase, Swiss-prot code Q07537) and in panel (b) the same information for GATA-binding factor A (Swiss-prot code P52168) i.e., a transcription factor. While the enzyme has an average value of general determinism (PC1DET = −0.045, MJ Determinism = 25.46), the transcription factor has an extremely high determinism (PC1DET = 2.74, MJ Determinism = 45.39).

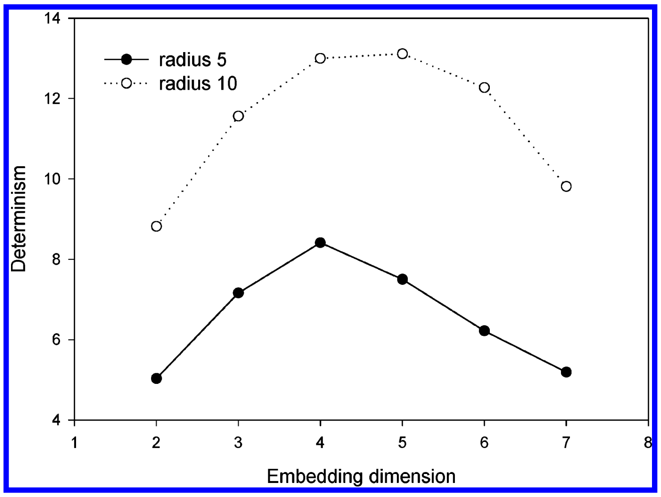

Figure 4. Scaling of Determinism with Embedding Dimension as a function of Radius in 1141 protein sequences. The RQA parameters are calculated on hydrophobicity profiles (MJ coding) averaged over the whole protein data set, at low values of Radius showing a maximum at Embedding Dimension = 4 (see the text for further explanations).

we tried to transcend the simple, naive detection of repeats by the simultaneous use of RQA and of different physicochemical codings of amino acids. The methodological novelty of the approach is in comparing three different hierarchical levels of proteins description: (i) the level of the physicochemical properties of monomers: this level is represented by the correlation between different physicochemical scales computed over the 20 natural amino acids (CODES); (ii) the level of representing proteins by the component residues through different physicochemical descriptions but with no reference to the residues ordering along the chain (MEAN); and (iii) the level of representing the specific distribution of residues along the sequence (REC, DET). While it turned out that the first two levels are fully superimposable, at the third level some peculiar features appeared, which we called syntactic constraints (regularities).

Such syntactic regularities, even if largely independent from the particular physicochemical coding of amino acids, are better evidenced by hydrophobicity, and we derived some hints about the structural consequences of their presence: the analysis of extremely deterministic sequences points to statistically singular, nucleation zones crucial for mutual recognition events. Such a conjecture is reinforced by the fact that the estimated length of 6 for the deterministic patch matches the 6.12 average length we calculated from the data by the Casadio group concerning approximately 800-folding “nucleation centers”.[31] Moreover, a relation between the deterministic peaks and aggregation properties of different proteins ranging from prion[32] to P53[33] and acylphosphatase[7] has been also demonstrated.

Which type of rules may be present in the juxtaposition of amino acids along protein sequences? More importantly, do these rules influence the sequence/structure/activity relations? To answer such questions, a Principal Component Analysis was applied to a data matrix having as rows the proteins and as variables the values of the RQA descriptors for the various codings. The same procedure was carried out after a random shuffling of each protein sequence, so to discriminate the order-dependent properties from the pure compositional features. The distribution of proteins along the most important RQA descriptor (DET) was investigated for its relation to protein structural (and possibly functional) features. The coding with the highest sensitivity in identifying syntactic rules (MJ hydrophobicity) was finally submitted to a scaling procedure to check the existence of a privileged scale at which the effect is maximized.

Figure 4 reports the embedding dimension scaling of average determinism over the 1141 proteins set for MJ coding at very low radius values (5% and 10% meandist): a maximum of determinism at an embedding dimension of 4 can be detected. In other words, using four-letter epochs of the primary structures allows the extraction of maximal information from the amino acid patterning. This is in agreement with the conclusions of other groups[1,17] who identified tetrapeptides as carriers of maximal Shannon entropy values by applying a classical information theory method to a large set of proteins. Since we used a minimal length of 3 consecutive recurrences to score determinism, the maximum of determinism at embedding dimension 4 corresponds to a characteristic length of deterministic patches of 6: thus, 4 and 6 appear as crucial numbers for identifying meaningful words, in the form of “quasi-repeats”, along protein sequences.

An interesting paper by Dokholyan[34] showed that the 20 element alphabet corresponding to the symbolic code is highly redundant in describing protein sequences. This redundancy stems from the physicochemical similarities of amino acid residues that drastically lower the dimensionality of the protein alphabet. This is in line with our findings that physicochemical codes are much more efficient than symbolic code in picking up syntactic regularities in protein sequences. Moreover, we complement the Dokholyan results by showing that correlations in amino acid properties exert an effect not only at the level of single residues (letters, alphabet) but also at the level of short patches of consecutive residues (words).

As for the practical impact of our study, the analysis of RQA descriptors of protein sequences, being not dependent on homology, could allow for (i) detection of unexpected “neighbors” of query structures, thus enlarging the possibility of both function assignment and protein engineering and (ii) possible classification of newly discovered sequences only on the basis of their primary structure. The latter observation is made statistically significant by the presence of only one enzyme subunit in the list of extremely high DET proteins (Table 4) and of very few enzymes at intermediate PC1DET values (0 ÷ 2). In this sense, it is worth mentioning that, while our unbiased set of 1141 proteins scored a DET equal to 27.46 ± 9.49, the analysis of four independent samples from SWISS−PROT, 20 sequences each, corresponding to (i) proteins of known 3D structures, (ii) coiled coils, (iii) collagen proteins, and (iv) silk fibroin, showed the following %DET ranking (MJ code): (i) 24.94 ± 3.23; (ii) 31.69 ± 9.65; (iii) 43.66 ± 8.95; and (iv) 48.24 ± 22.07, which is equivalent to the ranking obtained by the Dunker group in terms of their sequence complexity entropy criterion.[35] These results are consistent with the idea of a link between DET and structural disorder and open the way for a systematic structural and/or even functional characterization of proteins by RQA.

ACKNOWLEDGMENT

The work was supported in part by NSF grant #0240230, principal investigator J. P. Zbilut and by a MURST grant to A. Colosimo. The collaboration of Elisabetta Pizzi for both the choice of the data set and the bioinformatic assistance is gratefully acknowledged.

REFERENCES AND NOTES

(1) Weiss, O.; Jimenez-Montano, M. A.; Herzel, H. Information content of protein sequences. *J. Theor. Biol.* **2000**, *206*, 379−386.

(2) Popov, O.; Segal, D. M.; Trifonov, E. N. Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems* **1996**, *38*, 65−74.

(3) Gross, M. Linguistic analysis of protein folding. *FEBS Lett.* **1996**, *390*, 249−252.

(4) Trifonov, E. N.; Berezovsky, I. N. Proteomic Code. *Mol. Biol.* **2002**, *36*, 239−243.

(5) Taylor, W. R. A ‘periodic table’ for protein architecture. *Nature* **2002**, *416*, 657−660.

(6) Giuliani, A.; Benigni, R.; Zbilut, J. P.; Webber, C. L., Jr.; Sirabella, P.; Colosimo, A. Nonlinear signal analysis methods in the elucidation of protein sequenze structure relationships. *Chem. Rev.* **2002**, *102*, 1471−1491.

(7) Zbilut, J. P.; Colosimo, A.; Conti, F.; Colafranceschi M.; Manetti, C.; Valerio, M. C.; Webber, C. L., Jr.; Giuliani, A. Protein aggregation and folding: the role of deterministic singularities of sequenze hydrophobicity as determined by signal analysis approach. *Biophys. J.* **2003**, *85*, 3544−3557.

(8) Schneier, B. *Applied Cryptography*; John Wiley and Sons: New York, 1996.

(9) Kajava, A. V. Review: Proteins with Repeated Sequence-Structural Prediction and Modeling. *J. Struct. Biol.* **2001**, *134*, 132−144.

(10) Tsonis, P. A.; Tsonis, A. Linguistic Features in Eukaryotic Genomes. *Complexity* **2002**, *7*, 13−15.

(11) Sedgwick, S. G.; Smerdon, S. J. The ankyrin repeat: a diversity of interactions on a common structural framework. *TiBS* **1999**, *24*, 311−316.

(12) Webber, C. L., Jr.; Zbilut J. P. Dynamical assessment of physiological systems and states using recurrence plot strategies. *J. Appl. Physiol.* **1994**, 965−973.

(13) Rustici M.; Caravati, C.; Petretto, E.; Branca, M.; Marchettini, N. Transition scenarios durino the evolution of the Belousov−Zhabotinsky reaction in an unstirred batch reactor. *J. Phys. Chem. A* **1999**, *103*, 6564−6570.

(14) Leary, R. H.; Rosen, J. B.; Jambeck, P. An optimal structure-discriminative amino acid index for protein fold recognition. *Biophys. J.* **2004**, *86*, 411−419.

(15) Moelbert, S.; Emberly, E.; Tang, C. Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. *Protein Sci.* **2004**, *13*, 752−762.

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

(17) Strait, B. J.; Dewey, T. G. Strait, B. J.; Dewey, T. G. The Shannon information entropy of protein sequences. *Biophys. J.* **1996**, *71*, 741−742.

(18) Menne, K. M. L.; Hermjakob, H.; Apweiler, R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **2000**, *16*, 741−742.

(19) Chothia, C. The Nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1976**, *105*, 1−14.

(20) Kyte, J.; Doolitle, R. F. A simple method for displaying the hydiopathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(21) Miyazawa, S.; Jernigan, R. L. Estimation of Effective Interresidue contact energies from protein crystal structure: quasi-chemical approximations. *Macromolecules* **1985**, *18*, 534−552.

(22) Zimmermann, J. M.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170−201.

(23) Grantham, R. Aminoa Acid Difference Formula to Help explain protein evolution. *Science* **1974**, *185*, 862−864.

(24) Jones, D. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. *J. Theor. Biol.* **1975**, *50*, 167−183.

(25) Amodeo, P.; Fraternali, F.; Lesk, A. M.; Pastore, A. Modularity and homology: modelling a titin type I modules and their interfaces. *J. Mol. Biol.* **2001**, *311*, 283−296.

(26) Wang, J.; Wang, W. Grouping of residues based on their contact interaction. *Phys. Rev. E* **2002**, *65*, 41911-1−41911-5.

(27) Tang, H. C. H.; Wingreen, N. S. Nature of driving force for protein folding: a result from analyzing statistical potential. *Phys. Rev. Lett.* **1997**, *79*, 765−768.

(28) Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **2002**, *11*, 739−756.

(29) Romero P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; Dunker, K. Sequence complexity of disordered proteins. *Proteins Struct., Funct., Genet.* **2001**, *42*, 38−48.

(30) Dunker, K.; Brown, C. J.; Lawson, D.; Iakoucheva, L. M.; Obradovic, Z. Intrinsic Disorder and protein function. *Biochemistry* **2002**, *41*, 6573−6582.

(31) Compiani, M.; Fariselli, P.; Martelli, P. L.; Casadio, R. An entropy criterion to detect minimally frustrated intermediates in native proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9290−9294.

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

(33) Porrello, A.; Soddu, S.; Zbilut, J. P.; Crescenzi, M.; Giuliani, A. Discrimination of single amino acid mutations of the P53 protein by means of Recurrence Quantification Analysis. *Proteins: Struct., Funct., Bioinform.* **2004**, 55, 743−755.

(34) Dokholyan, N. V. What is the protein design alphabet? *Proteins: Struct., Funct., Bioinform.* **2004**, *54*, 622−628.

(35) Romero P.; Obradovic, Z.; Dunker, K. Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Lett.* **1999**, *462*, 363−367.

CI049838M