PRIMARY STRUCTURES OF PROTEINS AS SPACE-DEPENDENT SIGNALS

M. COLAFRANCESCHI[(1)], A. GIULIANI[(2)] AND A. COLOSIMO[(1)],

[(1)] Physiology and Pharmacology Dept & C.I.S.B. Research Center-Sapienza Univ. di Roma
[(2)] Environment and Health Dept., Istituto Superiore di Sanita' , Roma

Contents

1. Introduction

Proteins occupy a unique position in the hierarchy of natural systems, since they lie in the twilight zone between chemistry and biology. Proteins are linear heteropolymers that, unlike most synthetic polymers, consist of nonperiodic sequences of 20 different monomers (aminoacids). The majority of proteins fold as self-contained structures determined by the sequence of monomers. Thus, we can consider the particular linear arrangement of amino acids as a sort of "recipe" for making a water-soluble polymer with a well-defined three-dimensional architecture (54; 42; 8). It is important to stress this dynamical perspective. "Well defined three-dimensional structure" should not be intended as "fixed architecture": many proteins appear as partially or even totally disordered when analyzed with spectroscopic methods (1) in spite of the high efficiency in their physiological functions.

Understanding the link (if any) between sequence-embedded information and folding behavior is currently a crucial problem of both theoretical and applied physical biochemistr4y of proteins. More specifically, the problem deals with: i) sequence-based functional predictions, ii) 3D structure-based functional predictions, and iii) folding mechanism elucidation.

An amazing observation made by several authors concerns the existence of only weak departures of real protein sequences from random strings (42; 48) , namely from series whose autocorrelation structure remains substantially invariant after random shuffling the positions of its constituent elements. (61) An obvious consequence of that is the notion that the "code" linking a sequence to a particular structure (2) is not emerging from simple periodicities in the amino acids' occurrence (26).

There is voluminous literature dealing with theoretical models following ab initio approaches to the sequence-structure puzzle (15; 49). Most theoretical models adopt a statistical physics perspective based on proteins considered as lattices, i.e., squared grids in which each residue is considered as interacting with the same number of neighbours. Other groups instead of looking for general laws linking sequence and structure across all protein families, apply a purely local statistical approach (11; 46). In the last 15 years or so, investigating new algorithms to discover even relatively remote homologues of a given leader sequence allowed the development of new sequence alignment techniques, and represents a "leitmotif" in bioinformatics and computational biology (14; 3).

The present minireview instead is devoted to a scarcely populated but potentially quite interesting field of computational biochemistry: the use of signal analysis methods to describe protein sequences as mono-dimensional series. The protein sequences are described by means of a vector of numerical variables that summarize their autocorrelation structures. Thus, the simplest level of protein sequences description shifts from the pairwise alignment of structures to a self-consistent numerical description of the *single* sequence.

The main feature of the methods described in what follows is the production of numerical indexes that parametrize the protein sequences as a whole in terms of amount and profile of periodicities in the hydrophobicity distribution along the chain. This is similar to the quantitative structure activity relationships (QSAR) analyses widely used in medicinal chemistry (27; 28). By analogy with QSAR, the investigated molecules (proteins in this case, organic compounds in the case of QSAR) are described by means of an array of numerical features parametrizing various chemico-physical properties. These properties act as regressors (independent variables) for modeling a given biological activity, which in turn acts as a dependent variable. The biological activities most often modeled by QSAR are pharmacological or toxicological potencies, while the properties modeled so far for proteins are protein/peptide interactions, folding behavior, and thermal stability. From the theoretical side comes the consideration of protein sequence as a unitary system embedded into a global force field based on hydrophobicity (4), since the analysis ends with one number deriving from a computation extended over the whole sequence). From the statistical side comes the local approach and the use of soft data analysis methods with no peculiar distributional constraints (5). The main steps of the method can be summarized as follows: a) use of hydrophobic code for primary structures (52; 35); b) treatment of the hydrophobicity distribution along the sequence like a time series, with the corresponding use of nonlinear signal analysis techniques to underpin position-dependent properties of the hydrophobicity profiles (50; 43); c) adoption of a local approach for both inter-sequence (within homologous series of proteins) comparisons and intra-sequence (among short patches along the same sequence) analyses as a starting point for periodicity detection (47).

2. Analytical strategy

2.1. **The signal analysis perspective.** Protein sequences can be considered as discrete series equivalent to time series, with the aminoacid order playing the role of subsequent time steps. Thus, on a purely formal viewpoint, any technique used for signal and time series analysis could be successfully applied to protein primary structures. From a practical viewpoint, however, the fact that protein sequences are very short and basically non stationary signals drastically

limits the range of signal analysis techniques usable in this context(14). Thus, the ideal method for approaching signal analysis of protein sequences should be nonlinear, independent of any stationary assumptions, and able to deal with very short series. (10) Methods satisfying these constraints are those adopting a purely correlative point of view, with no a priori distributional and/or physical assumption. The only aim of such methods is looking for autocorrelation patterns along the series, i.e., for the recurrences of particular short motifs (like in recurrence quantification analysis, RQA) or for periodicities of no predefined functional form spanning all the studied sequences (like in principal component analysis, PCA). At the basis of these methods is the transformation of the original series into an "embedding matrix" with the method of delays. (19)

An n-dimensional embedding (deconvolution) procedure consists of building an n-column matrix out of the original linear array by shifting the series by a fixed lag. In the example below n = 4, lag = 1.

15	12	27	39
12	27	39	31
27	39	31	65
39	31	65	22
31	65	22	12
65	22	12	42
22	12	42	11
12	42	11	33
42	11	33	
11	33		
33			

The rows of the *embedding matrix (EM)* correspond to subsequent windows of length 4 (embedding dimension) along the sequence. Notice that the last (n-1) values are eliminated from the analysis as an obvious consequence of shifting the series. The choice of the embedding dimension corresponds to the choice of the scale at which the autocorrelation structure of the series is estimated.

2.2. **Recurrence Quantification Analysis (RQA).** The application of RQA is based on the calculation of the Euclidean distance between all the pairs of rows of an embedding matrix (47; 10). If the distance between two generic rows (i.e. windows of predefined length along the sequence) falls below a predefined 'radius', we get a recurrence. The concept of recurrence is straightforward: for any ordered series (time or spatial), a recurrence is a point which repeats itself. Recurrences are strictly local and independent of any mathematical assumption (23; 36). Furthermore, calculation of recurrences requires no transformation of the data and can be used for both linear and nonlinear systems (60; 22). The concept of recurrence can be expressed as follows: given a reference point, Xo, and a ball of radius r, a point X is said to recur (with reference to Xo) if :

$$X : \|X - Xo\| \leqslant r \tag{1}$$

In the case of a time series, i.e., of a system occupying in different times different positions along a trajectory in a suitable state space, the recurrences correspond to the time points where the system passes nearby to already visited states. In the case of protein sequences, time corresponds to the amino acid order and the recurrences are patches, with a length equal to the embedding dimension, sharing their profile with other patches along the chain. The number and relative

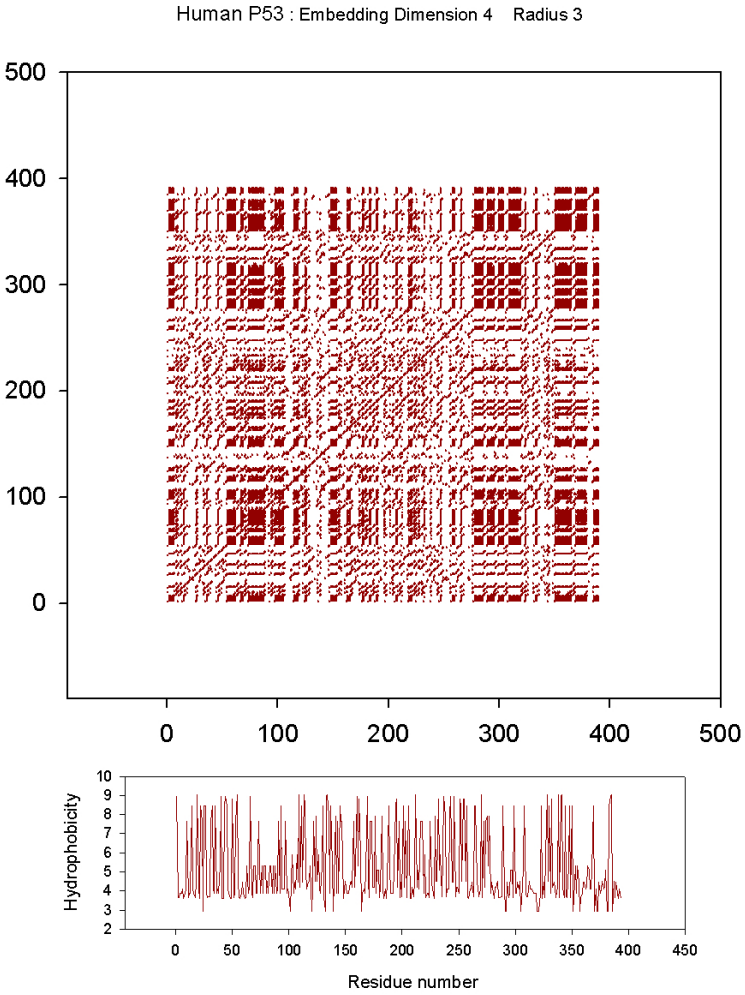

FIGURE 1. *Recurrence plot and hydrophobicity profile of human P53 protein.* The presence of an extremely deterministic ordering of amino acids between residues 61 and 98 is clearly evident in the figure in terms of its consequences on the recurrence plot. This highly deterministic portion is "resembled" by other segments along the sequence. This observation is not clear by the simple inspection of the hydrophobicity plot but is made evident by the recurrence plot: the "resemblances" correspond to linear (or alternatively horizontal given the symmetrical character of recurrence plot) banding of the plot. For the RQA descriptors meaning, see the text. (Modified from (67)).

positions of recurrences are expressed by recurrence plots (RP), that are symmetrical NxN arrays in which a point is placed at (i, j) whenever a point X_i on the trajectory is close to another point X_j.

The closeness between X_i and X_j is expressed by calculating the Euclidian distance between these two normed vectors, i.e., by subtracting one from the other obtaining the expression $\|Xi -$

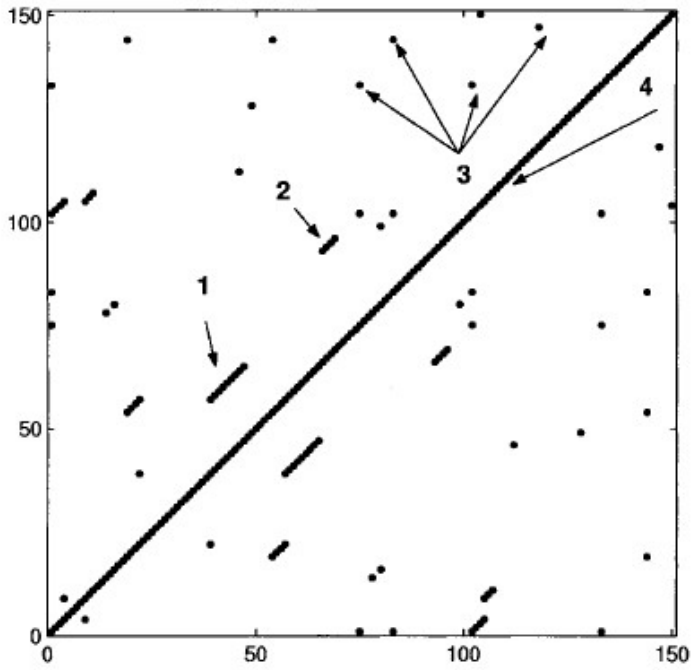

FIGURE 2. *Quantification of the main Recurrence Plots' features.* 1: a line segment composed of 8 recurrent points. 2: a 4-point line segment. 3: several recurrent points that are not part of a line segment. 4: the identity line (i.e., where $X_i = X_j$). Modified from (68)

$Xj\| \leqslant r$, where r is the predefined radius. If the distance falls within this radius, the two vectors are considered to be recurrent, and graphically this can be indicated by a dot (Figure 1).

An important feature of recurrence plots is the existence of short line segments parallel to the main diagonal (Figure 2), which correspond to sequences $(i, j), (i+1, j+1), ..., (i+k, j+k)$ such that the fragment:

$$X_j, X_{j+1}, ..., X_{j+k}$$

is close to:

$$X_i, X_{i+1}, ..., X_{i+k}$$

The absence of such patterns suggests randomness (18). For protein sequences these deterministic lines correspond to contiguous patches of similar hydrophobic/ hydrophilic patterns.

Several strategies to quantify features of such plots have been developed and led to the generation of the following variables (for an exhaustive definition, see (60; 39)):
- Recurrence (%REC) : % of recurrence points in a RP.
- Determinism (%DET) : % of recurrence points which form diagonal lines.
- Laminarity (%LAM) : % of recurrence points which form vertical lines.
- Maximum line (MAXL) : length of the longest diagonal line.
- Trapping time (TT) : average length of vertical lines.
- Entropy (ENT) : Shannon entropy of the distribution of the diagonal line lengths.
- Trend (TREND) : Paling of the RP towards its edges.

2.3. **Principal Component Analysis (PCA).** In contrast to RQA, principal component analysis (PCA) is a well-established method frequently used in physical as well as in social and biological sciences(5).

When applied to a time (or spatial) series that is originally monodimensional, PCA requires that the original series is represented on a multidimensional space by the agency of the embedding procedure. The original data can be projected into a new set of coordinates given by linear combinations of the original variables, and no original information is lost. The new coordinates are orthogonal by construction (i.e., statistically independent), each representing an independent aspect of the data set. The number of principal component is equal to the number of original variables, but principal components have the fundamental property of explaining the system variability in a hierarchical way. This implies that we can save the meaningful (signal-like) part of the information retained by the first principal components and discard the (noisy) last ones. In other words, the most correlated portion of information is retained by the first components, while all the singularities are discarded. Therefore, by the use of a threshold for the cumulative percentage of explained variance, PCA allows for the reduction of a complex system of correlations into a lower-dimensional one.

The application of PCA to a data set having as statistical units different proteins and as variables the RQA descriptors corresponding to each sequence, highlights the presence of regularities in the data set as a whole. This the allows a fully statistical investigation of protein structures, without being confined to few, specific cases. In what follows the application of this approach will be exemplified by discussing two problems of general relevance.

3. EXAMPLE 1: Studying a large protein dataset

In this example we show how the statistical information present in the 1141 protein sequences extracted from the Swiss-Prot data-set by Menne et al. (40) can be exploited, according to an approach outlined in (65). From the whole data set, available at:

ftp://ftp.ebi.ac.uk/pub/contrib/swissprot/testsets/signal

we selected the non-secreted, eukaryotic proteins, namely an ensemble that can be considered as a random pick from the entire protein niverse, without the bias of the initial hydrophobic, short signal typical of secreted proteins. By means of this relativelay large ensemble we tackled two relevant problems, namely: i) identifying an optimal physico-chemical code for protein sequences, and ii) correlating hydrophobicity patterns distribution and functional properties.

3.1. **An optimal physico-chemical code for aminoacid residues.** Table 1 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 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 posteriori was it recognized as a hydrophobicity scale (59; 53). Since it has been specifically tailored to protein structures, this may explain its performance in detecting amino acid patterning along polypeptide chains.

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

TABLE 1. *%REC (a) and %DET (b) statistics on Protein Sequences for Different Codings (column 1).* Recurrenc and Determinism are calculated on the whole data set of 1141 proteins used in this work. Ch: Chothia hydrophobicity (12); KD : Kyte and Doolittle hydrophobicity (32); MJ : Miyazawa-Jernigan hydrophobicity (41); mw : molecular weight; Vo : volume; Po : polarity; mr : molar refractivity.

If our hypothesis of a basic code-independent structure is true, when submitting the RQA-based representations of proteins of the various codings to a Principal Component Analysis, we should obtain as the main mode (first Principal Component) a consensus axis collecting all the codings and representing the degree of code-independent autocorrelation structure. 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 2). 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.

In the aim to separate order dependent from pure compositional effects, we repeated the above analyses on the shuffled sequences, 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

Code	PC1REC	PC2REC	PC3REC
(a)			
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
% of expl.variance	69.9	8.9	6.9
(b)			
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
% of expl.variance	50	9.4	8.9

TABLE 2. *Principal Component Analysis on the recurrence descriptors of protein sequences.* 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 the data set and, as columns, the %REC and %DET descriptors of each protein calculated from the profiles in the various amino acid codings (see Table 1). Sy refers to the symbolic, one-letter code. The last rows in both panels contain the % of total variance explained by each component.

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 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.

3.2. **Hydrophobicity dynamical patterns and structure/function features.** Proteins Distribution in a Principal Component Space. To find the consequences of amino acid patterning alon the primary structure in terms of protein 3D structure 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 3 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 3, 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 DNAprotein 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 (55; 45) 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) (17) 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 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.

In other words, 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". (13) Moreover, a relation between the deterministic peaks and aggregation properties of different proteins ranging from prion (63) to P53 (44) has been also demonstrated.

3.3. **The essential dimensions of the protein alphabet.** An interesting paper by Dokholyan (16) 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. Thus, 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.

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
P22238	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

TABLE 3. *Highly deterministic proteins in the examined data set (see the text).* List (in decreasing order of %Det) of elements in the 'High Determinism Tail' of the Distribution along the First Determinism Component (PC1DET)

.

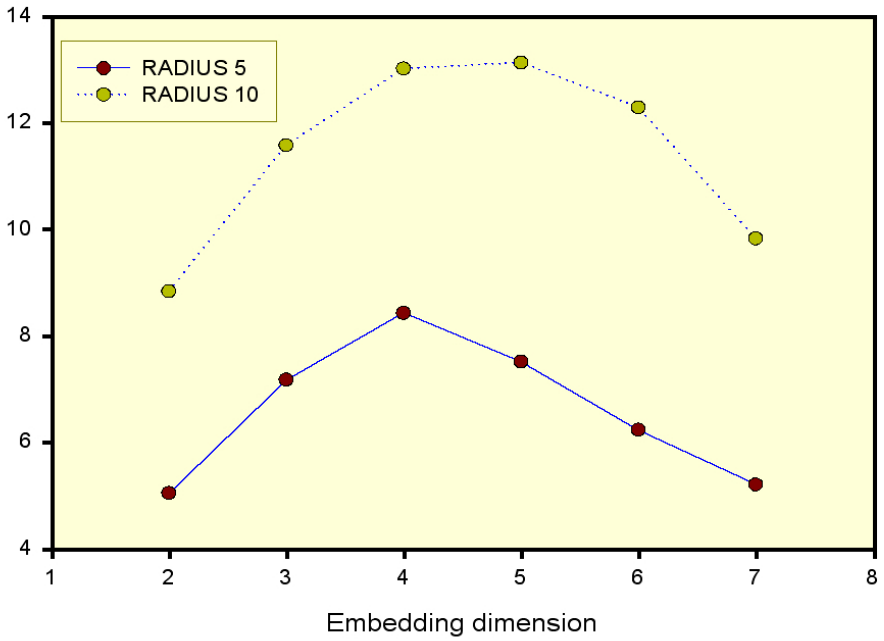

FIGURE 3. *Scaling of Determinism with Embedding Dimension as a function of Radius in 1141 protein sequences.* The RQA parameters are calculated on hydrophobicity profiles (MJ coding) averaged over the whole protein data set, at low values of Radius showing a maximum at Embedding Dimension = 4 (see the text for further explanations).

In order to provide a further check to Dokholyan results, a Principal Component Analysis was applied to a data matrix having as rows the 1141 proteins of our data-set 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 (62; 51) 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.

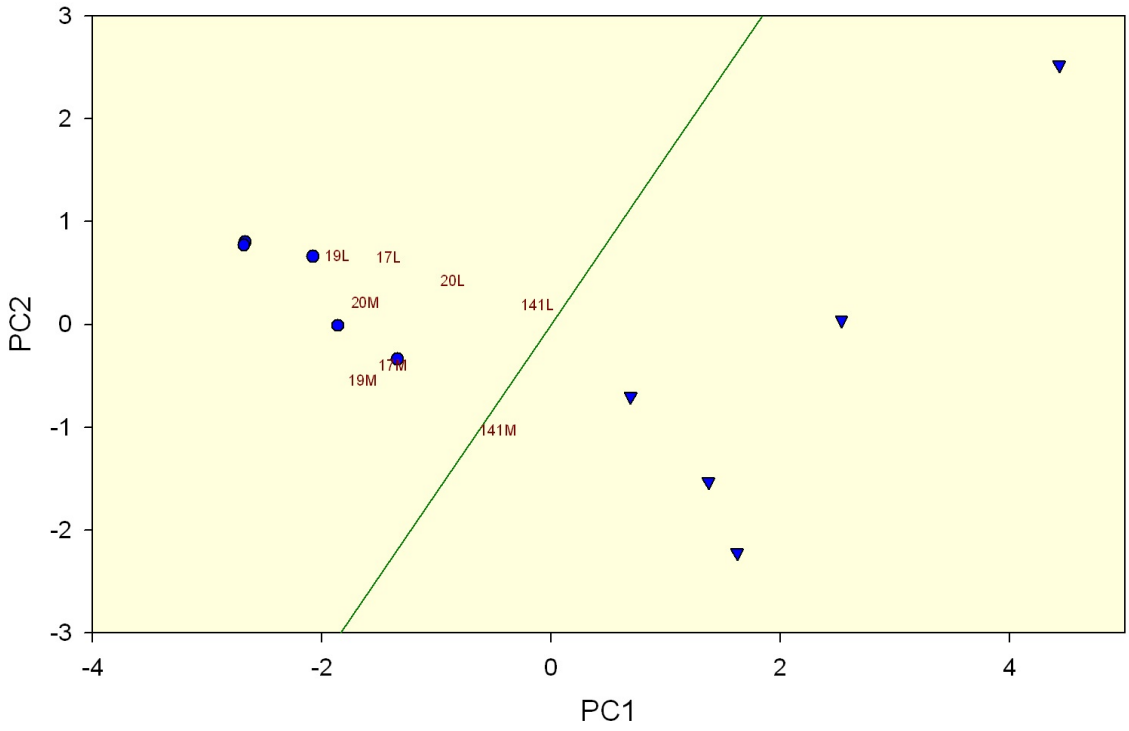

FIGURE 4. *Distribution of point mutants of the Aα-chain of human fibrinogen in a principal component space.* Circles and triangles refer, respectively, to the hemorrhagic and nonhemorrhagic natural point mutants listed in table 4. The straight line separates the hemorrhagic from the nonhemorrhagic group according to the results of a discriminant analysis. The location of simulated mutants (see the text) containing a lysine or a methionine in the position of the naturally occurring hemorrhagic mutants, indicated by self explanatory symbols, were calculated by projecting them in the PC1, PC2 plane on the basis of the appropriate set of RQA parameters according to the method in reference (34).

4. EXAMPLE 2 : STUDYING A SINGLE PROTEIN

In this example, we analyze by RQA the α, β and γ chains in the wild type and in a number of both natural and simulated mutants of human fibrinogen, summarizing the results obtained in a recent paper by Colafranceschi et al. (66). A structural basis for distinguishing between silent and pathological mutants (**For more details see the Supplementary Material**) was found in the case of mutations of the α chain, thanks to the peculiar features of this chain as

compared to the other two. Moreover, we could show that: a) the RQA-based classification of such mutants is in good agreement with the clinical classification based upon hemorrhagic and nonhemorrhagic (or thrombotic) mutants; b) the location of the mutated residues plays a role more relevant than their hydrophobic features; c) the artificial point mutants in the terminal zone (600-866 residues) of the extended isoform of the α-chain cluster together with the natural haemorrhagic mutants of the first 1-207 residues, suggesting a similar role for initial and final portions of the sequence.

To improve the sensitivity of the discrimination between the various type of mutants, we applied a local RQA analysis protocol (44), focusing over a window of 36 residues sliding along the hydrophobicity profile with lag = 1. The RQA variables are computed at each sliding step, and the sum of differences (in absolute value) between the mutant and the wild type calculated according to the following function:

$$f = \sum_i |V_N^{(i)} - V_M^{(i)}| \tag{2}$$

where V = any of the RQA parameters listed in table 1; N and M indicate the native and the mutant, respectively; i = the ith epoch in the polypeptide chain, corresponding to a sliding window of 36 residues along the sequence, lag = 1.

4.1. **Natural Mutants.** In table 4, the point mutants on the Aalpha chains used in the present work are listed together with their clinical impact, as reported on the Web (http://databases.biomedcentral.com/browsesubject/?sub-id = 2010, under the Fibrinogen Variants Database, 28/02/2005 update). A selected subset was obtained from the database excluding the ambiguous results. The score plot (PC1 vs. PC2) relative to the Aα-chain highlights a good separation between hemorrhagic and nonhemorrhagic mutants on the PC1 axis (fig. 4), at odds with the same type of plot relative to β and γ chains, where no significant separation of different groups appears (not shown).

brem17, E	[Gly17Val]	(58)
canter20, E	[Val20Asp]	(9)
detr19, E	[Arg19Ser]	(7)
lima141, E	[Arg141Ser]	(33)
munich19, E	[Arg19Asp]	(29)
cha3_554, NE	[Arg554Cys]	(57)
car5_532, NE	[Ser532Cys]	(37)
car2_434, NE	[Ser434Asn]	(34)
indi_554, NE	[Arg554Leu]	(6)
chri_526, NE	[Glu526Val]	(21)

TABLE 4. *Natively occurring mutants of the human fibrinogen (Aα chain).* The first column contains the names of the hemorrhagic (E) and nonhemorrhagic (NE) mutants; the second and the third columns contain, respectively, the mutation type and the corresponding reference.

4.2. **Simulated Mutants.** By simulating the effect of point mutations in selected locations of the Aα-chain we addressed the question whether the distinction between hemorrhagic and thrombotic mutants depends upon the substitution type or upon the substitution location. To answer the question we replaced one at a time, in the hemorrhagic mutants, the mutated residue with the two residues characterized by the largest difference in hydrophobicity according to the Miyazawa-Jernigan (41) scale, namely Met (hydrophobicity = 8.95) and Lys (hydrophobicity = 2.95). Thus, we produced two sets of artificial mutants, and the result of clustering in a PC1, PC2 plane the recurrence variables is reported in figure 5. Notice that, irrespective of the type of substituted residue, all the artificial mutants lie in the sector spanned by the hemorrhagic natural mutants, with the only partial exception of point 141M.

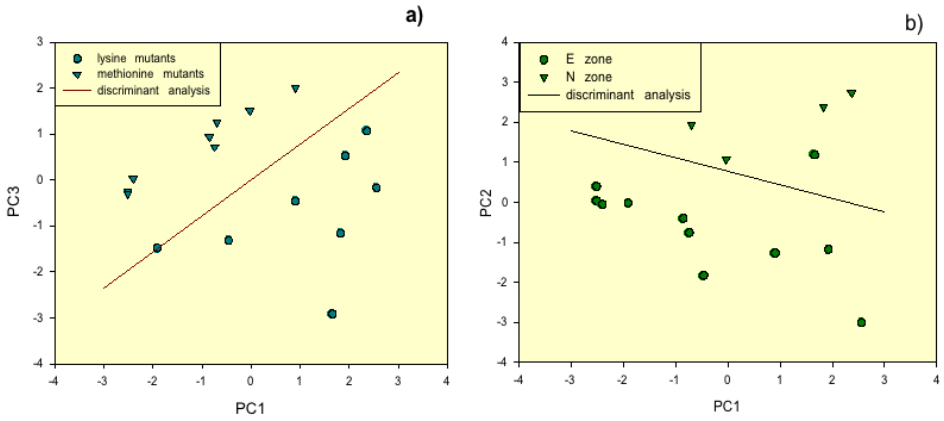

FIGURE 5. *Distribution of simulated point mutants of human fibrinogen (Aα-chain) in a principal component space.* A,B The results of substituting each of the wild-type cysteine residues (c47, c55, c64, c68, c180, c184, c461, c491) with lysine or methionine, obtaining as a total 16 artificial mutants. The PCA was carried out on the whole set of RQA variables listed in table 1. The principal components used in each panel were selected on the basis of a T test /F test indicating the significance of the discrimination between the different groups of mutants represented in the panel. The following zones of the Aα-chain were considered: E (residues 0-200), D (residues 200-400), and N (residues 400-600). Concerning substitutions of cysteine residues, a good separation between mutants different for the hydrophobicity of the substitution and for the substitution location along the chain is highlighted in a) and b), respectively. In the latter case, only E and N zones were considered because of the lack of cysteine residues in the D (200-400) zone.

Besides those of the naturally occurring mutants, other locations for the Lys and Met substitutions were chosen and, in particular, those corresponding to the Cys residues present in the Aα-chain at location 28, 36, 45, 49, 161, 165, 442, and 472. In this case, the PCA filter applied to the RQA variables provides 3 components (PC1, PC2 and PC3) which altogether explain 81% of the total variability. If PC1 and PC3 are used to define the plane, the different hydrophobicities of Lys and Met emerge (fig. 5A). However, substitutions of the Cys in the N zone (residues 0-200), namely c442 and c472, cluster in a different group from all the others (fig. 5B). In the case of a quite longer isoform of the Aα-chain (866 residues instead of 600), the Cys632,

Cys663, Cys799 and Cys812, were substituted in the last zone of the chain (residues 600-end) according to the same above criteria. Figure 7 shows that both Met and Lys substitutions of the above mentioned Cys residues cluster in the same group as the mutants simulated in the E zone (residues 400-600) typical of the hemorrhagic natural mutants, at difference with the mutants simulated in the intermediate section (residues 400-600).

The above result points to: i) a similar role of hydrophobic patterns in the first and in the last regions of the Aα-chain, and ii) a predictive usage of the RQA parameters in correlating structural features and potential functional properties. Since the recurrence patterns of both the beta and gamma chains were substantially different from those of the Aα chain, it appeared of interest to check the recurrence patterns of the former two chains against each other by means of a "cross Recurrence Plot (figure 7). This plot, analogous to a cross-correlation analysis, represents the sequences on the two axes and the dots indicate recurrences between stretches of residues in the position of the corresponding coordinates. A pattern of short, although clearly distinguishable, deterministic lines (three of which are indicated by arrows) appears in this plot, and it is noticeable that they are perfectly aligned with the position of the two Cys bridges connecting the two

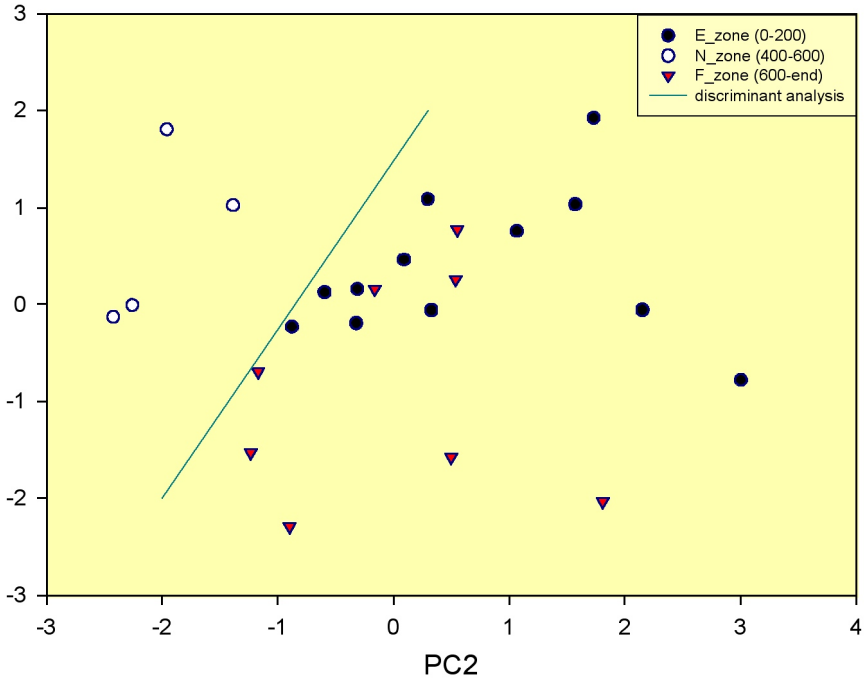

FIGURE 6. *Distribution of simulated point mutants of human fibrinogen (Aα E-chain, 847 residues) in a principal component space.* The plot was drawn following the same criteria as in figure 4. Besides the 16 substitutions in figure 5c,d, 8 more substitutions, namely those corresponding to c632, c663, c799 and c812, are included for a total of 24 mutants. The following zones were considered: E (residues 0-200, N (residues 400-600) and F (residues 600-end). The line marking the result of the discriminant analysis indicates a quite good separation between mutants in the N zone and in the E, F zones.

chains, namely Bβ c110 - γ c45 and Bβ c227 - γ c161 . Besides reinforcing the importance of Cys residue location in the formation of the Bβ-γ aggregate, this seems to indicate a combined set of chemical bonds and hydrophobic forces contributing to the overall stability of the interaction between thebeta and gamma chain in the fibrinogen trimers.

Altogether, the reported results: a) support a remarkable role of the αC domain in determining the structure and the properties of the clots, b) reinforce the hypothesis that the αC domain is an intrinsically unstructured entity, and c) uphold the view that the clinical symptoms observed in Aalpha-chain dysfibrinogenic patients can be predicted from calculations based on RQA. Accordingly, it seems fair to state that RQA represents a precious tool for the development of a database of point mutations in fibrinogen Aα-chain, in which epidemiological as well as structural and functional data could find a proficient integration.

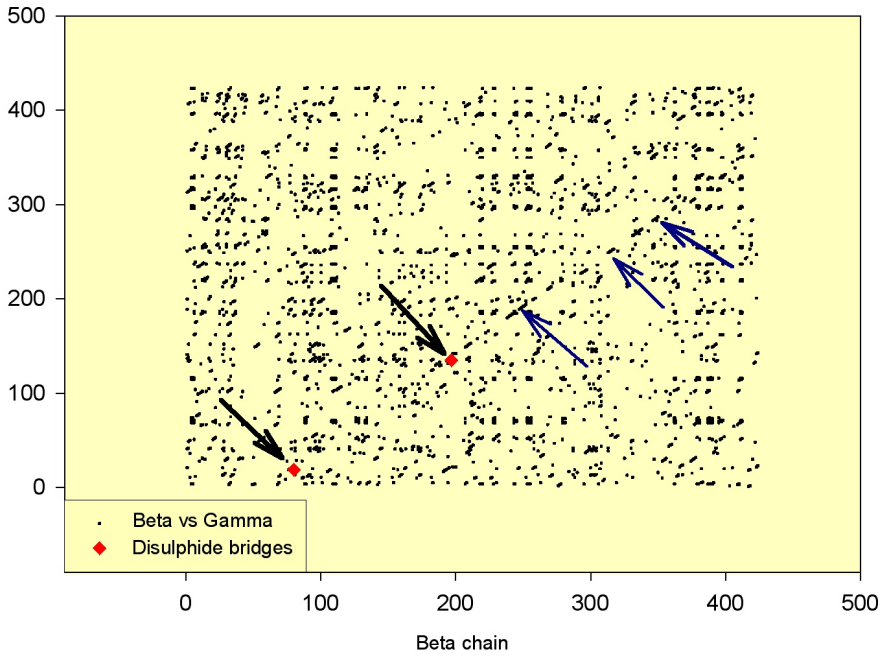

FIGURE 7. *Cross recurrence plots of Bβ-and γ chains of human fibrinogen.* The grey arrows mark the S-S bridges connecting the two chains, while the black arrows indicate some of the deterministic lines of recurrent points in perfect alignment with the S-S bridges (see the text). Notice that the residue numbering for both proteins refers to chains depleted of the signal peptides (30 and 26 amino acids for Bbeta and gamma respectively).

5. CONCLUSIONS

Since 1994 Hans Frauenfelder and Peter Wolynes (20) focused on the peculiarity of the sequence-structure relation in proteins and on the need to have very accurate physics principles of microscopic "simple" systems (like atoms, small molecules) cooperatively interacting to produce macroscopic principles describing complex systems (like protein structure). While we

do have an accurate knowledge of potentials (hydrophobic interactions, hydrogen bonding, size constraints, etc.) acting at microscopic levels, the "mesoscopic" principles needed to predict the 3D structure of proteins remain essentially unknown. This blend of microscopic principles and macroscopic consequences has been a typical feature of chemical sciences in the last century and also inspired the present review.

The reported examples are indicative of an analytical strategy located half-way between the purely empirical 'sequence homology' and 'theoretically intensive' ab initio approaches in protein science. The intensive use of statistics is made possible by the systematic reckoning of a set of dynamical descriptors characterizing each sequence, so that ensemble-based correlations can be sketched on large and heterogeneous data sets without the limitation of dealing with sequences similar enough to compute a meaningful homology score . On the other hand, coding the aminoacid residues by chemico-physical properties allows to link theoretically sound considerations with the obtained results.

The basic assumption of the proposed method, namely the consideration of a protein primary sequence as an ordered numerical series analogous to a time series, has a relevant impact for the prediction of structural and physiological properties. While this assumption is surely reasonable in general (given the already stated importance of the aminoacid sequence in the determination of protein function), the success of the approach in the specific cases, as for any empirical investigation, strictly depends upon both the selection of an appropriate data set and the formulation of a pertinent questions, consistent with the data at hand. As usual, the boundary conditions play a much more relevant role than basic principles when dealing with complex system behaviour.

6. Appendix

Fibrinogen structure (low resolution: see Figure 8).

Fibrinogen is a protein made up of two copies of three different polypeptide chains named Aα, Bβ and γ (69) composed of 644, 491 and 453 amino acids, respectively . It is assembled in hepatocytes to form an elongated (45 nm), tri-nodular molecule organized into dimers. The molecule includes a central disulfide-bridged E domain (or central nodule) containing the amino termini of all six chains and connected by two α-helical coiled-coil segments that extend in opposite directions from the center to two outer disulfide-bridged D domains (or external nodules).

Plasma-derived human fibrinogen shows a high degree of heterogeneity in healthy individuals (70), being typically made up of a predominant (50 to 70%) circulating component having a molecular mass of 340 kDa, and a component with one or two degraded Aα chains (20 to 50 %). Two splice variants, called γ (5 to 8%), and Aα extended (1 to 3 %), may also be present. The latter variant (also termed fibrinogen-420 for the high molecular mass of 420 kDa) contains Aα isoforms named αE (where subscript E means extended) which include a unique 236-residue C-terminal extension (αEC). The extension constitutes an additional globular domain containing a recognition site for leucocyte 2-integrins that mediate the recruitment of leucocyte to the site of inflammation.

References

[1] Adkins, J.N.; Lumb, K. J. Proteins: Struct. Funct. Genet. (2002), 1:1.
[2] Anfinsen, C.B. (1973), Science , 181:223.
[3] Baker, D.; Sali, A. Science (2001), 294:93.
[4] Banavar, J.R.; Maritan, A. Proteins: Struct. Funct. Genet. (2001), 42:433.
[5] Benigni, R.; Giuliani, A. Am. J. Physiol. (1994), 266:R1697.
[6] Benson M.D, Liepnieks J., Uemichi T., Wheeler G. and Correa R. (1993) Nat Genet 3: 252-255.
[7] Blomback M., Blomback B., Mammen E.F., Prasad A.S. (1968) Nature 218:134.
[8] Branden, C.I. and Tooze, J.(1991), Introduction to Protein Structure; Garland: New York.
[9] Brennan S.O., Hammonds B., George P.M. (1995), J. Clin. Invest. 96:2854.

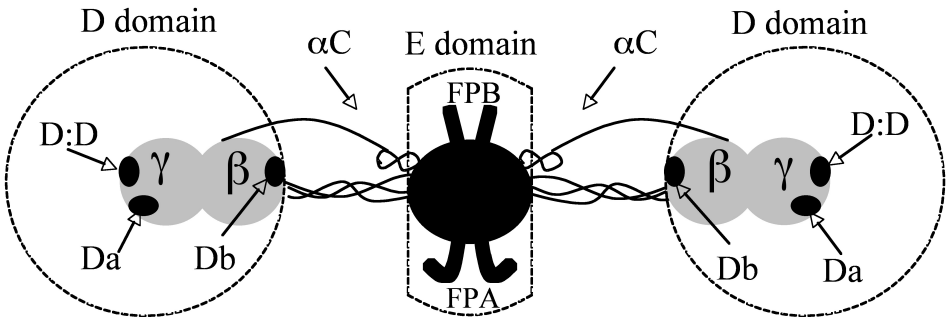

FIGURE 8. The major domains (E and D) and the coiled-coil segments (αC) connecting them in fibrinogen and fibrin monomer. Catalyzed conversion of fibrinogen to fibrin results in release of fibrinopeptides A (FpA, sequence Aα1-16) and B (FpB, sequence Bβ1-14) from the amino termini of the Aα- and Bβ-chains of fibrinogen, and exposure of two knobs, the EA and EB association sites, respectively. .

[10] Broomhead D.S. and King G.P. (1986), Physica D 20:217.
[11] Chasman D., Adams, R.M. (2001) J. Mol. Biol. , 307:683.
[12] Chothia C. (1976), J. Mol. Biol. 105:1.
[13] Compiani M,. Fariselli P., Martelli P.L. and Casadio R. (1998), Proc. Natl. Acad. Sci. U.S.A. 95:9290.
[14] Computational Methods in Molecular Biology (1998), Salzberg, S. L., Searls, D.B., Kasif, S., Eds.; Elsevier: Amsterdam.
[15] Dobson C.M., Karplus M. (1999), Curr.Opin. Struct. Biol. , 9:92.
[16] Dokholyan, N.V. (2004), Proteins: Struct., Funct., Bioinform. 54:622.
[17] Dunker K., Brown C.J., Lawson D., Iakoucheva L.M. and Obradovic (2002), Biochemistry 41:6573.
[18] Eckmann J.P., Kamphorst S.O. and Ruelle D. (1987), Europhys. Lett. 4:324.
[19] Feller W. (1968) An introduction to Probability Theory and Its Applications, Wiley: New York, Vol. 1.
[20] Frauenfelder H., Wolynes P. (1994), Phys. Today 47:58.
[21] George P.M., et al : www.geht.org/fr/pages/pratiqueBase UK B.html .
[22] Giuliani A., Piccirillo G., Marigliano V. and Colosimo A. (1998) Am. J. Physiol. 275:H1455.
[23] Giuliani A., Sirabella P., Benigni R. and Colosimo A. (2000) Protein Eng. 13:671.
[24] Grantham R. (1974) Science 185:862.
[25] Grigoriev I.V. and Kim S.H. (1999), Proteins: Struct. Funct. Genet., 96:14318.
[26] Guharay S., Hunt B.R., Yorke J.A. and White O.R. Physica D, 2000, 146:388.
[27] Hansch, C. Acc. Chem. Res. 1993, 26:147.
[28] Hansch C., Hoekman D. and Gao H. (1996), Chem. Rev., 96:1045.
[29] Henschen A., Kehl M. and Deutsch E. (1983), Hoppe Seylers Z Physiol Chem 364:1747.
[30] Irback A. and Sandelin E. (2000), Biophys. J., 79:2252.
[31] Jones D. (1975) J. Theor. Biol. 50:167.
[32] Kyte J. and Doolitle R.F. (1982), J. Mol. Biol. 157:105.
[33] Maekawa H, Yamazumi K, Muramatsu S, Kaneko M, Hirata H et al. (1992), J. Clin. Invest. 90:67.
[34] Maekawa N., de Bosch N.B., Carvajal Z., Ojeda A., Arocha-Pinango C.L., et al (1991) J. Biol. Chem. 266:11575.
[35] Makhadatze G.I. and Privalov P.L. (1995), Adv. Protein Chem. , 47:307.
[36] Manetti C., Ceruso M.A., Giuliani A., Webber C.L. and Zbilut J.P. (1999), J. Phys. Rev. E 59:992.
[37] Marchi R, Lundberg U, Grimbergen J, Koopman J, Torres A, et al. (2000) Thromb Haemost 84:263.

[38] Marti-Renom, M.A., Stuart A.C., Fiser A., Sanchez R., Melo F. and Sali, A.(2000), Ann. Rev. Biophys. Biomol. Struct., 29:291.
[39] Marwan N., Wessel N., Meyerfeldt U., Schirdewan A. and Kurths J. (2002), Phys. Rev. E 66(2):026702.
[40] Menne K.M.L., Hermjakob H. and Apweiler R. (2000), Bioinformatics 16:741.
[41] Miyazawa S. and Jernigan R.L. (1985) Macromolecules 18:534.
[42] Pande V. S., Grosberg A. Y. and Tanaka T. (2000), Rev. Mod. Phys. , 72:259.
[43] Pearson, W.R. and Lipman D.J. (1988), Proc. Natl. Acad. Sci. U.S.A., 85:2444.
[44] Porrello A., Soddu S., Zbilut J.P., Crescenzi M. and Giuliani A. (2004), Proteins: Struct. Funct. Bioinform., 55:743.
[45] Romero P.; Obradovic, Z.; Li, X.; Garner, E. C.; Brown, C. J.; (2001), Dunker, Proteins Struct., Funct., Genet. 42:38.
[46] Sander C. and Schneider R. (1991), Proteins: Struct. Funct. Genet., 9:56.
[47] Schreiber T. (1999), Phys. Rep. 308:1.
[48] Senno C.F., Micheletti A., Maritan A., Banavar J.R. (1998), Phys. Rev. Lett., 80:2237.
[49] Simons K.T., Strauss C. and Baker D.J. (2001), J. Mol. Biol., 306:1191.
[50] Sinha N. and Nussinov R. (2001), Proc. Natl. Acad. Sci. U.S.A. , 98:3139.
[51] Strait B.J., Dewey T.G., Strait B.J. and Dewey T.G. (1996), Biophys. J. 71:741.
[52] Sweet R.M. and Eisenberg D. (1983), J. Mol. Biol. , 171:479.
[53] Tang H.C.H., Wingreen, N. S. (1997) Phys. ReV. Lett. 79, 765-768.
[54] Taylor, W. R.; May A.C., Brown N.P. and Aszodi A. (2001), Rep. Prog. Phys. , 64:517.
[55] Uversky, V. N. (2002), Protein Sci. 11:739.
[56] von Heijne, G. J. Mol. Biol. 1982, 159, 537.
[57] Wada Y, Lord S.T., Dusart and Chapel Hill III (1994). Blood 84:3709.
[58] Wada Y., Niwa K., Maekawa H., Asakura S., Sugo T., et al. (1993) Thromb Haemost. 70:397.
[59] Wang J. and Wang W. (2002), Phys. ReV. E 65:41911.
[60] Webber C.L. and Zbilut J.P. (1994) J. Appl. Physiol. 76:965.
[61] Weiss O. and Herzel H.J. (1998), Theor. Biol. , 190:341.
[62] Weiss O., Jimenez-Montano M.A. and Herzel H. (2000), J. Theor. Biol. 206:379.
[63] Zbilut J.P., Webber C.L. Jr., Colosimo A. and Giuliani A. (2000), Protein Eng. 13:99.
[64] Zimmermann J.M., Eliezer N. and Simha R. (1968) J. Theor. Biol. 21:170.
[65] Colafranceschi M., Colosimo A., Zbilut J.P., Uversky V.M. and Giuliani A. (2005), J.Chem.Inf.Model, 45:183.
[66] Colafranceschi M., Papi M., Giuliani A: and Colosimo A., Pathophysiol. Haemost. Thromb. (2006), 35:417.
[67] Zbilut J.P., Giuliani A., Colosimo A., Mitchell J.C., Colafranceschi M., Marwan N., Charles L. Webber C.L. Jr., Uversky V.N. (2004), J. Proteo. Res., 3:1243.
[68] Zbilut J.P., Sirabella P., Giuliani A., Manetti M., Colosimo A. and Webber C.L. Jr. (2002), Cell Biochem. Biophys., 36:70.
[69] Mosesson M.W., (2004), J. Thromb. Haemost. 3:1894
[70] De Maat M. and Verschuur M., (2005), Curr. Opin. Hematol. 12:377.