www.elsevier.com/locate/yjtbi

Novel 3D Bio-Macromolecular Bilinear Descriptors for Protein Science: *Predicting Protein Structural Classes*

Yovani Marrero-Ponce, Ernesto Contreras-Torres, César R. García-Jacas, Stephen J. Barigye, Néstor Cubillán, Ysaías J. Alvarado

PII: S0022-5193(15)00141-1
DOI: http://dx.doi.org/10.1016/j.jtbi.2015.03.026
Reference: YJTBI8128

To appear in: *Journal of Theoretical Biology*

Received date: 17 September 2014
Revised date: 23 February 2015
Accepted date: 20 March 2015

Cite this article as: Yovani Marrero-Ponce, Ernesto Contreras-Torres, César R. García-Jacas, Stephen J. Barigye, Néstor Cubillán, Ysaías J. Alvarado, Novel 3D Bio-Macromolecular Bilinear Descriptors for Protein Science: *Predicting Protein Structural Classes*, *Journal of Theoretical Biology*, http://dx.doi.org/10.1016/j.jtbi.2015.03.026

Novel 3D Bio-Macromolecular Bilinear Descriptors for Protein Science: *Predicting Protein Structural Classes*

Yovani Marrero-Ponce (✉),[1-3] Ernesto Contreras-Torres,[4] César R. García-Jacas,[4] Stephen J. Barigye,[5] Néstor Cubillán[6] and Ysaías J. Alvarado[7]

[1]*Facultad de Química Farmacéutica, Universidad de Cartagena, Cartagena de Indias, Bolívar, Colombia.*
[2]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.*
[3]*Grupo de Investigación en Estudios Químicos y Biológicos, Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar, Cartagena de Indias, Bolívar, Colombia.*
[4]*Grupo de Investigación de Bioinformática, Centro de Estudio de Matemática Computacional (CEMC), Universidad de las Ciencias Informáticas (UCI), La Habana, Cuba.*
[5]*Departamento de Química, Universidade Federal de Lavras, CP 3037, 37200-000, Lavras, MG, Brazil.*
[6]*Laboratorio de Electrónica Molecular, Universidad del Zulia, Facultad Experimental de Ciencias, Departamento de Química. Maracaibo, República Bolivariana de Venezuela.*
[7]*Laboratorio de Caracterización Molecular y Biomolecular, Departamento de Investigación en Tecnología de los Materiales y el Ambiente (DITeMA), Instituto Venezolano de Investigaciones Científicas (IVIC), Avenida 74 con calle 14A, Maracaibo, República Bolivariana de Venezuela.*

Corresponding author (✉):

Y. Marrero-Ponce

ymarrero77@yahoo.es or ymponce@gmail.com

URL: http://www.uv.es/yoma or http://ymponce.googlepages.com/home

ABSTRACT

In the present study, we introduce novel 3D protein descriptors based on the bilinear algebraic form in the $\mathbb{R}^n$ space on the *coulombic* matrix. For the calculation of these descriptors, macromolecular vectors belonging to $\mathbb{R}^n$ space, whose components represent certain amino acid side-chain properties, were used as weighting schemes. Generalization approaches for the calculation of inter-amino acidic residue spatial distances based on Minkowski metrics are proposed. The simple- and double-stochastic schemes were defined as approaches to normalize the coulombic matrix. The local-fragment indices for both *amino acid-types* and *amino acid-groups* are presented in order to permit characterizing fragments of interest in proteins. On other hand, with the objective of taking into account specific interactions among amino acids in global or local indices, geometric and topological *cut-offs* are defined. To assess the utility of global and local indices a classification model for the prediction of the major four protein structural classes, was built with the *Linear Discriminant Analysis* (LDA) technique. The developed LDA-model correctly classifies the 92.6% and 92.7% of the proteins on the training and test sets, respectively. The obtained model showed high values of the *generalized square correlation coefficient* (GC^2) on both the training and test series. The statistical parameters derived from the internal and external validation procedures demonstrate the robustness, stability and the high predictive power of the proposed model. The performance of the LDA-model demonstrates the capability of the proposed indices not only to codify relevant biochemical information related to the structural classes of proteins, but also to yield suitable interpretability. It is anticipated that the current method will benefit the prediction of other protein attributes or functions.

***Keywords*:** 3D protein descriptor, bilinear form, coulombic matrix, Protein structural classes, LDA.

Running Head: *Novel 3D Algebraic Bio-Macromolecular Descriptors.*

1. INTRODUCTION

Molecular structural codification research has continued to attract the attention of several scholars in the present times, evidenced with the ever increasing amount of molecular descriptors (MDs) proposed (Barigye et al., 2013; García-Jacas et al., 2014; Todeschini and Consonni, 2009). These MDs can accordingly be employed to develop models that link the chemical structure with some activity/property (QSAR/QSPR) of interest and select candidate structures for new drugs using several statistical or machine learning techniques. A lot of efforts have been placed on the characterization of small-to-medium-sized molecules, and large number of MDs have been proposed in the literature (Barigye et al., 2013; García-Jacas et al., 2014; Todeschini and Consonni, 2009). However, the same cannot be claimed for macromolecules (e.g. proteins) in that a few molecular parameters have been proposed to encode protein sequences (Rao et al., 2011) and to a much lesser extent to account for the protein spatial structure (Estrada, 2002; González-Díaz and Uriarte, 2005; González Díaz et al., 2004; Gromiha and Selvaraj, 2001; Plaxco et al., 1998; Ruiz-Blanco et al., 2010; Zhou and Zhou, 2002).

It is well-known that a single descriptor or a small number of descriptors cannot wholly represent the molecular complexity or model all physicochemical responses and biological interactions, because only a portion of the chemical information is encoded from a given molecular structure representations schemes (Randić et al., 2009; Todeschini and Consonni, 2009). Thus, there is an emerging need in protein science to develop novel representations of proteins and novel protein descriptors, able to provide new information and better characterization of macromolecular structures (Randic et al., 2010). A general strategy followed to define new topological (2D)-protein descriptors is to extend the MDs used in classic QSAR studies to describe polypeptide chains (González et al., 2002; Moreau and Broto, 1980; Ramos

de Armas et al., 2004b).This intuitive idea was also applied by Marrero-Ponce *et al.* to define several 2D-algebraic-based protein descriptors; these MDs are based on the quadratic-, linear- and bilinear- algebraic forms to obtain graph-theoretical invariants from the biopolymer structure codified by using a graph-theoretical model called the macromolecular pseudograph α-carbon atom adjacency matrix (Marrero-Ponce et al., 2005b; Marrero-Ponce et al., 2004; Ortega-Broche et al., 2010). Moreover, these indices used the macromolecular vectors to codify biochemical information by means of several properties of the amino acid side-chain (R group), in analogy to the well-known molecular vector to represent organic molecules (Marrero-Ponce et al., 2005b; Marrero-Ponce et al., 2004; Ortega-Broche et al., 2010). The utility of the above-mentioned indices was assessed in the prediction of the biological stability of a set of Arc mutants, obtaining quantitative models with straightforward interpretability, good predictability, stability and favorable performance in comparison with several bio-macromolecular descriptors (Marrero-Ponce et al., 2005b; Marrero-Ponce et al., 2004; Ortega-Broche et al., 2010). An in-house comparison of the algebraic forms revealed that bilinear indices exhibited comparable-to-superior performance than the quadratic and linear indices, respectively (Ortega-Broche et al., 2010).

On other hand, in a recent report, Marrero-Ponce and coworkers introduced the novel 3D-QSAR alignment-free MDs known as [QuBiLS-MIDAS (acronym for Quadratic, Bilinear and *N*-Linear Maps based on *n*-Tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic WeightingS)] to codify the 3D chemical structure of organic compounds. These indices are based on the multi-linear algebraic forms on the *N-Tuple Spatial-(Dis) Similarity Matrix* (Marrero-Ponce et al., 2014, Accepted for publication). Several preliminary studies with the QuBiLS-MIDAS 3D-MDs demonstrated satisfactory behavior, suggesting that this algebraic strategy

yields information-rich indices of relevance in chemoinformatic tasks (Marrero-Ponce et al., 2014, Accepted for publication). In regard to the overall performance of the three algebraic forms used in the definition of these MDs, the bilinear form-based indices yielded superior performance than the quadratic- and linear-based analogues, respectively, in the QSAR studies performed (Marrero-Ponce et al., 2014, Accepted for publication).

Taking into account the suitable performance of the 2D-protein bilinear indices (Ortega-Broche et al., 2010) and the encouraging results obtained with the QuBiLS-MIDAS 3D-MDs (Marrero-Ponce et al., 2014, Accepted for publication), particularly the bilinear algebraic form, the extension of the QuBiLS-MIDAS MDs for the characterization of the 3D structure of proteins seems to be a promissory undertaking.

Different aspects have been of interest in protein structures and functions, research including protein subcellular location prediction (Chou and Shen, 2007), protein remote homology detection (Liu et al., 2012; Liu et al., 2013; Liu et al., 2014b), predicting membrane proteins and their types (Cai and Chou, 2006), protein structural class prediction (Chou, 2005) and so on. Of particular interest is the structural class identification, which is useful in enhancing the prediction accuracy of the tertiary structure of a given protein (Chou, 1992), and has played an important role in the development of prediction methods for other protein features (Chou, 2005). Due to its importance in protein science, many computational methods have been proposed to address this challenge and these are classified into three main groups according to the approaches often used to represent the protein sample: the amino acid (AAC)-, pseudo amino acid (PseAAC)- and functional domain (FunD)- composition, respectively (Chou, 2005). The main drawback of the AAC-based methods (Chou, 1995; Chou and Zhang, 1994; Liu and Chou, 1998) is the lack of information on sequence order-effects, thus in general sense, prediction

quality was very limited. In order to codify more sequence-order information and hence improve the prediction quality, the PseAAC was introduced (Chou, 2001) and subsequently different kinds of PseAAC were built for enhancing the prediction accuracy (Chen et al., 2006; Ding et al., 2007; Xiao et al., 2008a; Xiao et al., 2008b; Xiao et al., 2006; Zhang et al., 2008). In addition to the AAC and the PseAAC, the FunD approach aimed at formulating the sample of a protein has been proposed and this probably constitutes one of the most significant progress in this field (Chou and Cai, 2004). Recently, besides the conventional PseAAC approach (Kong et al., 2014; Li et al., 2009), other studies incorporating evolutionary information in the representation of the protein sample have been proposed (Chen et al., 2008b; Zhang et al., 2014). For comprehensive reviews on the progress of prediction methods see (Chou, 2005; Chou, 2011; Chou, 2000).

The core objective of the present report is to introduce a new class of 3D-protein indices based on the bilinear algebraic forms. To evaluate the utility of these indices in the description of the proteins' macromolecular structure, LDA models to predict the protein structural classes are built.

2. THEORETICAL FRAMEWORK

2.1. Bilinear Coulombic Indices for Amino Acid-Level and Total (Global) Definitions.

Proteins are polymers of amino acids, with each amino acid residue linked to its neighbor by a peptide bond (Lehninger et al., 2005). The 20 amino acids commonly found as residues in proteins are α-amino acids and differ from each other in their side chains(R groups), which vary in structure, size, electric charge and these factors influence the solubility of the amino acids (Lehninger et al., 2005).

If each amino acid is considered as a *"pseudo-vertex"*, that is, a vertex composed of several vertices (atoms), then the physicochemical properties of each pseudo-vertex (α-amino acid) can be weighted according to the nature of its R group. On this basis, the k^{th} bilinear coulombic indices for amino acid "*a*" (${}_bL_{aZ}$) are calculated as bilinear forms (maps) in $\mathbb{R}^n$, on a canonical basis set, and are defined as:

$$ {}_bL_{aZ} = {}^{a,p,k}b_z^{O-B}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{a,p,k}z_{ij}x_m^i y_m^j = [X]^T \ {}^aZ_k^p [Y] \quad (\mathbf{1}) $$

where, *n* is the number of amino acids (α-amino acids) in the protein, ${}^{a,p,k}z_{ij}$ are the elements of the k^{th} power (see next subsection) *amino acid-level coulombic matrix* (representing a single amino acid "*a*") ${}^aZ_k^p$ and are calculated from the coefficients ${}^{p,k}z_{ij}$ of the global (whole-protein) k^{th} *coulombic matrix* Z_k^p as follows:

$$ {}^{a,p,k}z_{ij} = {}^{p,k}z_{ij} \text{ if } (i = a \wedge j = a) $$

$$ {}^{a,p,k}z_{ij} = \tfrac{1}{2}\,{}^{p,k}z_{ij} \text{ if } (i = a \vee j = a) \quad (\mathbf{2}) $$

$$ {}^{a,p,k}z_{ij} = 0 \text{ otherwise.} $$

On other hand, x_m^i and y_m^j are the components of the macromolecular vectors $\bar{x}_m$ and $\bar{y}_m$, respectively, in the canonical basis set. Accordingly, [X] and [Y] are column vectors (*n*×1 matrices) of the coordinates of macromolecular vectors $\bar{x}_m$ and $\bar{y}_m$, respectively, $[X]^T$ (a 1×*n* matrix) is the transpose of the vector of properties [X] and *O-B* is the combination of properties of amino acids. The use of amino acid-based macromolecular vectors for codifying polypeptide sequences is explained in detail in (Marrero-Ponce et al., 2005b; Marrero-Ponce et al., 2004; Ortega-Broche et al., 2010). The components (coordinates) of these macromolecular vectors are numerical values, which represent certain amino acid side-chain property (Marrero-Ponce et al.,

2005b; Marrero-Ponce et al., 2004; Ortega-Broche et al., 2010). In the present report, the following properties are used as weighting schemes : molecular mass (MM) (Mathews et al., 2000), side-chain volume (MV) (Zamyatnin, 1972), z-values (Hellberg et al., 1987), atomic charge (ECI) (Collantes and Dunn III, 1995), isotropic surface area (ISA) (Collantes and Dunn III, 1995), Hoop-Woods hydropathy index (HWS) (Hopp and Woods, 1981), Kyte-Doolittle hydropathy index (KDS) (Kyte and Doolittle, 1982), isoelectric point (PIE) (Hellberg et al., 1987); relative frequencies with which an amino acid appears forming α-helices (PAH), β-sheets (PBS) and reverse turns (PTT) , respectively (Mathews et al., 2000); geometric compatibility parameters (L19, ξ) and heat of formation (EPS) (Sak et al., 1999), (see Table 1 for details). Thus, a peptide (or protein) having 5, 10, 15, . . ., *n* amino acids can be represented by means of vectors, with 5, 10, 15, . . ., *n* components, belonging to the spaces $\mathbb{R}^5$, $\mathbb{R}^{10}$, $\mathbb{R}^{15}$, . . . , $\mathbb{R}^n$ (Ortega-Broche et al., 2010). For instance, if one wants to encode the bradykinin-potentiating pentapeptide *VKWAA* (Collantes and Dunn III, 1995), using the weighting scheme defined by the z_1-scale and z_3-scale, then the following macromolecular vectors $\bar{x}_m = [-2.69\ 2.84 - 4.75\ 0.07\ 0.07]$ and $\bar{y}_m = [-1.29 - 3.14\ 0.85\ 0.09\ 0.09]$ are obtained and both belong to the product space $\mathbb{R}^5$.

Table 1 comes about here

If a protein is partitioned into "A" amino acids, then the global matrix Z_k^p can be partitioned into "A" amino acid-level matrices (${}^aZ_k^p$), and thus the k^{th} power of Z_k^p is exactly the sum of the k^{th} power of the "A" amino acid-level matrices. As can be noticed from (Eq.1), each amino-acid level matrix ${}^aZ_k^p$ determines an amino acid-level bilinear coulombic index ${}_bL_{aZ}$, [designated by the acronym: LOVI (LOcal Vertex Invariant)] (Balaban, 1994; Todeschini and Consonni, 2009; Todeschini, 2010) for amino acid "*a*". In this way, the total (whole-protein)

bilinear coulombic indices are calculated from the contribution of each amino acid and thus can be represented as a vector of size *n* (denoted here as ${}_b\bar{L}_Z$), where each component ${}_bL_{aZ}$ of ${}_b\bar{L}_Z$ corresponds to the bilinear coulombic amino acid-level index for amino acid "*a*". Therefore, the total (whole-protein) k^{th} bilinear coulombic indices are calculated as the sum of each k^{th} amino acid-level bilinear coulombic index ${}_bL_{aZ}$, in way similar to the approaches proposed in (Marrero-Ponce et al., 2014, Accepted for publication; Ortega-Broche et al., 2010):

$$ {}^{p,k}b_Z^{O-B}(\bar{x}_m, \bar{y}_m) = \sum_{a=1}^{n} {}_bL_{aZ} = [X]^T Z_k^p [Y] \qquad (\mathbf{3}) $$

The matrix Z_k^p can be classified as a *generalized reciprocal matrix* $M^{-\lambda}$ (Todeschini and Consonni, 2009). Generalized reciprocal matrices are a type of matrices obtained by raising the non-diagonal matrix elements to some negative exponent, where λ is usually an integer positive parameter (Todeschini and Consonni, 2009). One of the most popular reciprocal matrices obtained for λ=1 is the *reciprocal geometry matrix* $\mathbb{G}^{-1}$. The *reciprocal geometry matrix* is an *n*×*n* symmetric matrix, where *n* is the number of atoms in a molecule, each entry $(r_{ij})^{-1}$ is calculated by raising the non-diagonal elements r_{ij} of the *geometry matrix* $\mathbb{G}$ to the power -1 (Todeschini and Consonni, 2009). On other hand, the elements r_{ij} of the *geometry matrix* are calculated as the *Euclidean* distance between atoms *i* and *j* and diagonal entries are always zero (Todeschini and Consonni, 2009). Recently, several approaches were proposed as generalizations of the aforementioned *geometry matrix* in the definition of the QuBiLS-MIDAS 3D-MDs (Marrero-Ponce et al., 2014, Accepted for publication). One of such approaches consists in the generalization of inter-atomic spatial distances through the Minkowski distance norm (Marrero-Ponce et al., 2014, Accepted for publication). An approach based on this metric

is adopted in the present report to codify information on the 3D structure of proteins (discussed in the next subsection).

2.2 Coulombic Matrix for the Representation of the 3D Structure of Proteins.

The protein spatial structure is a complex three-dimensional object, defined by the 3D distribution of its constituent atoms. As is well-known, the protein tertiary structure depends mainly on a complex network of inter-residue interactions, which play an important role in the processes of stabilizing and maintaining the macromolecular structure (Lehninger et al., 2005). Hence, these interactions are a suitable starting point to codify information on the macromolecular structure (Di Paola et al., 2012).On other hand, using graphical approaches to study biological problems can provide an intuitive picture or useful insights in the analysis of complicated relations in these systems (Lin and Lapointe, 2013), as demonstrated in previous studies on a series of important biological processes, such as enzyme-catalyzed reactions (Zhou and Deng, 1984), inhibition of HIV-1 reverse transcriptase (Althaus et al., 1993), drug metabolism systems (Chou, 2010), sequence evolution (Wu et al., 2010), and protein-protein interactions (Zhou, 2011) studied using the wenxiang diagram or graph (Chou et al., 1997; Chou et al., 2011) .

The representation of proteins as molecular graphs, where the amino acids are considered as *pseudo-vertices* and the covalent interactions between amino acids (peptide bonds) and non-covalent interactions between the side chains of amino acids as *pseudo-edges*, permits a matrix based representation of its bio-macromolecular structure, which in turn serves as a valuable source for protein MDs. Here, the coulombic matrix Z_k^p is defined.

Formally, the coulombic matrix Z_k^p is an $n \times n$ square matrix, where n is the number of amino acids within the protein, and its entries ${}^{p,k}z_{ij}$ are defined as follows:

$$ {}^{p,k}z_{ij} = \begin{cases} \dfrac{1}{\left(d_{ij}^{\,p}\right)^{k}}, if\ i \neq j \\ 0, if\ i = j \end{cases} \qquad \textbf{(4)} $$

where, $d_{ij} = \left(\left|x^i - x^j\right|^p + \left|y^i - y^j\right|^p + \left|z^i - z^j\right|^p\right)^{\frac{1}{p}}$, d_{ij}^p is the distance between two vectors in $\mathbb{R}^3$ (x^i, y^i, z^i) and (x^j, y^j, z^j) that correspond to the spatial coordinates for the α-carbon atoms (C_α) of the amino acids *i* and *j*, respectively, and *p* is a Minkowski distance-based metric, $1 \leq p \leq 3$, thus for *p=1* or *p=2* the distance between pairs of C_α is computed as the well-known Manhattan or Euclidean distances, respectively. It is worth noting that when no normalizing procedure is performed (see next subsection) for the elements of Z_k^p, this matrix is designated as the k^{th} *non-stochastic coulombic matrix* ${}_{ns}Z_k^p$ (NS^k-CM) and its entries are denoted as ${}_{ns}^{p,k}z_{ij}$. In addition, the matrix ${}_{ns}Z_k^p$ determines the k^{th} total (global) *non-stochastic bilinear coulombic indices* ${}_{ns}^{p,k}b_Z^{O-B}(\bar{x}_m, \bar{y}_m)$, which are calculated by replacing the general coulombic matrix Z_k^p by ${}_{ns}Z_k^p$ in (Eq. 3). The term *coulombic* is inspired in the relation between the distance and the magnitude of non-covalent interactions of diverse nature. In ref (Kar, 2007), it is demonstrated that the relation between the distance and the strength of non-covalent interactions contributes in a greater or lesser extent to the maintenance of the 3D structure of a macromolecule according to the distance at which the functional groups are interacting. Therefore, with the aim of modeling the functional relationship between the distance and the strength of non-covalent interactions among the functional groups of amino acids in a given protein, the parameter *k* is used, e.g. for *k* = 1, *k* = 2, ${}_{ns}Z_k^p$ reflects Coulombic-like and/or gravitational-like interactions. The maximum *k* value of 12 is related to the non-bonded (mainly steric) interactions associated with the functional form of the Lennard-Jones 6-12 potential. On other hand, it is important to remark that

the *reciprocal geometry matrix* $\mathbb{G}^{-1}$coincides with the specific ${}_{ns}Z_1^2$ constructed by using *p=2* (Euclidean distance) and *k=1*.

2.3. Normalization Formalisms based on Simple-Stochastic and Double-Stochastic Schemes.

Probabilistic transformations for matrices that encode information on the molecular structure have been previously performed in the calculation of both molecular and bio-macromolecular descriptors (Carbo-Dorca, 2000; González-Díaz and Uriarte, 2005; González et al., 2002; Marrero-Ponce et al., 2005a; Marrero-Ponce et al., 2008; Marrero-Ponce et al., 2014, Accepted for publication; Ramos de Armas et al., 2004a). These methods utilize simple stochastic scaling, where the sum of the elements of each row is used as a scaling factor, generating alternative non-symmetric matrices, whose columns can be interpreted as discrete probability distributions.

With the purpose of normalizing the k^{th} *non-stochastic coulombic matrix*, two approaches are defined: the simple- and double-stochastic coulombic matrices, respectively, (see Figure 1). Firstly, the k^{th} *simple-stochastic coulombic matrix*, ${}_{ss}Z_k^p$ (SS^k-CM) is an *n*×*n* square non-symmetric matrix and its elements ${}_{ss}^{p,k}z_{ij}$ are defined as follows:

$$ {}_{ss}^{p,k}z_{ij} = \frac{{}_{ns}^{p,k}z_{ij}}{\sum_{j=1}^{n} {}_{ns}^{p,k}z_{ij}} \qquad \textbf{(5)} $$

An *n*×*n* square matrix is considered to be *stochastic* if has the property that the sum of the elements in each row or each column is 1, that is, the row elements or the column elements are non-negative real numbers that can be interpreted as probabilities (Edwards and Penney, 1988).

The non-symmetrical property of the matrix ${}_{ss}Z_k^p$ is due to the fact that the probability for amino acid *i* to interact with an amino acid *j*, is different from the probability for the amino acid *j*

to interact with the amino acid *i*. With the aim of equalizing the probabilities in both senses, Marrero-Ponce *et al.* introduced the double-stochastic matrix as an alternative normalization strategy (Marrero-Ponce et al., 2014, Accepted for publication). In the same spirit, we employed in this study the k^{th} *double-stochastic coulombic matrix*, ${}_{ds}Z_k^p$ (DS^k-CM) as a normalization approach computed through the double stochastic transformation of ${}_{ns}Z_k^p$, for details on this procedure refer to (Sinkhorn and Knopp, 1967).

Consequently, in analogy to the k^{th} *total non-stochastic bilinear coulombic indices*, the k^{th} *total simple-stochastic-* ${}_{ss}^{p,k}b_Z^{O-B}(\bar{x}_m, \bar{y}_m)$ and *double-stochastic-* ${}_{ds}^{p,k}b_Z^{O-B}(\bar{x}_m, \bar{y}_m)$ *bilinear coulombic indices* are calculated from the k^{th} *simple-stochastic-* ${}_{ss}Z_k^p$ and *double-stochastic-* ${}_{ds}Z_k^p$ *coulombic* matrices, respectively.

Figures 1 and 2 come about here

2.4. Local-Fragment (amino acid-type, groups) Bilinear Coulombic Indices.

The proposed matrices (${}_{ns}Z_k^p$, ${}_{ss}Z_k^p$, ${}_{ds}Z_k^p$) could be employed for codifying information on certain fragments *F* of the protein. Therefore, the k^{th} *local-fragment coulombic matrix* Z_{kF}^p can be obtained from the global matrix Z_k^p. This matrix Z_{kF}^p contains information on distances among C_α of α-amino acids belonging to specific polypeptide fragments (*F*) and its elements ${}^{p,k}z_{ijF}$ are defined as follows:

$$
\begin{aligned}
&{}^{p,k}z_{ijF} = {}^{p,k}z_{ij} \text{ if } (i \wedge j) \in F \\
&{}^{p,k}z_{ijF} = \tfrac{1}{2}\,{}^{p,k}z_{ij} \text{ if } (i \vee j) \in F \text{ but not both} \\
&{}^{p,k}z_{ijF} = 0 \text{ otherwise.}
\end{aligned}
\qquad \textbf{(6)}
$$

It should also be pointed out that for every partitioning of a protein into *R* polypeptide fragments there will be *R* polypeptide local-fragment matrices. Analogous to the k^{th} amino acid-

level indices, the k^{th} *local-fragment amino acid-level indices* are calculated as bilinear forms using the following expression:

$$ {}_{bF}L_{aZ} = {}^{a,p,k}b_{ZF}^{O-B}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n}\sum_{j=1}^{n} {}^{a,p,k}z_{ijF}\, x_m^i y_m^j = [X]^T \; {}^{a}Z_{kF}^{p}[Y] \qquad (7) $$

where, ${}^{a,p,k}z_{ijF}$ is the k^{th} element of the row "*i*" and column "*j*" of the k^{th} *local-fragment amino acid-level matrix* ${}^{a}Z_{kF}^{p}$ according to the amino acid *"a"*. This matrix is extracted for each amino acid of the protein from the local-fragment matrix Z_{kF}^{p}. Note that similar to their total analogues, when no normalizing procedure is carried out over the entries ${}^{p,k}z_{ij}$ of the matrix Z_k^p, the resulting local-fragment matrix Z_{kF}^{p} is designated as the k^{th} *non-stochastic coulombic local-fragment matrix* ${}_{ns}Z_{kF}^{p}$. It follows that the simple-stochastic ${}_{ss}Z_{kF}^{p}$ and double-stochastic ${}_{ds}Z_{kF}^{p}$ local-fragment matrices of order *k*, can be computed from the ${}_{ns}Z_{kF}^{p}$ in the same manner as described in Subsection 2.3. These local analogues can also be expressed in matrix form for each macromolecular vector $\bar{x}_m \in \mathbb{R}^n$ and $\bar{y}_m \in \mathbb{R}^n$. Similar to the total indices, the local-fragment analogues may be represented as a vector ${}_{bF}\bar{L}_Z$ of size *n*, where each component ${}_{bF}L_{aZ}$ of ${}_{bF}\bar{L}_Z$ corresponds to the local-fragment bilinear amino acid-level index (LOVI) for amino acid "*a*". Therefore, the k^{th} *local-fragment bilinear coulombic indices* ${}^{p,k}b_{ZF}^{O-B}(\bar{x}, \bar{y})$, are calculated as a summation over vector of LOVIs ${}_{bF}\bar{L}_Z$.

It is important to remark that a local-fragment (*F)* can be a sequence of consecutive residues as well as groups of residues distant in the sequence. In this paper, the local indices can be calculated by using the following local-fragments (*amino acid-type*): apolar (RAP), polar positively charged (R+), polar negatively charged (R-), polar uncharged (RPU), aromatic (ARG) and aliphatic (ALG). In the *amino acid-type* formalism, each α-amino acid in the protein is

classified into amino-acid-type (fragment), depending on the nature of its R group. Also we defined *groups* that include the amino acids that do not favor the folding and/or cannot be commonly found in proteins as part of α-helices or β-sheets (UFG), α-helices favoring amino acids (FAH), β-sheets favoring amino acids (FBS) and β-turns favoring amino acids (AFT) (Mathews et al., 2000). Additionally, groups composed of amino acids of the same kind (R amino acids) in the protein were defined, that is, 20 groups with one for each natural amino acid, (e.g. F=Ala, F=Arg,…, F=Val). Table 2 shows the amino acidic composition of the local-fragments that are pre-defined in the TOMOCOMD-*CAMPS* software (acronym of TOpological MOlecular COMputational Design- *Computer-Aided Modelling in Protein Science*). However, there is an option for the users to define their own local-fragments.

Table 2 comes about here

Similar to the global indices, the k^{th} *(local-fragment) non-stochastic* ${}^{p,k}_{ns}b^{O-B}_{ZF}(\bar{x}_m, \bar{y}_m)$, *simple-stochastic* ${}^{p,k}_{ss}b^{O-B}_{ZF}(\bar{x}_m, \bar{y}_m)$ and *double-stochastic* ${}^{p,k}_{ds}b^{O-B}_{ZF}(\bar{x}_m, \bar{y}_m)$ *bilinear coulombic indices* are introduced and are computed from the local-fragment matrices ${}_{ns}Z^{p}_{kF}$, ${}_{ss}Z^{p}_{kF}$, ${}_{ds}Z^{p}_{kF}$, respectively.

2.5. Geometric and Topological Constraints-based Approach

Non-covalent interactions have an important influence on the final structure of macromolecules, their specific binding modes and in the process of self-organizing of macromolecular and cellular structures, among other structural and functional aspects (Mathews et al., 2000). The relationship between the distance and the magnitude of the non-covalent interactions of diverse nature demonstrates that these contribute to the maintenance of the 3D structure of the protein, depending on the distance that the interacting groups are found. In this way, some of these interactions are only important when the functional groups are so close

among themselves or distant in the sequence but sterically close (large-contacts). On other hand, the relationship between the topology and the folding of biopolymers has been elucidated in diverse studies, where significant correlation between simple structural parameters and the speed of protein folding has been found (Gromiha et al., 2004; Gromiha, 2003; Gromiha and Selvaraj, 2001; Plaxco et al., 1998). Among these parameters are: RCO (acronym for Relative Contact Order) and LRO (acronym for Long-range order) proposed by Plaxco *et al.*(Plaxco et al., 1998) and Gromiha *et al.* (Gromiha and Selvaraj, 2001), respectively. In this way, sometimes it may be useful to build matrices with information on the contact (interaction) between amino acid residues found at a certain distance (or distance range) in the sequence with the objective of studying possible relations between a specific property and topological features of the native state of the protein.

Bearing all this in mind, with the purpose of taking into account only some type of non-covalent interactions and thus consider only significant interactions in global or local indices, two different approaches are applied:

1) Geometric cut-off (*l*), based on Euclidean distance at lag *l*, termed as "*length cut-off*".
2) Graph-theoretical cut-off (*p*) based on topological distance at lag *p*, designated as "*path cut-off*".

The application of one or both cut-offs over ${}_{ns}Z_k^p$ generates the non-stochastic coulombic matrix at the lags *l* and/or *p* and their entries are calculated from the ${}_{ns}Z_k^p$ as follows:

$$
\begin{aligned}
{}^{p,k}_{ns}z_{ij} &= {}^{p,k}_{ns}z_{ij} \times \delta^{ij} \\
where,\ \delta^{ij} &= 1\ if\ l_{\min} \le l_{ij} \le l_{\max}\ \ and/or\ \ p_{\min} \le p_{ij} \le p_{\max} \\
&= 0\ otherwise
\end{aligned}
\qquad \textbf{(8)}
$$

where, l_{min} and l_{max} are the lower and upper bounds for Euclidean distance, respectively, and l_{ij} is the geometric Euclidean distance between the α-amino acids *i* and *j*; p_{min} and p_{max} are the pre-

defined topological distance thresholds, p_{ij} is the topological distance between the amino acids *i* and *j*. It is important to note that when the length and/or path thresholds are applied to the computation of the ${}_{ns}Z_k^p$, a sparse matrix (a matrix with relatively few nonzero elements) is obtained, where each entry ${}_{ns}^{p,k}z_{ij}$ coincides with its original definition (the term δ^{ij}=1, see (Eq.4), only if the Euclidean (l_{ij}) and/or topological (p_{ij}) distances between amino acids *i* and *j* lie(s) within the pre-defined geometric (l_{min}-l_{max}) and/or topological (p_{min}-p_{max}) intervals and is zero otherwise.

The constraints approach (both length and path thresholds) permit unifying geometric and topological information in the same matrix and they also allow us to consider the most relevant interactions and at the same time excluding irrelevant chemical information due to long-range interactions. It is not is mandatory to use any constraints for calculations. However, incorporating this approach may be beneficial as the "cut-offs" permits the discrimination of the interaction types.

For instance, the use of the *length* criterion (together with exponent *k*) permits taking into account only those non-covalent interactions, among the functional groups of the amino acids, which significantly contribute to the maintenance of the 3D protein structure. In addition, the *k* exponent in the term $(d_{ij}^p)^k$ of (Eq. 4) models the functional relationship that exists between the distance and the strength of the interaction between the functional groups of the amino acids *i* and *j*.

On the other hand, the *path* criterion permits the selection of the non-covalent interactions for amino acids within a given topological distance. It should be noted that the topological distance between two amino acids *i* and *j* is determined by the *shortest path* between vertices *i*

and j ($C_\alpha^{\,i}$,$C_\alpha^{\,j}$) of the graph whose i^{th} vertex represents the $C_\alpha^{\,i}$ of the peptide backbone and the edges are the peptide bonds between the amino acids i and j.

Illustrations of the application of the length, path or both constrains to the computation of entries of the non-stochastic coulombic matrix of order 1 at the lags l and/or p ($_{ns}Z_1^p$) to characterize the 3D structure of a sample peptide could be found in the Figure 3.

Figure 3 comes about here

Lastly, the k^{th} simple- and double-stochastic matrices at the lags l and/or p can be computed from the k^{th} non-stochastic matrix at the lags l and/or p, in the same way as described in Subsection 2.3. Thus, the k^{th} (global or local) NS-, SS- and DS- bilinear coulombic indices at the lags l and/or p are calculated from the k^{th} (global or local) NS-, SS- and DS- coulombic matrices at the lags l and/or p, respectively.

3. APPLICATION OF THE BILINEAR COULOMBIC INDICES TO THE PREDICTION OF PROTEIN STRUCTURAL CLASSES.

3.1 Benchmark Dataset.

The prediction of protein structural classes is of relevance in protein science, and it generally consists of classifying a protein into one of the major structural classes (*All-α*, *All-β*, *α/β, α+β*) (Levitt and Chothia, 1976). The development of a classification model for the prediction of the major protein structural classes is a key aspect in the present study. To this end, the widely used dataset proposed in (Chou, 1999) was selected; it consists of 204 proteins of which: 52 are *All-α*, 61 *All-β*, 45 *α/β* and 46 *α+β*. In the construction of this benchmark dataset, a cutoff threshold of 30% was used (Lin and Li, 2007), a value considerably stringent to guarantee low homology bias and redundancy in this dataset. The validation of the obtained models

represents one of the most important steps in the QSAR/QSPR model development process as it provides criteria on the true predictive ability of the generated models (Todeschini and Consonni, 2009). In statistical analysis, the following three cross-validation methods are often used to examine a model for its effectiveness in practical applications: independent dataset test, subsampling test, and jackknife test (Chou, 2001). On one hand, of the three test methods, the *jackknife* test is deemed by some authors as the least arbitrary and most objective that can always yield a unique result for a given benchmark dataset as elaborated in (Chou, 2011). In the context of structural class prediction the jackknife test, that is, the leave-one-out cross-validation approach is often employed for assessing the predictive power of the models (Cai et al., 2002a; Cai et al., 2002b; Cai et al., 2006; Chen et al., 2008a; Chen et al., 2006; Chen et al., 2008b; Chou, 1999; Ding et al., 2007; Lin and Li, 2007; Shen et al., 2005; Xiao et al., 2008a; Xiao et al., 2008b; Xiao et al., 2006; Zhang et al., 2008).

On the other hand, as discussed in (Golbraikh and Tropsha, 2002; Gramatica, 2007; Tropsha et al., 2003), when the number (N) of instances (compounds) is high (e.g. N>100), the leave-one-out approach performs very similar to the fit, due to small perturbation of the data when one instance is left out; thus it should be considered as a measure of the goodness-of-fit (internal performance) of the model, rather than a measure of its predictive ability, hence high performance in internal cross-validation can be regarded as a necessary, but insufficient condition for the models to have a high predictive power (Golbraikh and Tropsha, 2002; Gramatica, 2007; Tropsha et al., 2003).

Actually, as the real utility of a QSAR/QSPR model relies in its ability to accurately predict the modeled (activity/property) for new compounds, a realistic evaluation of the model true predictive power must be determined in the most appropriate and rigorous way possible

(Tropsha et al., 2003).Therefore, the external validation should be seen as a useful complement to internal validation, rather than as a substitute or superior alternative (Gramatica, 2007). Following the above statement, we adopted the independent dataset test in this study to validate the classification model.

Therefore, the original dataset was split into the training and test sets, for the calibration and external validation of the classification model (see Figure 4). However, there exists the issue concerning to the splitting method employed as the models' results are strongly dependent on the splitting of the data (Todeschini and Consonni, 2009). As can be seen in (Todeschini and Consonni, 2009), the splitting into training and test sets can be used reliably only if the splitting of the data is performed by a well-stated criterion, such as a criterion based on experimental design or cluster analysis or other deterministic approaches. In this sense, the original dataset was split into the training and test sets using the cluster analysis method. Specifically, the *k*-means method with Euclidean distance as the similarity/dissimilarity measure was applied to the 4 sets of 52 (*All-α*), 61 (*All-β*), 45 (*α/β*) and 46 (*α+β*) proteins, respectively, which were distributed separately in 3 (*All-α*), 7 (*All-β*), 3 (*α/β*) and 4 (α+β) clusters. Random stratified sampling was performed as strategy to guarantee a good representativity in each case, where each one of the clusters in each structural class was taken as a stratum. As a result, a training set composed of 149 proteins was obtained, having: 38 (*All-α*), 48 (*All-β*), 29 (*α/β*) and 34 (α+β), and a test set containing 55 proteins of which: 14 were (*All-α*), 13 (*All-β*), 16 (*α/β*) and 12 (α+β); proteins within the test set were never used to build a model.

Figure 4 comes about here

3.2 Development of the Classification Model.

The classification model was built with Linear Discriminant Analysis (LDA) implemented in the STATISTICA 6.0 software package (http://www.statsoft.com/) using total and local 3D-protein bilinear indices, which were calculated with the TOMOCOMD-CAMPS software. The choice of LDA to generate the classification model is based on its simplicity (McFarland and Gans, 1990). The canonical transformation of LDA, [i.e. Canonical Discriminant Analysis (CDA)] was used in the present study to derive the discriminant functions. The CDA is a dimension-reduction technique that derives linear combinations of the variables (MDs) also known as *canonical components*, in a decreasing order of possible multiple correlation with the groups (classes) of observations (proteins). The number of derived canonical components by CDA is equal to the number of original variables or the number of classes minus one, depending on which parameter (variables or classes) is smaller. In this case, we have four structural classes of proteins, hence three canonical components or canonical discriminant functions (CDF) can be obtained. The CDF of the best classification model obtained are given below together with the corresponding statistical parameters:

$$\begin{aligned}\boldsymbol{CDF1} = {} & -2.3\times10^{-10}\,{}^{1,1}_{ds}b_Z^{MM-L19}(\bar{x}_m,\bar{y}_m) - 6.2\times10^{-6}\,{}^{3,1}_{ns}b_{Zval}^{MM-PIE}(\bar{x}_m,\bar{y}_m) \\ & +1.2\times10^{-8}\,{}^{1,1}_{ds}b_Z^{ECI-MM}(\bar{x}_m,\bar{y}_m) - 1.1\times10^{-6}\,{}^{2,1}_{ds}b_Z^{ECI-PAH}(\bar{x}_m,\bar{y}_m) \\ & +3.0\times10^{-10}\,{}^{3,1}_{ds}b_Z^{ISA-PIE}(\bar{x}_m,\bar{y}_m) - 5.0\times10^{-8}\,{}^{1,1}_{ds}b_Z^{PIE-ECI}(\bar{x}_m,\bar{y}_m) \\ & -3.3\times10^{-10}\,{}^{3,1}_{ds}b_Z^{PIE-EPS}(\bar{x}_m,\bar{y}_m) + 5.3\times10^{-7}\,{}^{2,1}_{ds}b_Z^{PAH-PBS}(\bar{x}_m,\bar{y}_m) \\ & -3.9\times10^{-9}\,{}^{3,1}_{ds}b_Z^{PBS-MV}(\bar{x}_m,\bar{y}_m) + 1.6\times10^{-9}\,{}^{2,2}_{ds}b_Z^{PBS-PAH}(\bar{x}_m,\bar{y}_m) \\ & +1.7\times10^{-11}\,{}^{3,2}_{ds}b_Z^{PTT-MV}(\bar{x}_m,\bar{y}_m) + 6.0 \end{aligned} \quad \textbf{(9)}$$

$$N = 149 \quad \lambda = 0.01 \quad \chi^2(33) = 660.03 \quad p < 0.01$$

$$\boldsymbol{CDF2} = 1.0\times10^{-9}\,{}^{1,1}_{ds}b_Z^{MM-L19}(\bar{x}_m,\bar{y}_m) - 6.2\times10^{-6}\,{}^{3,1}_{ns}b_{Zval}^{MM-PIE}(\bar{x}_m,\bar{y}_m)$$

$$-5.1 \times 10^{-8}\ {}^{1,1}_{ds}b_Z^{ECI-MM}(\bar{x}_m,\bar{y}_m) + 4.4 \times 10^{-6}\ {}^{2,1ds}b_Z^{ECI-PAH}(\bar{x}_m,\bar{y}_m)$$

$$-2.9 \times 10^{-9}\ {}^{3,1}_{ds}b_Z^{ISA-PIE}(\bar{x}_m,\bar{y}_m) + 2.2 \times 10^{-7}\ {}^{1,1}_{ds}b_Z^{PIE-ECI}(\bar{x}_m,\bar{y}_m) \qquad \textbf{(10)}$$

$$+1.2 \times 10^{-9}\ {}^{3,1}_{ds}b_Z^{PIE-EPS}(\bar{x}_m,\bar{y}_m) - 3.6 \times 10^{-6}\ {}^{2,1}_{ds}b_Z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$$

$$+2.4 \times 10^{-8}\ {}^{3,1}_{ds}b_Z^{PBS-MV}(\bar{x}_m,\bar{y}_m) + 1.8 \times 10^{-8}\ {}^{2,2}_{ds}b_Z^{PBS-PAH}(\bar{x}_m,\bar{y}_m)$$

$$-1.5 \times 10^{-10}\ {}^{3,2}_{ds}b_Z^{PTT-MV}(\bar{x}_m,\bar{y}_m) - 1.3$$

$N = 149$ $\lambda=0.27$ $\chi^2(20) = 185.99$ $p < 0.01$

$$\boldsymbol{CDF3} = -5.7 \times 10^{-10}\ {}^{1,1}_{ds}b_Z^{MM-L19}(\bar{x}_m,\bar{y}_m) - 1.4 \times 10^{-6}\ {}^{3,1}_{ns}b_{Zval}^{MM-PIE}(\bar{x}_m,\bar{y}_m)$$

$$-6.4 \times 10^{-8}\ {}^{1,1}_{ds}b_Z^{ECI-MM}(\bar{x}_m,\bar{y}_m) - 7.6 \times 10^{-6}\ {}^{2,1}_{ds}b_Z^{ECI-PAH}(\bar{x}_m,\bar{y}_m)$$

$$+9.2 \times 10^{-10}\ {}^{3,1}_{ds}b_Z^{ISA-PIE}(\bar{x}_m,\bar{y}_m) - 2.2 \times 10^{-7}\ {}^{1,1}_{ds}b_Z^{PIE-ECI}(\bar{x}_m,\bar{y}_m) \qquad \textbf{(11)}$$

$$+3.1 \times 10^{-10}\ {}^{3,1}_{ds}b_Z^{PIE-EPS}(\bar{x}_m,\bar{y}_m) + 5.4 \times 10^{-6}\ {}^{2,1}_{ds}b_Z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$$

$$-3.3 \times 10^{-8}\ {}^{3,1}_{ds}b_Z^{PBS-MV}(\bar{x}_m,\bar{y}_m) + 4.4 \times 10^{-10}\ {}^{2,2}_{ds}b_Z^{PBS-PAH}(\bar{x}_m,\bar{y}_m)$$

$$-2.6 \times 10^{-11}\ {}^{3,2}_{ds}b_Z^{PTT-MV}(\bar{x}_m,\bar{y}_m) - 0.01$$

$N = 149$ $\lambda=0.96$ $\chi^2(9) = 185.99$ $p = 0.80$

where, N is the number of cases (proteins), λ is Wilks statistic, $\chi^2(d.f)$ is Chi-square statistic, $d.f$ degrees of freedom and p is the associated signification level. The quality of the discriminant functions was assessed by means of the Wilks' λ. The Wilks' λ statistic takes values from 0 (perfect discrimination) to 1 (no discrimination). Wilks' λ global value of the LDA-model approximates to 0.01 and the F (33,398) statistic associated to λ (47.4) was very significant at a p-level<0.001. These statistics suggest the rejection of the null hypothesis, which enunciates the equality of multivariate means. It is thus possible significantly discriminate among the four

classes of the considered proteins using linear combinations of the total and local bilinear coulombic indices.

Statistical signification tests of the Wilks' λ for each one of the functions reveal that only CDF1 and CDF2 allow discriminating significantly among the groups' means. As shown in Table 3, the values for centroids of each group (class) in CDF3 are more proximate than the values for centroids in CDF1 and CDF2. In addition, the relative magnitude of the eigenvalue associated to CDF3 indicates that the percentage of variance explained for this function is approximate to 0.13%. Unlike Wilks' λ and the associated Chi-square test, the percent of variance explained is indicative of the practical rather than the statistical significance of the functions for group discrimination. Therefore, in comparison to the proportions of variance explained by CDF1 (91.4%) and CDF2 (8.5%), the variance explained by means of the CDF3 is not significant and it does not relevantly contribute to the LDA-model.

A comparison among the centroids in each CDF (see Table 3), indicates that CDF1 mainly discriminates the classes *All-α* and *α+β* as a whole from the classes *All-β* and *α/β*, respectively, as well as *All-β* from *α/β*. On other hand, the CDF2 discriminates the *All-β* proteins from the *All-α*, *α/β* and *α+β*, respectively. In addition, the CDF2 discriminates between the classes *All-α* and *α+β* better than any other discriminant functions. A joint analysis of centroids and standardized coefficients (see Tables 3-4), on the CDF1 and CDF2 indicates that since the only negative centroid in CDF1 corresponds to α/β proteins, then proteins with higher values for ${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m, \bar{y}_m)$ and/or ${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m, \bar{y}_m)$ and lower values for ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m, \bar{y}_m)$ and/or ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m, \bar{y}_m)$ tend to be classified as *α/β* proteins, whereas those proteins with higher values for ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m, \bar{y}_m)$ and/or ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m, \bar{y}_m)$ and lower values for ${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m, \bar{y}_m)$ and/or ${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m, \bar{y}_m)$ are classified as *All-α* or *α+β*.

On other hand, *All-β* is the only class with a positive centroid in CDF2, therefore proteins with higher values, particularly in variables ${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m,\bar{y}_m)$ and ${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m,\bar{y}_m)$, and lower values in ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m,\bar{y}_m)$ and/or ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$ are predicted as *All-β*, whereas those proteins that present high values, mainly for variables, ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m,\bar{y}_m)$ and/or ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$ are classified as *All-α* or *a/β*. Although the centroids of the classes *All-α* and *a/β* in CDF2 have the same sign, this function is the one that most contributes to the discrimination between these classes. Consequently, proteins with mean values of ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m,\bar{y}_m)$ and/or ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$ and lower values for ${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m,\bar{y}_m)$ and ${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m,\bar{y}_m)$ are predicted more frequently as *All-α*. Nevertheless proteins that are more commonly classified as *α+β* have, in general, mean values for ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m,\bar{y}_m)$, ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$, ${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m,\bar{y}_m)$ and ${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m,\bar{y}_m)$ .In general, the variables ${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m,\bar{y}_m)$, ${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m,\bar{y}_m)$, ${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m,\bar{y}_m)$, and ${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m,\bar{y}_m)$ are the most relevant in the separation of the four classes since they have the standardized coefficients of greater magnitude in the canonical discriminant functions with higher explained variance.

Tables 3 and 4 comes about here

The relevance of these descriptors in the structural classification of proteins may be due in part to the type of biochemical information codified in the macromolecular vectors $(\bar{x}_m,\bar{y}_m)$ used in their calculation. For instance, the properties PAH and PBS codify conformational information of amino acids *i.e.* they describe the frequency with which an amino acid appears forming part of segments of α-helices and β-sheets, respectively (Levitt, 1978). Furthermore, it is known that the trend of an amino acid to favor protein folding depends on the volume of its side-chain [(MV) and (MM)] and its polarity (ECI) (Collantes and Dunn III, 1995). Thus, for

instance, amino acids that promote the formation of β-sheets have on average, more voluminous side chains (MV) [thus their molecular mass (MM) is greater] and also they have lower polarity than those that favor the formation of α-helices (Levitt, 1978). Therefore, it is not surprising that the MDs calculated using the properties: PAH, PBS, MV, MM and ECI have relevant discriminatory power in the structural classification of proteins.

3.3 Assessing the Accuracy of the Classification Model.

The quality of the LDA-model was preliminary determined by examining the rates of correct classification on the training and test series. As can be observed from Table 5, the LDA-model (Eqs.9-11) yields a rate of correct classification of 86.8% (33/38) for the class *All-α*, 97.9% (47/48) for *All-β*, 100% (29/29) for *α/β* and 85.3% (29/34) for *α+β*, for an overall accuracy of 92.6% (138/149) on the designed training set. On other hand, the LDA-model correctly classifies the 92.7% (51/55) of the proteins within the test set. These results indicate that this model is suitable for discriminating among the four protein structural classes (Eriksson et al., 2003; Golbraikh and Tropsha, 2002).

The classification of proteins was performed by means of *a posteriori* classification probability; it represents the probability, with which a protein belongs to a particular class, and it was calculated from the Mahalanobis distance (D) and the associated distribution (Hotelling's T^2). The classification results on the training and test sets are consistent with the previous analysis of the discriminatory power in each of the discriminant functions, in that the lowest percentages of correct classification are obtained for the classes *All-α* and *α+β* (see Table 5). These classes are precisely those which have the nearest centroids in CDF1 and CDF2. In addition, it should be pointed out that most misclassified proteins for the class *All-α* truly belong to the class *α+β* and vice versa, (see SI1 for details).

To perform a deeper evaluation of the quality of the LDA-model, other performance measures were considered (Baldi et al., 2000). On one hand, the LDA-model exhibited high values for the percentage measures sensitivity and specificity (see Table 5), where sensitivity is the probability of correctly predicting a positive case and specificity is the probability that a positive prediction is correct (Baldi et al., 2000). Additionally, the LDA-model showed high values of the generalized square correlation GC^2; this parameter quantifies the linear correlation degree between the classification of the model and the experimental class, where a value of +1 represents a complete linear correlation among the variables in consideration (Baldi et al., 2000). In this case, the LDA-model yield 0.91 and 0.83 on the training and test sets, respectively. This means that there exists a high linear correlation between the protein bilinear indices and the structural classes of proteins. The parameter (GC^2) was considered on the basis of two main factors: firstly the capacity of providing a more balanced assessment of the prediction than percentage measures and secondly the difficulty in generalizing the Matthews correlation coefficient to consider more than two classes (Baldi et al., 2000).

Table 5 comes about here

In order to assess the internal validity of the model, we performed *bootstrapping cross-validation* on the training set. The basic requirement for this method is that the data must be representative of the population from which it is drawn (Eriksson et al., 2003; Tropsha et al., 2003). During the application of this strategy, *K* groups of *n* elements are randomly selected from the original dataset. Some of these elements may be included more than once in the extracted sample, whereas others are never selected (Tropsha et al., 2003). Like other internal validation methods, high values for the average global accuracy is a proof of the robustness of the model (Tropsha et al., 2003). To apply the *bootstrapping* validation method the training set

was divided ten times in two subsets, one of them containing the 75% of the cases (about 112 proteins) to fit the model and the other with the remaining 25% of the cases (about 37 proteins) for its validation. Random stratified sampling with replacement was employed as strategy for the selection of the validation samples, where the four classes of proteins represent the stratums. The amount of selected cases (proteins) in each stratum corresponds approximately to the proportion of that stratum (class) in the original dataset. In each one of the ten experiments of cross-validation: the global accuracy, Wilk's λ and Fisher ratio (F) were calculated; subsequently the average for each parameter was computed, (see Table 6 for details). The results attained in the bootstrapping cross-validation demonstrate the robustness and stability of the model in presence of perturbations in the data caused by the procedure. In addition, the fitting parameters: (λ, D^2, F) and the global accuracy on the training (Q_{Total}^{a}) and test (Q_{Total}^{b}) sets yield acceptable values from a statistical point of view.

Table 6 comes about here

4. CONCLUSIONS

Novel 3D bio-macromolecular descriptors relevant to protein QSPR studies were proposed. We have demonstrated that the use of linear combinations of the novel 3D-protein bilinear indices is able not only to significantly discriminate among the four protein structural classes, but also permits the interpretation of the model obtained. The bootstrapping and the external validation tests established the robustness, stability and the high predictive power of the proposed LDA-model. Therefore, it may be suggested that the proposed MDs constitute a suitable tool to count on in protein research.

5. FUTURE OUTLOOKS

In forthcoming studies we will develop sequence-based (2D) protein descriptors, which could be used to build 2D-prediction methods for several protein attributes such as: protein structural classes (Chou, 2005), protein subcellular location (Chou and Shen, 2007), DNA binding proteins (Liu et al., 2014c; Liu et al., 2014d) and so on. Additionally, efforts will be made to provide web-servers for these new sequence-based predictors as is suggested in (Chou, 2011) and followed through in a series of recent publications (Chen et al., 2013; Guo et al., 2014; Liu et al., 2014a; Liu et al., 2014d; Liu et al., 2014e).

Supplementary Information Available: The results of classification of the 204 proteins according to the LDA-model.

***Conflict of Interest**:* The authors confirm that this article content has no conflict of interest.

Acknowledgement: Marrero-Ponce, Y. thanks the program *'International Professor'* for a fellowship to work at *Cartagena University* and *Universidad Tecnológica de Bolívar* in 2013 and 2014, respectively. Barigye, S. J. acknowledges support from CNPq.

5. REFERENCES

Althaus, I. W., Chou, J. J., Gonzales, A. J., Deibel, M. R., Chou, K. C., Kezdy, F. J., Romero, D. L., Palmer, J. R., Thomas, R. C., 1993. Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548-6554, doi:10.1021/bi00077a008.

Balaban, A., T., 1994. Local versus Global (i.e. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. J. Chem. Inf. Comput. Sci. 34, 398.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412-424.

Barigye, S. J., Marrero-Ponce, Y., Pérez-Giménez, F., Bonchev, D., 2013. Trends in Information Theory Based Chemical Structure Codification. Chem. Rev.

Cai, Y.-D., Chou, K.-C., 2006. Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. J. Theor. Biol. 238, 395-400, doi:http://dx.doi.org/10.1016/j.jtbi.2005.05.035.

Cai, Y.-D., Hu, J., Liu, X., Chou, K.-C., 2002a. Prediction of protein structural classes by neural network method. J. Mol. Des. 1, 332-338.

Cai, Y.-D., Liu, X.-J., Xu, X.-b., Chou, K.-C., 2002b. Prediction of protein structural classes by support vector machines. Comput. Chem. (Oxford, U. K.) 26, 293-296.

Cai, Y.-D., Feng, K.-Y., Lu, W.-C., Chou, K.-C., 2006. Using LogitBoost classifier to predict protein structural classes. J. Theor. Biol. 238, 172-176.

Carbo-Dorca, R., 2000. Stochastic Transformation of Quantum Similarity Matrixes and Their Use in Quantum QSAR (QQSAR) Models. Int. J. Quantum Chem. 79, 163-177.

Collantes, E. R., Dunn III, W. J., 1995. Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogs. J. Med. Chem. 38, 2705-2713.

Chen, C., Chen, L.-X., Zou, X.-Y., Cai, P.-X., 2008a. Predicting protein structural class based on multi-features fusion. J. Theor. Biol. 253, 388-392.

Chen, C., Tian, Y.-X., Zou, X.-Y., Cai, P.-X., Mo, J.-Y., 2006. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243, 444-448.

Chen, K., Kurgan, L. A., Ruan, J., 2008b. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Comput. Chem. 29, 1596-1604.

Chen, W., Feng, P.-M., Lin, H., Chou, K.-C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res., gks1450.

Chou, K.-C., 1992. Energy-optimized structure of antifreeze protein and its binding mechanism. J. Mol. Biol. 223, 509-517, doi:http://dx.doi.org/10.1016/0022-2836(92)90666-8.

Chou, K.-C., 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Struct., Funct., Bioinf. 21, 319-344, doi:10.1002/prot.340210406.

Chou, K.-C., 1999. A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. 264, 216-224.

Chou, K.-C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics 43, 246-255, doi:10.1002/prot.1035.

Chou, K.-C., 2005. Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr. Protein Pept. Sci. 6, 423-436.

Chou, K.-C., 2010. Graphic Rule for Drug Metabolism Systems. Curr. Drug Metab. 11, 369-378, doi:10.2174/138920010791514261.

Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236-247, doi:http://dx.doi.org/10.1016/j.jtbi.2010.12.024.

Chou, K.-C., Zhang, C.-T., 1994. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J. Biol. Chem. 269, 22014-22020.

Chou, K.-C., Cai, Y.-D., 2004. Predicting protein structural class by functional domain composition. Biochem. Biophys. Res. Commun. 321, 1007-1009, doi:http://dx.doi.org/10.1016/j.bbrc.2004.07.059.

Chou, K.-C., Shen, H.-B., 2007. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1-16.

Chou, K.-C., Zhang, C.-T., Maggiora, G. M., 1997. Disposition of amphiphilic helices in heteropolar environments. Proteins: Struct., Funct., Genet. 28, 99-108.

Chou, K.-C., Lin, W.-Z., Xiao, X., 2011. Wenxiang: a web-server for drawing wenxiang diagrams. Natural Science 3, 862.

Chou, K. C., 2000. Prediction of Protein Structural Classes and Subcellular Locations. Curr. Protein Pept. Sci. 1, 171-208, doi:10.2174/1389203003381379.

Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Giuliani, A., 2012. Protein Contact Networks: An Emerging Paradigm in Chemistry. Chem. Rev. 113, 1598-1613, doi:10.1021/cr3002356.

Ding, Y.-S., Zhang, T.-L., Chou, K.-C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept. Lett. 14, 811-815.

Edwards, C. H., Penney, D. E., 1988. Elementary linear algebra. Prentice Hall, Englewoods Cliffs.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environ. Health Perspect. 111, 1361.

Estrada, E., 2002. Characterization of the folding degree of proteins. Bioinformatics 18, 697-704.

García-Jacas, C. R., Marrero-Ponce, Y., Barigye, S. J., Valdés-Martiní, J. R., Rivera-Borroto, O. M., Verbel, J. O., 2014. N-Linear Algebraic Maps to Codify Chemical Structures: is a suitable generalization to the atom-pairs approaches? Curr. Drug Metab. 15, 441-469.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! J. Mol. Graphics Modell. 20, 269-276.

González-Díaz, H., Uriarte, E., 2005. Proteins QSAR with Markov average electrostatic potentials. Bioorg. Med. Chem. Lett. 15, 5088-5094.

González, D., De Armas, R. R., Uriarte, E., 2002. In silico Markovian bioinformatics for predicting 1Ha-NMR chemical shifts in mouse epidermis growth factor (mEGF). Online J. Bioinformatics 1, 83-95.

González Díaz, H., Molina, R., Uriarte, E., 2004. Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies. Polymer 45, 3845-3853.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR & Combinatorial Science 26, 694-701, doi:10.1002/qsar.200610151.

Gromiha, M., Saraboji, K., Ahmad, S., Ponnuswamy, M., Suwa, M., 2004. Role of non-covalent interactions for determining the folding rate of two-state proteins. Biophys. Chem. 107, 263-272.

Gromiha, M. M., 2003. Importance of native-state topology for determining the folding rate of two-state proteins. J. Chem. Inf. Comput. Sci. 43, 1481-1485.

Gromiha, M. M., Selvaraj, S., 2001. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. 310, 27-32.

Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., Chou, K.-C., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics, btu083.

Hellberg, S., Sjoestroem, M., Skagerberg, B., Wold, S., 1987. Peptide quantitative structure-activity relationships, a multivariate approach. J. Med. Chem. 30, 1126-1135.

Hopp, T. P., Woods, K. R., 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U. S. A. 78, 3824-3828.

Kar, A., 2007. Medicinal chemistry. New Age International (P) Ltd., Publishers, New Delhi.

Kong, L., Zhang, L., Lv, J., 2014. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 344, 12-18, doi:http://dx.doi.org/10.1016/j.jtbi.2013.11.021.

Kyte, J., Doolittle, R. F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105-132.

Lehninger, A., Nelson, D. L., Cox, M. M., 2005. Lehninger's Principles of Biochemistry. WH Freeman and Company, New York.

Levitt, M., 1978. Conformational preferences of amino acids in globular proteins. Biochemistry 17, 4277-4285.

Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552-558.

Li, Z.-C., Zhou, X.-B., Dai, Z., Zou, X.-Y., 2009. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415-425, doi:10.1007/s00726-008-0170-2.

Lin, H., Li, Q.-Z., 2007. Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J. Comput. Chem. 28, 1463-1466.

Lin, S.-X., Lapointe, J., 2013. Theoretical and experimental biology in one-A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. Journal of Biomedical Science and Engineering 6, 435-442, doi:10.4236/jbise.2013.64054.

Liu, B., Wang, X., Chen, Q., Dong, Q., Lan, X., 2012. Using Amino Acid Physicochemical Distance Transformation for Fast Protein Remote Homology Detection. PLoS ONE 7, e46633, doi:10.1371/journal.pone.0046633.

Liu, B., Wang, X., Zou, Q., Dong, Q., Chen, Q., 2013. Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation. Mol. Inf. 32, 775-782.

Liu, B., Liu, F., Fang, L., Wang, X., Chou, K.-C., 2014a. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics, btu820, doi:10.1093/bioinformatics/btu1820.

Liu, B., Xu, J., Zou, Q., Xu, R., Wang, X., Chen, Q., 2014b. Using distances between Top-n-gram and residue pairs for protein remote homology detection. BMC Bioinformatics 15, S3.

Liu, B., Xu, J., Fan, S., Xu, R., Zhou, J., Wang, X., 2014c. PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. Mol. Inf., doi:10.1002/minf.201400025.

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., Chou, K.-C., 2014d. iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. PLoS ONE 9, e106691, doi:10.1371/journal.pone.0106691.

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., Dong, Q., Chou, K.-C., 2014e. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30, 472-479.

Liu, W.-m., Chou, K.-C., 1998. Prediction of Protein Structural Classes by Modified Mahalanobis Discriminant Algorithm. Journal of Protein Chemistry 17, 209-217, doi:10.1023/a:1022576400291.

Marrero-Ponce, Y., Huesca-Guillén, A., Ibarra-Velarde, F., 2005a. Quadratic indices of the molecular pseudograph's atom adjacency matrix and their stochastic forms: a novel approach for virtual screening and in silico discovery of new lead paramphistomicide drugs-like compounds. J. Mol. Struct.: THEOCHEM 717, 67-79, doi:DOI: 10.1016/j.theochem.2004.11.027.

Marrero-Ponce, Y., Castillo-Garit, J. A., Castro, E. A., Torrens, F., Rotondo, R., 2008. 3D-chiral (2.5) atom-based TOMOCOMD-CARDD descriptors: theory and QSAR applications to central chirality codification J. Math. Chem. 44, 755-786.

Marrero-Ponce, Y., Medina-Marrero, R., Castillo-Garit, J. A., Romero-Zaldivar, V., Torrens, F., Castro, E. A., 2005b. Protein linear indices of the 'macromolecular pseudograph α-carbon atom adjacency matrix' in bioinformatics. Part 1: Prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor. Bioorg. Med. Chem. 13, 3003-3015, doi:DOI: 10.1016/j.bmc.2005.01.062.

Marrero-Ponce, Y., Marrero, R., Castro, E., Ramos de Armas, R., González-Díaz, H., Romero Zaldivar, V., Torrens, F., 2004. Protein Quadratic Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability. Molecules 9, 1124-1147.

Marrero-Ponce, Y., García-Jacas, C. R., Barigye, S. J., Valdés-Martiní, J. R., Rivera-Borroto, O. M., Pino-Urias, R. W., Cubillán, N., Alvarado, Y. J., 2014, Accepted for publication. Optimum Search Strategies or Novel 3D Molecular Descriptors: is there a Stalemate? Curr. Bioinf.

Mathews, C. K., van Holde, K. E., Ahern, K. G., 2000. Biochemistry. BenjaminCummings, San Francisco.

McFarland, J., Gans, D., 1990. Linear Discriminant Analysis and Cluster Significance Analysis. Compr. Med. Chem. 4, 667-689.

Moreau, G., Broto, P., 1980. The auto-correlation of a topological-structure-a new Molecular Descriptor. Nouveau Journal De Chimie-New Journal of Chemistry 4, 359-360.

Ortega-Broche, S. E., Marrero-Ponce, Y., Díaz, Y. E., Torrens, F., Pérez-Giménez, F., 2010. Tomocomd-camps and protein bilinear indices–novel bio-macromolecular descriptors for protein research: I. Predicting protein stability effects of a complete set of alanine substitutions in the Arc repressor. FEBS J. 277, 3118-3146.

Plaxco, K. W., Simons, K. T., Baker, D., 1998. Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 277, 985-994.

Ramos de Armas, R., González Díaz, H., Molina, R., Uriarte, E., 2004a. Markovian Backbone Negentropies: Molecular Descriptors for Protein Research. I. Predicting Protein Stability in Arc Repressor Mutants. Proteins: Struct., Funct., Bioinf. 56, 715–723.

Ramos de Armas, R., González Díaz, H., Molina, R., Pérez González, M., Uriarte, E., 2004b. Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides. Bioorg. Med. Chem. 12, 4815-4822.

Randic, M., Zupan, J., Balaban, A. T., Vikić-Topić, D. e., Plavšić, D., 2010. Graphical Representation of Proteins†. Chem. Rev. 111, 790-862.

Randić, M., Mehulić, K., Vukičević, D., Pisanski, T., Vikić-Topić, D., Plavšić, D., 2009. Graphical representation of proteins as four-color maps and their numerical characterization. J. Mol. Graphics Modell. 27, 637-641.

Rao, H., Zhu, F., Yang, G., Li, Z., Chen, Y., 2011. Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. 39, W385-W390.

Ruiz-Blanco, Y. B., García, Y., Sotomayor-Torres, C., Marrero-Ponce, Y., 2010. New set of 2D/3D thermodynamic indices for proteins. A formalism based on the Molten Globule theory. Phys. Procedia 8, 63-72.

Sak, K., Karelson, M., Järv, J., 1999. Modeling of the amino acid side chain effects on peptide conformation. Bioorg. Chem. 27, 434-442.

Shen, H.-B., Yang, J., Liu, X.-J., Chou, K.-C., 2005. Using supervised fuzzy clustering to predict protein structural classes. Biochem. Biophys. Res. Commun. 334, 577-581.

Sinkhorn, R., Knopp, P., 1967. Concerning nonnegative matrices and doubly stochastic matrices. Pacific J. Math. 21, 343-348.

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics. WILEY-VCH, Weinheim.

Todeschini, R., Consonni V., 2010. New Local Vertex Invariants and Molecular Descriptors Based on Functions of the Vertex Degrees. MATCH Commun. Math. Comput. Chem 64, 359-372.

Tropsha, A., Gramatica, P., Gombar, V. K., 2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69–77.

Wu, Z.-C., Xiao, X., Chou, K.-C., 2010. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267, 29-34, doi:http://dx.doi.org/10.1016/j.jtbi.2010.08.007.

Xiao, X., Wang, P., Chou, K.-C., 2008a. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J. Theor. Biol. 254, 691-696.

Xiao, X., Lin, W.-Z., Chou, K.-C., 2008b. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J. Comput. Chem. 29, 2018-2024.

Xiao, X., Shao, S.-H., Huang, Z.-D., Chou, K.-C., 2006. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J. Comput. Chem. 27, 478-482.

Zamyatnin, A., 1972. Protein volume in solution. Prog. Biophys. Mol. Biol. 24, 107-123.

Zhang, L., Zhao, X., Kong, L., 2014. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 355, 105-110, doi:http://dx.doi.org/10.1016/j.jtbi.2014.04.008.

Zhang, T.-L., Ding, Y.-S., Chou, K.-C., 2008. Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J. Theor. Biol. 250, 186-193.

Zhou, G.-P., 2011. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. J. Theor. Biol. 284, 142-148, doi:http://dx.doi.org/10.1016/j.jtbi.2011.06.006.

Zhou, G., Deng, M., 1984. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem. J. 222, 169.

Zhou, H., Zhou, Y., 2002. Folding rate prediction using total contact distance. Biophys. J. 82, 458-463.

ANNEXES

(Tables and Figures to be inserted in the Main Text)

Table 1. Amino acid side-chain labels.

Amino acid	Code		MM[a]	MV[b]	z-scale[c]			ECI[d]	ISA[e]	HWS[f]	KDS[g]
					z_1	z_2	z_3				
Alanine	ALA	A	89	88.6	0.01	-1.73	0.09	0.05	62.90	-0.5	1.8
Arginine	ARG	R	174	173.4	2.88	2.52	-3.44	1.69	52.98	3.0	-4.5
Asparagine	ASN	N	132	114.1	3.22	1.45	0.84	1.31	17.87	0.2	-3.5
Aspartate	ASP	D	133	111.1	3.64	1.13	2.36	1.25	18.46	3.0	-3.5
Cysteine	CYS	C	121	108.5	0.71	-0.97	4.13	0.15	78.51	-1.0	2.5
Glutamate	GLU	E	146	143.8	3.08	0.39	-0.07	1.31	30.19	0.2	-3.5
Glutamine	GLN	Q	147	138.4	2.18	0.53	-1.14	1.36	19.53	3.0	-3.5
Glycine	GLY	G	75	60.1	2.23	-5.36	0.30	0.02	19.93	0.0	-0.4
Histidine	HIS	H	155	153.2	2.41	1.74	1.11	0.56	87.38	-0.5	-3.2
Isoleucine	ILE	I	131	166.7	-4.44	-1.68	-1.03	0.09	149.77	-1.8	4.5
Leucine	LEU	L	131	166.7	-4.19	-1.03	-0.98	0.01	154.35	-1.8	3.8
Lysine	LYS	K	146	168.6	2.84	1.41	-3.14	0.53	102.78	3.0	-3.9
Methionine	MET	M	149	162.9	-2.49	-0.27	-0.41	0.34	132.22	-1.3	1.9
Phenylalanine	PHE	F	165	189.9	-4.92	1.30	0.45	0.14	189.42	-2.5	2.8
Proline	PRO	P	115	112.7	-1.22	0.88	2.23	0.16	122.35	0.0	-1.6
Serine	SER	S	105	89.0	1.96	-1.63	0.57	0.56	19.75	0.3	-0.8
Threonine	THR	T	119	116.1	0.92	-2.09	-1.40	0.65	59.44	-0.4	-0.7
Tryptophan	TRP	W	204	227.8	-4.75	3.65	0.85	1.08	179.16	-3.4	-0.9
Tyrosine	TYR	Y	181	193.6	-1.39	2.32	0.01	0.72	132.16	-2.3	-1.3
Valine	VAL	V	117	140.0	-2.69	-2.53	-1.29	0.07	120.91	-1.5	4.2

[a]Molecular Mass (Mathews et al., 2000), [b]Side-chain amino acid volume (Zamyatnin, 1972), [c]Z-scale (Hellberg et al., 1987),[d]Atomic charge (Collantes and Dunn III, 1995), [e]Side-chain isotropic surface area (Collantes and Dunn III, 1995), [f]Hoop-Woods hydropathy index (Hopp and Woods, 1981), [g]Kyte-Doolittle hydropathy index (Kyte and Doolittle, 1982).

Table 1. Amino acid side-chain labels (*continued*).

Amino acid	Code		PIE[h]	PAH[i]	PBS[j]	PTT[k]	L19[l]	ξ[m]	EPS[n]
Alanine	ALA	A	6.01	1.29	0.90	0.78	19.20	-77.85	-433.66
Arginine	ARG	R	10.76	0.96	0.99	0.88	17.80	108.86	-403.21
Asparagine	ASN	N	5.41	0.90	0.76	1.28	21.72	-55.42	-466.61
Aspartate	ASP	D	2.77	1.04	0.72	1.41	17.14	47.89	-518.10
Cysteine	CYS	C	5.07	1.11	0.74	0.80	18.83	160.13	-425.69
Glutamate	GLU	E	3.22	1.44	0.75	1.00	18.55	134.68	-479.54
Glutamine	GLN	Q	5.65	1.27	0.80	0.97	17.31	53.27	-531.69
Glycine	GLY	G	5.97	0.56	0.92	1.64	19.48	-148.03	-420.86
Histidine	HIS	H	7.59	1.22	1.08	0.69	13.97	-4.57	-378.92
Isoleucine	ILE	I	6.02	0.97	1.45	0.51	20.76	-104.80	-449.27
Leucine	LEU	L	5.98	1.30	1.02	0.59	17.65	-148.50	-448.27
Lysine	LYS	K	9.74	1.23	0.77	0.96	17.05	47.61	-446.97
Methionine	MET	M	5.74	1.47	0.97	0.39	17.88	46.37	-435.34
Phenylalanine	PHE	F	5.48	1.07	1.32	0.58	16.81	47.67	-376.77
Proline	PRO	P	6.48	0.52	0.64	1.91	18.55	169.73	-422.17
Serine	SER	S	5.68	0.82	0.95	1.33	18.91	30.24	-479.75
Threonine	THR	T	5.87	0.82	1.21	1.03	17.15	46.04	-483.37
Tryptophan	TRP	W	5.89	0.99	1.14	0.75	20.94	178.69	-365.49
Tyrosine	TYR	Y	5.66	0.72	1.25	1.05	16.86	49.11	-446.32
Valine	VAL	V	5.97	0.91	1.49	0.45	17.88	-106.50	-434.30

[h]Isoelectric point (Hellberg et al., 1987); [i,j,k]Relative frequencies with which an amino acid appear forming α-helices, β-sheets and reverse turns, respectively (Mathews et al., 2000); [l,m]Geometric compatibility parameters (Sak et al., 1999); [n]Heat of formation (Sak et al., 1999).

Table 2. Amino acidic composition of the local fragments pre-defined in the 3D module of the TOMOCOMD-CAMPS software.

Local-Fragment	Amino acids
RAP[a]	PRO, ILE, ALA, VAL, LEU, PHE, TRP, MET.
R+[b]	LYS, HIS, ARG.
R-[c]	ASP, GLU.
RPU[d]	ASN, CYS, GLY, SER, THR, TYR, GLN.
ARG[e]	PHE, TYR, TRP.
ALG[f]	GLY, ALA, PRO, VAL, LEU, ILE, MET.
UFG[g]	GLY, PRO.
FAH[h]	ALA, CYS, LEU, MET, GLU, GLN, HIS, LYS.
FBS[i]	VAL, ILE, PHE, TYR, TRP, THR.
AFT[j]	GLY, SER, ASP, ASN, PRO.

[a]Apolar; [b]Polar positively charged; [c]Polar negatively charged; [d]Polar uncharged; [e]Aromatic; [f]Aliphatic; [g]Unfolding amino acids; [h]Helix favoring amino acids; [i]Beta-sheets favoring amino acids; [j]Beta-turn favoring amino acids.

Table 3. LDA-model's canonical discriminant functions at group centroids.

Class	CDF1	CDF2	CDF3
All-α	3.94	-1.47	0.24
All-β	0.25	2.28	0.04
α/β	-10.1	-1.0	-0.02
α+β	3.9	-0.7	-0.3

Table 4. Standardized coefficients of the canonical discriminant functions 1 and 2.

Variable	Function	
	1	2
${}^{1,1}_{ds}b_z^{MM-L19}(\bar{x}_m, \bar{y}_m)$	-4.65	20.5
${}^{3,1}_{ns}b_{zval}^{MM-PIE}(\bar{x}_m, \bar{y}_m)$	-0.45	0.4
${}^{1,1}_{ds}b_z^{ECI-MM}(\bar{x}_m, \bar{y}_m)$	8.21	-34.3
${}^{2,1}_{ds}b_z^{ECI-PAH}(\bar{x}_m, \bar{y}_m)$	-5.13	20.7
${}^{3,1}_{ds}b_z^{ISA-PIE}(\bar{x}_m, \bar{y}_m)$	1.04	-10.1
${}^{1,1}_{ds}b_z^{PIE-ECI}(\bar{x}_m, \bar{y}_m)$	-1.51	6.8
${}^{3,1}_{ds}b_z^{PIE-EPS}(\bar{x}_m, \bar{y}_m)$	-0.59	2.1
${}^{2,1}_{ds}b_z^{PAH-PBS}(\bar{x}_m, \bar{y}_m)$	4.14	-28.5
${}^{3,1}_{ds}b_z^{PBS-MV}(\bar{x}_m, \bar{y}_m)$	-3.89	24.1
${}^{2,2}_{ds}b_z^{PBS-PAH}(\bar{x}_m, \bar{y}_m)$	0.67	7.3
${}^{3,2}_{ds}b_z^{PTT-MV}(\bar{x}_m, \bar{y}_m)$	0.80	-7.2

Table 5. Statistical parameters for the LDA-model.

Set	*Global accuracy Q(%)*	*Sensitivity*		*Specificity*	
		Class		***Class***	
Training	92.6	All-α	86.8	All-α	86.8
		All-β	97.9	All-β	100.0
		α/β	100.0	α/β	100.0
		α+β	85.3	α+β	82.9
		Class		***Class***	
Test	92.7	All-α	85.7	All-α	85.7
		All-β	100.0	All-β	100.0
		α/β	100.0	α/β	100.0
		α+β	83.3	α+β	83.3

Table 6. Statistical parameters for the bootstrapping cross-validation of the LDA-model.

Group	Q_{Total}[a]	λ	D^2_{12}	D^2_{13}	D^2_{14}	D^2_{23}	D^2_{24}	D^2_{34}	F	Q_{Total}[b]
1.	92.24	0.007	25.29	226.80	1.04	154.02	20.78	233.32	39.56	90.91
2.	90.40	0.009	25.90	186.24	0.77	111.11	22.08	186.70	38.56	91.67
3.	90.74	0.01	27.63	170.07	0.75	95.14	23.65	169.21	31.04	92.68
4.	93.10	0.007	26.73	202.30	0.92	137.77	21.29	205.05	38.73	90.91
5.	93.86	0.01	25.65	191.66	0.78	116.07	21.27	191.76	34.16	91.43
6.	93.75	0.007	29.27	202.31	1.20	129.80	22.94	200.89	37.91	89.19
7.	93.64	0.009	27.49	214.02	0.65	128.75	23.17	212.85	34.58	92.31
8.	91.38	0.01	26.51	183.70	1.06	112.16	20.21	182.25	34.51	90.91
9.	92.24	0.009	28.33	220.64	0.74	134.93	24.03	220.67	38.00	93.40
10.	88.79	0.009	27.59	233.83	0.81	187.94	23.34	234.35	36.56	90.91
Av	92.01	0.01	27.04	203.16	0.87	130.77	22.28	203.71	36.36	91.43
Std	1.68	0.001	1.25	20.59	0.18	26.03	1.33	21.81	2.70	1.17
LL	86.99	0.004	23.70	140.51	0.37	77.25	18.24	141.84	28.69	89.05
UL	97.41	0.01	29.97	266.07	1.39	172.94	26.33	264.84	44.26	94.01

[a,b]Global accuracy of the LDA-model on the training (75% of the proteins) and test (25% of the proteins) sets, respectively, Av: Arithmetic mean, Std: Standard deviation, LL and UL: are calculated by mean of subtract 1.5 times the difference between first and third quartiles to the first and third quartiles, respectively.

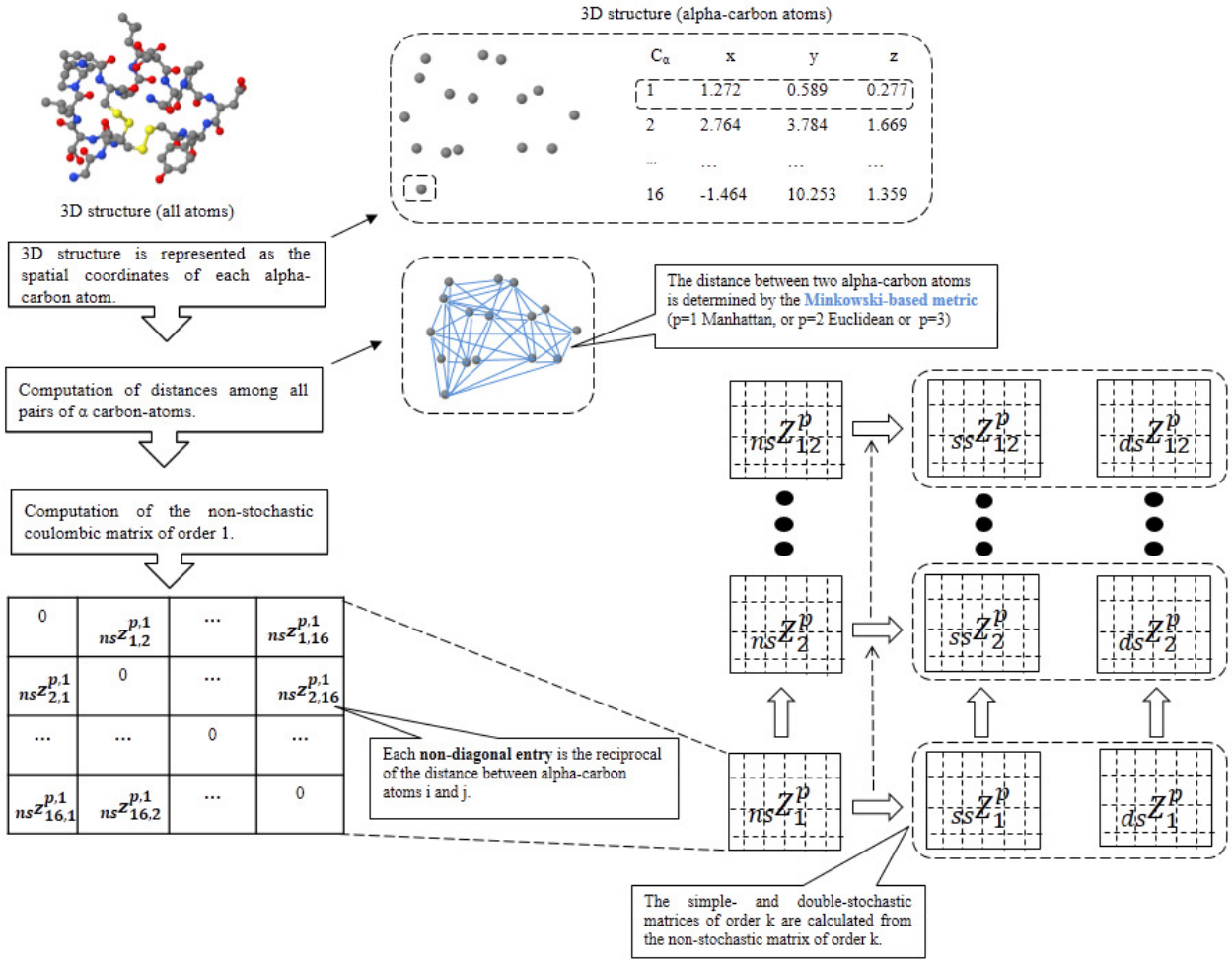


Figure 1. Major steps in the computation of the non-stochastic, simple-stochastic and double-stochastic coulombic matrices.

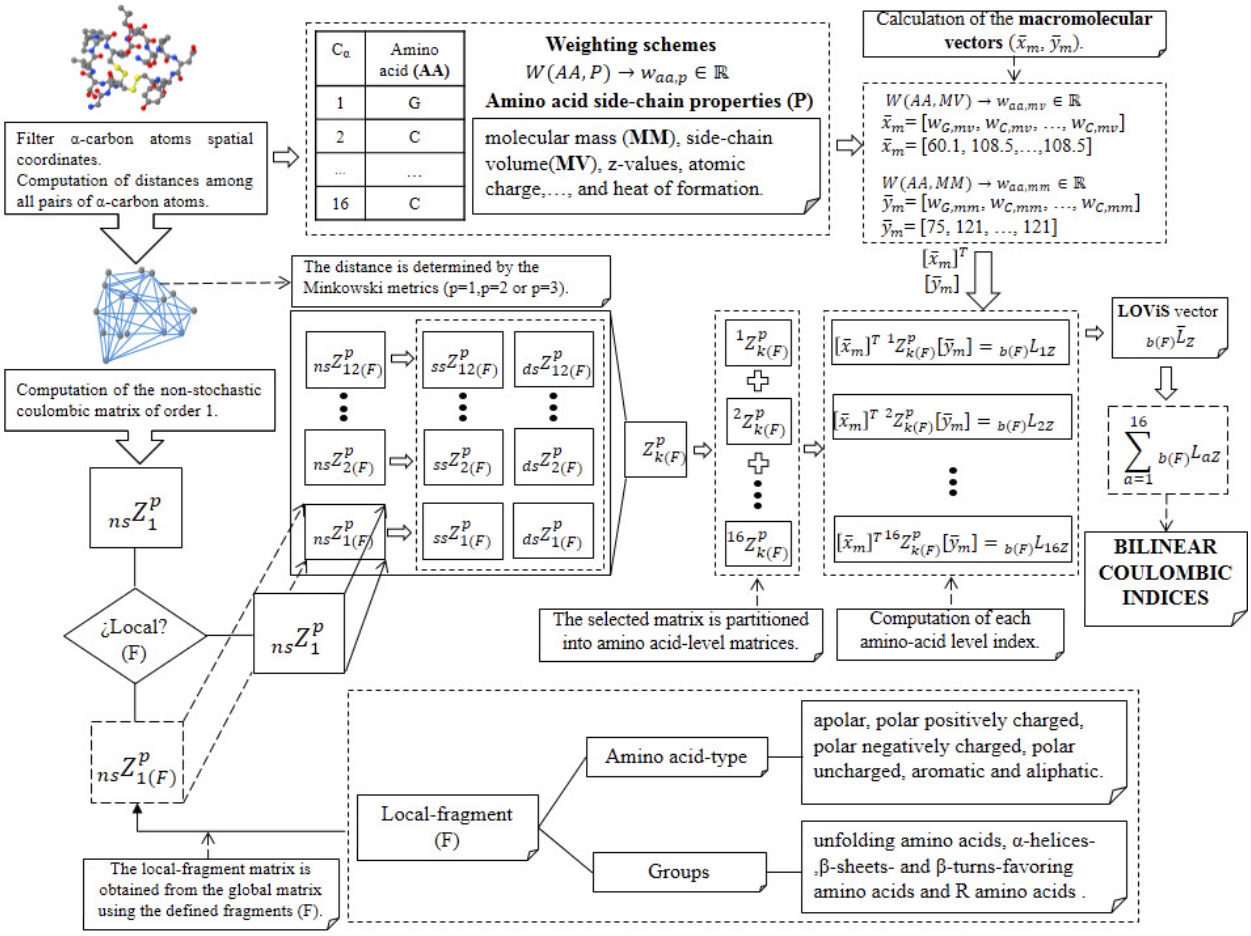


Figure 2. Workflow followed in the calculation of the 3D-protein bilinear MDs.

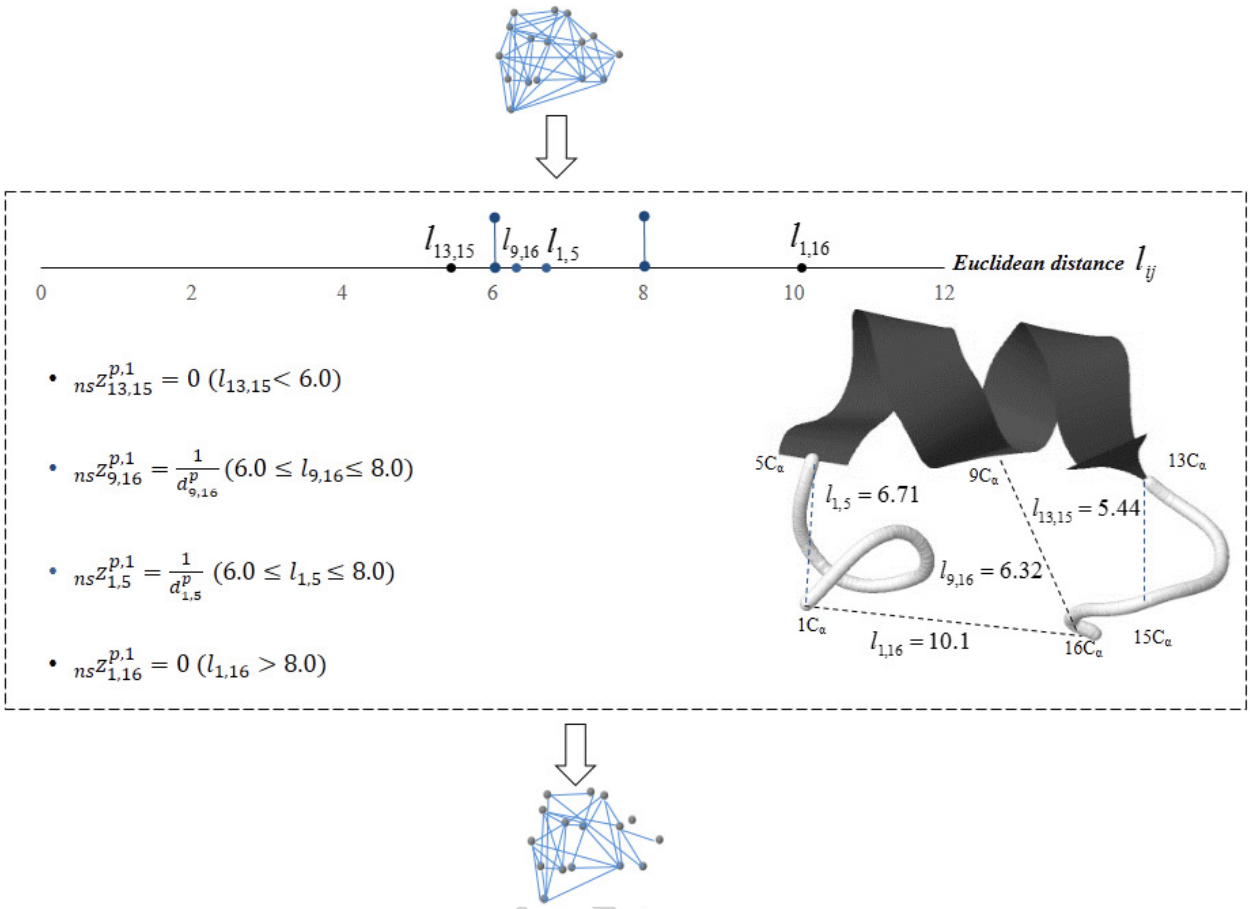


Figure 3A). Application of the geometric lag l, cut-off interval lag l [6; 8], in the calculation of entries ${}^{p,1}_{ns}\boldsymbol{z_{ij}}$ of the non-stochastic coulombic matrix of order 1, ${}_{ns}\boldsymbol{Z^{p}_{1}}$.

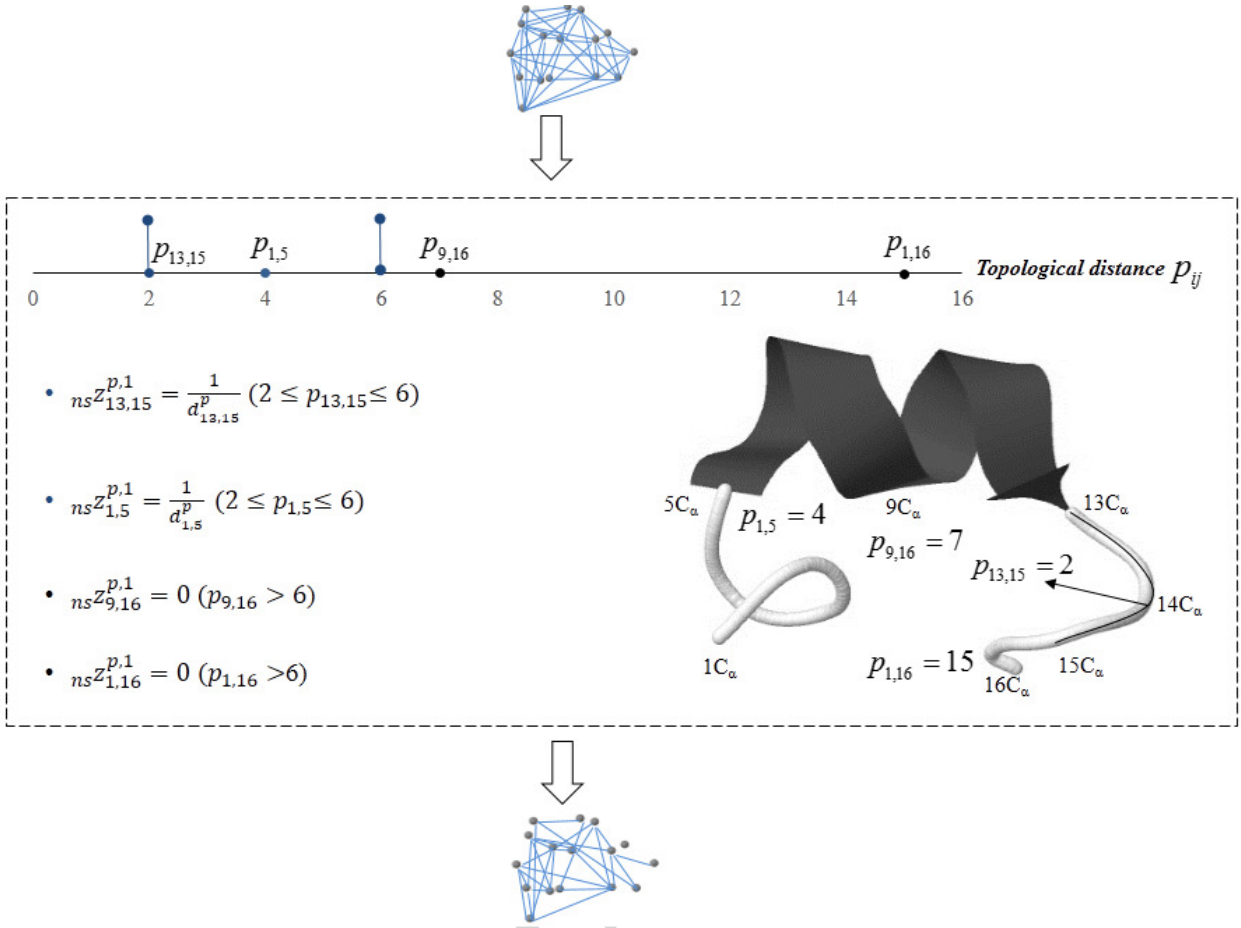


Figure 3B). Application of the topological lag p, cut-off interval lag p [2; 6], in the calculation of entries ${}^{p,1}_{ns}\boldsymbol{z_{ij}}$ of the non-stochastic coulombic matrix of order 1, ${}_{ns}\boldsymbol{Z^{p}_{1}}$.

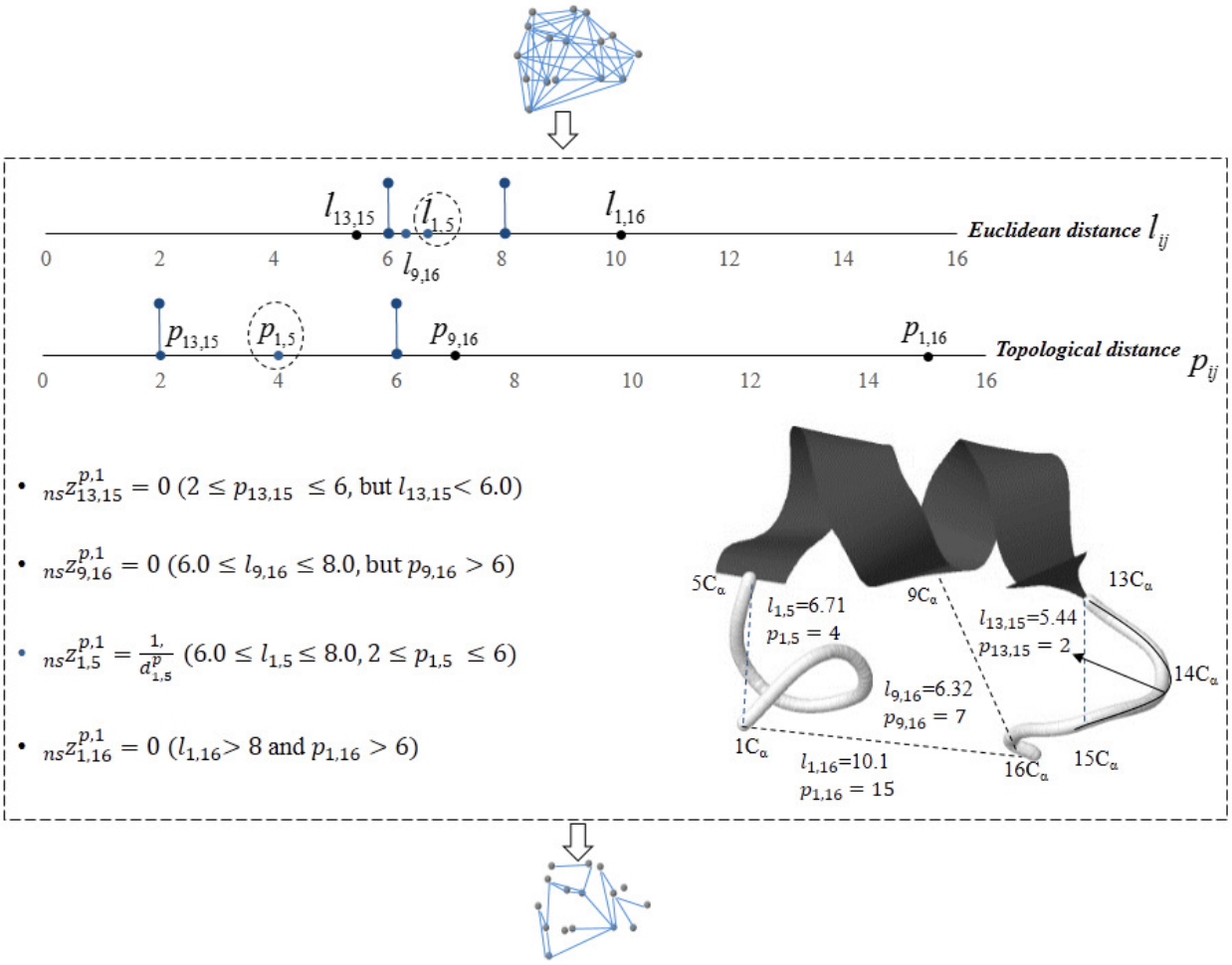


Figure 3C). Application of the geometric lag l and topological lag p, cut-off intervals (lag l [6; 8]), lag p [2; 6]) in the calculation of entries $^{p,1}_{ns}\boldsymbol{z_{ij}}$ of the non-stochastic coulombic matrix of order 1, $_{ns}\boldsymbol{Z^{p}_{1}}$.

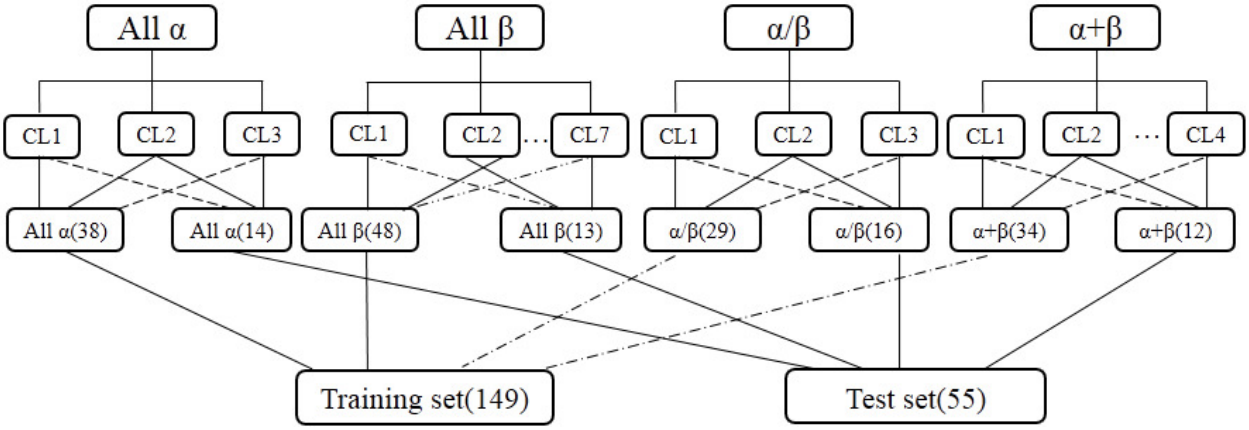
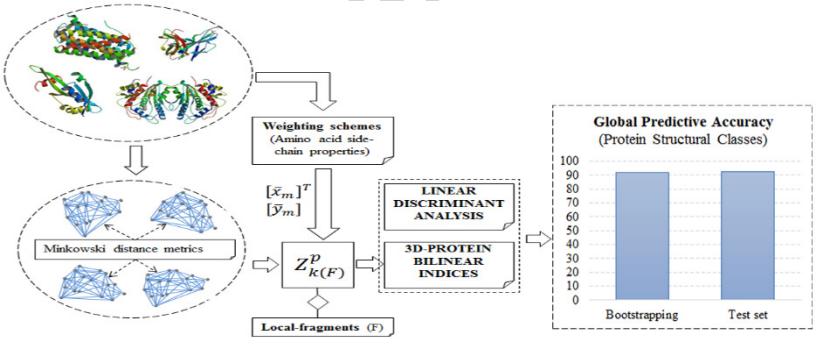


Figure 4. Selection of the training and test sets, for the calibration and external validation of the LDA-model.

Graphical Abstract

Novel 3D Bio-Macromolecular Bilinear Descriptors for Protein Science: Predicting Protein Structural Classes

Yovani Marrero-Ponce (✉), Ernesto Contreras-Torres, César R. García-Jacas, Stephen J. Barigye, Néstor Cubillán, and Ysaías J. Alvarado.



- New 3D protein descriptors based on the bilinear algebraic form are proposed.
- We define the coulombic matrix to codify the 3D structure of proteins.
- Normalization approaches for the coulombic matrix are employed.
- Local-fragment indices and constrains approach are defined.
- We built a model that showed high accuracy predicting protein structural classes.