"DISTRIBUTION OF HYDROPHOBICITY PATTERNS IN PROTEINS PRIMARY STRUCTURES: A STATISTICAL STUDY"

DOTTORANDO: MAURO COLAFRANCESCHI

DOTTORATO DI RICERCA IN BIOFISICA (XVII CICLO)
Dipartimento di Fisiologia Umana e Farmacologia "V. Erspamer" – Università degli Studi di Roma "La Sapienza"

DIRETTORE SCUOLA DI DOTTORATO: **PROF. ALFREDO COLOSIMO**
Dipartimento di Fisiologia Umana e Farmacologia "V. Erspamer" – Università degli Studi di Roma "La Sapienza"

TUTORE SCIENTIFICO: **PROF. ALFREDO COLOSIMO**

DOCENTI ESAMINATORI:

PROF. MARIO COMPIANI – Dip. Scienze Chimiche – Università di Camerino
PROF. STEFANO PASCARELLA – Dip. Scienze Biochimiche – Università degli Studi di Roma "La Sapienza"
PROF. ENRICO STADERINI – Dip. Biopatologia e Diagnostica per Immagini – Università degli studi di Roma "Tor Vergata"

Acknowledgements:

I would like to take the opportunity of thanking my Professor, the colleagues and friends, who have contributed to the preparation of this thesis. Especially Prof. Alfredo Colosimo and Dr. Alessandro Giuliani, for their crucial role in my scientific preparation; Prof. Joseph P. Zbilut for the collaboration and the individual responses to my professional requests; Dr. Paolo Sirabella and Dr. Giovanna Zimatore for computational supports.

INDEX

1. AMINOACIDS AND PROTEINS

1.1 Introduction

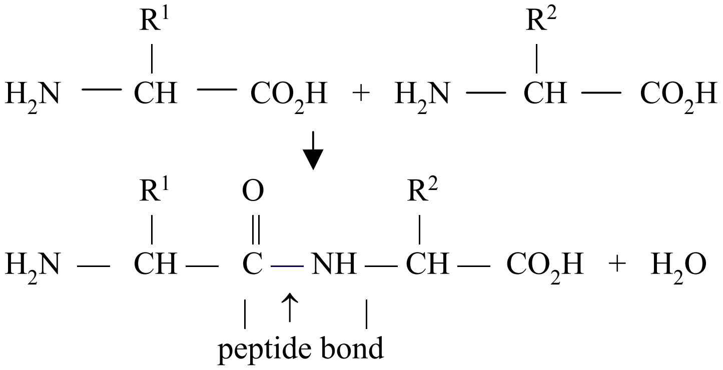
The biological consequences of the genetic information encoded by nucleic acids depends almost completely on proteins. Proteins play a variety of crucial roles in biological systems and affect many properties characterizing a living organism: the rates of chemical reactions in organisms are dependent upon enzyme proteins, the structures of cells and tissues rely upon structural proteins, the integrity of the genome is maintained by both DNA repair proteins and histones.

In spite of their diverse biological functions, proteins are a relatively homogeneous class of molecules. They are linear polymers constituted by various combinations of the same 20 aminoacids, and they differ only in the sequence in which the amino acids are assembled into polymeric chains.

Protein functional diversity is partly due to the chemical diversity of the aminoacids, but primarily is due to the fact that the building blocks (aminoacids) linked in different sequences form different three-dimensional structures: the 3-D structure is in turn responsible for the functional role of proteins. The basic aim of this thesis is to try and demonstrate the emerging of brand new properties, not directly derivable from physico-chemical features of aminoacids, due to the juxtaposition of different residues into linear chains.

1.2 Protein primary structure

The 20 aminoacids are linked each other to form protein sequences by the **peptide bond** (see below).


R1
R2
H2N — CH — CO2H + H2N — CH — CO2H
R1 O R2
H2N — CH — C — NH — CH — CO2H + H2O

peptide bond

Generally, between 50 and 3000 such aminoacids are linked in this way to form a **polypeptide chain**. The polypeptide backbone is a repetition of the basic unit common to all aminoacids. When the side chain is included, this unit is described as an aminoacid residue. Some properties of the aminoacid residues are listed in table 1.1.

Residue	Abbreviations	Hydrophobicity	Charge	Frequency in
		(MJ scale)		proteins (%)
Alanine	Ala (A)	5.33	0	8.3
Arginine	Arg (R)	4.18	+1	5.7
Asparagine	Asn (N)	3.71	0	4.4
Aspartic acid	Asp (D)	3.59	-1	5.3
Cysteine	Cys (C)	7.93	0	1.7
Glutamine	Gln (Q)	3.87	0	4
Glutamic acid	Glu (E)	3.65	-1	6.2
Glycine	Gly (G)	4.48	0	7.2
Histidine	His (H)	5.1	+1	2.2
Isoleucine	Ile (I)	8.83	0	5.2
Leucine	Leu (L)	8.47	0	9
Lysine	Lys (K)	2.95	+1	5.7
Methionine	Met (M)	8.95	0	2.4
Phenylalanine	Phe (F)	9.03	0	3.9
Proline	Pro (P)	3.87	0	5.1
Serine	Ser (S)	4.09	0	6.9
Threonine	Thr (T)	4.49	0	5.8
Tryptophan	Trp (W)	7.66	0	1.3
Tyrosine	Tyr (Y)	5.89	0	3.2
Valine	Val (V)	7.63	0	6.6

Table 1.1 Properties of individual aminoacid residues.

Sequences of aminoacid residues in proteins are usually written with either the three-letter or the one-letter abbreviations, starting with the N-terminal residue. The aminoacid sequence, appropriately called the **primary structure**, identifies a protein unambiguously.

1.3 Non covalent interactions

Protein primary structures are covalent structures. Knowledge of this structure is usually adequate for characterizing the chemistry of small molecules, but not for proteins. The large size of polypeptide chains enables them to fold back on themselves so that many simultaneous interactions take place among different parts of the molecule.

A complex, three dimensional structure results, which provides the unique environments and orientations of the functional groups that give proteins their special properties. The biological activities of proteins are also mediated by their interactions with the environment: water, salts, membranes, other proteins, nucleic acids and the numerous other molecules in living systems. All of these interactions arise from a limited set of fundamental non covalent forces (electrostatic forces, hydrogen-bond, Van der Waals interactions, etc.).

1.4 Hydrophobic interaction

Water is a very poor solvent for non polar molecules compared with most organic liquids. Non polar molecules cannot participate in the hydrogen bonding that appears to be so important in liquid water. This relative absence of interactions between non polar molecules and water causes interactions among the non polar groups themselves to be much more favorable than would be the case in other solvents.

This preference of non polar atoms for non aqueous environments has come to be known as the "hydrophobic interaction" and it is a major factor in the stability of proteins. Hence the hydrophobic interaction results in a tendency of non polar atoms to interact with each other rather than with water. This tendency stems from the decrease of entropy caused by the structuring of water coming from the exposition of hydrophobic residues to the aqueous environment.

The hydrophobicity of the individual aminoacid side chains have been measured experimentally by a variety of methods. Usually, the numeric value of hydrophobicity is expressed in terms of partition coefficients between non polar solvent (octanol, ethanol, etc.) and water. Another important class of hydrophobicity indexes are the so called ‘statistical potentials’, in which the relative hydrophobicity of residues is derived by their mutual propensities to occupy spatially close positions in actual 3D protein structures.

2. THE FOLDING PROBLEM

2.1 The basic problem

In order to become biologically active, the vast majority of protein sequences must fold to a unique stable structure. The question of how individual protein sequences efficiently and reliably achieve their native state following synthesis on the ribosome is one of the most intriguing problems in structural biology. In a cell, folding takes place within a complex environment containing high concentrations of a wide variety of molecules and ions.

It is well established that many factors are associated with the cellular folding process, including molecular chaperones and folding catalysts. The various factors are involved in a wide range of control and localization processes, but do not provide conformational information for the polypeptide chains with which they interact.

The evidence gathered over many years supports the fundamental principle, formulated initially by Anfinsen and others, that the code for folding resides within the aminoacid sequence[1]. The fundamental question is, therefore, how the sequence codes for the fold.

From a chemical viewpoint, proteins are linear heteropolymers that, unlike most synthetic polymers, are formed of basically non periodic sequences of 20 different monomers. While artificial polymers are generally very large extended molecules forming a matrix, the majority of proteins fold as self-contained structures determined by the sequence of monomers [2]. 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 structure [3].

Well defined 3-D structure should not be intended as “fixed architecture”. Many proteins appear as partially or even totally disordered when analyzed with spectroscopic methods [4]. However, this apparent disorder corresponds to an efficient organization as for protein physiological function. The task of being water soluble while maintaining the structural specificity necessary for a physiologically motivated activity is not easy, and only a relative minority of linear amino acid arrangements can actually accomplish this.

Thus, the most basic problem is "what particular linear arrangement of aminoacids makes a real protein ?”. An operational definition of this question can be given in the following terms: “Do the analysis of linear arrangements of aminoacids corresponding to real proteins show some peculiarities not present in random sequences ?”. In order to approach the above problem in a quantitative manner protein sequences were considered as numerical series whose elements are the subsequent residues coded by different physico-chemical properties.

2.2 Protein aggregation and misfolding

As a matter of fact, proteins interacting with other proteins of the same or different kind are not endowed with qualitatively different features. Hence, in principle, the interaction between different portions of the same molecule (protein folding) and the interaction between several molecules (protein aggregation) represent two faces of the same coin.

Because of the central role played by protein folding in cell biology, an aberrant folding, or ‘misfolding’ will give rise to malfunctioning of biological processes. The number of diseases known to be associated with misfolding is large and increasing all the time. Such diseases are often familial, as impairment of the ability of a protein to fold can result from even single mutations in the chain.

Particular attention has, however, been focused on those diseases in which misfolding results in aggregation, particularly when these aggregates occur in the highly organized form known as amyloid fibrils [5]. This class of conditions includes Alzheimer's disease and Creutzfeldt-Jakob disease, as well as a range of other neurological or systemic diseases.
Although the different diseases have many different features, their molecular origins may have much in common. The aggregation of incompletely folded, misfolded or even partially degraded proteins is a complex process that progresses through a series of small oligomers to more organized structures such as protofilaments before well defined fibrils are formed.

As clearly stated by Chiti et al. [6] , the possibility of forming aggregates is intrinsic to any protein. Consequently, it may be difficult to delineate a clear cut border separating aggregating and non aggregating proteins. Any modification of environmental conditions (PH, temperature, ionic strenght, etc.) could in principle drive any protein structure to shift from an isolated globular existence in solution to the formation of multimeric aggregates.

This possibility is implicit in the character of the hydrophobic interaction. The main driving force shaping protein tertiary structures is the need to be soluble in the water. For this task to be accomplished, the protein must fold in such a way as to hide hydrophobic residues, while exposing polar residues [7] . On the contrary, the exposition to the solvent of corresponding hydrophobic residues in several molecules will favour protein-protein interaction, since it is energetically favourable than their exposition to the water.

Hence protein-protein interaction can be considered as another aspect of the same phenomenon. In general, the search for aggregation cores is not basically different from the search for folding cores, and aggregation can be simply considered an alternative folding. The choice between correct (autonomous) and incorrect (multimeric) folding is a matter of relative preponderance (in energetic terms for given boundary conditions) of the two possible ways. In particular, the relative probability is driven by the balance of hydrophobic charge and steric effects of sequence/environment interaction [5] . This implies the possibility of recognizing the relative propensity for aggregation by means of an efficient physico-chemical representation of proteins.

A second crucial element for protein aggregation is the charge. A net charge different to zero exposed on the surface of different molecules can result in two conflicting effects according to the reciprocal sign of the charge. Charge having opposite sign cause a reciprocal attraction between protein molecules (pro-aggregation effect), while charges with the same sign cause a reciprocal repulsion between molecules (anti-aggregation effect).

3. AIM OF THE THESIS

The idea of this thesis originates in a recent work [8] in which the possibility to obtain an unambiguous and self-consistent measure of complexity for any time (or spatial) series has been demonstrated. This assertion makes it possible to virtually compare each type of numerical series in terms of complexity and it opens the way to a wider application of dynamic systems concepts in empirical sciences.

The above considerations point out the possibility to apply dynamic analysis techniques to protein primary structures (spatial series), in the aim to give a contribution to the clarification of the sequence/structure/function puzzle (see chapter 4). In recent times some studies about this topic have been carried out [9, 10] . In these papers the use of an opportune coding of primary structures (based on the hydrophobicity profiles of aminoacid side chains), associated to the calculation of dynamic (order-dependent) descriptors, provided brand-new information compared to the classical analysis of sequence homology. This result points to an alternative method to sequence alignment for studying sequence/function relationships.

Recently it has been proposed that some key features of protein hydrophobicity patterns analyzed by non linear signal processing techniques might determine necessary conditions for aggregation [11]. In this frame the charge acts through modulation of the repulsive/attractive electrostatic forces between nearby molecules. This observation may form the basis for the construction of a "charge/hydrophobicity" model of protein aggregation [12] (see the conclusions).

On the other hand, since the crucial role of hydrophobicity in protein folding is well known, the aim of the thesis is to check if the 'extra information' provided by the hydrophobicity can result in some general syntactic rule in the aminoacid distribution along protein sequences (see chapter 5), i.e. a decreasing of complexity of hydrophobicity profiles compared to those of other aminoacid codings (polarity, bulkiness, etc..). The possibility of comparing the complexity values of different aminoacid 'translations' arises from the existence of an unambiguous complexity scale for different numerical series [8].

A second basic aim of this thesis is to check a possible relation between complexity descriptors calculated on the numerical transformations of primary structures and structural/functional features of proteins (see chapter 6). The usefulness of such an analysis is due to the fact that sometimes proteins with a low sequence homology (< 20%) can adopt similar folds. These similarities are not detectable by means of the classical methods based on sequence alignment. On the other hand, the analysis of RQA descriptors, being not dependent on sequence homology, could allow for the detection of unexpected “neighbours” of query structures, so enlarging and refining the performance of function assignment methods.

The importance of acquiring new effective analytical tools in this field is based on the firm belief that the ability in predicting the 3-D structure exclusively starting from the aminoacid sequence will largerly improve the full exploitation of the huge information coming from the human genome sequence.

4. MATHEMATICAL METHODS

4.1 Signal analysis perspective

When coded as monodimensional numerical arrays of some physico-chemical property (hydrophobicity, volume, polarity, etc.) of their aminoacid residues, protein sequences can be considered as numerical discrete series equivalent to time series, with the aminoacid order playing the role of subsequent time intervals. Thus, on a purely formal point of view, any technique commonly used for signal and time series analysis could be successfully applied to protein primary structures. From a practical viewpoint, 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.

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 [13]. Methods satisfying these constraints are those approaching the analyzed series from a purely correlative point of view, with no a priori distributional and/or physical assumption. The only aim of these methods is to look for autocorrelation patterns along the series, i.e., for the recurrences of particular short motifs along the chain (like in recurrence quantification analysis, RQA) or for periodicities of no predefined functional form spanning all the studied sequences (like in singular value decomposition, SVD). At the basis of all these methods is the transformation of the original series into its “embedding matrix” with the method of delays [14].

4.2 Embedding procedure

The embedding procedure consists of building an *n*-column matrix (in the example below *n* = 4) out of the original linear array by shifting the series by a fixed lag. For example, given the series 15, 12, 27, 39, 31, 65, 22, 12, 42, 11, 33..., the corresponding 4-dimensional embedding space at lag = 1 (the discrete character of amino acid sequences dictates this choice) is :

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 for the embedding. The choice of the embedding dimension corresponds to the choice of the scale at which the autocorrelation structure of the series is estimated.

All the signal analysis techniques used in research on proteins give a global picture of the series in terms of degree of complexity (relative order/ disorder of hydrophobicity distribution along the series), presence of singularities (regions within the sequence strongly different in terms of hydrophobicity pattern), and specific periodicities. In general, they quantitatively describe the shape of the hydrophobicity profiles by appropriate numerical indicators. In the present work this approach has been extended by the analysis of the distributions of various different physico-chemical properties of aminoacid residues along protein sequences (see Strategy of Analysis).

4.3 Algorithms used for the analysis

4.3.1 Recurrence Quantification Analysis (RQA)

Recurrence quantification analysis is a relatively new nonlinear technique, originally developed by Eckmann et al.[15] as a purely graphical method and then made quantitative by Webber and Zbilut [16] . It was successfully applied to different fields ranging from physiology [17] to molecular dynamics [18] and the study of chemical reactions [19]. Only in relatively recent times RQA was investigated by our group for its ability to deal with protein sequences [9, 10].

The application of RQA is based on the calculation of the Euclidean distance between all the pairs of rows of the embedding matrix. If the distance between two generic rows (i.e. windows of predefined length along the sequence) falls below the radius, we obtain a recurrence.

The concept of recurrence is straightforward: for any ordered series (time or spatial), a recurrence is a point which repeats itself. In this respect, the statistical literature points out that recurrences are the most basic of relations [20] shaping a given system, since they are strictly local and independent of any mathematical assumption regarding the system itself. Furthermore, it is worth stressing that calculation of recurrences requires no transformation of the data and can be used for both linear and nonlinear systems [17] . The concept of a recurrence can be expressed as follows: given a reference point, X_0, and a ball of radius r, a point X is said to recur (with reference to X_0) if : $\{ X : \| X - X_0 \| \leq r \}$.

In the case of a time series, i.e., of a system occupying in different times different positions along a trajectory in a suitable state space, the recurrences correspond to the time points where the system passes nearby to already visited states. In the case of protein sequences, time corresponds to the amino acid order and the recurrences are patches, with a length equal to the embedding dimension, sharing their profile with other patches along the chain. The number and relative positions of recurrences are expressed by recurrence plots (RP), that are symmetrical $N \times N$ arrays in which a point is placed at (i, j) whenever a point X_i on the trajectory is close to another point X_j. The closeness between X_i and X_j is expressed by calculating the Euclidian distance between these two normed vectors, i.e., by subtracting one from the other obtaining the expression $\| X_i - X_j \| \leq r$, where r is a fixed radius.

If the distance falls within this radius, the two vectors are considered to be recurrent, and graphically this can be indicated by a dot (Figure 4.1).

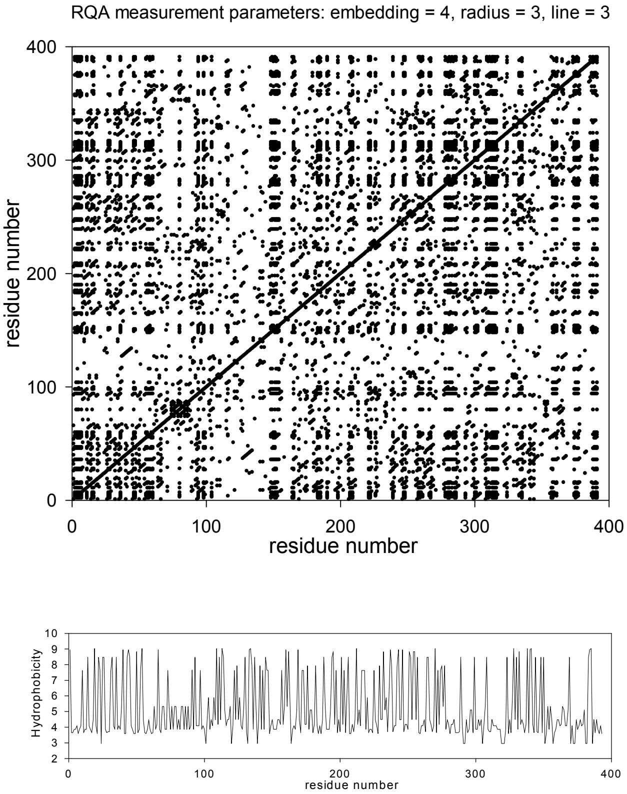


Figure 4.1 Recurrence plot of human P53 protein. The recurrence plot of human P53 protein is reported together with the corresponding hydrophobicity plot (bottom). The presence of an extremely deterministic ordering of amino acids between residues 61 and 98 is clearly evident in the figure in terms of its consequences on the recurrence plot. This highly deterministic portion is "resembled" by other segments along the sequence. This observation is not clear by the simple inspection of the hydrophobicity plot but is made evident by the recurrence plot: the "resemblances" correspond to linear (or alternatively horizontal given the symmetrical character of recurrence plot) banding of the plot. The RQA numerical descriptors corresponding to the plot have been reported together with the chosen measurement settings (see text for further details).

Thus, recurrence plots correspond to the distance matrix between the different epochs (rows of the embedding matrix) filtered, by the action of the radius, to a binary 0/1 matrix having a 1 (dot) for distances falling below the radius and a 0 for distances greater than radius. Distance matrixes are demonstrated [21] to convey all the relevant information for the global reconstruction of a given system. An important feature of such matrixes is the existence of short line segments parallel to the main diagonal, which correspond to sequences (i, j), $(i + 1, j + 1)$, ..., $(i + k, j + k)$ such that the fragment $X(j)$, $X(j + 1)$, $X(j + k)$ is close to $X(i)$, $X(i + 1)$, ..., $X(i + k)$. The absence of such patterns suggests randomness [15]. For protein sequences these deterministic lines correspond to contiguous patches of similar hydrophobic/ hydrophilic patterns. The "line" is a measurement parameter which state the minimum number of adjacent recurrent points required to define a deterministic line.

Because graphical representations may be difficult to evaluate, several strategies to quantify features of such plots have been developed [16, 22]. Hence, the quantification of recurrences leads to the generation of the following variables:

- Recurrence (REC) : percentage of recurrence points in an RP.
- Determinism (DET) : percentage of recurrence points which form diagonal lines.
- Laminarity (LAM) : percent 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.

These recurrence variables quantify the deterministic structure and complexity of the plot (Figure 4.2).

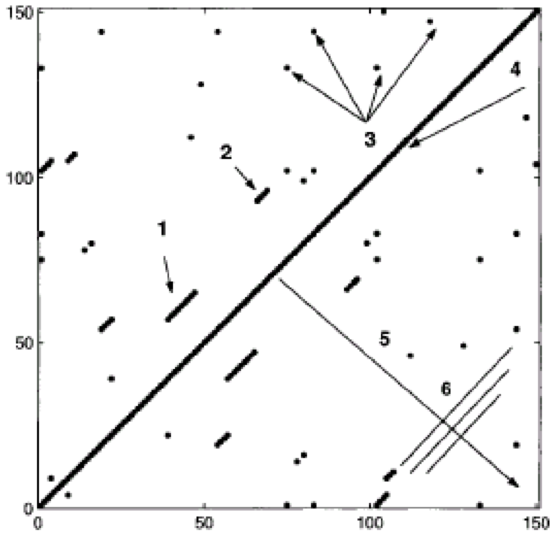


Figure 4.2. Example of an RP with typical features forming the basis for its quantification. (1) A line segment composed of 8 recurrent points; (2) a 4-point line segment; (3) several recurrent points that are not part of a line segment; (4) the identity line (i.e., where Dij = 0); (5) this feature points to a line perpendicular to the identity line which indicates the paling of the RP towards its edges (i.e. the TREND descriptor); (6) several perpendicular lines along which recurrences may fall. The position of these line segments away from the identity line provide the basis for the index used to calculate TREND.

The application of n statistical indexes to the recurrence plots gives rise to an n - dimensional representation of the studied series. This n - dimensional representation gives a summary of the autocorrelation structure of the series and has been demonstrated [20] to correlate with the visual impression a set of unbiased observers derive from the inspection of an ensemble of recurrence plots.

4.3.2 Principal Component Analysis (PCA)

In contrast to RQA, singular value decomposition (SVD) is a well-established method frequently used in physical as well as in social and biological sciences [23]. SVD roughly corresponds to PCA (principal component analysis): the term SVD is preferred to the term PCA in physical applications and, in general, when dealing with dynamical phenomena. As in PCA, the aim of SVD is to project an originally multidimensional phenomenon onto a reduced set of new orthogonal axes, representing the basic modes explaining the analyzed data set.

When applied to a time (or spatial) series that is originally monodimensional, SVD necessitates that the original series is represented on a multidimensional space by the agency of the embedding procedure. This “expansion” of the original mono-dimensional series on a multidimensional support made by the time-lagged copies of the original series allows for the autocorrelation structures of the series to be appreciated [14]. The EM can be thought of as a multivariate matrix having subsequent patches of amino acids of length equal to the embedding dimension as statistical units (rows) and the whole sequence lagged by subsequent delays as variables (columns).

The original data can be projected into a new set of coordinates US (principal component scores or eigenfunctions) such that no original information is lost. The new coordinates are linear combinations of the original variables: they 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 PCA has an optimal property which has made this method one of the most widespread modeling techniques in diverse science fields: the projection of the original data on the new component space spanned by a smaller number of dimensions ($A < N$, where N is the embedding dimension) is optimal in a least-squares sense.

As a matter of fact principal components have the foundamental property of explaining the system variability in an 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 noise in the error term. In other words, the most correlated portion of information (in terms of coordinated variation of any pattern along the chain) is retained by the first components, while all the singularities are discarded in the minor components. Therefore, by the use of a threshold as for the cumulative percentage of explained variance, the Principal Component Analysis allows for the reduction of a complex system of correlations in a less-dimensional one.

5. STRUCTURE-RELATED SINGULARITIES ALONG PROTEIN SEQUENCES

5.1 Premise

Real protein sequences show only weak departures from random strings [24]. A "random string" should be intended, in the information theoretic sense [25], as a series whose autocorrelation structure remains substantially invariant after random shuffling of the positions of its constituent elements. This fact means that the "code" linking a sequence to a particular structure is not emerging from simple periodicity in the aminoacid occurrence.

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 [26] . 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. Aminoacid sequences, on the other hand, are translated into biologically meaningful messages in the form of proteins by the physico-chemical environment (e.g., ionic strength, relative hydrophobicity, temperature, pressure).

A focus on the numerical series of physico-chemical properties of aminoacid residues has provided interesting results in the study of specific protein behavior [11] . 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.

On the basis of these considerations in the present study the following assumption has been adopted: given the 'perfect code' allowing the sequence-structure problem to be solved is still elusive, different physico-chemical codings of aminoacid residues have been considered as masking codes with respect to the 'folding message' conveyed by the primary structure. In this way the analysis of protein sequences can highlight both 'code dependent' (imposed by the masking code) and 'code independent' (linked to the real message) regularities in the juxtaposition of residues.

The approach, inspired to the QSAR models, is based on the selection of two elements:

1. A set of aminoacid physico-chemical descriptors;
2. A statistical technique (Recurrence Quantification Analysis, RQA [16]) adequate to infer biological activity features from the analysis of physico-chemical properties [9, 10].

5.2 Material and methods

A data set composed of 1141 protein sequences extracted from the Swiss-Prot repository by Menne et al. [27] was analyzed. The extraction was made in the aim of discriminating, on a pure sequence basis, the secreted and non-secreted eukariotic proteins. In this work the negative (non secreted) subset was selected, in order to avoid any selection bias due to a (mostly hydrophobic) N-terminal signal peptide in the secreted proteins. The data set is available at: ftp://ftp.ebi.ac.uk/pub/contrib/swissprot/testsets/signal.

Each protein sequence was transformed into 7 numerical profiles by means of the following physico-chemical properties of aminoacid residues (Figure 5.1):

1. Chothia hydrophobicity scale [28];
2. Kyte and Doolittle hydrophobicity scale [29];
3. Miyazawa-Jernigan hydrophobicity scale [30];
4. molecular weight;
5. polarity [31];
6. molar refractivity [32];
7. bulkiness [33].

As an extra coding, the standard one-letter symbolic code was also used.

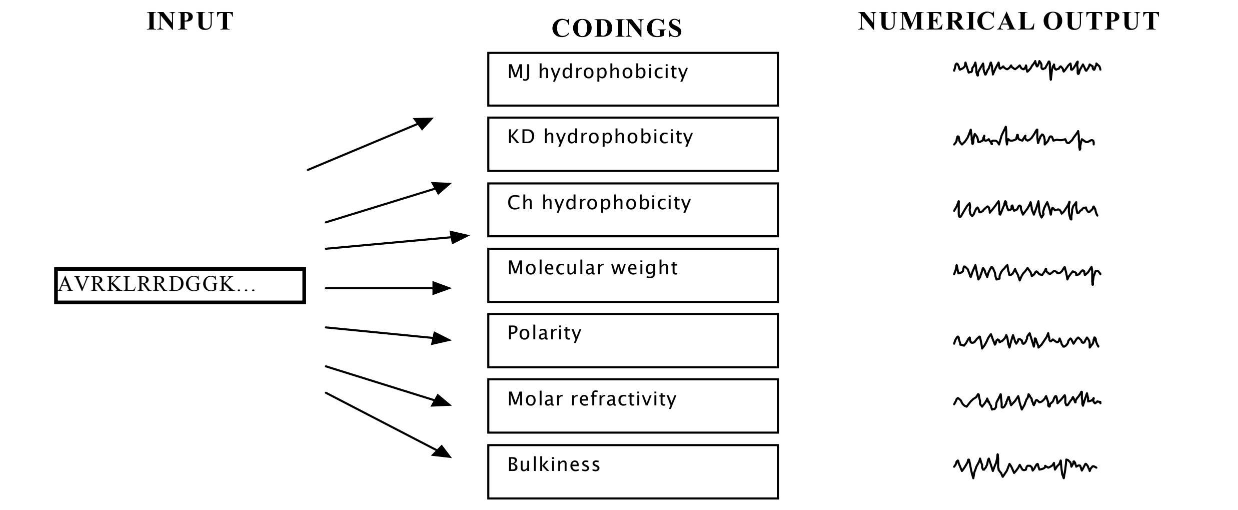


Figure 5.1 Overview of the data analysis strategy used in this work

The resulting numerical translations of protein sequences were submitted to RQA. In this phase, two RQA variables were considered: % Recurrence (REC) and % Determinism (DET). The setting parameters for the application of RQA were:

- radius equal to 20% of the mean Euclidean distance between all the patches of length equal to the embedding dimension; for the symbolic series the radius was obviously imposed equal to zero, due to the impossibility of measuring recurrent relations other than identity.
- embedding dimension (ED) equal to 4; in the case of the symbolic series the ED was set to 3, in order to obtain a reasonable percentage of recurrences.

The setting of these parameters was motivated by previous works aimed at defining the information content of protein primary structures [34], and was further substantiated by a scaling procedure.

5.3 Analytical details

5.3.1 Correlation analysis

The analysis was based upon correlating the different translations of protein sequences as well as the physico-chemical properties used for these translations. A metric to estimate the *a priori* similarities between different properties was defined. Since proteins are consituted by 20 different units (aminoacids), the Pearson correlation coefficients between pairs of codes computed on the space of the 20 natural aminoacids (Table 5.1) unambiguously 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 conveyed information.

a)

Aminoacid	Chothia hyd	KD hyd	MJ hyd	MW	Polarity	M. Refr	Bulkiness
Ala	0.38	1.8	5.33	89	8.1	4.34	11.5
Glu	0.18	-3.5	3.65	147	12.3	17.56	13.57
Met	0.4	1.9	8.95	149	5.7	21.64	16.25
Tyr	0.15	-1.3	5.89	181	6.2	31.53	18.03
Arg	0.01	-4.5	4.18	174	10.5	26.66	14.28
Gly	0.36	-0.4	4.48	75	9	0	3.4
Phe	0.5	2.8	9.03	165	5.2	29.4	19.8
Val	0.54	4.2	7.63	117	5.9	13.92	21.57
Asn	0.12	-3.5	3.71	132	11.6	12	12.82
His	0.17	-3.2	5.1	155	10.4	21.81	13.69
Pro	0.18	-1.6	3.87	115	8	10.93	17.43
Asp	0.15	-3.5	3.59	133	13	13.28	11.68
Ile	0.6	4.5	8.83	131	5.2	18.78	21.4
Ser	0.22	-0.8	4.09	105	9.2	6.35	9.47
Cys	0.5	2.5	7.93	121	5.5	35.77	13.46
Leu	0.45	3.8	8.47	131	4.9	19.06	21.4
Thr	0.23	-0.7	4.49	119	8.6	11.01	15.77
Gln	0.07	-3.5	3.87	146	10.5	17.26	14.45
Lys	0.03	-3.9	2.95	146	11.3	21.29	15.71
Trp	0.27	-0.9	7.66	204	5.4	42.53	21.67

b)

Code pairs	r
Chothia hyd / KD hyd	**0.964**
Chothia hyd / MJ hyd	**0.852**
Chothia hyd / MW	-0.294
Chothia hyd / Polarity	**-0.789**
Chothia hyd / M. Refractivity	0.064
Chothia hyd / Bulkiness	0.356
KD hyd / MJ hyd	**0.868**
KD hyd / MW	-0.266
KD hyd / Polarity	**-0.87**
KD hyd / M. Refractivity	0.075
KD hyd / Bulkiness	0.458
MJ hyd / MW	0.191
MJ hyd / Polarity	**-0.905**
MJ hyd / M. Refractivity	0.481
MJ hyd / Bulkiness	0.621
MW / Polarity	-0.089
MW / M. Refractivity	**0.838**
MW / Bulkiness	0.546
Polarity / M. Refractivity	-0.414
Polarity / Bulkiness	-0.606
M. Refractivity / Bulkiness	0.579

Table 5.1 a) Numerical values of the physico-chemical properties of aminoacids. b) Pearson correlation coefficients between properties calculated on the space of 20 natural aminoacids.

The above procedure was repeated on the space of the 1141 protein sequences. After the application of RQA on the various numerical translations of each protein, two RQA descriptors were analyzed: % of recurrences (REC) and % of determinism (DET). The mean value of each numerical series was also considered. For both RQA descriptors and Mean a matrix constituted by 1141 rows (protein sequences) and 8 columns (7 physico-chemical properties of aminoacids plus the standard one-letter symbolic coding) was obtained. The Pearson correlation coefficients between pairs of columns of these matrixes were computed (Table 5.2).

Code pairs	**r (Mean)**	**r (REC)**	**r (DET)**
Chothia hyd / KD hyd	0.70	0.70	0.46
Chothia hyd / MJ hyd	0.67	0.64	0.44
Chothia hyd / MW	0.12	0.79	0.41
Chothia hyd / Polarity	0.63	0.52	0.50
Chothia hyd / M. Refractivity	0.09	0.76	0.43
Chothia hyd / Bulkiness	0.37	0.55	0.36
KD hyd / MJ hyd	0.90	0.66	0.45
KD hyd / MW	0.23	0.63	0.41
KD hyd / Polarity	0.92	0.50	0.63
KD hyd / M. Refractivity	0.04	0.71	0.37
KD hyd / Bulkiness	0.53	0.68	0.50
MJ hyd / MW	0.17	0.65	0.34
MJ hyd / Polarity	0.92	0.40	0.44
MJ hyd / M. Refractivity	0.40	0.58	0.33
MJ hyd / Bulkiness	0.69	0.53	0.34
MW / Polarity	0.03	0.51	0.43
MW / M. Refractivity	0.82	0.86	0.43
MW / Bulkiness	0.56	0.59	0.42
Polarity / M. Refractivity	0.26	0.57	0.42
Polarity / Bulkiness	0.60	0.62	0.51
M. Refractivity / Bulkiness	0.53	0.70	0.37

Table 5.2 Pearson correlation coefficients between pairs of physico-chemical properties calculated on the RQA-based representations of proteins.

The correlation coefficients between properties calculated on the RQA-based representations of proteins (Table 5.2) can be compared with those calculated on the space of 20 natural aminoacids (table 5.1, panel b). If the RQA-based relations resemble the similarities between the respective codes, the analysis would simply return the structure of the translation rules (code-dependent features), with no information on the possible existence of general syntactic rules along the sequences.

It is worth noting that the correlations coefficients calculated on the space of 20 aminoacids simply reflect the nature of the physico-chemical codes, with the hydrophobicity and polarity scales closely related to each other and independent from the other codes (table 5.1, b). These purely physico-chemical relations are almost exactly maintained when the average value of each coding is considered (Mean), while for REC and DET descriptors a new information seems to emerge. In fact, the Pearson correlation coefficients between the r column reported in table 5.1 (b) and each column reported in table 5.2 were computed.

For the Mean descriptor the correlation coefficient scores 0.95, for DET it scores only 0.46, while for REC it drops to –0.20. This means that, while the linear descriptor of protein sequences (MEAN) still reflects the original relations between codes scored at the aminoacid level (r = 0.95), the non linear descriptions (REC and DET) are no more related to the original physico-chemical meaning of the codes. Through the agency of a non linear tool (RQA), the features of the protein systems do not simply derive from those of the physico-chemical codes, but may reflect some new, higher-level property.

The above result might seem to be a trivial consequence of the fact that two segments are scored as recurrent because they are made by exactly the same residues: if this is the case, each code will invariably give the same result. This, however, is not the case. In fact, Table 5.3 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 or not protein sequences are made of repeated patches, due to evolutionary events [35] influencing structural and functional features. 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 physico-chemical profiles and relaxing the strict identity in favour of the weaker similarity requirement, the ability to detect recurrent "words" is enhanced.

3.3.2 Miyazawa-Jernigan hydrophobicity scale

Table 5.3 reports the average value of recurrence (REC) and determinism (DET) in the 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 even underestimates the real difference since the symbolic coding is analyzed with a shorter segment length (Embedding Dimension = 3) as compared to other codings (ED = 4).

a) REC

CODE	Mean	Std. Dev.	min.	max.
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
bulk	0.63	0.52	0.21	9.88
Sy	0.08	0.39	0	8.91

b) DET

CODE	Mean	Std. Dev.	min.	max.
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
bulk	18.19	10.14	0	89.81
Sy	17.07	21.47	0	100.00

Table 5.3 Descriptive statistics of REC (panel a) and DET (panel b) variables of protein sequences for different codings. The last rows in both panels report the symbolic one-letter code.

MJ hydrophobicity scores an average recurrence and determinism much higher than the other physico-chemical 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 aminoacid residues in a large ensemble of 3-D protein structures: it was designed as a kind of statistical potential for aminoacid interactions, and only *a posteriorii* was it recognized as an hydrophobicity scale [36, 37]. Since MJ index has been specifically tailored to protein structures, this may explain its performance in detecting aminoacid patterning along polypeptide chains.

In a recent work by Leary et al. [38] , 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 aminoacid index profile of a test sequence to a profile centroid for each class. A mathematical optimization procedure was then applied to determine an aminoacid index of maximal structural discriminatory power. Figure 5.2 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.

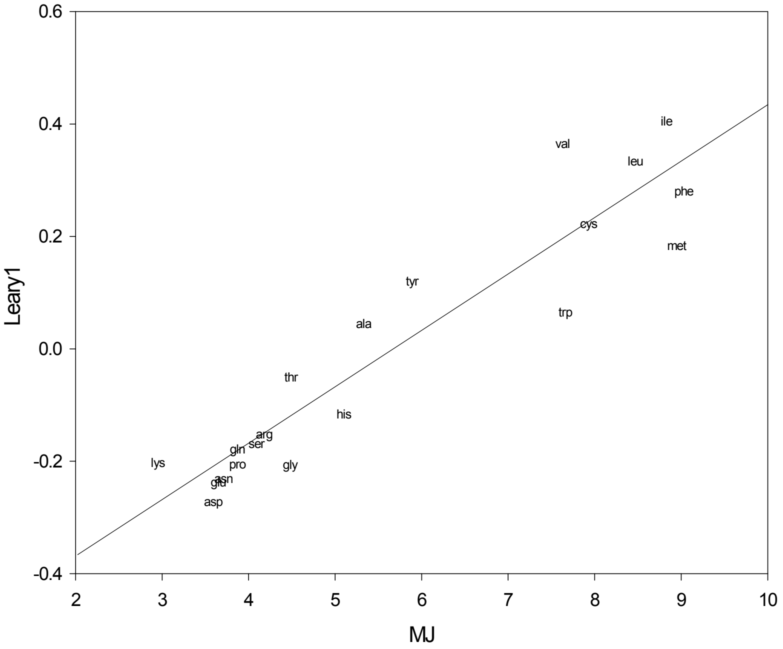


Figure 5.2 Codings of the natural aminoacids by the MJ hydrophobicity scale and the Structural Discriminatory Index by Leary et al. (2004).
The figure reports the relation between the MJ hydrophobicity scale and an index having the maximal structural discriminatory power in a set of 3876 protein sequences. 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.

5.3.3 Percentage of determinism

In the aim to identify which type of rules may be present in the juxtaposition of aminoacids along protein primary structures, the RQA-based representations of proteins (both REC and DET matrixes) were submitted to a Principal Component Analysis (PCA). The PCA produces as main mode (first Principal Component, PC1) a consensus axis collecting all the codings. PC1, both for REC and DET, was by far the most important source of information explaining, respectively, 70% and 50% of the total variability (Table 5.4). It is worth noting that the symbolic coding is highly correlated with the first component, as a further indication of the role of code-independent autocorrelation measure played by PC1 for the RQA descriptors.

Thus, through the application of the PCA, the presence of syntactic rules in the aminoacid distribution along protein sequences, pointed out by the correlation analysis, is confirmed. The percentage of variance explained the PC1 represents the relevance of the code-independent autocorrelation structure of the system.

a) REC

Code	PC1	PC2	PC3
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
Bulk	**0.79**	0.39	0.21
Sy	**0.92**	-0.14	-0.25
% variance	69.9	8.9	6.9

b) DET

Code	PC1	PC2	PC3
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
% variance	50	9.4	8.9

Table 5.4 Factor loadings of the aminoacid properties on the first three Principal Components from RQA filtered proteins. 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 having as variables the REC (panel a) and DET (Panel b) descriptors of each protein calculated from the different physico-chemical profiles. 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, the above analyses has been repeated on the shuffled texts, looking for possible invariants after a random scrambling of aminoacid 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 determinism no correlation was detected by DET(Native) and DET (Shuffled) (r = -0.14).

This result suggests that:

i) the percentage of recurrences in each protein sequence is strongly dependent on the aminoacid composition;

ii) the percentage of determinism only depends on the order of aminoacids along the chain and it is not significantly affected by compositional effects.

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 relative position and not the number of recurrent patches.

Since any peculiar syntactic rule of aminoacid 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.

5.3.4 Scaling procedure

The coding with the highest sensitivity in identifying syntactic rules (MJ hydrophobicity) was eventually submitted to a scaling procedure in the aim to check for the existence of a privileged scale at which the effect is maximized.

In general, a too small embedding dimension leads to false recurrences, while an overembedding should theoretically not distort the reconstructed phase trajectory. Concerning the radius, smaller values would lead to a better distinction of small variations. However, the recurrence point density decreases in the same way and thus the statistics of continuous structures in the RP soon becomes insufficient. On the other hand, larger values cause a higher recurrence point density, but a lower sensitivity to small variations [18].

Figure 5.3 reports the embedding dimension scaling of average determinism over the 1141 proteins set for *MJ* coding at very low radius values (5% and 10% of the mean distance). 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 aminoacid patterning. Since a minimal length of 3 consecutive recurrences was used 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.

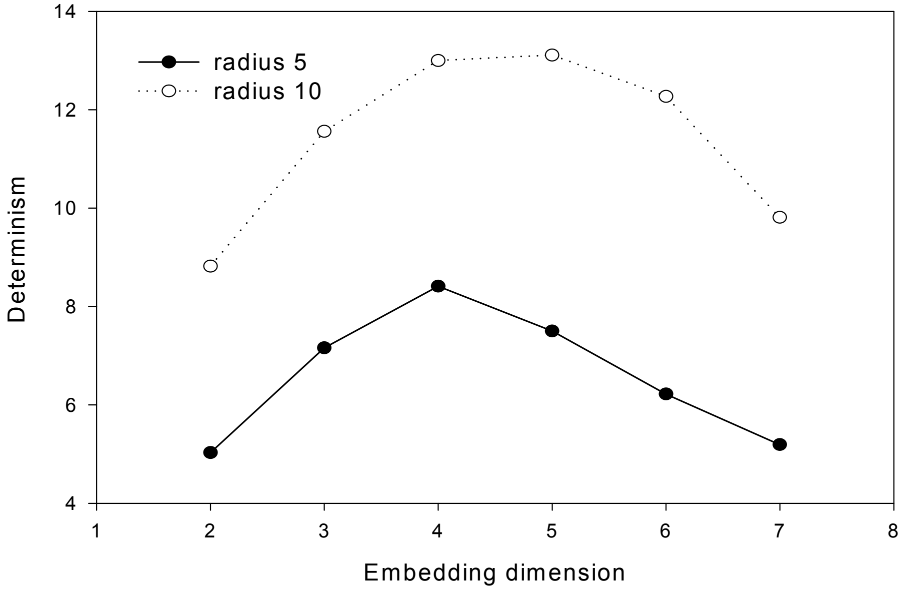


Figure 5.3 Scaling of Determinism with Embedding Dimension as a function of Radius in 1141 protein sequences. At low values of radius a maximum of determinism at Embedding Dimension equal to 4 is evident (see the text for further explanations).

6. LOOKING FOR SEQUENCE/FUNCTION RELATIONSHIPS

6.1 Protein distribution in a Principal Component space

In the aim to find the consequences of the aminoacid patterning in terms of structural or functional features, proteins endowed with exceedingly high values for the first determinism Principal Component (PC1DET) were inspected. Protein distribution along this Component is quite asymmetric with very small but long tail (PC1DET > 2) made of extremely deterministic sequences (Figure 6.1).

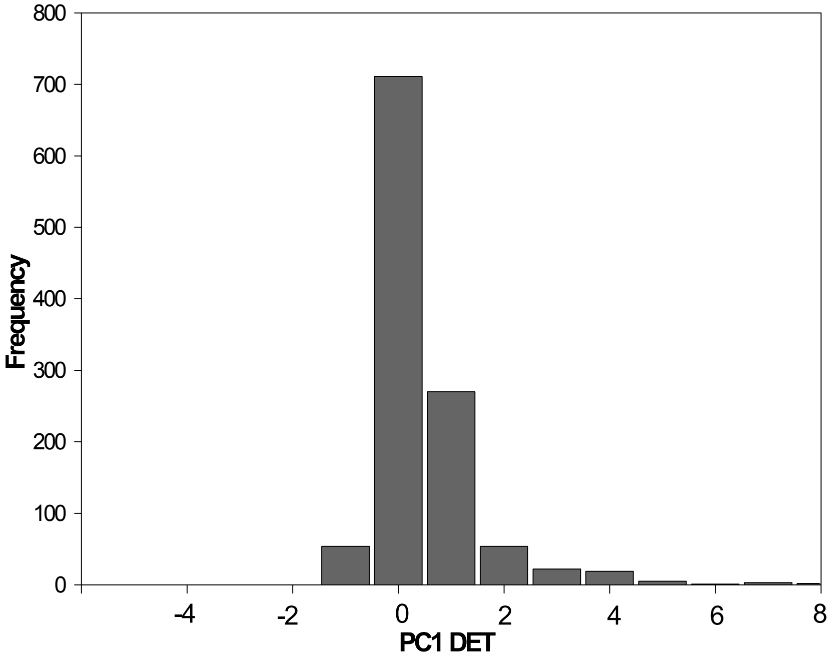


Figure 6.1 Protein distribution of the 1141 protein sequences along the first determinism component

Table 6.1 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 included 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.

All the extremely deterministic proteins share the property of being involved in protein-protein interactions, both for regulatory and structural purposes (e.g., protamines and trascription factors) as well as of forming polymeric assemblies (cornifin, myosin, keratin).

No enzyme or enzyme subunit is present in Table 6.1, with the only exception represented by the protein Q34522, i.e. NADH–Ubiquinone oxydoreductase, chain 3 (reported in bold). 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. This is a remarkable asymmetry given enzymes are by far the most represented functional class in the 1141 set.

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 6.1 Elements of the “high determinism tail” in the distribution along the first determinism component (PC1DET) of the protein data set.

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 [39, 40] suggested a check for the presence of an excess of natively unfolded zones in the deterministic tail of the data set.

The ten most deterministic sequences scored a percentage of computationally estimated disorder (computed by the PONDR predictor [41]) of 66.46% against the 27.27% of the ten proteins located at the low determinism tail (significance of $p < 0.001$). Calculation of a foldability coefficient [39] for the highly deterministic sequences listed in Table 6.1 indicates that more than 75% may be classified as natively unfolded.

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 analogy with "rhymes" within otherwise prosodically unstructured texts can be attempted. Such rhymes make two different texts (or two different portions of the same text) mutually recognizable and interacting. Their different and possibly function-related patterning in protein sequences is shown in Figure 6.2, 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 6.2, panel a), while in the case of a "strongly interacting" protein (transcription factor, Figure 6.2, panel b) the rhymes are clearly defined by block structures in the RPs.

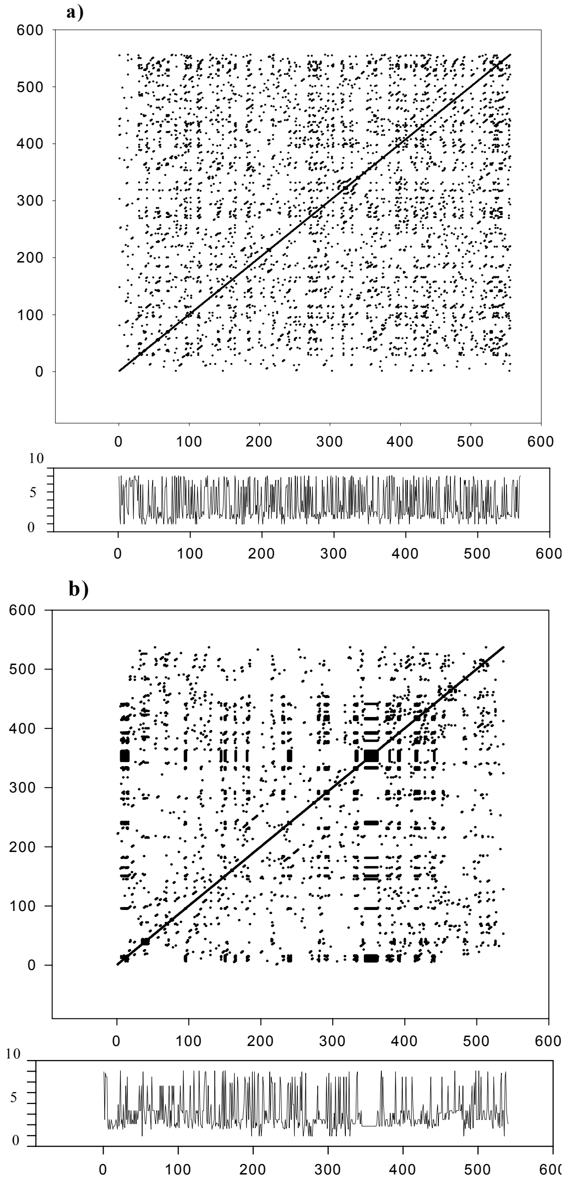


Figure 6.2. Recurrence Plots of an enzymatic molecule in comparison with an interacting protein (transcription factor). The figure reports in panel a) the recurrence plot and hydrophobicity series (MJ) of an enzymatic protein (Polypeptide N-acetylgalactosaminyl transferase, Swiss-prot code Q07537), and 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).

6.2 Protein aggregation: analysis and results

The application of RQA on protein primary structures allows to single out some periodicity in the aminoacid distribution, correlated with the propensity to protein-protein interaction. On the basis of the previous results it has been decided to concentrate on the problem of protein aggregation.

Protein dataset was analyzed with regard to the hydrophobicity distribution and to the net charge, in the aim to check some peculiar features linking protein primary structures and biological features. For such a study only the most efficient hydrophobicity coding (MJ scale) was considered. The net charge for each sequence was simply calculated by adding the electric charges (0, +1, -1) of the single residues (see table 1.1). The matrix having as variables the RQA descriptors calculated on the MJ hydrophobicity profiles, net charge and charge normalized by length was submitted to a Principal Component Analysis (PCA).

The PCA produces a solution with two significant components. PC1 and PC2 together explain about the 59% of the variability of the system (see table 6.2). PC1 is highly correlated with all the RQA dynamic (order dependent) descriptors, with the only exception represented by the TREND, while PC2 is correlated with Length, TREND, Charge, and especially with the Charge normalized by Length (r = 0.801).

	PC1	PC2	PC3	PC4	PC5
LENGTH	0.397	**-0.606**	-0.443	-0.069	0.351
MEAN	-0.359	-0.423	0.282	**0.742**	0.057
REC	**0.837**	0.234	0.257	0.003	0.202
DET	**0.851**	0.06	0.134	0.112	0.025
MAXL	**0.707**	-0.409	-0.341	0.119	0.2
ENT	**0.801**	-0.117	0.056	0.282	-0.001
TREND	0.172	**0.648**	0.491	0.018	0.357
LAMIN	**0.842**	0.227	0.053	-0.197	-0.192
TRAPT	**0.641**	0.039	-0.073	0.162	-0.573
CHARGE	-0.17	**0.74**	-0.441	0.347	-0.072
Ch/Length	0.028	**0.801**	-0.439	0.145	0.198
% variance	36.65	22.23	10.15	7.9	6.8

Table 6.2 Factor loadings of the Principal Component Analysis applied on the whole dataset

The following step was the ordering of the dataset by means of PC1 and PC2 values. Proteins with extremely high values of PC1 essentially coincide with the extremely deterministic proteins reported in table 6.1 (i.e. proteins involved in protein-protein interaction), given that DET is the most correlated variable with PC1 (r = 0.851, see table 6.2). The analysis of the higher extreme of PC2 highlight an exclusive presence of histones and RNA-binding proteins (with only one exception), both involved in interactions with nucleic acids (see table 6.3).

Code	PC2	Class
P42132	17.932	I
P42129	10.802	I
P13275	8.880	I
P19757	6.851	I
P07978	6.795	I
P17502	6.719	I
P40631	5.747	I
P11020	5.528	I
Q05831	5.464	I
P06144	4.972	I
Q09821	4.934	I
P22974	4.712	I
P02259	4.596	I
P10922	4.570	I
P43278	4.370	I
P06894	4.339	I
P26377	4.304	I
P15870	4.269	I
P07305	4.241	I
P02254	4.156	I
P29258	4.121	I
P14798	4.014	R
P79781	3.898	R
P06895	3.839	I
P06146	3.766	I
P40262	3.737	I
P55947	3.714	A
P06145	3.664	I
P02285	3.645	I
P17268	3.605	I
P15796	3.569	I
Q02877	3.548	R

Table 6.3 Proteins with extremely high values of PC2. I = histones, R = RNA-binding proteins, A = other proteins.

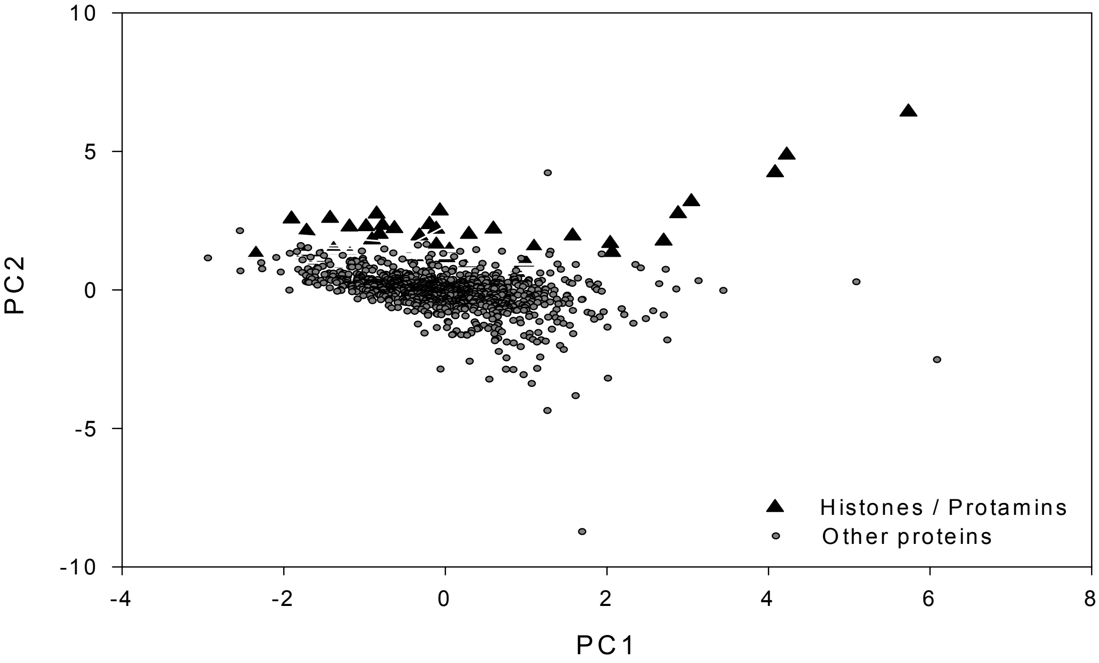
The lower extremes of both PC1 and PC2 components don’t highlight any particular tendency concerning the presence of specific classes of proteins.

Since nucleic acids are negatively charged owing to the presence of phosphate groups, proteins interacting with them must have a strong positive charge in order to form stable aggregates with DNA or RNA. Both histones and RNA-binding proteins are quite short and charged molecules. Consequently, they present highly positive values of Charge/Length, i. e. the most correlated descriptor with PC2. This fact leads to their clear-cut separation on the PC2 axis.

The association of TREND with PC2 is due to the particular characteristic of this RQA descriptor, which is strongly dependent from the length of the sequence: short sequences tend to have higher values of TREND compared to the long ones. The presence in the dataset of histones and RNA-binding proteins leads to the association of Length (and, indirectly, of TREND) with PC2.

Plotting PC1 vs. PC2 histones are outliers on the PC2 axis (see figure 6.3). An analogous behaviour on PC2 axis, even if in minor degree, is shown by the RNA-binding proteins. These classes press a lot the dataset on PC2 axis, so making difficult to single out other significant clusters. In order to avoid this problem, PCA was repeated on the dataset lacking in histones and RNA-binding proteins. The factor loadings of the PCA are reported in Table 6.4. These loadings are roughly in agreement with those calculated on the whole dataset.

Figure 6.3



For the above discussed reasons, the removal of histones and RNA-binding proteins, causes a decreasing in the correlation coefficients of Length and TREND with PC2. Hence in this case PC2 is essentially driven by the electric charge. The meaning of the two principal components can be summarized in this way: PC1 represents the regularities in the hydrophobicity distribution along the sequence, while PC2 principally represents the electric charge.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
LENGTH	0.469	**-0.594**	0.371	-0.228	0.219
MEAN	-0.458	-0.223	0.303	**0.762**	0.138
REC	**0.832**	0.247	-0.187	0.087	0.187
DET	**0.831**	0.21	-0.049	0.186	0.191
MAXL	**0.741**	-0.314	0.359	-0.046	0.156
ENT	**0.796**	0.051	0.105	0.262	0.111
TREND	-0.136	**0.594**	-0.504	0.124	0.304
LAMIN	**0.83**	0.264	-0.193	-0.101	-0.158
TRAPT	**0.604**	0.044	0.072	0.244	**-0.669**
CHARGE	-0.219	**0.712**	0.545	-0.01	-0.105
Ch/Length	-0.033	**0.65**	**0.632**	-0.16	0.129
% variance	37.43	17.85	12.72	7.78	6.97

Table 6.4 Factor loadings of the Principal Component Analysis applied on the dataset lacking in histones and RNA-binding proteins

6.3 Protein classes 'reference table'

The plot PC1/PC2 of the dataset lacking in histones and RNA-binding proteins was used as a reference picture on which small samples belonging to particular protein classes were projected. The objective is to check the possible presence of remarkable clusters of proteins with common function or fold, in order to build a kind of 'reference table' on which the biological functions derive from purely sequential properties.

Relative to this aim a set of small classes, each one constituted by 10 proteins endowed with the same Swiss-Prot keyword, were selected. The single classes were inserted, one at a time, in the reference picture previously described, checking if the 10 proteins grouped together in a particular zone of the PC1/PC2 plane. The classes which not satisfied the above requirement were eliminated from the analysis, while for the other classes the mean values of PC1 and PC2 into the class were calculated. These mean values represented a kind of cartesian coordinates by means of which each class was projected of the PC1/PC2 plane.

A summarizing table reporting the projections of the classes on the space of the components of the 'hydrophobicity patterning' (PC1) vs. 'electric charge' (PC2) was obtained (see figure 6.4).

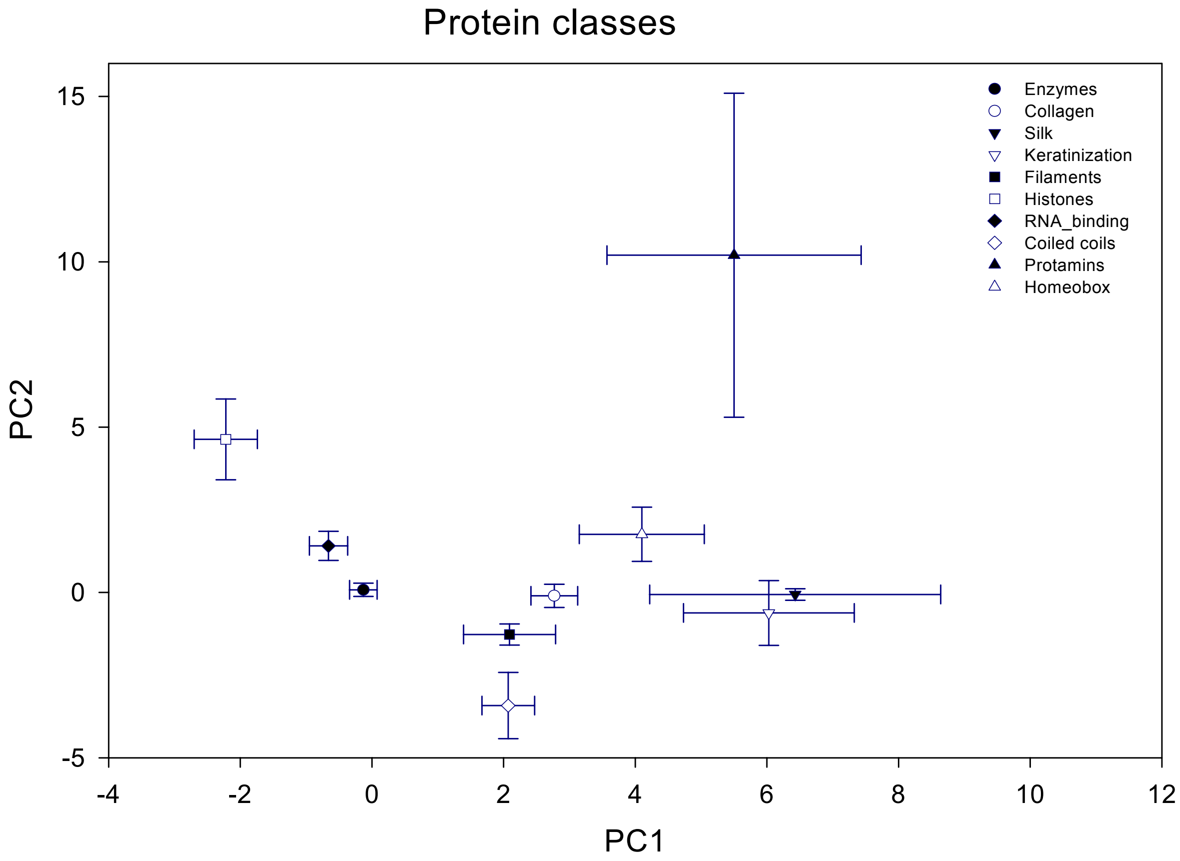


Figure 6.4. Reference table of protein classes on PC1 / PC2 plane. The coordinates are represented by the mean values of PC1 and PC2 into each class. The standard errors are also reported.

The great majority of the examined classes shows positive values on the PC1 axis. On the contrary enzymes, the most numerous class of the dataset, have a mean value of PC1 slightly lower than zero. The positive extreme of PC1 is characterized by protein involved in protein-protein interaction. In this case the aggregation feature correlated to the highly repetitive hydrophobicity pattern is clearly evident.

As a matter of fact structural proteins are almost completely located on the positive side of the PC1 axis. At the positive extreme of the axis are located silk fibroin and proteins involved in the process of keratinization, namely long extracellular polymers more similar to syntetic polymers than to soluble proteins, as well as protamins, which are characterized by high mean values for both PC1 and PC2. A peculiar position on the table is occupied by the ‘Filaments’ class, endowed with positive values of PC1 and very negative values of PC2. The ‘Filaments’ class consists of intermediate filaments and neurofilaments, which aggregate forming long polymers. This feature explains their highly regular distribution of hydrophobicity (positive values of PC1).

Nevertheless their mechanism of polymerization is not only driven by hydrophobic interactions, but also by electrostatic interactions. For example neurofilaments remain separate by means of the exposition of identical negative charges (repulsive) along the chains of adjacent filaments, in the aim to keep open the width of the channel formed by themselves. In order to carry out this function they need to have a strong negative net charge, which produces their highly negative values on PC2 axis.

6.4 Histones

The histone class was analyzed in more detail. Histones can be divided in 5 subclasses: H1, H2A, H2B, H3 and H4. In eukariotic cells, histones are the principal protein constituents of the nucleosome. The nucleosome is the basic subunit of chromatin lacking in non histonic proteins. The nucleosome is consituted by 3 components: the nucleosome core, the histone h1 (in higher eukariotes) and linker DNA (see figure 6.5).

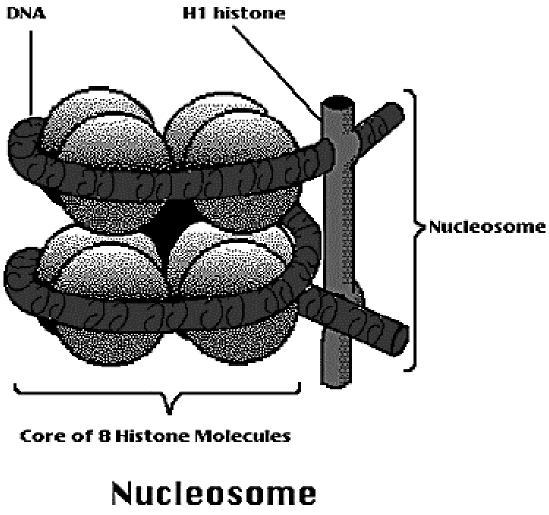


Figure 6.5 Nucleosome structure

The nucleosome core contains an octamer of 2 each of the core histones (H2A, H2B, H3 and H4) and 146 bp of DNA wrapped 1.75 turns. Core histones dimerize through their histone fold motifs generating H3/H4 dimers and H2A H2B dimers.

The first order of chromatin compaction consists of 146 bp of DNA wound in two superhelical turns around a core histone octamer. A single histone H1 polypeptide interacts with an additional 20 bp of DNA. The linker histone H1 serves to stabilize a higher order chromatin fiber that is fundamental to the structural organization of chromosomes. The linker histone binds to each nucleosome, and by self affinity, links these nucleosomes together. Modulation of linker histone binding is thought to be an additionally important element in the ordering of chromatin structure accompaning activation and inactivation of gene transcription.

Hence the H1 histone plays a different role compared to the other histone subunits. While other subunits are especially involved in protein-protein interactions, the H1 subunit is especially a DNA-binding protein. On the basis of this consideration the histone class of the dataset has been divided in two subclasses, H1 subunit and other subunits, in order to check a possible different location on the reference table (see Figure 6.6).

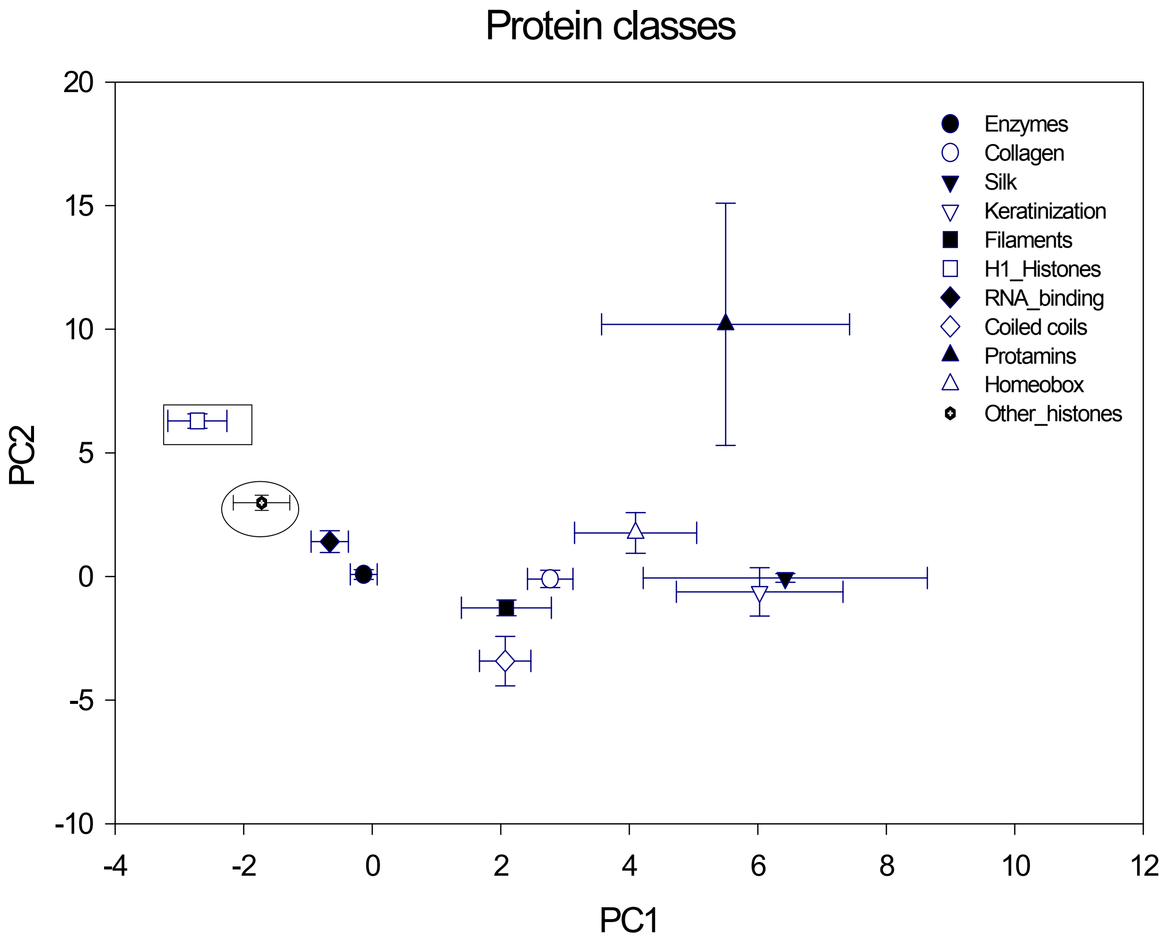


Figure 6.6 Protein table with the histone class divided in 2 subclasses: H1 subunit (reported into the rectangle) and other subunits (reported into the circle).

As expected, H1 histones show highly positive values on the PC2 axis, but don't show any periodic hydrophobicity distribution along the sequence (PC1 < 0). The other histone subunits (H2A, H2B, H3 and H4) show less positive values of PC2 but a slightly more periodic hydrophobicity distribution.

6.5 Discussion

The classification by means of the 2 concepts of 'regularities in the hydrophobicity distribution' along protein sequences and electric net charge allows to single out some gross scale functional features of proteins. This features can be summarized in this way:

1) Proteins involved in the construction of supermolecular complexes (usually structural proteins) show strong regularities in the hydrophobicity distribution along the primary structure.
2) Globular proteins (usually enzymes) present a disordered (quasi random) hydrophobicity distribution along the primary structure.
3) Proteins interacting with charged molecules (DNA, RNA, metals, etc.) have an high electric net charge. The most relevant examples of this situation are represented by histones and RNA binding proteins.

7. GENERAL CONCLUSIONS

The simultaneous use of RQA and of different physico-chemical properties of amonoacids allows to achieve a general picture of the aminoacid distribution in protein primary structures. The analysis goes beyond the simple detection of repeats and leeds to the location of almost a grammar of the aminoacid patterning, based upon strong order-dependent regularities.

In particular the existence of some general 'syntactic rules' shaping the juxtaposition of aminoacid residues was demonstrated: the above rules are destroyed by the random shuffling of protein sequences (sequence dependence) but are generally maintained by changing the physico-chemical coding of residues (code independence). This behavior points to the existence of an emergent level of protein architecture different from the physical chemistry of the aminoacid residues.

The seven properties considered in this thesis constitute a representative sample of all the aminoacid physico-chemical properties. The high correlation among the considered codings points to a general superimposition of the RQA based description of protein sequences whatever the code. This implies that the use of some other property in place of those considered in this context will essentially provide the same information.

Protein constraints emerging in this work can be summarized by the following statements:

i) Protein sequences (at least the eukariotic ones) are constituted by 'imperfect repetitions' of modules of 4-6 residues length. These modules have probably changed their composition due to point mutational events, most often introducing residues similar to the original ones in terms of physico-chemical properties. Thus the above modules are very difficult to recognize by means of the pure symbolic coding of aminoacids, while they appear much more invariant if considered in terms of the relative hydrophobicity of their constituents residues.

ii) Classifying protein primary structures by means of their relative determinism (repetition of self-similar modules in specific locations along the sequence) allows the identification of proteins undergoing a lot of protein-protein interactions. The supplement of the hydrophobicity based determinism with global net charge improves the recognition of some protein functional classes.

The link between the general syntactic invariants and the propensity to interact is consistent with a recent work by the Janet Thornton group [42] in which, analyzing a huge data base of point mutations for a lot of different protein systems, the authors state that the most critical portions of the proteins as for the biological consequences of the mutations are the 'quaternary interfaces' i.e. the portions of the sequences where protein systems interact. The crucial value of interactions is highlighted by the so called 'misfolding diseases' provoked by the precipitation of protein complexes due to the anomalous self-interaction of different protein systems.

Some hints about the structural consequences of the discovered regularities were also derived. The analysis of extremely deterministic sequences points to statistically singular, interaction zones crucial for mutual recognition events. It is worth stressing that proteins undergoing specific interactions with other proteins of the same or different kind are not endowed with qualitatively different features. This allows for a hypothesis of deterministic spots playing the role of both folding "initiators" and aggregation hot points, depending on whether the boundary conditions promote intramolecular or intermolecular interactions.

Such a conjecture is reinforced by the fact that the above demonstrated estimated length of 6 for the typical "deterministic patch" matches the average 6.12 length calculated in the approximately 800 folding “nucleation centres” collected by the Casadio group [43] . Moreover, a relation between the deterministic peaks and aggregation properties of different proteins ranging from prion [44] to P53 [45] has been also demonstrated. Such a finding included a correspondence between short deterministic patches of hydrophobicity distribution along the sequence, with 3-D "unstructured" portions of acylphosphatase (AcP) [11] .

Concerning the question of protein aggregation, it is possible to reach a more sophisticated level of description compared to the analysis of the simple electric net charge in addition to the single RQA descriptors calculated on the hydrophobicity profiles. In particular the two basic elements involved in the aggregation process (i.e hydrophobic and electrostatic interactions) can be simultaneously considered by combining the net charge with some selected RQA descriptors, in order to obtain empirical formulae which can be tested about their ability in predicting the aggregation propensity of protein molecules.

The link between the propensity of a given protein to form partially fold intermediates and the propensity to aggregate is well known. In a recent paper [12] , an attempt to unify "charge/hydrophobicity" and "partially folded intermediate" models of protein aggregation has been presented. In this work an empirical formula derived for the prediction of aggregation propensity of AcP has been successfully tested in discriminating between two protein sets: proteins that are able to adopt equilibrium partially folded conformations and proteins which have been shown to unfold without the formation of any equilibrium intermediate.

The demonstration of the ability of the same empirical formula to model both the aggregation propensity of a specific system and the existence of partially folded intermediates represents some noteworthy evidence of the overlapping of the two phenomena: aggregation and the formation of partially folded intermediates.

It's worth to note that in the context of the data set used for this thesis, any empirical formula combining charge and hydrophobicity doesn't provide any substantial improvement in the ability of discriminating between different functional classes, compared to the simple net charge.

This is due to the fact that the 1141 protein data set has been randomly extracted and for this reason it is extremely heterogeneous. The calculation of the electric net charge represents a more coarse-grain level of representation compared to the RQA description. Consequently the differences between protein sequences in terms of electric charge compress the differences in terms of RQA descriptors.

From a practical point of view, the demonstration of a non-trivial and non purely stochastic autocorrelation structure of aminoacid distribution along protein sequences opens the way to an alternative method to sequence alignment for comparing proteins. The long term objective is the possibility to classify newly discovered sequences only on the basis of their primary structure, stemming from the much lower DET of a typical enzyme as compared to an aggregation-prone chain.

REFERENCES

1) Dobson C.M. and Karplus M. (1999) Current Opinion in Structural Biology 9, 92.
2) Giuliani A., Benigni R., Zbilut J.P., Webber C.L., Sirabella P. and Colosimo A. (2002) Chem. Rev. 102, 1471.
3) Taylor W.R., May A.C., Brown N.P. and Aszodi A. (2001) Rep. Prog. Phys. 64, 517.
4) Adkins J.N. and Lumb K.j. (2002) Proteins: Struct. Funct. Genet. 1, 1.
5) Dobson C.M. (2003) Nat. Rev. Drug Discov. 2, 154.
6) Chiti F., Taddei N., Baroni F., Capanni C., Stefani M., Ramponi G. and Dobson C.M. (2002) Nat. Str. Biol. 9, 137.
7) Bryngleson J. D., Onuchic J.N., Socci N.D. and Wolynes P.G. (1995) Proteins 21, 167.
8) Giuliani A., Colafranceschi M., Webber Jr. C.L. and Zbilut J.P. (2001) Physica A. J. 301, 567.
9) Giuliani A., Sirabella P., Benigni R. and Colosimo A. (2000) Protein Eng. 13, 671.
10) Webber C.L., Giuliani A., Zbilut J.P. and Colosimo A. (2001) Proteins: Struct. Funct. Genet. 44, 292.
11) Zbilut J.P., Colosimo A, Conti F., Colafranceschi M., Manetti C., Valerio M.C. and Giuliani A. (2003) Biophys.J. 85, 3544.
12) Zbilut J.P., Giuliani A., Colosimo A., Mitchell J.C., Colafranceschi M., Marwan N., Webber Jr C.L. and Uversky V.N. Accepted for publication in Journal of Proteome Research.
13) Schreiber T. (1999) Phys. Rep. 308, 1.
14) Broomhead D.S. and King G.P. (1986) Physica D 20, 217.
15) Eckmann J.P., Kamphorst S.O. and Ruelle D. (1987) Europhys. Lett. 4, 324.
16) Webber C.L. and Zbilut J.P. (1994) J. Appl. Physiol. 76, 965.
17) Giuliani A., Piccirillo G., Marigliano V. and Colosimo A. (1998) Am. J. Physiol. 275, H1455.
18) Manetti C., Ceruso M.A., Giuliani A., Webber C.L. and Zbilut J.P. (1999) J. Phys. Rev. E 59, 992.
19) Rustici M., Caravati C., Patretto E., Branca M. and Marchettini N. (1999) J. Phys. Chem. A. 103, 6564.
20) Feller W. (1968) An introduction to Probability Theory and Its Applications; Wiley: New York, Vol. 1.
21) Rao C.R. and Suryawanshi S. (1996) Proc. Natl. Acad. Sci. USA 93, 12132.
22) Marwan N., Wessel N., Meyerfeldt U., Schirdewan A. and Kurths J. (2002) Phys. Rev. E 66 (2), 026702.
23) Benigni R. and Giuliani A. (1994) Am. J. Physiol. 266, R1697.
24) Senno C.F., Micheletti A., Maritan A. and Bonaver J.R. (1998) Phys. Rev. Lett. 80, 2237.
25) Shannon C.E. (1948) Bell System Technical Journal.
26) Popov O., Segal D.M. and Trifonov E.N. (1996) Biosystems 38, 65.
27) Menne K.M.L., Hermjakob H. and Apweiler R. (2000) Bioinformatics 16, 741.
28) Chothia C. (1976) J. Mol. Biol. 105, 1.
29) Kyte J. and Doolitle R.F. (1982) J. Mol. Biol. 157, 105.
30) Miyazawa S. and Jernigan R.L. (1985) Macromolecules 18, 534.
31) Grantham R. (1974) Science 185: 862.
32) Jones D. (1975) J. Theor. Biol. 50, 167.
33) Zimmermann J.M., Eliezer N. and Simha R. (1968) J. Theor. Biol. 21, 170.
34) Strait B.J. and Dewey T.G. (1996) Biophys. J. 71, 741.
35) Kajava A.V. (2001) J. Struct. Biol. 134, 132.
36) Wang J. and Wang W. (2002) Phys. Rev. E 65, 41911-1.
37) Tang H. and Wingreen N.S. (1997) Phys. Rev. Lett. 79, 765.
38) Leary R. H., Rosen J. B. and Jambeck P. (2004) Biophys. J. 86, 411.
39) Uversky V.N. (2002) Protein Science 11, 739.
40) Romero P., Obradovic Z., Li X., Garner E., Brown C.J. and Dunker K. (2001) Prot. Struct. Funct. Genet. 42, 38.
41) Dunker K., Brown C.J., Lawson D., Iakoucheva L.M. and Obradovic Z. (2002) Biochemistry 41, 6573.
42) Steward R.E., MacArthur M.W., Laskowski R.A. and Thornton J.M. (2003) Trends in Genetics 19, 505.
43) Compiani M., Fariselli P., Martelli P.L. and Casadio R. (1998) Proc. Natl. Acad. Sci. USA 95, 9290.
44) Zbilut J.P., Webber Jr. C.L., Colosimo A. and Giuliani A. (2000) Protein Eng. 13, 99.
45) Porrello A., Soddu S., Zbilut J.P., Crescenzi M. and Giuliani A. (2004) Proteins: Struct. Funct. and Bioinf. 55, 743